Udviklingsopgave 1.7
Fastlæggelse og dokumentation af headerstruktur

Jørg Asmussen

Version 1.1
20. februar 2008

Indhold

Generelt

I dette dokument beskrives opbygningen af korpusheadere for korpustekster, som skal bruges til DOT/*ordnet*-formål. Da der allerede blev defineret header-strukturer i forbindelse med opbygningen af Den Danske Ordbogs korpus (DDOC) og Korpus 2000, er det vigtigt, at oplysninger, som er indeholdt i disse headere, i et acceptabelt omfang kan konverteres til den header-struktur, der skal gælde for *ordnet*. Mere om konverteringen af tekstmateriale fra Den Danske Ordbogs korpus, Korpus 2000 og Infomedia kan læses i afsnit 2 på side 10. Det er ikke alle oplysninger, der vil kunne bevares i den her beskrevne headerstruktur. Disse oplysninger

vil dog stadig være knyttet til de oprindelige tekststykkers headere i leverandørformatet.[1] Således kan man i forbindelse med en evt. fremtidig udvidelse af headerstrukturen vende tilbage til disse teksters oprindelige headere og ekstrahere og konvertere yderligere relevant information fra dem.

DOT/*ordnet*-headeren er stærkt inspireret af Korpus 2000-headeren, som igen var stærkt inspireret af DDOC-headeren. Korpus 2000-headeren søgte at fjerne nogle af de uhensigtsmæssigheder, DDOC-headeren havde, først og fremmest en for fin detaljeringsgrad af mulige oplysninger – fx brugte man 131 forskellige genrer. Tilsvarende prøver DOT/*ordnet*-headeren at komme nogle af Korpus 2000-headerens uhensigtsmæssigheder til livs, især ved at eliminere en række oplysningstyper, der i praksis ikke blev brugt/udfyldt alligevel.

En header kan opfattes som en liste over attribut-værdi-par, hvor attributterne er bestemte tekstuelle parametre og værdierne en nøje fastlagt udfyldning af dem. Hver tekstenhed (i hver sin fil) i tekstbank- og korpusformat begynder med en sådan header i XML-format. Attributnavnene (dvs. XML-elementbetegnelserne) og værdimængderne fremgår af tabellerne i denne dokumentation. Ingen attributter må udelades i headerne,[2] og de skal stå i den rækkefølge, som fremgår af tabellerne ndf. Attributterne falder i fire typer, nemlig

- administrative (`adminInfo`)
- tekstinterne (`tekstIntern`)
- teksteksterne om teksten (`omTeksten`)
- teksteksterne om værket (`omVærket`)
- sprogbrugerrelaterede[3] (`sprogbrugerInfo`)

Headerstrukturen er formelt skitseret i figur 1. Strukturen er med vilje holdt så flad som mulig for at sikre en forholdsvis problemfri processering også med simple midler.

```
header ::= adminInfo tekstIntern omTeksten omVærket sprogbrugerInfo+
adminInfo ::= (attribut værdi)+
tekstIntern ::= (attribut værdi)+
omTeksten ::= (attribut værdi)+
omVærket ::= (attribut værdi)+
sprogbrugerInfo ::= (attribut værdi)+
```

Figur 1: Formel beskrivelse af headerstruktur

DOT/*ordnet*-headeren svarer både med hensyn til attributter, attributnavne og udfyldning med få undtagelser til den obligatoriske del af Korpus 2000-headeren. I forbindelse med Korpus 2000-projektet blev der oprettet en header-database *etb*, der bruger den fulde Korpus 2000-header, selvom den fakultative del af oplysningerne (nok) aldrig er blevet udfyldt. Dokumentation af den fulde Korpus 2000-header findes under `univers/3_korpus/01.007_headerstruktur/k2000-baggrund/`.

[1] For DDOC's vedkommende betragtes dets oprindelige korpusformat som leverandørformatet (jf. dokumentationen om korpusformater), så den oprindelige header er bevaret i dette format.

[2] Dog gælder det, at den del af headeren, hvor sprogbrugeroplysningerne beskrives, gentages for hver enkelt sprogbruger, der har medvirket ved tekstfrembringelsen, jf. afsnit 1.5 på side 10.

[3] Parallelt med *sprogbruger* anvendes betegnelserne *afsender* og *tekstproducent*.

1 Headerens enkelte bestanddele

Headeren består af én del med administrative, tre dele med tekstrelaterede og en med sprogbrugerrelaterede oplysninger. De enkelte dele fremgår ikke af selve XML-strukturen, der som nævnt bevidst er holdt så flad som mulig, men af elementbetegnelserne, der alle er præfigeret med et bogstav, der indikerer oplysningsgruppen. Udfyldningen af de enkelte oplysningstyper sker så vidt muligt med talværdier, der i en egentlig visning kan omsættes til de tekststrenge, man ønsker i en given kontekst.

Default og udefineret

Generelt gælder der for udfyldningen, at det for oplysningstyper, der skal udfyldes med en talværdi, er *0* – *nul* –, som er default-værdien, mens *-1* angiver, at pågældende værdi er udefineret. For oplysningstyper, der udfyldes med tekst, er default - (*bindestreg* med betydningen 'findes ikke', 'er ikke relevant'), mens den udefinerede værdi er *?*. Disse konventioner vil medføre en række konverteringer af udfyldningerne i Korpus 2000-materialet, jf. afsnit 2 på side 10.

1.1 Administrative oplysninger (`adminInfo`)

Tabel 1 på side 4 lister obligatoriske administrative oplysninger og beskriver dem og deres udfyldningsmuligheder.

1.2 Tekstinterne oplysninger (`tekstIntern`)

De oplysninger, som direkte kan udledes af selve teksten, kaldes tekstinterne. De forskellige oplysninger, som skal registreres i headeren, fremgår af tabel 2 ndf.

Kommentarer

<iGenre>: Det er problematisk at definere en værdimængde for oplysningstypen *genre*. I Den Danske Ordbogs Korpus brugte man 131 forskellige genrer, der i praksis ikke lod sig administrere. I Korpus 2000 havnede man – efter en del diskussion – i den modsatte grøft med en (lidt uortodoks) genreklassifikation med kun tre værdier, nemlig *ikke-skønlitteratur*, *skønlitteratur* og *privat*. I DOT/*ordnet*-sammenhæng er denne klassifikation udvidet igen med inspiration fra wikipedia[4], som har et ganske fornuftigt bud på en nogenlunde differentieret, men alligevel administrerbar klassifikation. Overordnet skelnes stadig mellem fiktive og ikke-fiktive tekster, og det er muligt kun at nøjes med en angivelse af det uden videre differentiering.[5] Hvis der kan differentieres yderligere, bruges følgende klassifikation:

Orienterende: lærebog, nyhedsartikel, afhandling, monografi, biografi, selvbiografi, nekrolog, rejseskildring, reportage osv.

Vurderende, debatterende eller argumenterende: essay, partiprogram, kronik, anmeldelse, pamflet, partsindlæg, politisk tale osv.

Påvirkende eller adfærdsregulerende: lov, bekendtgørelse, regulativ, aftale, reklame, annonce, brugsanvisning, girokort osv.

Lyrisk: salme, digt, lejlighedsdigt, sang, vise osv.

Episk: roman, novelle, eventyr, saga, børnebog, ungdomsroman osv.

Dramatisk: tragedie, komedie, skuespil, parodi, revy osv.

[4] Jf. `http://da.wikipedia.org/w/index.php?title=Genre_%28litteratur%29&oldid=1360185`.
[5] Det vil under konverteringen fx være tilfældet for K2000-tekster samt for Infomedia-materialet.

Navn	**Beskrivelse**	**Udfyldning**
<aTextid>	Tekstenhedens id	Værdimængden for tekstid'er er fastlagt i dokumentationen om korpusformater
<aFile>	Tekstens filnavn	Det filnavn, der gælder for teksten i tekstbankformatet; oplyses uden ekstension. Jf. i øvrigt dokumentation om korpusformater
<aSrcfile>	Kildefilens navn	Navnet på den fil, som teksten oprindelig blev leveret i; oplyses *med* ekstension. Kun selve filnavnet anføres, ikke stien, som vil kunne erueres ud fra de øvrige headeroplysninger
<aCollec>	Korpussamlingen, som kan være projektorienteret eller materialeorienteret	Tilladte talværdier: 0: løbende DOT/*ordnet*-indsamling 1: løbende via Infomedia 2: DDO, talesprog 3: DDO, skriftsprog 4: Korpus 2000
<aAcqy>	Indsamlingsår for pågældende tekstenhed	Tilladte talværdier er firecifrede årstal og *0*, hvis det nøjagtige årstal ikke kendes.
<aSuppl>	Tekstleverandørens id	Koder i form af talværdier er fastlagt i dokumentationen under `univers/3_korpus/01.007_headerstruktur/dokumentation/leverandoeroplysninger/`. Registrerede leverandøroplysninger findes i tabellen *leverandoer* i databasen *leverandoer*
<aRestr>	Restriktioner på brugen af teksten	Tilladte talværdier: 0: kun udsnit ('citater') må vises 1: hele teksten må vises
<aAno>	Anonymiseringer i forbindelse med visning	Tilladte talværdier: 0: anonymisering ikke nødvendig 1: personnavne anonymiseres 2: stednavne anonymiseres 4: forfatternavn anonymiseres 8: teksttitel anonymiseres Kombinationer af anonymiseringsmuligheder udtrykkes ved at lægge pgl. koder sammen, fx 3: person- og stednavne anonymiseres

Tabel 1: Administrative headeroplysninger

Navn	Beskrivelse	Udfyldning
<iTxttit>	Tekstens titel	Angives fuldt ud, hvis teksten har en titel; hvis den ikke har, indsættes en bindestreg
<iGenre>	Tekstens genregruppe. Under hver af de nævnte genregrupper henregnes flere genrer, jf. kommentar til tabellen	Overordnet skelnes mellem fiktive, ikke-fiktive og genremæssigt ubestemmelige tekster: 0: ubestemmelig 1: fiktiv 2: ikke-fiktiv I de tilfælde, hvor genren kan specificeres nærmere, kan følgende talværdier komme til anvendelse for fiktive tekster: 11: lyrisk 12: episk 13: dramatisk 14: diverse Og følgende talværdier kan komme til anvendelse for ikke-fiktive tekster: 21: orienterende 22: vurderende etc. 23: påvirkende etc.
<iDomain>	Fagområdet, som teksten vedrører	Når *Drejebogens* opgave 1.8 om automatisk domænetilordning er løst, bestemmes værdierne automatisk. Der vil blive anvendt samme værdimængde som for DDO's korpus, enten den fulde med 66 værdier eller den reducerede *emnegruppe* med 12, afhængig af hvilken af dem der giver de bedste tilordningsresultater. Indtil da vil værdien være udefineret *-1*
<iTokens>	Antal løbende ord, som tekststykket består af	Beregnes automatisk, når tekstbankversionen af en tekst (og dermed headeren) oprettes
<iTypes>	Antal *forskellige* ord, som tekststykket består af	Beregnes automatisk sammen med beregningen af *token*-værdien

Tabel 2: Tekstinterne headeroplysninger

Diverse litteræreː anekdote, vittighed, gåde osv.

Yderligere oplysninger fremgår af tabel 2 på side 5.

1.3 Teksteksterne oplysninger (`omTeksten`)

Oplysninger, som ikke umiddelbart kan udledes af selve teksten (derfor tekst*eksterne* oplysninger), vises i tabel 3 på side 7.

Navn	Beskrivelse	Udfyldning
<ePrody>	Tekstens produktionsår	Udfyldes med et firecifret årstal. Hvis produktionsåret ikke er kendt, oplyses et skøn og <eProdyc> sættes til *1*
<eProdyc>	Sikkerhed vedr. produktionsåret	Tilladte talværdier: 0: sikker 1: usikker
<eExpr>	Udtryksmediet (skrift, tale el. mellemstadier)	Tilladte talværdier: 0: skrift 1: tale 2: papirtale 3: talepapir
<eAspect>	Kommunikativt aspekt	Tilladte talværdier: 0: reception ('professionel') 1: produktion ('ikke-professionel')
<eAgerel>	Aldersrelation mellem afsender og modtager	Tilladte talværdier: 0: ukendt 1: voksen–voksen 2: voksen–barn 3: voksen–ung 4: voksen–ældre 5: barn–voksen 6: barn–barn 7: barn–ung 8: barn–ældre 9: ung–voksen 10: ung–barn 11: ung–ung 12: ung–ældre 13: ældre–voksen 14: ældre–barn 15: ældre–ung 16: ældre–ældre *barn* er i alderen 6–14 *ung* er i alderen 15–25 *voksen* er i alderen 26–60 *ældre* er i alderen 61–

Fortsættes på næste side

Tabellen fortsat fra foregående side

Navn	Beskrivelse	Udfyldning
<eMedium>	Primært el. oprindeligt udgivelsesmedium	Tilladte talværdier: 0: ukendt 1: avis 2: blad/ugeblad 3: bog 4: internet 5: småtryk 6: tidsskrift/fagblad Der kan etableres en yderligere differentiering, hvis der er behov for det. Den laves i lighed med oplysningstypen <iGenre> ved at de givne grundkategorier underinddeles, fx *41* for mail, *42* for blog etc.
<eLang>	Oplysning om tekstens sprog. Det vil som regel sige dansk. Men med denne oplysning er der åbnet mulighed for også at samle tekster på andre sprog, fx som led i etableringen af parallelkorpora	Tilladte talværdier: 0: dansk 1: engelsk 2: tysk 3: fransk 4: spansk Der kan tænkes en yderligere uddifferentiering i sprogtrin eller dialekter efter de samme principper, som gør sig gældende for oplysningstyperne <iGenre> og <eMedium>
<eOlang>	Oplysning om tekstens originalsprog. Til markering af oversatte tekster	Tilladte talværdier: 0: dansk 1: engelsk 2: tysk 3: fransk 4: spansk
<eLoc>	Lokalisering af teksten. Hvis teksten er del af et værk, oplyses hvor i værket teksten er at finde	Udfyldes afhængig af værkets type på flg. måde (sidetal oplyses altid som den side, hvor teksten eller tekstudsnittet starter): For aviser: *3:11* betyder *3. sektion, side 11* For blade: *3:11* betyder *3. nummer, side 11* For bøger: *311* betyder *side 311* For internet: hele URL'en oplyses For småtryk: *2-3* betyder side *2–3* For tidsskrifter gælder det samme som for blade Hvis der ikke kan gives nogen lokaliseringsoplysning, indsættes bindestreg

Fortsættes på næste side

Tabellen fortsat fra foregående side

Navn	Beskrivelse	Udfyldning
<eUrl>	Tekstens (eller tekstleverandørens) URL. Kan i søgeværktøjer siden bruges som link	Hvis teksten findes på www, anføres tekstens URL, ellers anføres en URL, der kommer så tæt på teksten som muligt, fx indgang i en bogdatabase på et forlag, en nyhedstjeneste (fx Infomedia) e.l. Hvis der ikke findes nogen URL, indsættes bindestreg

Tabel 3: Teksteksterne headeroplysninger om teksten

Kommentarer

<eAgerel>: Denne oplysning vil i over 99% af tilfældene have værdien *0* (*voksen-voksen)*; den vil derfor aldrig være egnet til distributionsundersøgelser. Imidlertid kan den med tiden være interessant i forbindelse med udtræk af bestemte subkorpora. Derfor bibeholdes den, selvom den kan være vanskelig at administrere.

1.4 Værkoplysninger (`omVærket`)

Denne gruppe af oplysninger vedrører det evt. værk, teksten stammer fra. Værket kan være en avis, en antologi eller en anden type tekstsamling. Headeroplysningerne i denne gruppe fremgår af tabel 4 på side 9.

Navn	Beskrivelse	Udfyldning
<vTit>	Værktitel, fx en avis' eller en antologis navn eller titel	Skrives fuldt ud, fx *Berlingske Tidende*. Der bør føres en fortegnelse over skrivekonventioner. Hvis der ikke er nogen værktitel, indsættes bindestreg
<vPubly>	Værkets udgivelsesår	Oplyses som et firecifret tal, fx *2007*. Hvis udgivelsesåret ikke kendes nøjagtigt, oplyses et omtrentligt udgivelsesår og <vPublyc> sættes til *1*
<vPublyc>	Sikkerhed vedr. udgivelsesåret	Tilladte talværdier: 0: sikker 1: usikker
<vPublm>	Værkets udgivelsesmåned; kun relevant for aviser og blade	Skrives som et tal mellem *1* og *12*. Hvis irrelevant, udfyldes med *0*
<vPubld>	Værkets udgivelsesdato; kun relevant for aviser og blade	Skrives som et tal mellem *1* og *31*. Hvis irrelevant, udfyldes med *0*

Tabel 4: Headeroplysninger om værket

1.5 Sprogbrugeroplysninger (`sprogbrugerInfo`)

Da en tekst kan være produktet af flere sprogbrugere, oprettes der i headeren om muligt en sprogbrugerdel for hver sprogbruger. I disse tilfælde sættes <uType> for hver sprogbruger til *sprogbrugerkollektiv*. Da korpussøgesystemer og tekstbanksystemer ikke nødvendigvis kan håndtere tekster med flere afsendere, bør man altid sørge for at give oplysningerne for den formodede 'primære' sprogbruger først, som så går ind som 'stedfortræder' for alle de øvrige. Kan der ikke fastlægges en 'primær' sprogbruger (fx ved wikipedia-tekster e.l.), sættes sprogbrugerens <uName1> og <uName2> til - (bindestreg) og <uType> sættes til *1* (sprogbrugerkollektiv). For oversatte tekster gælder, at oversætteren betragtes som primær sprogbruger. Sprogbrugeroplysninger fremgår af tabel 5 på side 11.

2 Konvertering af headere fra eksisterende materiale

2.1 Konvertering fra Infomedia

Infomediateksterne i det originale leverandørformat findes under `/home/filebase/dot/korpus/tekstmateriale/leverandoerformat/ordnet-indsamling/ej_i_tekstbanken/infomedia/`. De konverterede tekster (med headere, i tekstbankformat) lægges i `/home/filebase/dot/korpus/tekstmateriale/teksbankformat/[årstal]/periodika/`, mens de originale tekster i leverandørformatet efter konverteringen flyttes til `/home/filebase/dot/korpus/tekstmateriale/leverandoerformat/ordnet-indsamling/i_tekstbanken/` efter de regler, der er beskrevet i dokumentationen om korpusformat og tekstflowet.

Et eksempel på hvordan en Infomedia-fil er opbygget, ses i bilaget på side 13.

2.1.1 Administrative oplysninger

Udfyldningen af headeren med administrative oplysninger er beskrevet i tabel 6 på side 12.

2.1.2 Tekstinterne oplysninger

Udfyldningen af headeren med tekstinterne oplysninger er beskrevet i tabel 7 på side 13.

2.1.3 Teksteksterne oplysninger

Udfyldningen af headeren med teksteksterne oplysninger er beskrevet i tabel 8 på side 14.

2.1.4 Værkoplysninger

Udfyldningen af headeren med værkoplysninger er beskrevet i tabel 9 på side 15.

2.1.5 Sprogbrugeroplysninger

Udfyldningen af headeren med sprogbrugeroplysninger er beskrevet i tabel 10 på side 15. Det er uvist, hvordan tekster med flere sprogbrugere er mærket op i Infomedia. Dette bør der tages hensyn til under konverteringen: Når/hvis der dukker en tekst op med flere forfattere, opstilles der med baggrund i denne tekst en regel for, hvordan flere forfattere skal håndteres under header-konverteringen.

Navn	Beskrivelse	Udfyldning
<uName1>	Sprogbrugerens fornavn(e)	Angives så nøjagtigt som muligt. For mere kendte forfattere bør der føres en fortegnelse over skrivekonventioner. OBS! Hvis <uType> er et afsenderkollektiv sættes <uName1> og <uName2> til - (bindestreg)
<uName2>	Sprogbrugerens efternavn	Kun ét efternavn er tilladt. For mere kendte forfattere bør der føres en fortegnelse over skrivekonventioner. OBS! Hvis <uType> er et afsenderkollektiv sættes <uName1> og <uName2> til - (bindestreg)
<uType>	Sprogbrugertype. Bruges til at skelne mellem individuelle og kollektive afsendere, som man har, hvis teksten er produceret af flere ligeværdige afsendere (fx wikipedia-tekster).	Tilladte talværdier: 0: individuel afsender 1: afsenderkollektiv OBS! Ved et afsenderkollektiv sættes <uName1> og <uName2> til - (bindestreg) og <uSex> til *0*
<uRole>	Sprogbrugerens rolle under tekstproduktionen. Bruges til at skelne mellem tekstens evt. oversætter og tekstens oprindelige afsender	Tilladte talværdier: 0: ophavsmand 1: oversætter
<uSex>	Sprogbrugerens køn	Tilladte talværdier: 0: blandet/usikker 1: mand 2: kvinde OBS! Hvis teksten er produceret af et afsenderkollektiv (af formodentligt blandet køn) sættes <uSex>-værdien til *0*
<uBorny>	Sprogbrugerens fødselsår	Oplyses som et firecifret tal, fx *1963*
<uBornyc>	Sikkerhed vedr. fødselsåret	Tilladte talværdier: 0: sikker 1: usikker

Tabel 5: Headeroplysninger om sprogbrugeren

Navn	Beskrivelse af udfyldningen med Infomedia-data
<aTextid>	Tekst-id'er består af ticifrede tal. Alle Infomedia-tekster skal begynde med cifrene *49*, de næste to cifre er leveringsårstallet, altså fx *05*, *06* eller *07*; de sidste seks cifre er en løbende nummerering af teksterne inden for pgl. leveringsår; tællingen begynder med *000000* og ender med *999999*. Den anden tekst fra 2005-leveringen har således id'et *4905000001*
<aFile>	Filnavnet konstrueres som beskrevet i dokumentationen om korpusformat og tekstflow. De tre første bogstaver i filnavnet tages fra de tre første bogstaver i attributet *FormalName* i elementet `/NewsML/NewsItem/NewsComponent/AdministrativeMetadata/Source/Party` i Infomedia-filen. Samtidig registreres denne bogstavskode og hele kildenavnet i `/home/filebase/dot/korpus/tekstmateriale/tekstbankformat/_filnavnekoder/koder.txt` i det format, der er beskrevet i *_koder.txt* sammesteds. OBS! Under opbygningen af denne kodeliste skal der sikres et 1-til-1-forhold mellem koder og navne! Det midterste led af filnavnet er tekst-id'et som beskrevet ovf., mens sidste led er *00*. For ovennævnte tekst kunne udfyldningen således være *pol-4905000001-00*, hvis teksten stammede fra *Politiken*. Filnavnet oplyses i headeren uden ekstension, mens den konkrete fil har ekstensionen *.tbt*
<aSrcfl>	Navnet tages fra den oprindelige InfoMedia-fil, sådan som den hed, da den blev leveret. Udfyldes incl. *xml*-ekstensionen
<aCollec>	Udfyldes generelt med talværdien *1*
<aAcqy>	Udfyldes med det årstal, der er navnet på mappen *[årstal]* i `/home/filebase/dot/korpus/tekstmateriale/leverandoerformat/ordnet-indsamling/ej_i_tekstbanken/infomedia/[årstal]/...`, hvorfra den respektive konverteringstekst stammer
<aSuppl>	Koderne oplyses som talværdier, der kan slås op i `univers/3_korpus/01.007_headerstruktur/dokumentation/leverandoeroplysninger/`, som bør koordineres med listen over filnavnekoder ovf. For *Politiken* er leverandørkoden eksempelvis *11002*
<aRestr>	Udfyldes generelt med talværdien *0*
<aAno>	Udfyldes generelt med talværdien *0*

Tabel 6: Administrative oplysninger fra Infomedia

Navn	Beskrivelse af udfyldningen med Infomedia-data
<iTxttit>	Titlen hentes fra elementet `/NewsML/NewsItem/NewsComponent/NewsComponent/NewsLines/HeadLine` i Infomedia-filen; mangler dette element, eller er det tomt, indsættes bindestreg (-) i stedet for. OBS! Hvis overskriften ikke er oplyst som en del af selve teksten, dvs. står et sted i *hedline*-elementet, typisk som *hl1*, skal den under tekstkonverteringen kopieres hertil, idet en overskrift altid betragtes som en del af selve teksten også
<iGenre>	Infomedia-materialet vil indeholde både orienterende, vurderende og påvirkende tekster. Da det (p.t.) ikke er muligt at skelne mellem disse genrer i materialet på en ikke-manuel måde, markeres alle groft som *ikke-fiktive* tekster, dvs. header-elementet udfyldes altid med talværdien *2*
<iDomain>	Udfyldes generelt med talværdien *-1*
<iTokens>	Alle tokens i den konverterede tekst tælles, og det samlede antal oplyses her
<iTypes>	Alle types i den konverterede tekst tælles, og det samlede antal oplyses her

Tabel 7: Tekstinterne oplysninger fra Infomedia

2.1.6 Eksempel

Som bilag ses på side 13 en Infomedia-fil, hvor selve teksten er fjernet, men hvor samtlige metaoplysninger er bibeholdt. Oplysningerne i denne fil omsættes til den DOT/ordnet-header, som er gengivet som bilag på side 18.

2.2 Konvertering fra Korpus 2000

Her beskrives kun konverteringen af K2000-headerne. K2000-teksterne skal også tilpasses DOT/*ordnet*-formatet, jf dokumentationen af korpusformatet (udviklingsopgave 1.5) samt beskrivelsen af postprocesseringsopgaver (udviklingsopgave 1.22).

2.2.1 Administrative oplysninger

2.2.2 Tekstinterne oplysninger

2.2.3 Teksteksterne oplysninger

2.2.4 Værkoplysninger

2.2.5 Sprogbrugeroplysninger

2.2.6 Eksempel

Bilag

Bilag 1: Metaoplysninger i Infomedia-fil

```
<NewsML Duid="e02f1b9a">
  <NewsEnvelope>
    <DateAndTime>2004-12-26T03:55:12</DateAndTime>
    <NewsService FormalName="Infomedia" />
    <NewsProduct FormalName="Avisdatabasen Mediearkivet" />
    <Priority FormalName="" />
```

Navn	Beskrivelse af udfyldningen med Infomedia-data
<ePrody>	Udfyldes med de fire første cifre i elementet `/NewsML/NewsItem/Identification/NewsIdentifier/DateId` i Infomedia-filerne
<eProdyc>	Udfyldes generelt med talværdien *0*
<eExpr>	Udfyldes generelt med talværdien *0*
<eAspect>	Udfyldes generelt med talværdien *0*
<eAgerel>	Udfyldes generelt med talværdien *1*
<eMedium>	Listerne over filnavnekoder og leverandørkoder, som der referes til i tabel 6, og som bør samles til én liste, bør udvides med en oplysning om mediet. For Infomedia-materialet vil indtil videre to medier være sandsynlige, nemlig overvejende *avis* og måske også *blad/ugeblad*. Headeroplysningen udfyldes afhængig af, hvilket medie den identificerede kilde tilhører, enten med talværdien *1* eller *2*. OBS! Det er ikke sikkert, om der p.t. overhovedet leveres andet end avismateriale
<eLang>	Udfyldes generelt med talværdien *0*. Frasorterede tysksprogede Flensborg Avis-tekster får her værdien *2*
<eOlang>	Udfyldes generelt med talværdien *0*. Frasorterede tysksprogede Flensborg Avis-tekster får her værdien *2*
<eLoc>	Hvis mediet er *avis* udfyldes på følgende måde: Oplysning om både sektionen og sidetallet hentes fra elementer med stien `/NewsML/NewsItem/NewsComponent/Metadata/Property`. Sektionsnumret står i attributet *Value* i en sådan *Property*, hvor attributet *FormalName* er *PSEK*, mens sidetallet står som *Value*-attributet i et søskende-*Property*-element, hvis *FormalName*-attribut er sat til *PSID*. Således kan en konkret udfyldning komme til at se således ud: *1:7*. Hvis mediet er *blad/ugeblad*, er det tænkeligt, at den relevante information skal hentes et andet sted fra Infomedia-filen. Det er dog ikke sikkert, at der aktuelt leveres andet end avismateriale fra Infomedia. Hvis den rette værdi ikke kan udledes af materialet, indsættes *-1*
<eUrl>	Udfyldes generelt med `infomedia.dk`

Tabel 8: Teksteksterne oplysninger fra Infomedia

Navn	Beskrivelse af udfyldningen med Infomedia-data
<vTit>	Værktitlen er for avisers og blades vedkommende avisens eller bladets navn. Oplysningen hentes fra elementet `/NewsML/NewsItem/NewsComponent/AdministrativeMetadata/Source/Party` i Infomedia-filerne
<vPubly>	Udfyldes med de fire første cifre i elementet `/NewsML/NewsItem/Identification/NewsIdentifier/DateId` i Infomedia-filerne
<vPublyc>	Udfyldes generelt med talværdien *0*
<vPublm>	Udfyldes med 5. og 6. ciffer i elementet `/NewsML/NewsItem/Identification/NewsIdentifier/DateId` i Infomedia-filerne. OBS! Hvis 5. ciffer er *0*, udfyldes kun med 6. ciffer
<vPubld>	Udfyldes med 7. og 8. ciffer i elementet `/NewsML/NewsItem/Identification/NewsIdentifier/DateId` i Infomedia-filerne. OBS! Hvis 7. ciffer er *0*, udfyldes kun med 8. ciffer

Tabel 9: Værkoplysninger fra Infomedia

Navn	Beskrivelse af udfyldningen med Infomedia-data
<uName1>	Udfyldes med alle tokens (ord) undtagen det sidste i attributet *FormalName* i elementet `/NewsML/NewsItem/NewsComponent/AdministrativeMetadata/Creator/Party` i Infomedia-filen. Hvis navnet ikke kan bestemmes ud fra materialet, indsættes *?*.
<uName2>	Udfyldes med det sidste token i attributet *FormalName* i elementet `/NewsML/NewsItem/NewsComponent/AdministrativeMetadata/Creator/Party` i Infomedia-filen. Hvis navnet ikke kan bestemmes ud fra materialet, indsættes *?*
<uType>	Udfyldes generelt med talværdien *0*
<uRole>	Der skal rettes opmærksomhed mod en evt. særlig opmærkning af oversatte tekster i Infomedia-materialet. Som udgangspunkt sættes dette header-attribut dog til talværdien *0*
<uSex>	Udfyldes indtil videre med talværdien *-1*
<uBorny>	Sættes til talværdien *-1* (dvs. 'udefineret')
<uBornyc>	Sættes til talværdien *-1* (dvs. 'udefineret')

Tabel 10: Sprogbrugeroplysninger fra Infomedia

```
</NewsEnvelope>
<NewsItem>
  <Identification>
    <NewsIdentifier>
      <ProviderId></ProviderId>
      <DateId>20041226</DateId>
      <NewsItemId>e02f1b9a</NewsItemId>
      <RevisionId Update="N" PreviousRevision="0">1</RevisionId>
      <PublicIdentifier>urn:newsml:infomedia.dk:
                        20041226:e02f1b9a</PublicIdentifier>
    </NewsIdentifier>
  </Identification>
  <NewsManagement>
    <NewsItemType FormalName="News" />
    <FirstCreated>355 AM</FirstCreated>
    <ThisRevisionCreated>355 AM</ThisRevisionCreated>
    <Status FormalName="final" />
  </NewsManagement>
  <NewsComponent>
    <TopicSet FormalName="Infomedia Topics">
      <Topic Duid="ts_thes_emballage" Details="Auto">
        <TopicType FormalName="THES" />
        <Description>EMBALLAGE</Description>
      </Topic>
      <Topic Duid="ts_thes_netlist" Details="Auto">
        <TopicType FormalName="THES_NETLIST" />
        <Description>EMBALLAGE</Description>
      </Topic>
    </TopicSet>
    <AdministrativeMetadata>
      <FileName>e02f1b9a.xml</FileName>
      <Provider>
        <Party FormalName="Jyllands-Posten" />
      </Provider>
      <Creator>
        <Party FormalName="Tea Krogh Sørensen" />
      </Creator>
      <Source>
        <Party FormalName="Jyllands-Posten" />
      </Source>
    </AdministrativeMetadata>
    <RightsMetadata>
      <UsageRights>
        <UsageType>0</UsageType>
      </UsageRights>
    </RightsMetadata>
    <Metadata>
      <MetadataType FormalName="InfomediaMetadata" />
      <Property FormalName="KKOD" Value="JYP" />
```

```
  <Property FormalName="PSEK" Value="1" />
  <Property FormalName="TYPE" Value="" />
  <Property FormalName="PSID" Value="7" />
  <Property FormalName="PIND" Value="JP..jp.Indland7.byttegave-regler.
                                     ART.1103990200" />
  <Property FormalName="PSNA" Value="" />
  <Property FormalName="OFIL" Value="\\10.45.16.200\y\
                                     Original-Repository\JYP\2004\12\26\
                                     JYP_35516.txt.035501" />
  <Property FormalName="PUBX" Value="JYP" />
  <Property FormalName="PUGE" Value="Søndag" />
  <Property FormalName="STATUS" Value="MA" />
  <Property FormalName="PMOD" Value="355 AM" />
  <Property FormalName="FOED" Value="2004-12-26T03:55:12" />
  <Property FormalName="PAFS" Value="" />
  <Property FormalName="UDGA" Value="" />
</Metadata>
<NewsComponent>
  <NewsLines>
    <HeadLine>Julegaver: Husk byttemærket og emballagen</HeadLine>
    <SubHeadLine>Julegaver: Husk byttemærket og emballagen</SubHeadLine>
    <DateLine>20041226</DateLine>
    <ByLine>Tea Krogh Sørensen</ByLine>
  </NewsLines>
  <ContentItem>
    <Characteristics>
      <SizeInBytes>365</SizeInBytes>
    </Characteristics>
    <DataContent>
      <nitf>
        <head>
          <revision-history comment="Formated article"
                            name="Camilla.Balling"
                            function="writer-author"
                            norm="20041226T095111+01:00" />
        </head>
        <body>
          <body.head>
            <hedline>
              <hl1></hl1>
              <hl2>Hjælp med gode råd og hård jura før det store
                   bytteræs sætter ind.</hl2>
            </hedline>
          </body.head>
          <body.content>
            <block>
            <p id="p1"> [SELVE TEKSTEN]</p>
            </block>
            <media media-type="PDF">
```

```
                    <media-reference>
                      JP..jp_26-12-2004_Falles_1_1_7_Indland7.pdf
                    </media-reference>
                    <media-caption></media-caption>
                  </media>
                  <media media-type="image">
                    <media-reference></media-reference>
                    <media-caption></media-caption>
                  </media>
                </body.content>
              </body>
            </nitf>
          </DataContent>
        </ContentItem>
      </NewsComponent>
    </NewsComponent>
  </NewsItem>
</NewsML>
```

Bilag 2: Infomedia-oplysninger konverteret til DOT/*ordnet*-header

OBS! Udfyldninger, som er markeret med en afsluttende asterisk, tjener blot til illustration af en *mulig* udfyldning. Efter en faktisk konvertering vil de sandsynligvis være lidt anderledes. Headereksemplet ses på den følgende side.

<aTextid>4905000017*</aTextid>
<aFile>jyl-4905000017-00*</aFile>
<aSrcfl>e02f1b9a.xml*</aSrcfl>
<aCollec>1</aCollec>
<aAcqy>2005</aAcqy>
<aSuppl>11051</aSuppl>
<aRestr>0</aRestr>
<aAno>0</aAno>
<aTxttit>Julegaver: Husk byttemærket og emballagen</aTxttit>
<iGenre>2</iGenre>
<iDomain>-1</iDomain>
<iTokens>418*</iTokens>
<iTypes>178*</iTypes>
<ePrody>2004</ePrody>
<eProdyc>0</eProdyc>
<eExpr>0</eExpr>
<eAspect>0</eAspect>
<eAgerel>1</eAgerel>
<eMedium>1</eMedium>
<eLang>0</eLang>
<eOlang>0</eOlang>
<eLoc>1:7</eLoc>
<eUrl>infomedia.dk</eUrl>
<vTit>Jyllands-Posten</vTit>
<vPubly>2004</vPubly>
<vPublyc>0</vPublyc>
<vPublm>12</vPublm>
<vPubld>26</vPubld>
<uName1>Tea Krogh</uName1>
<uName2>Sørensen</uName2>
<uRole>0</uRole>
<uSex>0</uSex>
<uBorny>-1</uBorny>
<uBornyc>-1</uBornyc>