Kvalitetssikring i DK-CLARIN

Dette dokument samler bidragene fra hver enkelt arbejdspakke, som har beskrevet hvordan de sikrer kvaliteten af deres egne afleveringer. I DK-CLARIN projektet produceres forskellige typer af afleveringer:
Ressourcer: data og værktøjer
Rapporter: specifikationer, dokumentation, statusrapporter

Kvalitetssikring generelt

Det er den enkelte delarbejdspakkes ansvar at kvalitetssikre afleveringer.
Det sker for dokumenter ved at de korrekturlæses mv., for data og software ved testning, validering mv.

Ideelt set bør kvalitetssikringen bygge på anerkendte metoder, fx de metoder til validering af dataressourcer som er udviklet i ELRA, men da projektet ikke råder over de nødvendige økonomiske ressourcer til at valideringen kan udføres i fuldt omfang for alle afleveringer, er der tale om reduceret validering/selvvalidering.

Alle ressourcer og andre afleveringer skal dokumenteres, og dokumentationen skal lægges på DK-CLARIN's interne hjemmeside under Afleveringer på styregruppens del af hjemmesiden.

Kvalitetssikring af de enkelte arbejdspakkers afleveringer

Arbejdspakke 2.1

Arbejdet vil blive udført efter de bedste internationale, praktisk gennemførlige korpuslingvistiske standarder, der skal sikre, at det resulterende korpus' kvalitet fuldt ud svarer til state of the art inden for feltet.
En egentlig formaliseret kvalitetskontrol er ikke forudset i planen, da de forhåndenværende ressourcer er utilstrækkelige. En sådan kontrol kan evt. gennemføres som et særskilt projekt, efter at WP2.1 er afsluttet.

Arbejdspakke 2.2

Kvalitetssikring af de forskellige dele af korpusindsamlingsprocessen er inkorporeret i planlægningen fra starten, og dele heraf står nøjere specificeret i første udkast af specifikationerne for korpus (30-09-08). Kvalitetssikring i denne sammenhæng består af en lang række forskellige procedurer og er en løbende del af korpusindsamlingen og –opmærkningen. Derfor skilles kvalitetssikring ikke ud som en selvstændig arbejdsopgave.

- **Domænetaksonomi:** For de 2 store domæner vil der blive udarbejdet en taksonomi der sammenholdt med et klassifikationssystem vil belyse domænets sammensætning. Således sikres det at de væsentligste aspekter af domænet er dækket. Dette menes ikke at være nødvendigt for de mindre fagområder. Hvad angår det mellemstore domæne, it, er det endnu ikke afklaret om der bør udarbejdes en taksonomi.

- **Kvalitetskontrol af tekstformat/konvertering:**
  Arbejdsgruppen har planlagt en række automatiske/semiautomatiske kvalitetstjek:
  1. ortografisk kontrol vha. frekvenslister

Der genereres komplette frekvenslister for et meget begrænset antal tekster og top X tokens screenes manuelt for usædvanligheder. En effektiv værdi for X fastsættes i forhold til cost-benefit og denne værdi anvendes i et antal stikprøver.
OCR-genkendt tekst vil blive gennemset manuelt for formateringsfejl.

2. om en teksts sprog er dansk vha. sammenligninger med BNC
3. fagsproglighed. Ved at måle henholdsvis læsbarhed, $Kval_{lix}$, og fagsproglighed/termtæthed ($Kval_{fag}$) kan man sikre sig mod tekster der i virkeligheden er almensproglige.

De to sidste kvalitetstjek vil blive udført på alle tekster i korpus, og tekster der falder uden for en grænseværdi vil blive kasseret.

- **Kvalitetskontrol af metadata:** Af de automatisk genererede metadata vil kun en delmængde som især emnekategori og kommunikativ kontekst kræve kvalitetskontrol, hvorimod oplysninger om antal tokens, antal types, teksttype etc. ikke behøver at blive kontrolleret. Kontrollen vil bestå af en manuel skimning af dokumentindholdet og eventuel revision af de automatisk genererede værdier. Antallet af stikprøver kan først specificeres efter at en indledende indsamling og analyse har fundet sted.

- **Kvalitetskontrollen af dokumentrengøring og tokenisering**: Da tokenisering forventes at være en forholdsvis ukompliceret proces, vil der ikke blive brugt tid på at kvalitetskontrollere denne. Hvad angår automatisk udtræk /omformatering af tabelindhold, identifikation af tekster til figurer/grafer, formler, overskrifter etc. er kvalitetskontrol helt afgørende. Der skal kontrolleres mindst én fra hver kilde, og det vil forekomme at enkelte stykker af en tekst der er for vanskelige at rengøre, må fjernes, evt. må hele teksten opgives.

- **Kvalitetskontrol af tekstannotation:** Kvalitetskontrol af den automatiske PoS-tagging og lemmatisering vil foregå ved at der udtrækkes et antal ord fra hver tekst hvorpå præcisionen af lemma og PoS vil blive målt. Kvalitetskontrol af termtagging er meget vanskelig da den kræver indgående domænekendskab og kun kan foretages manuelt. I den afsluttende dokumentation vil det blive understreget at de taggede termer kun er term*kandidater* der kan være behæftet med vis usikkerhed.

Arbejdspakke 2.3

WP2.3. har ud fra erfaringer fra tidligere arbejde med ældre danske tekster afledt en arbejdsgang der indeholder kvalitetskontrol. Denne kontrol består af omhyggelig korrektur af alt arbejde. Korrekturen udføres på flere niveauer i arbejdsprocessen og af flere forskellige medarbejdere. Dermed opfanges evt. fejl og mangler langt bedre end hvis arbejdet fx blev gennemført vha. automatisk annoterende software. Den omhyggelige specifikation af arbejdsprocesserne og protokoller hvor det indføres hvem der har udført hvilke processer hvornår, er et led i kvalitetssikringen.

Arbejdspakke 2.4

Arbejdet vil blive udført efter de bedste internationale, praktisk gennemførlige tekstfilologiske og sprogteknologiske standarder, der skal sikre, at den resulterende tekstsamlings kvalitet fuldt ud svarer til *state of the art* inden for feltet.
En egentlig formaliseret kvalitetskontrol er ikke forudset i planen, da de forhåndenværende ressourcer er utilstrækkelige. En sådan kontrol kan evt. gennemføres som et særskilt projekt, efter at WP2.1 er afsluttet.

Arbejdspakke 2.5

Der foreligger ingen skriftlig redegørelse for den anvendte kvalitetssikringsmetode.

Arbejdspakke 2.6

Det falder ikke inden for projektets resursemæssige rammer at lave *gold standard corpora,* dvs. perfekt opmærkede korpora, der er gennemgået manuelt. Vi vil derimod lave stikprøver og rapportere om fejlfrekvens og fejltyper i de givne stikprøver. På baggrund af resultaterne fra stikprøverne vil justeringer i metoder og værktøjer blive vurderet. For alle korpora vil der blive udarbejdet beskrivelser af hvordan de er processeret, sammen med en kvalitetsvurdering for opmærkning og alignering.

Arbejdspakke 3.1

Data base

With respect to the multimodal database, the project group has identified the following quality issues and ways to control quality for each of them. Transcription will be done in the CLAN tool (developed by the CHILDES and TALKBANK projects) and the data will be available according to the TALKBANK conventions.

Selection of relevant material

Quality control means that the database should achieve a maximal distribution of data in face-to-face and remote interaction (typically phone calls) as well as in the number of speakers. (cf. file on database construction).

Audio and video recording

There is no formalized quality control on the data themselves. If data can be transcribed (i.e. if the speech is intelligible), and if the data are rich in interactional details, they can be part of the corpus.

Transcription

Transcription consistency is a quality issue with will be dealt with in the following way: Transcriptions will most often be produced by one or more transcribers. The resulting raw-transcriptions will be checked subsequently by two independently working researchers.
Issues of timing will be especially in focus.
Prosodic information is notoriously difficult to agree upon and to mark and inconsistencies will be found.
Technical consistency in the transcription will be checked by the CHAT editor on two levels: When data are exported to XML by the CHATTER program, the editor checks for consistency with respect to XML.

Annotation of gestures at CST

Selection of relevant material

In this part of the project a subset of the transcribed videos provided by SDU will be annotated with features concerning non-verbal behavior, according to the MUMIN/CLARIN specifications which have been defined in the first year of the project.
The criteria for the selection of this subset are the following:
it must be possible to see hand and face gestures and body postures of the conversation participants; speech and gesture must interact in the video.

Annotation

The annotation of gestures will be made in the ANVIL tool following the MUMIN/CLARIN specifications which describe the allowed attributes and values for each annotation category. The ANVIL tool ensures that only these attributes and values are used in the annotation process, and that the resulting annotation is a correct XML file with respect to the given MUMIN/CLARIN specifications.
In the MUMIN network the MUMIN annotation was evaluated measuring interannotator agreement in terms of kappa-score (Carletta et al. 1996). The results of this evaluation showed satisfactory results for both the recognition of gestures and the assignment of the various categories (Allwood et al. 2007).
Because only two annotators are working at the transcription of gestures and because of the limited amount of time assigned to this task in the project, only a small part of the annotation will be made by both annotators. We will indicate in a comment the part of the annotation produced by both annotators.

Arbejdspakke 3.2

CLARIN-afleveringen består af 16 enkeltinterviews og fire gruppesamtaler foretaget med studerende på henholdsvis stx og htx. Der foreligger videooptagelser med forskellige antal kameraer, jf. afleveringspapirerne. Der er ikke grund til at specificere kvalitetssikring af videooptagelserne.
For CLARIN-afleveringen er indført særlig stramme kvalitetssikringsprocedurer som involverer følgende skridt:

Udskrivning af optagelserne sker i overensstemmelse med den manual som er udviklet til brug for LANCHARTs data. Manualen sikrer at alle ord der genfindes i Retskrivningsordbogen (RO), staves sådan som de er anført i RO. Udskrivning sker med brug af programmet Transcriber. Alle filer findes således som Transcriber-filer. Der læses korrektur på filerne således at det sikres at alle manualens konventioner overholdes strikt. For CLARIN-afleveringen er der indført en anden og sidste korrektur som sikrer at de løsninger som er overladt til et skøn, følger en og kun en persons skøn. Denne anden korrekturlæsning udføres for samtlige filers vedkommende af Louise Gad, vores mest erfarne korrekturlæser og i øvrigt leder af udskrivningen. I og med at filerne derefter lægges i det korpus som kodes, konverteres de til Praat textgrids. Alle filer findes således også som Praat text grids.

Samtlige filer *kodes for diskurskontekst* i overensstemmelse med manualen for denne analysetype (se Sprogforandringscentrets hjemmeside under 'manualer og rapporter'). Det betyder at det for alle filerne er muligt at se hvilken *samtaletype* der er tale om, hvilke *aktivitetstyper* der indgår, hvilke *makro-talehandlinger* som findes i filen, hvilke *interaktionstyper* der findes, hvilke *genrer* der er manifesteret og hvilke *skift i udsigelse* der kan konstateres, alt sammen på baggrund af udskriften. Al kodning sker af to kodere, hvoraf den første foretager den egentlige kodning, mens den anden kontrollerer overensstemmelsen med kodningsmanualen og retter eventuelle fejl.

For en enkelt fils vedkommende gælder at den kodes for en række grammatiske variable og en række fonetiske variable. Dette sker for at muliggøre en demonstration af hvordan filerne kan anvendes til at belyse en række sprogforandringsprocesser som for øjeblikket manifesterer sig som variation.

Søgemaskinen

Kvalitetssikring af *søgemaskinen*: Disse oplysninger skønnes ikke nødvendige i denne forbindelse før der er taget stilling til integration af dette *tool* i CLARIN.

Arbejdspakke 3.3

Kvalitetssikring af de ressourcer der udvikles af WP3.3. kvalitetssikres ved

1. kollegaevaluering (uvildige, kvalificerede datalingvister på CBS)
2. publikation af resultater i faglige tidsskrifter
3. data valideres og testes efter god, videnskabelig praksis
4. dokumentationen lægges på DK-CLARIN's interne hjemmeside
5. kritisk CLARIN-intern evaluering ved de øvrige WP3x parter

Arbejdspakke 4.1

Kvalitetskontrol/Validering
Opgaven varetages af DSL og vil tage form af stikprøvevis kontrol af 2 % af de kodede synsets, dvs. ca. 1400 synsets. Valideringen indebærer retning af deciderede fejl og identificering af delområder der kræver særlig behandling.

Arbejdspakke 4.2.1

Der foreligger ingen skriftlig redegørelse for den anvendte kvalitetssikringsmetode.

Arbejdspakke 4.2.2

Opgave 5: Kvalitetssikring i form af stikprøvebaseret validering af ca. 2 % af det sammenkædede ordforråd; forkerte links rettes. Dokumentation af validering inkl. eventuelle problemtyper. Allokeret tid: 0,25 (VIP)

Arbejdspakke 5.1

For all tasks indicating software development, such as "implement", "adapt", etc., the activities are assumed to include tests, all relevant documentation, source code repository, bug fixing logs etc. to provide a final product.

Arbejdspakke 5.2

Arbejdet i arbejdspakke 5.2 baseres på anvendelse af anerkendte standarder for metadataformater og dataformater. Anvendelse af standarder og udmøntningen af disse standarder fastlægges i dialog med projektets forskere. Alle delarbejdspakker afleverer testdata som evalueres i forhold til de aftalte specifikationer, og resultaterne for disse evalueringer meddeles leverandørerne. For værktøjsleverandører er der dialog om integrationen, hvor der tages højde for at implementere robuste løsninger. I web-grænsefladearbejdet er der inddraget en følgegruppe, der løbende giver feedback ang. funktionalitet og design. Arbejdspakkens aktiviteter foregår således i dialog med leverandører og brugere. Der implementeres desuden valideringsprocesser for alle ressourcetyper, sådan at alle ressourcer, der deponeres i infrastrukturen, gennemgår en valideringsproces inden deponering. Arbejdet udføres i tæt samarbejde med arbejdspakke 5.1. De juridiske aspekter (aftaler og licenser) håndteres af arbejdspakke 1.