Udviklingsopgave 1.5
Fastlæggelse og dokumentation af korpusformat og beskrivelse tekstflowet under korpusopbygningen

Jørg Asmussen

Version 2.0
7. maj 2008

Generelt om tekstformater

Tekstmaterialet, som DSL's Afdeling for Digitale Ordbøger og Tekstkorpora (DOT) modtager til korpusformål, opbevares i fire forskellige formater:

Leverandørformat – de formater, som DOT modtager materialet i, incl. evt. metadata

Simpelt standardformat – et forenklet, standardiseret format i en simpel XML-struktur med standardiserede metadata i form af en *header*

Ordopdelt standardformat – i princippet det samme som simpelt standardformat, dog er selve teksten her delt op i sætninger (perioder) og ord ('*tokenized*'), hvert i sit eget element i XML-strukturen

Tekstbankformat – ordopdelt standardformat omsat til en MySQL-tabel

Mens leverandørformater kan være vidt forskellige, ligger det simple standardformat fuldstændig fast og er derfor et entydigt udgangspunkt for al videre processering. Konverteringen fra leverandørformat til simpelt standardformat sker vha. specielt udviklede transducere, jf. særskilt dokumentation.

Det ordopdelte standardformat fås ved at processere materialet vha, en *tokenizer*. Tokenizeren forventer, at inputtet er i simpelt standardformat, jf. særskilt dokumentation.

Det ordopdelte standardformat er det format, som kræves for at kunne importere materialet i tekstbanken. Tekstbanken er i princippet en MySQL-database med et særligt interface *TeCoLex*, som er under opbygning, jf. særskilt dokumentation. Al videre processering af tekstmaterialet, fx tagging, parsing, kompilering af søgbare korpora sker i tekstbanken vha. *TeCoLex*-værktøjet.

Tekstmaterialet opbevares fysisk i filkataloget `/DOT/textrepository/`på hosten `ja-korpus.dsl.lan`, hvor det er fordelt på en række underkataloger, hvis navne afspejler leverandører, enkeltleverancer m.m. I hvert underkatalog giver en metafil (hvis

navn indledes med en underscore[1]) nærmere oplysninger om katalogets indhold. Hvert underkatalog findes i op til tre forskellige varianter som skelnes ved forskellige ekstensioner. Således indeholder `leverance\_xyz.src` materialet i leverandørformat, mens `leverance\_xyz.std` indeholder samme materiale konverteret til simpelt standardformat og `leverance\_xyz.tok` i ordopdelt standardformat. Efter at indholdet fra fra et *.tok*-katalog er blevet importeret i tekstbanken og dermed også foreligger i tekstbankformat, skifter det ekstension til *.tok-imported* og der lægges en import-log som en yderligere metafil ned i det.

1 Leverandørformat

Leverandørformatet opbevares af 'arkæologiske' grunde:[2] Det sikrer, at man til enhver tid kan vende tilbage til udgangspunktet, hvor flest mulige af de oprindelige informationer er bevaret. Det kan fx være typografiske informationer i form af forskellige koder i teksten, det kan være en særlig opmærkning af ord eller længere passager i teksten, fx stednavne, personnavne, eller det kan være detaljerede metadata.

2 Simpelt standardformat

2.1 Beskrivelse

I forhold til leverandørformatet vil simpelt standardformat som regel være en forenkling. Repræsentationsformatet er en simpel XML-struktur. Den består af en header, som er beskrevet i dokumentationen til udviklingsopgave 1.7, samt selve teksten: Teksten står mellem taggene `<text id="x">` og `</text>`, teksten er opdelt i typografiske afsnit, omgivet af `<p>` og `</p>`. Overskrifter betragtes også som afsnit. Anden form for opmærkning, end hvad der er nævnt her, er ikke tilladt. Under konverteringen fra leverandørformatet fjernes evt. andre typer opmærkning, fx `<em>` og `<href>` i Infomedia-materialet.[3] Tekster i simpelt standardformat skal være kodet i tegnsættet *ISO Latin 1*.[4] Et eksempel på en tekst i simpelt standardformat, dog uden headeroplysninger, ses i figur 1.

Da tekster i simplet standardformat er de mindste tekstenheder, et korpus normalt kan sammensættes af, er det vigtigt, at de hver især ikke bliver for lange. På denne måde sikres det, at et korpus kan komponeres ret præcist. Længere tekster, typisk tekster i bogform, skal

[1] Filnavne på andre filer end metafiler, dvs. filer med korpustekster, må derfor aldrig indledes med en underscore.

[2] I visse tilfælde er dette oprindelige format ikke længere tilgængeligt, fx for Den Danske Ordbogs korpus. I disse tilfælde er leverandørformatet et allerede konverteret format, der har været brugt til formål forud for *ordnet*-projektet.

[3] I Infomedia-materialet står den korpusrelevante tekst mellem et eller flere `<block>`-tags. Dog skal man være opmærksom på, at overskriften altid betragtes som en del af selve korpusteksten, selvom den står uden for `<block>`-taggene. Derfor skal den hentes fra den relevante del i strukturen og indsættes som et afsnit i starten af hver korpustekst, jf. endvidere korpusheaderdokumentationen.

[4] UTF-8 (e.l.) bruges ikke, da korpussøgeværktøjet CQP, som teksterne i sidste ende skal gøres søgbare i, endnu ikke understøtter det.

```
<textunit>
<header>...</header>
<text id="2000104133" >
<p>Skribent: Cindy</p>
<p>Artiklen er læst 1607 gange</p>
<p>Cheerleaders</p>
<p>Så er der dømt Cheerleaders over hele linien.
Connery.dk har været et smut på Brøndby stadion,
og har mødt pigerne.</p>
<p>Se artiklens galleri</p>
<p>Brøndby Cheerleaders</p>
<p>Rygterne om hvorvidt cheerleaders helst
skal være blondiner og dummere end dumme kan
bestemt ikke bekræftes fra min side. Før en
hjemmekamp på Brøndby stadion mødte jeg en flok
meget veltalende og søde piger.</p>
</text>
</textunit>
```

Figur 1: Eksempel på opmærkning i simpelt standardformat

derfor deles op i mindre enheder, når de omsættes fra leverandørformatet til det simple standardformat. Således omsættes fx en roman, der originalt måske foreligger i én fil, til én fil for hvert kapitel. En tekst (eller rettere en *tekstenhed*) i simpelt standardformat bør normalt ikke være på mere end ca. 50 sider.

2.2 Placering og filnavne

Tekster i simpelt standardformat placeres som tekstfiler i filkataloger med ekstensionen *.std*, som bør ligge parallelt (som søskende)[5] til de tilsvarende kataloger med teksterne i leverandørformat.

En fil med tekst i simpelt standardformat indeholder maksimalt én tekst eller – ved længere tekster – ét tekstudsnit, fx et kapitel. Filnavnene bør svare til dem, teksterne har i leverandørformatet, dog skal de have ekstensionen *.std*. Filnavne bør såvidt muligt være opbygget efter følgende struktur: *kildekode-id-tekstdel*.[6]

Kildekoden er en kode på tre små karakterer fra intervallet *a–z*, der for periodika indikerer avisens eller bladets navn, fx *pol* (Politiken), *ber* (Berlingske Tidende), og for nonperiodika leverandøren, fx navnet på forlaget, fx *gyl* eller *gad*. Der føres en særskilt liste over disse koder i kataloget `/DOT/textrepository/_codes/`.

Tekst-id'et, som er det midterste led i filnavnet, er et ticifret tal. I Korpus 2000-materialet forekommer der to typer tekst-id'er, nemlig ticifrede, der begynder med *1* for materiale fra Po-

[5] Der afviges fra denne parallelitet ved tekster leveret fra Infomedia.

[6] For Infomedia-teksternes vedkommende er det kun teksterne i det simple tekstbankformat, der er opbygget på denne måde. Filnavnene på teksterne i leverandørformat bevares som de er.

litiken og Jyllands-Posten, og sekscifrede for resten af materialet. De sekscifrede opgraderes til ticifrede, der begynder med *2*. DDO-materialet bruger et system med bogstavkoder, som skal omsættes til ticifrede tal, der begynder med *3*. *Ordnet*-relateret materiale har ticifrede id'er, der begynder med *4*, herunder ligger Infomedia-tekster i serien *4900000000 – 4999999999*.

Filnavnets sidste led er en tocifret talkode, der især er relevant for opdelte tekster (fx bøger), og som angiver den pgl. del af teksten. For ikke-opdelte tekster er talkoden *00*, for opdelte ligger den for de enkelte dele mellem *01* og *99*. Ekstensionen *.std* på filnavet oplyser blot, at der er tale om en tekstfil med en en tekst i simpelt standardformat[7]. Et eksempel på et konkret filnavn er `pol-1000105670-00.std`, som er en tekst fra Korpus 2000's materiale.

3 Ordopdelt standardformat

3.1 Beskrivelse

Det ordopdelte standardformat adskiller sig fra det simple standardformat ved at teksterne er delt op i sætninger og ord vha. en *tokenizer*, som er beskrevet i særskilt dokumentation. Det er dette format, der er udgangspunkt for indlæsningen i tekstbanken.

De enkelte sætninger er omgivet af taggene `<s id="x">` og `</s>`. Sætnings-id'et kan enten være tomt eller indeholde en uforanderlig kode. Koden giver mening, hvis der entydigt skal kunne refereres til en bestemt sætning, fx hvis den indgår i forskellige typer specialkorpora, fx træbanker. Tekster fra Korpus 90 og Korpus 2000 har alle faste sætnings-id'er.

En sætning består af en række tokens. Et token er markeret med tagget `<t>`, hvortil der knytter sig følgende attributter:

word indeholder ordet i en normaliseret ortografisk form, dvs. at store bogstaver er omsat til små og diakritiske tegn er fjernet; derudover gælder bl.a. transformationerne *ä→æ*, *ö→ø*[8], *ß→s*, *µ→m* og *þ→t*. Forkortelsespunktummer betragtes som en del af tokenet og bevares i den normaliserede form, mens bindestreger fjernes. Jf. i øvrigt funktionen *normString* i modulet *convert.py*.

space indeholder mellemrummet mellem det aktuelle og det forudgående token samt evt. andre tegn, der er hægtet på begyndelsen af ordet (fx indledende anførselstegn). OBS! Første ord i en sætning vil altid blive indledt af (mindst) et mellemrum, også i et afsnits første sætning.

ortho indeholder ordet i præcis den ortografiske form, som det har i den originale tekst.

punct indeholder interpunktionstegn og andre tegn, som er hægtet på slutningen af ordet. For de fleste tokens vedkommende er dette attribut tomt. OBS! Tokenizeren forsøger at skelne mellem interpunktionstegn og forkortelsespunktummer. Forkortelsespunktummer betragtes som en del af strengen i attributterne *word* og *ortho*. En sætning, der

[7] Kodet i *ISO Latin 1*.

[8] Det er vigtigt, at der i DOT's (webbaserede) søgeværktøjer etableres en mulighed for at indtaste ækvivalenter til bogstaverne *æ*, *ø* og *å*, så disse tegn også kan indtastes på ikke-danske tastaturer.

slutter på en forkortelse med afsluttende forkortelsespunktum (fx *osv.*) har dermed et tomt *punct*-attribut, da punktummet her betragtes som del af ordet.

Figur 2viser et udsnit af et teksteksempel i ordopdelt standardformat (uden header). NB! I eksemplet er der *ikke* taget hensyn til de generelle restriktioner, som gælder for tegnsættet i attributter, jf. appendiks 5.1. De generelle restriktioner gælder for tekster i ordopdelt standardformat og tekstbankformat.

3.2 Placering og filnavne

Tekster i ordopdelt standardformat placeres som tekstfiler i filkataloger med ekstensionen *.tok*, som bør ligge parallelt (som søskende)[9] til de tilsvarende kataloger med teksterne i leverandørformat og simpelt standardformat. Filnavnene svarer derudover til dem, der er beskrevet for det simple standardformat under 2.2. Et eksempel på et konkret filnavn for en tekst i ordopdelt standardformat er `pol-1000105670-00.tok`.

4 Tekstbankformat

En detaljeret beskrivelse af tekstbankformatet gives i dokumentationen om tekstbankens opbygning. Grundlæggende svarer tekstbankformatet til det ordopdelte standardformat omsat til et tabelformat, hvor hvert token svarer til en række og hvert tokenattribut til en kolonne. Der opereres med særlige typer af rækker for de strukturelle tags `<s>`, `</s>`, `<p>` og `</p>`. Headeroplysninger lægges i et antal særlige tabeller. Ud over de tokenattributter, som allerede er beskrevet under 3, er der i teksttabellen afsat felter til følgende ekstra attributter. Felterne får tildelt de rette værdier, dels under indlæsningen af tekster i tekstbanken, dels efter anvendelse af særlige værktøjer (taggere, parsere m.m.).

tix Tokenindeks på det aktuelle token. Tokenindekset er som udgangspunkt en fortløbende nummerering af tokenerne inden for en tekstenhed, dog en, der kun må foretages én gang. Slettes der sidenhen et token, må dette ikke have nogen indflydelse på de øvrige toeknindekser: De beholder deres oprindelige værdi. Tilføjes der sidenhen et token, får det et indeks, som svarer til det hidtil højeste tokenindeks i teksten plus én. Tokenindekset bruges til entydig identifikation af et token, bl.a. i forbindelse med brugerannotering på tokenniveau. Tildeles ved indlæsning af teksten.

ccode Oplyser, hvilke korpora pågældende token (egl. pågældende tekst) er med i. Koden er binært opbygget jf. tabel 1. Oplysningen kan bruges til at indsnævre en søgning til et bestemt korpus eller nogle bestemte korpora. Korpuskoden er egentlig en oplysning på tekstniveau, ikke på tokenniveau, og den bør med tiden afløses af et andet princip. Tildeles/modificeres, når teksten indlemmes i korpus.

lemma Lemmaform i en ortografisk normaliseret form. Der anvendes samme normaliseringsalgoritme, som for attributtet *word*, dog bevares bindestreger i lemmaformen. Tildeles af lemmatizer el. tagger.

[9] Der afviges fra denne parallelitet ved tekster leveret fra Infomedia.

```
<textunit>
<header>...</header>
<text id="2000104133" >
<p>...</p>
<p>...</p>
<p>
<s>
<t word="så" space=" " ortho="Så" punct=""/>
<t word="er" space=" " ortho="er" punct=""/>
<t word="der" space=" " ortho="der" punct=""/>
<t word="dømt" space=" " ortho="dømt" punct=""/>
<t word="cheerleaders" space=" " ortho="Cheerleaders"
punct=""/>
<t word="over" space=" " ortho="over" punct=""/>
<t word="hele" space=" " ortho="hele" punct=""/>
<t word="linien" space=" " ortho="linies" punct="."/>
</s>
<s>
<t word="connerydk" space=" " ortho="Connery.dk" punct=""/>
<t word="har" space=" " ortho="Så" punct=""/>
<t word="været" space=" " ortho="været" punct=""/>
<t word="et" space=" " ortho="et" punct=""/>
<t word="smut" space=" " ortho="smut" punct=""/>
<t word="på" space=" " ortho="på" punct=""/>
<t word="brøndby" space=" " ortho="Brøndby" punct=""/>
<t word="stadion" space=" " ortho="stadion" punct=","/>
<t word="og" space=" " ortho="og" punct=""/>
<t word="har" space=" " ortho="har" punct=""/>
<t word="mødt" space=" " ortho="mødt" punct=""/>
<t word="pigerne" space=" " ortho="Så" punct="."/>
</s>
</p>
<p>...</p>
<p>...</p>
<p>...</p>
</text>
</textunit>
```

Figur 2: Eksempel på opmærkning i ordopdelt standardformat

lemnum Betydningsnummerering, der ved polyseme lemmaformer kan udpege den rette. Bør i sidste instans bruges til at hægte korpus op på DDO og DanNet. Indtil da vil indeks-attributtet typisk have værdien *0*.[10] Tildeles i en særlig kørsel efter lemmatisering og tagging.

pos Tokenets ordklasse. Indtil videre bruges VISL-værdierne. Tildeles af tagger.

morph Tokenets bøjningsoplysninger. Indtil videre bruges VISL-værdierne. CQP-attributnavn *morph*. Tildeles af tagger.

xtag Udvidet opmærkning, der kan vedrøre forskellige forhold, fx specificere semantikken. Indtil videre overtaget fra VISL. CQP-attributnavn *xtag*. Tildeles fx af tagger.

syntax Syntaksoplysninger. Indtil videre bruges VISL-værdierne. CQP-attributnavn *syntax*. Tildeles af parser.

fix Metaoplysninger om taggingen, typisk rettelsesoplysninger, der hvor den oprindelige VISL-tagging blev rettet eller modificeret. CQP-attributnavn *fix*. Tildeles af forskellige værktøjer.

Kode	**Korpus**
1	DK 87-90
2	Den Danske Ordbogs Korpus (DDOC)
4	Parole
8	Korpus 90
16	Korpus 2000

Tabel 1: Anvendte korpuskoder. Er en tekst med i flere korpora, summeres pgl. koder.

En beskrivelse af det anvendte inventar i de forskellige attributter gives i forbindelse med dokumentationen af lemmatiseren og taggeren. jf. også relevante opgavebeskrivelser i *Drejebogen*. For ikke-opmærkede tekster vil en række attributter være udfyldt med en pladsholder *USPEC* (under- eller uspecificeret).

4.1 Placering og filnavne

Tekstbanken foreligger som MySQL-database på hosten `ja-korpus.dsl.lan`. Den tilgås med web-interfacet *TeCoLex* fra samme host. Jf. særskilt dokumentation om tekstbanken.

[10] I KDK-1 er betydningsindekserne fra VISL-opmærkningen flyttet til dette attribut. I VISL-materialet var de hægtet som indeks på selve lemmaformen.

5 Appendikser

5.1 Generelle restriktioner i tegnsættet for alle attributter

Af hensyn til korpussøgeværktøjet CQP gælder der visse tegnsætrestriktioner for tekser, der skal indekseres i CQP-systemet. CQP fordøjer ikke alle tegn ligegodt, fx bliver det tomme tegn, som ofte vil optræde i attributtet *punct*, erstattet af `__UNDEF__` under indekseringen. En skråstreg vil ganske vist blive indekseret rigtigt, men volder problemer i CQP-outputtet, når visning af flere attributter er slået til, idet skråstregen her bruges til markering af attributgrænser. Endvidere er det lidt uvist, hvordan CQP forholder sig til mellemrumstegnet under indekseringen, især hvis det optræder som eneste tegn, fx i attributtet *space*: Ignoreres det? Fortolkes det som kolonneadskiller? Sikkert er, at det, når det er slået til under visningen, kan medføre unhensigstmæssig formatering af CQP's konkordansoutput.

Derfor gælder for alle attributter i ordopdelt standardformat og tekstbankformat ud over det, der er anført under gennemgangen af de enkelte attributter ovenfor, følgende generelle regler for det anvendte tegnsæt:

- Underscoretegn erstattes af tilde. Vil man undgå sammenfald med tilde i det originale korpusmateriale, må man konvertere korpusmaterialet sådan, at det ikke indeholder tilder.[11]
- Mellemrumstegnet erstattes generelt af underscore. Initiale mellemrumstegn i attributter bliver slettet forinden.
- Nummertegnet (el. 'havelågen' `#`) erstattes af to procenttegn (`%%`). Vil man undgå sammenfald med evt. optrædende dobbelte procenttegn i originalmaterialet, bør man konvertere det til noget 'sikkert' forinden.[12]
- Nummertegnet tjener herefter som det tomme attributtegn: Er *punct* fx tomt, indeholder det `#`.
- Skråstreg erstattes af backslash. Vil man undgå sammenfald med evt. allerede forekommende backslashtegn i originalmaterialet, bør man konvertere sig ud af det forinden.[13]

Der er endnu ikke taget endelig stilling til, hvorvidt nogle af disse generelle tegnkonverteringer stilstiende skal foretages i brugerens søgeforespørgsler til CQP, eller om han skal have eksplicit kendskab til disse særtilfælde.

[11] En sådan konvertering er ikke gennemført for KDK-1.
[12] Dette er ikke sket for KDK-1's vedkommende.
[13] Dette er ikke tilfældet for KDK-1.