Arbejde med korpora i DK-CLARIN
(Tidligere titel: Korpus-workflow)

DK-CLARIN WP 5.1-arbejdspapir
Jørg Asmussen
med input fra WP 5.1, WP 5.2, WP 2.1 og WP 2.2

Version 3.0 – 27. november 2008

Resumé

I dette arbejdspapir beskrives, hvordan man ideelt set skal kunne arbejde med monolingvale korpora i en DK-CLARIN-kontekst, og hvad dette betyder for DK-CLARIN's tekniske infrastruktur. Der kan skelnes mellem to brugergrupper: Sprogforskere og sprogteknologer, hver med deres specifikke behov. Da det resursemæssigt er vanskeligt at implementere en komplet løsning, der tilgodeser alles behov, prioriteres sprogforskerens formodede behov højest.

1 Indledning

Dette arbejdspapir beskriver, hvordan man ideelt set skal kunne arbejde med monolingvale korpora i en DK-CLARIN-kontekst, og hvilke krav det stiller til DK-CLARIN's tekniske infrastruktur. Der skitseres en ret omfattende model, som inden for rammerne af det løbende DK-CLARIN næppe vil kunne realiseres i sin helhed. Derfor giver papiret også en prioritering af funktionaliteterne i modellen, og indkredser således hvad der som minimum bør være muligt i en første version af DK-CLARIN's tekniske infrastruktur.

Baggrunden for dette papir var forsøget på en beskrivelse af følgende arbejdsscenario:

1. En bruger finder frem til og udvælger bestemte tekster (et brugerdefineret korpus) vha. DK-CLARIN's portalsite.
2. Brugeren annoterer efterfølgende de udvalgte tekster efter brugerens selvdefinerede annotationssystem.
3. Brugeren gemmer sit korpus sammen med egne annotationer, hvorved de også kan komme andre brugere til gode.

Selvom scenariet synes ligetil, viste en udmøntning af det sig at være mere kompleks end umiddelbart antaget. Hvis scenariet skal realiseres, kræves i princippet alle de under afsnit 2 beskrevne funktionaliteter implementeret i infrastrukturen – en opgave, som DK-CLARIN's nuværende WP 5-resurser næppe står mål med. Derfor blev det nødvendigt at underkaste de enkelte funktionaliteter en relevansvurdering i relation til de begrænsede resurser, WP 5 faktisk råder over.

Papiret begrænser beskrivelsen til monolingvale, ikke-parallelle, ikke-alignerede skriftsprogskorpora. Håndtering af multilingvale, parallelle, alignerede korpora stiller særlige krav til konkordansinterfaces m.v., og det er uvist, hvorvidt de umiddelbart vil kunne håndteres af de samme konkordansværktøjer som de monolingvale korpora.[1] Endelig udgør talesprogskorpora en yderligere kategori med særlige kravsspecfikationer, som heller ikke er genstand for dette arbejdspapir.

En nærmere beskrivelse af den anvendte korpus-terminologi i dette dokument findes i [Asmussen og Offersgaard, 2008].

2 Funktionaliteter

DK-CLARIN's portalsite bør ideelt tilbyde følgende funktionaliteter for arbejdet med tekstkorpora. Brugerne skal have mulighed for at

1. få et overblik over tilgængelige tekstsamlinger og tekstmaterialet heri via de karakteristika (metadata), der er knyttet til både samlingerne som helhed og hver enkelt tekstenhed i dem. Dette skal være muligt, uanset om tekstmaterialet fysisk befinder sig i DK-CLARIN's centrale repository eller lokalt hos en eller flere af DK-CLARIN's partnere

2. sammensætte et udvalg af tekster (tekstenheder) på baggrund af bestemte tekstkarakteristika (metadata) og få adgang til dem, gerne i deres helhed og uanset deres fysiske placering i DK-CLARIN's infrastruktur (centralt og/eller distribueret)

3. arkivere et sådant udvalg af tekster »virtuelt« (som en række pointere til de egentlige tekstenheder) på portalsitet sammen med en beskrivelse (metadata), så også andre (gennem rettighedsstyring) kan få adgang til netop dette udvalg af tekster. Et sådant udvalg af tekster udgør i princippet allerede et korpus, dog ét, man ikke umiddelbart vil kunne foretage avancerede konkordanssøgninger og statistiske undersøgelser i

4. indeksere et brugerdefineret korpus fremkommet som resultat af arbejdsgangene under pkt. 1 til 3, så det bliver muligt at foretage avancerede konkordanssøgninger i det, og arkivere dette indekserede korpus i det centrale repository sammen med metadata, så andre (gennem rettighedsstyring) også kan få adgang til konkordanssøgninger i det[2]

5. søge i et brugerdefineret korpus via et konkordansinterface på DK-CLARIN's portalsite

6. søge i prædefinerede korpora, som blev eller bliver tilvejebragt af særlige projekter tilknyttet DK-CLARIN eller arbejdspakker i DK-CLARIN, uanset om deres konkordansinterface er en integreret del af DK-CLARIN's portalsite, eller om det stilles til rådighed decentralt hos en eller flere af DK-CLARIN's partnere

[1] Det henstilles til WP 2.6 at udarbejde et særligt arbejdspapir med vurderinger og anbefalinger.

[2] Denne funktionalitet er betinget af brugen af IMS Open Corpus Workbench, jf. også 2.4.

7. annotere et brugerdefineret eller et prædefineret korpus på tokenniveau og siden redigere i annotationerne

8. arkivere brugerannotationer i det centrale repository, så andre (gennem rettighedsstyring) kan få adgang til dem

9. indlemme brugerannotationer i den indekserede version af et brugerdefineret korpus, så de bliver søgbare i et konkordansinterface

10. få adgang til korpusrelevante tools som tokenizer, lemmatizer, tagger og andre annoteringsværktøjer med mulighed for at tilpasse dem egne behov uanset deres fysiske placering i DK-CLARIN's infrastruktur

I det følgende beskrives hver enkelt af disse ti funktionaliteter nærmere i hver sit afsnit, og det skitseres, hvad der skal til for at realisere den. Afsnittene består derfor altid af en kortfattet beskrivelses- og en implementationsdel.[3] Herudover kunne man for hvert afsnit også tilføje en (eller flere) case-del(e).

En implementering af samtlige ti funktionaliteter ville være særdeles ressursekrævende og er derfor ikke realistisk inden for det nuværende DK-CLARIN-projekt, selvom den overordnede projektbeskrivelse ville kunne fortolkes sådan i sin yderste konsekvens. Derfor giver listen over funktionaliteter ovf. blot et ideelt billede af, hvad der på længere sigt kunne være ønskeligt.

Det er endvidere uvist, i hvor høj grad den komplette samling af funktionaliteter vil være lige meget efterspurgt blandt alle korpusbrugere. I to af DK-CLARIN's korpusarbejdspakker, WP 2.1 (referencekorpus) og WP 2.2 (fagsprogskorpus), gås der ud fra, at brugermålgruppen primært er sprogforskere, som ønsker at kunne lave sproglige undersøgelser vha. et prædefineret korpus, og kun sekundært er sprogteknologer, som vil være interesseret i selv at sammenstille egne korpora, annotere dem og stille dem til rådighed for andre. Det er dette syn, der vil ligge til grund for en prioriteringsdel i hvert af de følgende afsnit, hvori der gives en anbefaling af, hvorvidt den beskrevne funktionalitet synes påkrævet i en første version af DK-CLARIN, eller hvorvidt den kan undværes.

2.1 Overblik over tilgængelige tekstsamlinger

Beskrivelse

Denne arbejdsgang skal kunne give svar på,

- hvad og hvor meget der findes af tekstmateriale inden for fx forskellige perioder, genrer, fagområder, medier, kommunikationssituationer etc., med hvilke typer tokeniseringer[4] og med hvilke annotationer på tokenniveau

- og på hvilken måde man kan man få adgang til materialet. Fx gennem download, visning af teksterne på skærmen, simpel eller avanceret konkordanssøgning i dem, statistisk profilering af dem, håndtering af dem med andre værktøjer på DK-CLARIN-sitet

[3] Inden en egentlig implementering kan finde sted, må en detaljeret kravsspecifikation for hver enkelt funktionalitet udarbejdes. Det er ikke sigtet med dette arbejdspapir at give denne type specifikationer.

[4] Der bør stiles mod en standard-tokenisering i DK-CLARIN.

Implementering

På portalsitet implementeres en browser, der kan vise, hvilket materiale der findes, og hvilken form for adgang til det brugeren er berettiget til. Oversigterne i browseren genereres automatisk på baggrund af materialets metadata. Brugeren skal nemt kunne navigere blandt oversigterne. Metadata, som er grundlaget for materialebrowserens funktion, stilles til rådighed af DK-CLARIN-partnerne enten lokalt hos dem som en webservice eller centralt i portalsitets metadata-repository.

Prioritering

Denne funktionalitet er væsentlig for både sprogforskere og sprogteknologer.

Funktionaliteten betragtes som særdeles central også i relation til det samlede projekts målsætning, således som den kommer til udtryk i afsnittet *Technology* i annex 1 i DK-CLARIN's samlede projektbeskrivelse. Funktionaliteten har derfor højeste prioritet.

2.2 Sammensætning af et udvalg af tekster

Beskrivelse

Denne funktionalitet ligger i forlængelse af den, der blev beskrevet i afsnittet ovf. Når brugeren har dannet sig et overblik over, hvilket tekstmateriale der er tilgængeligt, kan han definere en profil, der på baggrund af metadata og evt. forekomst af bestemte ord i teksterne kan udpege præcist alle de tekster, som matcher den. I det omfang, det er teknisk og juridisk muligt, skal brugeren endvidere have en eller anden form for adgang til teksterne, fx ved at han kan få vist teksterne i deres helhed på skærmen, downloade dem eller få adgang til dem i et konkordansværktøj.

Implementering

På portalsitet implementeres et interface til definition af profiler og et værktøj, der på baggrund af en sådan profil[5] fremfinder de matchende tekster[6], således at brugeren har fornemmelsen af umiddelbart at råde over sit udvalg af tekster, uanset hvor teksterne ellers måtte ligge rent fysisk i den samlede infrastruktur. Brugeren kan fx blive præsenteret for en liste over relevante tekster. Et klik på et emne på listen kan fx føre til en fuldtekstversion af pågældende tekst og en fuldstændig oversigt over dens metadata, evt. endda download af alle eller nogle tekster i udvalget, hvis DK-CLARIN's ophavsretaftaler tillader det. Metadata stilles til rådighed af de enkelte tekstsamlingers udbydere i form af webservices eller afleveres til DK-CLARIN-portalsitets metadata-repository. En primitiv form for konkordanssøgning på en sådan gruppe tekster kan evt. også overvejes implementeret. Endelig kan man i en senere fase af DK-CLARIN (efter

[5] Endvidere kan en simpel søgning på forekomst af bestemte ord i tekstmaterialet overvejes som en del af profilen, fx på et niveau, der svarer til Googles.

[6] Skal et udvalg af tekster sammenstilles med materiale fra flere forskellige tekstsamlinger (måske placeret hos flere forskellige udbydere), kræves, at deres metadata er identisk opbyggede, hvorfor tekstsamlingerne så vidt muligt bør anvende en fælles metadatastandard: TEI P5-headere. Endvidere kræves ensartet tokenisering og – i det omfang, det giver mening – annotation.

2010) forestille sig, at et sådant tekstudvalg også kan blive behandlet af andre værktøjer på sitet.

Prioritering

Denne funktionalitet vil være interessant for sprogforskere, især hvis den kombineres med funktionaliteterne 3–5. Den vil være interessant for sprogteknologer, især hvis den kombineres med funktionaliteterne 7–10.

Funktionaliteten hænger tæt sammen med den, der er beskrevet i afsnit 2.1 ovenfor og er derfor ligeledes central for DK-CLARIN's overordnede målsætning. Den bør derfor være implementeret i den første version af infrastrukturløsningen.

2.3 Arkivering af et brugerdefineret udvalg af tekster

Beskrivelse

Da det kan være en omfattende opgave at oprette en profil, der præcist udvælger de tekstenheder fra tekstsamlingerne, som man som bruger måtte være interesseret i at beskæftige sig med, bør et brugerdefineret udvalg af tekster kunne arkiveres på portalsitet, således at brugeren selv kan få adgang til det igen, og således at han også kan give andre adgang hertil.

I forbindelse med arkiveringen bliver brugeren bedt om at give en beskrivelse for udvalget i form af fastlagte metadata. Brugeren kan endvidere fastlægge offentlighedsgraden for udvalget.

Implementering

På portalsitet implementeres et interface til input af brugerens beskrivelse af det valgte udtræk i form af fastlagte metadata. Tekstudvalget gemmes som en række pointere på portalsitet sammen med brugerens metadata for udvalget samt følgende oplysninger:

1. bruger-id
2. søgeprofilen, som ligger til grund for udvælgelsen af teksterne

Prioritering

Med hensyn til brugermålgruppen for denne funktionalitet, jf. afsnittet *Prioritering* under afsnit 2.2.

Funktionaliteten kan betragtes som hørende sammen med dem beskrevet i afsnit 2.1 og 2.2 og bør derfor implementeres. Funktionaliteterne 1–3 kan betragtes som et hele og bør derfor alle være implementeret i første version af infrastrukturen. Hvis en prioritering af disse tre funktionaliteter skulle vise sig nødvendig, anbefales de implementeret i rækkefølgen 1, 2, 3.

2.4 Indeksering af et brugerdefineret korpus

Beskrivelse

Et specifikt udvalg af tekster er i princippet et korpus. Når brugeren har udvalgt bestemte tekster, der tilsammen udgør et korpus, er det nærliggende, at

han også bør have mulighed for at udføre avancerede konkordanssøgninger i det og helst også forskellige former for statistiske undersøgelser. Derfor bør der stilles et avanceret konkordansværktøj til rådighed evt. kombineret med en række statistiske værktøjer, som brugeren kan læse sit korpus ind i.

Implementering

Indtil videre må *IMS Open Corpus Workbench*[7] http://cwb.sourceforge.net/ anses som det bedste bud på et avanceret konkordansværktøj. For at gøre et korpus søgbart i dette system kræves, at korpusset indekseres forinden. Selve indekseringen tager tid, afhængig af korpusstørrelse, annoteringsgrad og platform skal der typisk regnes med mellem 5 og 30 minutter. Indekseringsprogrammerne er udviklet til en korpusadministrator med særlige rettigheder: Der må derfor udvikles en særlig »indpakning«, hvis alle og enhver skal kunne bruge det over nettet til at indeksere egne korpora med. Indeksering kræver endvidere, at inputtet foreligger i et særligt format (non-XML) og tegnsæt (8 bit), så konverteringsprogrammel må også udvikles.

I stedet for en eksplicit bruger-igangsat indeksering kan man også overveje at få foretaget en stiltiende indeksering i forbindelse med gemningen af et tekstudvalg, jf. afsnit 2.3, måske især som en løsning for mindre korpora.

Indekserede brugerdefinerede korpora bør gemmes centralt på DK-CLARIN's portalsite og der bør gives adgang til dem via en specielt udviklet grænseflade, jf. følgende afsnit.

Prioritering

Med hensyn til brugermålgruppen for denne funktionalitet, jf. afsnittet *Prioritering* under afsnit 2.2.

Generelt er brugerstyret indlæsning af korpora i IMS Open Corpus Workbench teknisk og administrativt ikke-trivielt. Det anbefales derfor at nedprioritere implementeringen af denne funktionalitet i det nuværende DK-CLARIN, og i stedet for at prioritere funktionalitet 6, søgning i prædefinerede korpora – en funktionalitet, der også tilgodeser sprogforskeres behov.

2.5 Konkordanssøgning i brugerdefinerede korpora

Beskrivelse

Der skal kunne laves konkordanssøgninger i et brugerdefineret korpus via et interface à la det, som bruges til KorpusDK, jf. http://ordnet.dk/korpusdk/teksteksempler.

Implementering

En konkordansgrænseflade i DK-CLARIN kan tage udgangspunkt i KorpusDK's eksisterende grænseflade. Det kan undersøges, hvorvidt det er muligt at anvende en DK-CLARIN-konfigureret version af den. Alternativt kan DK-CLARIN udvikle sin egen.

[7] I daglig tale ofte kaldt *CQP*, som står for *Corpus Query Processor*, som er selve søgemaskinen i systemet.

KorpusDK's konkordansgrænseflade samt i øvrigt den langt mere simple konkordansvisning på http://sproget.dk kommunikerer med Jørg Asmussens korpusserver *PyCOCS*, udviklet til DSL, som beror på IMS Open Corpus Workbenchs CQP-komponent som søgemaskine. PyCOCS ville med overskuelige midler kunne rekonfigureres til brug inden for DK-CLARIN. Alternativt kan DK-CLARIN udvikle en egen løsning.

Et åbent spørgsmål er, om det skal være muligt at foretage konkordanssøgninger i en brugerfastlagt serie af korpora, idet dette ville øge implementationens kompleksitet betydeligt. En mulighed er, at man for en sådan række af korpora laver en selvstændig søgning i hvert korpus og får præsenteret søgeresultaterne som et hele i samme grænseflade.

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 end KorpusDK's. Statistiske beregninger i korpora kan være ret tidkrævende: For at nedbringe beregningstiden er en præprocessering af data som regel nødvendig, hvilket fx kunne ske som et særligt led før eller efter korpusindekseringen, men en sådan præprocessering er i sig selv igen tidkrævende.

Prioritering

Med hensyn til brugermålgruppen for denne funktionalitet, jf. afsnittet *Prioritering* under afsnit 2.2.

Udvikling af en grænseflade til konkordanssøgninger samt evt. statistikværktøjer er resursetunge opgaver, hvadenten der bygges videre på eksisterende løsninger, eller der gås egne veje. Derfor bør WP 5 afveje denne implementeringsopgave nøje i forhold til andre infrastrukturopgaver, der også skal løses.

Det anbefales derfor at nedprioritere implementeringen af denne funktionalitet i det nuværende DK-CLARIN, og i stedet for at prioritere funktionalitet 6, søgning i prædefinerede korpora – en funktionalitet, der også tilgodeser sprogforskeres behov.

2.6 Konkordanssøgning i prædefinerede korpora

Beskrivelse

Færdige, afsluttede korpora, som er opbygget i forskellige DK-CLARIN-arbejdspakker eller som stilles til rådighed af andre projekter/institutioner, kan indekseres og gøres søgbare en gang for alle. Der skal kunne foretages konkordanssøgninger i dem via portalsitet.

Implementering

Hvis portalsitet etablerer et eget, egnet konkordansinterface, kan dette bruges som fælles interface for alle konkordanssøgbare korpora under DK-CLARIN. Korpusserveren og de binære indeksfiler kan enten ligge på portalsitet eller lokalt hos en partner.

Alternativt kan et konkordansinterface ligge hos en partner og der må så henvises dertil fra portalsitet.

Prioritering

Denne funktionalitet henvender sig primært til sprogforskere. Den bør få høj prioritet, hvad enten den implementeres centralt på portalsitet eller lokalt hos en udbyder, fx DSL, der i forvejen hoster en sådan funktionalitet for KorpusDK og sproget.dk.

2.7 Annotering af et korpus

Beskrivelse

Både brugerdefinerede og prædefinerede korpora skal helt eller delvis kunne annoteres af brugerne på tokenniveau med annotationssystemer opstillet af brugeren og ved hjælp af annotationsværktøjer stillet til rådighed på portalsitet eller brugerens egne værktøjer.

Hvis brugeren ikke ønsker at opmærke et helt 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.

Implementering

Brugerannotering, som skal harmonere med allerede eksisterende annoteringer, forudsætter, at tekstenhederne i et korpus én gang for alle er blevet 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 tekstsamlingen (men ikke nødvendigvis på samme maskine som samlingen), bør tekstenheder tokenopdeles allerede i forbindelse med indlemmelsen af dem i en samling.[8] Det er hensigtsmæssigt, at DK-CLARIN definerer, hvordan tekst tokenopdeles.[9] Da hvert token via tekst-id'et og tokenindekset i så fald vil kunne udpeges entydigt på tværs af DK-CLARIN, kan en ny, brugerspecifik annotering på tokenniveau nemt linkes til de eksisterende tokens i en tekstsamling 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.

Der kan skelnes mellem to typer for annotation: manuel eller automatisk.

Til den manuelle annotation kræves et annoteringsinterface på portalsitet, hvori brugeren kan fastlægge sine annotationer, og som han kan anvende til at gå tekstmaterialet igennem med og for de enkelte tokens i det sætte de annotationer, han ønsker. Annotationen kan gemmes på portalsitet. Et sådant annoteringsinterface kan evt. bygges oven på en konkordansgrænseflade.

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 skal gemmes lokalt, ville denne løsning måske være interessant. Ellers må man forudse alvorlige ophavsretlige problemer.

[8] En samling tekster, hvor teksterne ligger i en fastlagt struktur bestående af metadata og header, betegnes også *tekstbank*. jf. [Asmussen og Offersgaard, 2008].

[9] Jf. under *token* i [Asmussen og Offersgaard, 2008].

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, hvis ikke det må downloades, kan der muligvis defineres et webservice-agtigt interface, hvorigennem annoteringen kan udføres.

Prioritering

Funktionaliteterne 7–9 kan betragtes som en helhed. Målgruppen for disse funktionaliteter er især sprogteknologer.

At implementere værktøjer til annotering, herunder administration af annoteringer, er en ikke-triviel opgave, og det er tvivlsomt, om den kan udføres som led i det løbende DK-CLARIN-projekt. Opgaven bliver ikke lettere af, at annotering kan ske enten manuelt eller maskinelt.

WP 5 magter ikke selv at udvikle værktøjer til maskinel, ofte lingvistisk, annotering, men bør i stedet koncentrere anstrengelserne om at offentliggøre eksisterende annoteringsautomater hhv. gøre dem til open source, jf. 2.10.

Det anbefales at nedprioritere denne funktionalitet og evt. se bort fra den i en første version af DK-CLARIN's infrastrukturløsning.

2.8 Arkivering af brugerens annotationer

Beskrivelse

Efter at annoteringen (eller en del af den) er udført, skal de brugerspecifikke annotationer kunne gemmes på (eller via) portalsitet.

Er der foretaget en lokal annotering på brugerens computer, skal annoteringsdata være defineret i en annoteringsfil i et fastlagt format, som så kan uploades til DK-CLARIN's portalsite. Der skal også afleveres metadata for annoteringen i et fastlagt format. Når annoteringen uploades til portalsitet, registrerer dette annoteringen som en ny annoteringsresurse og korpusbeskrivelsen lagres sammen med annoteringen. Portalsitet kan oplyse tekstbankværten om eksistensen af den nye annotering. Annoteringen behøver ikke at dække hele korpusset. Den del af korpusset, som ikke er annoteret, kan nul-annoteres, sådan at andre kan se, at brugeren ikke fik behandlet hele korpusset.

Implementering

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 et eksisterende korpus. Annotationen kan derfor betragtes 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.

På portalsitet 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.

Prioritering

Med hensyn til brugermålgruppen for denne funktionalitet, jf. afsnittet *Prioritering* under afsnit 2.7.

Det anbefales at nedprioritere denne funktionalitet og evt. se bort fra den i en første version af DK-CLARIN's infrastrukturløsning.

2.9 Konkordanssøgning med brugerannotationer

Beskrivelse

Brugerskabte annotationer på tokenniveau skal kunne gøres tilgængelige i konkordansværktøjet, og der skal kunne laves konkordanssøgninger på dem på linje med alle andre tilgængelige annotationer på tokenniveau.

Implementering

På baggrund af annoteringsresursen og det korpus, den vedrører, skal man kunne indeksere et nyt korpus, så det bliver søgbart i IMS Open Corpus Workbench, jf. afsnit 2.4.

Prioritering

Med hensyn til brugermålgruppen for denne funktionalitet, jf. afsnittet *Prioritering* under afsnit 2.7.

Det anbefales at nedprioritere denne funktionalitet og evt. se bort fra den i en første version af DK-CLARIN's infrastrukturløsning, ikke mindst fordi implementeringen af denne funktionalitet forudsætter en implementering af funktionalitet 4, jf. 2.4, samt muligheder for, at brugerne kan skabe og publicere deres egne annoteringsresurser.

2.10 Adgang til korpusrelevante tools

Beskrivelse

Brugerne kan via portalsitet få adgang til korpustools, fx tokenizer eller taggere, som er blevet stillet til rådighed af andre partnere, eller som er blevet brugt ved behandlingen af tilgængeligt tekstmateriale i infrastrukturen, og bruge og modificere dem på open source-lignende betingelser.

Implementering

Der bør etableres administrative faciliteter til opbevaring/download/webservice-anvendelse af disse tools incl. deres algoritmer, data og dokumentation.

Prioritering

Brugermålgruppen er primært sprogteknologer. Denne funktionalitet kan derfor ligeledes nedprioriteres. Tilstedeværelsen af denne funktionalitet er imidlertid en forudsætning, hvis implementeringen af funktionaliteterne 7–9 skal give mening og bør derfor prioriteres højere end disse.

Litteratur

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