# Digitaliseringen af den store danske ordbog – et kapitel i historien om *ODS på nettet*

I oktober 2005 udkom det femte og sidste suppleringsbind til *Ordbog over det danske Sprog* – ODS eller "den store danske ordbog" som den også kaldes. Dermed blev der sat punktum for et mere end hundredårigt projekt, og en epoke i dansk sprogbeskrivelse var slut. Helt slut? Nej, under en måned efter at det sidste supplementsbind blev udsendt, kunne Det Danske Sprog- og Litteraturselskab (DSL) lancere den første version af *ODS på nettet* til trøst for ordbogens mange trofaste fans (ordet *Fan* optræder i øvrigt i ODS-Supplementet med første belæg fra 1928). Det er nemlig lykkedes DSL at skaffe midler til et projekt der bl.a. omfatter digitalisering af ODS og publicering af en samlet netversion af ODS og Supplementet. Projektet hedder *ordnet.dk* og finansieres af Kulturministeriet og Carlsbergfondet med støtte fra Det Elektroniske Forskningsbibliotek. Ud over ODS vil *ordnet.dk* også komme til at omfatte en onlineordbog over det moderne sprog, baseret på data fra *Den Danske Ordbog*, og en opdateret version af tekstsamlingen *Korpus 2000.* Men her skal det handle om ODS.

"Nåja selvfølgelig" og "På høje tid", har mange nok tænkt da de stødte på siden første gang, og det er da ganske rigtigt også en sag der har ligget DSL på sinde gennem lang tid. At det overhovedet har kunnet lade sig gøre at have en funktionsdygtig netudgave af ODS klar efter kun to års arbejde, skyldes i høj grad at vi har kunnet nyde godt af erfaringerne fra andre, tilsvarende projekter. Den første beslutning der skulle træffes, var digitaliseringsmåden: scanning eller indtastning? Her har der virkelig været noget at lære fra ODS' søsterprojekter i landene omkring os. De er nemlig alle digitaliserede i dag, og nogle af dem var endda meget tidligt ude. I Sverige begyndte man allerede i 1983 at digitalisere *Svenska Akademiens Ordbok*, og den løsning redaktionen valgte, var optisk scanning. Der skulle imidlertid gå næsten 15 år før processen var gennemført, og den der besøger hjemmesiden i dag, vil hurtigt indse at resultatet langtfra er perfekt. I England begyndte man at digitalisere *Oxford English Dictionary* i 1984, og her valgte man den løsning at lade værket indtaste. Det tog væsentlig kortere tid, i alt fem år, men krævede til gengæld 120 indtastere, 50 korrekturlæsere og 13,5 mio. dollars før det digitale manuskript var etableret. Det er bl.a. disse – dyrekøbte – erfaringer der har gjort det muligt at gennemføre digitaliseringen af ODS både væsentlig hurtigere og for blot en brøkdel af de beløb der har været brugt af vores svenske og engelske kolleger.

Vi har valgt at følge den model der blev brugt af den senest digitaliserede af de store nationalordbøger, brødrene Grimms *Deutsches Wörterbuch*, DWB. Den blev digitaliseret i perioden 1998-2003/4, og modellen var indtastning i to versioner, den såkaldte double keying-metode. Digitaliseringen af ODS er foregået i samarbejde med *Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften* ved Universitetet i Trier, den institution som også var ansvarlig for det tyske projekt. Modellen med dobbeltindtastning kræver en nærmere forklaring.

## Dobbeltindtastning

I forbindelse med lanceringen af *ODS på nettet* hæftede flere journalister sig ved det paradoksale i at det ligefrem skulle være en fordel at indtastningen blev udført af kinesere der ikke forstod noget som helst af det de skrev. Det synes jeg nu ikke behøver være så svært at forstå. Enhver der har taget et kursus i blindskrift, ved at man bliver

dygtigst hvis man kan lade være med at læse tekstens indhold mens der skrives. Hvis det er en tekst på ens eget sprog, kan det nemt fjerne koncentrationen fra det egentlige, og man risikerer at komme til at skrive det man *tror* eller *forventer* der står, i stedet for det der faktisk står. Afskrift er rent kopieringsarbejde, og opmærksomheden bør derfor være på formen, ikke på indholdet.

Selve processen kan inddeles i to faser: en første fase hvor rådigitaliseringen finder sted, og en efterfølgende fase hvor det digitale manuskript fortolkes og opmærkes strukturelt. Skematisk kan det sammenfattes i følgende figur (der er en modificeret version af den tyske projektskitse).

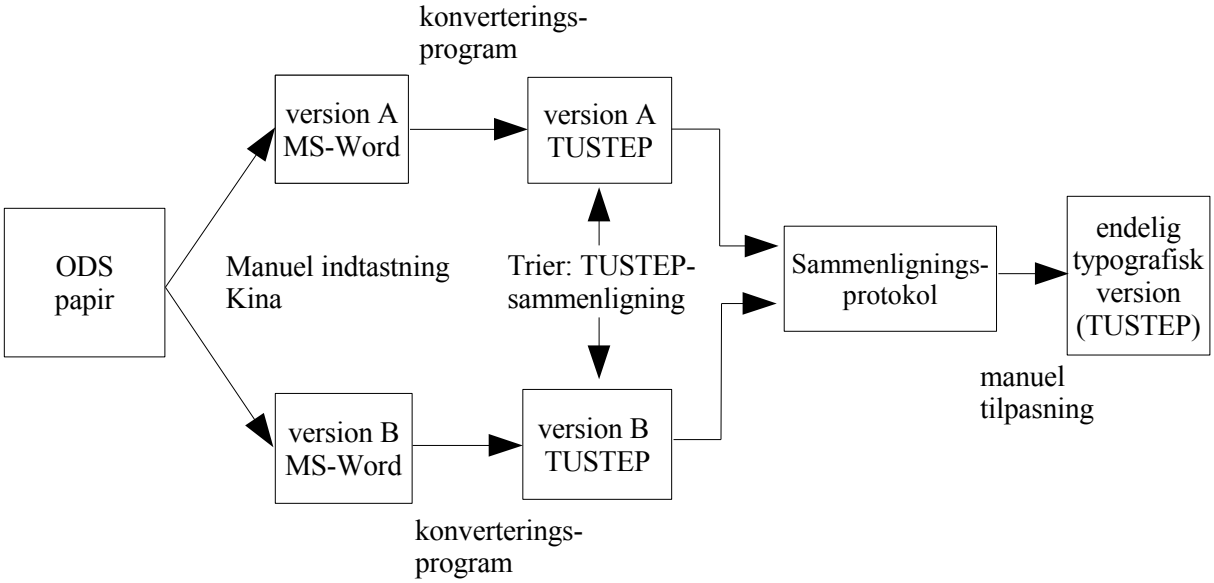


Fig. 1 Rådigitalisering (Kina, Trier)

Forud for selve indtastningsarbejdet analyserede vi de typografiske konventioner i ODS grundigt og udarbejdede en indtastningsmanual for at sikre at alle typografiske oplysninger fra den trykte udgave blev bevaret. Manualen indeholdt anvisninger på hvilke koder indtasterne skulle skrive for de forskellige specialtegn, symboler og andet der ikke er indeholdt i standardversioner af tekstprogram og tastatur. Den fysiske indtastning blev udført af firmaet TQY DoubleKey i Nanjing, Kina, på den måde at teksten blev indtastet i to versioner af to uafhængige grupper indtastere i et almindeligt tekstbehandlingsprogram (MS Word). Efterfølgende blev de to versioner konverteret til et særligt format, TUSTEP[1], og de to versioner sammenlignet automatisk. Forskellen mellem de to versioner blev udskrevet i en sammenligningsprotokol, og denne protokol blev til sidst bearbejdet manuelt indtil den endelige typografiske version var klar, en version der har form som tekstfiler hvor tekstens formatering er opmærket med de særlige, maskinlæsbare TUSTEP-koder. Fig. 2 giver et indtryk af hvordan filerne ser ud.

1 TUSTEP (= Tübinger System von Textverarbeitungsprogrammen) er et særligt tekstbehandlings- og satssystem der benyttes til filologiske publikationer.

```
$0303.01 <P>__<A+1>#F+afstr#.omme,#F-</A+1> #/+v.#/- [#P+!#P-
au#P+^!#P-sdr%::om'#P+^e#P-] #/+#s+str#.omme
$0303.02 bort.#/-#s- #/+vist kun som vbs.:#/- <A+1>{shu}
#F+Afstr#.omning,#F-</A+1>
$0303.03 en. Nedb#.orens . . Afstr#.omning gennem
$0303.04 #/+(dr#.^onledninger). LandbO.I.608.#/- #F+#./#F-
#/+(konkr.)#/-
$0303.05 Afstr#.omningen gennem Dr#.^anledningerne
$0303.06 #/+(udgjorde)#/- ca. 31 p. Ct. af Nedb#.oren. #/+smst.
$0303.07 jf. (overf.; sj.):#/- her #/+(#P+O#P-: i London)#/- er Moral,
$0303.08 der kan ikke dv#.^ales ved enkelte Udv#.^axter
$0303.09 og Afstr#.omninger, som altid findes i en
$0303.10 stor By. #/+HCAnd.XII.193.#/-</P>
$0303.11 <P>__<A+1>#F+afstudse,#F-</A+1> #/+v.#/- [#P+!#P-au#P+^!#P-
sdus#P+^e#P-] #F+-ede.#F- #/+vbs.#/- #F+-ning.#F-
$0303.12 #/+(l. br.) (fjerne ell.) #s+afkorte#/-#s- #/+ved
#s+studsning.#/-#s-
$0303.13 Afstudse . . Tr#.^aer. #/+vAph.(1759).#/- En Hund
$0303.14 med afstudsede #.Oren. #/+VSO.#/- Dykkerser
$0303.15 (S#.om med afstudsede Hoveder). #/+Hallager.
$0303.16 231.#/- kort Haar og afstudset Sk#.^ag. #/+Ing.
$0303.17 DM.196. H#.oyen.Moltke.4.#/-</P>
```

Fig. 2 Artiklerne *afstrømme* og *afstudse* i TUSTEP-format

Fordelen ved denne fremgangsmåde er dels at alle de typografiske oplysninger fra det trykte forlæg bevares – og det er meget vigtigt for den følgende strukturfortolkning – dels at dobbeltindtastningen sikrer at fejlprocenten bliver så lav at en efterfølgende menneskelig korrekturlæsning kan undværes. Det sidste er i sagens natur altafgørende for tids- og resurseforbruget for et værk af ODS' omfang. Stikprøver fra DWB har vist at det elektroniske manuskript stemmer 99,997 % overens med det trykte forlæg. Vi har endnu ikke gennemført systematiske stikprøver af materialet, men eftersom ODS med hensyn til både kompleksitet og typografisk tydelighed er noget klarere end DWB, er forventningen at nå en korrekthed på 99,999 %, dvs. gennemsnitlig 1 fejl pr. 100.000 typeenheder.

## Strukturopmærkning

Efter modtagelsen af den endelige typografiske version fra Kina og Tyskland begyndte DSL's del af arbejdet, nemlig fortolkningen af den typografiske opmærkning. Den kan i skematisk form anskues på følgende måde:

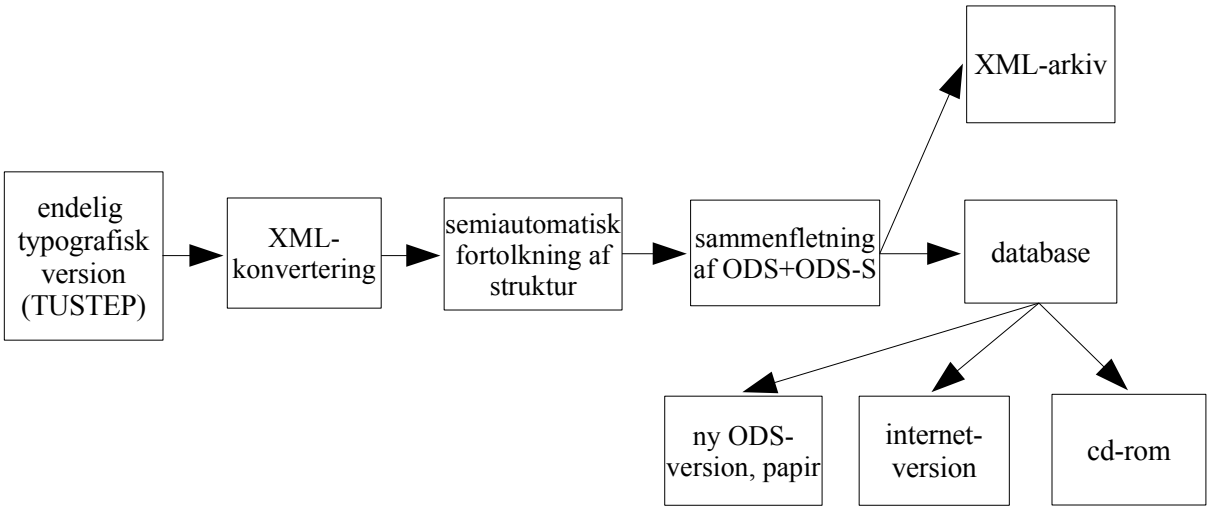


Fig. 3 Fra rådata til fulddigitaliseret version

Det siger næsten sig selv at den typografiske information er helt afgørende når man skal bevæge sig fra typografisk opmærkning til strukturopmærkning. Typografien er ganske enkelt det eneste holdepunkt man har, når man skal fortolke artiklernes strukturelle opbygning. Skriftsnit, skriftgrad, linjeskift, indrykning, ordinær, kursiv, fed el. spatieret skrift, ofte i kombination med interpunktionstegn, er de signaler den trykte bogside råder over til at angive hvornår én oplysningstype slutter og en anden begynder. Den første netudgave er blot en simpel html-version af ordbogen med en meget grov strukturopmærkning, og det er grunden til at det endnu ikke er muligt at foretage avancerede opslag i den: Indtil videre har vi alene isoleret opslagsordet og ordklasseangivelsen, mens resten af artiklen vises i sin helhed som løbende tekst. Der er altså ikke tale om den færdige, fulddigitaliserede version som man ser i figur 3; den er slutmålet og kan derfor først ventes realiseret når projektet afsluttes i 2009. Derimod er det planen at der i løbet af de næste år med jævne mellemrum vil komme nye og forbedrede versioner i takt med at søgefaciliterne forbedres og strukturopmærkningen bliver stadig finere. Når de enkelte oplysningstyper er blevet opmærket, vil det blive muligt at foretage mere specifikke opslag så man kan få svar på spørgsmål af typen: "Hvilke ord har vi indlånt fra arabisk i det 19. århundrede?", "Hvilke adjektiver optræder i citater af Herman Bang og J.P. Jacobsen?" eller "Hvordan afspejler artikler og citatmateriale synet på jøderne i mellemkrigstiden?" Derved kan ordbogen måske få en ny anvendelse inden for fx kultur- eller litteraturhistoriske studier fordi man får adgang til den store rigdom af oplysninger på måder der ikke har været mulige før.

Den anden hovedopgave består i at flette det elektroniske manuskript sammen med ODS-Supplementet. Supplementet foreligger allerede i elektronisk form og har derfor ikke skullet omkring Kina og Tyskland først. Af samme grund kan Supplementet også være en hjælp til den strukturelle fortolkning af typografien, men det kan ikke give det endelige svar. Supplementet er i nogen grad blevet ensrettet i sin struktur til det elektroniske format så den er mere homogen end man kan se det i ODS. Og strukturen i Supplementet er primært fastlagt med henblik på papirpublicering; den er ikke i et format der overholder en standardiseret dokumentbeskrivelse, fx i form af en DTD.

Det vil derfor være naturligt at begge værker i forbindelse med strukturopmærkningen konverteres til tidens standard for dokumentbeskrivelse, XML. Herefter kan de to manuskripter flettes sammen i en fælles base, dels i et XML-arkiv der sikrer langtidslagring af selve dokumentindholdet, dels i en relationel database som danner

udgangspunkt for såvel søgning i netudgaven som for eventuelle fremtidige udgivelser – digitale eller på papir.

Opgaven med at flette de to værker sammen er en udfordring der ikke skal undervurderes. Så længe det drejer sig om at tilføje hele artikler, kan det forholdsvis nemt gøres automatisk, men supplementsartiklerne dækker over en bred vifte af tilføjelser, ændringer og sletninger og kan gælde et hvilket som helst element i en eksisterende artikel. Fra vores kolleger på *Oxford English Dictionary* ved vi at det var en af de helt store udfordringer ved digitaliseringen af OED, og nogle af problemerne har forfulgt dem helt indtil i dag. Og det gælder generelt at de operationer der ikke kan udføres maskinelt, er ekstremt resursekrævende.

Om og hvor godt det lykkes at få flettet værkerne sammen, hører imidlertid til næste kapitel af historien, og det er ikke skrevet endnu.

## Forfatteroplysninger

Lars Trap-Jensen, f. 1960

cand.mag., Mphil

ledende redaktør ved projektet *ordnet.dk*

Det Danske Sprog- og Litteraturselskab