LÆSNING AF MASKINLÆSBARE TEKSTER

Ole Norling-Christensen
Gyldendals Ordbøger
Postboks 11
DK-1001 København K

1. Indledning

Så længe maskinlæsbare tekster kun læses af de maskiner, de er beregnet for, giver det sjældent problemer. Vanskeligere bliver det, når man ønsker at genbruge teksterne i andre maskiner eller til formål, som ikke var forudset, da teksterne blev til.

Hensigten med denne artikel er, at

- give en oversigt over vanskelighedernes art,
- afgrænse de tilfælde, hvor de kan overvindes med rimeligt enkle midler,
- orientere om de seneste års standardisering på området, og
- foreslå, at vi, der arbejder med disse ting, prøver at enes om et fælles udvekslingsformat for tekster som kunne være af interesse for flere.

Problemstillingen er væsentlig både for nogle forlag og for datalingvister. Jeg har arbejdet en del med den i begge sammenhænge og kan derfor måske kaste en smule tværfagligt lys over den.

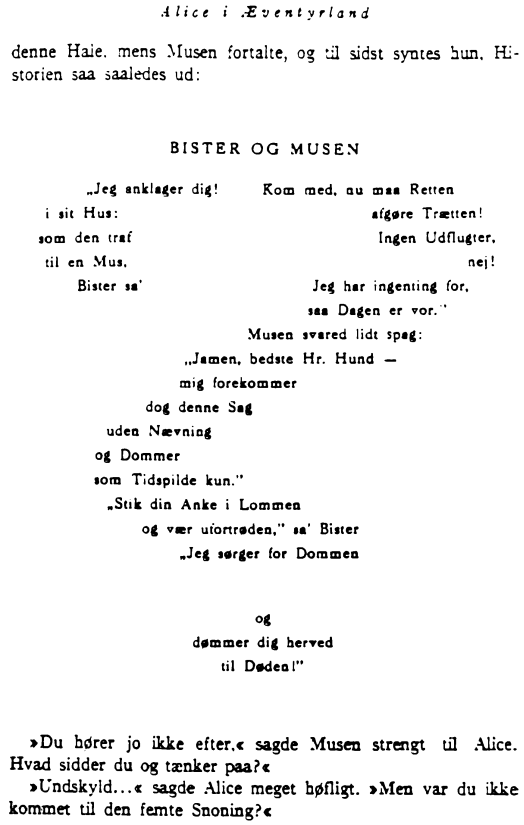
På forlaget møder vi problemet næsten daglig: en forfatter kommer med en diskette og siger: "Her er mit manuskript!" Skønt en fotosættemaskine kan modtage indtil 2000 tegn i sekundet, mens en tasteoperatør ikke kan præstere meget mere end 2 (Vail 1987), har vi - i hvert fald indtil for nylig - som regel måttet sige "Nej tak, vi vil hellere have en udskrift på papir". Det er galt nok; men værre er det, når det drejer sig om den slags tekster - typisk ordbøger og leksika - som over år eller tiår løbende skal revideres. Tidsintervallerne bliver her så lange, at de typografiske systemer i mellemtiden er blevet ændrede eller endda helt udskiftede.

For datalingvistikken måtte det være ideelt, at enhver tekst, som man ønskede at studere, blot kunne hentes på biblioteket i standardiseret form. En hindring herfor er, at det ikke altid er ganske klart, hvilke træk ved en trykt tekst den maskinlæsbare version skal afspejle. Herom mere nedenfor; lad mig foreløbig blot - uden yderligere kommentar - illustrere det med et citat (fig. 1, næste side).

2. Brug af maskinlæsbar tekst

Inden for den datamatstøttede leksikografi kan jeg umiddelbart pege på tre områder, hvor anvendelse af maskinlæsbare tekster kunne være et realistisk alternativ til egen (gen)indtastning: Datamatisk behandling af allerede eksisterende, trykte ordbøger; opbygning af korpuser m.m.; automatisk excerpering fra

Læsning af maskinlæsbare tekster
Ole Norling-Christensen
Proceedings of NODALIDA 1987, pages 237-258



Alice i Æventyrland

denne Hale, mens Musen fortalte, og til sidst syntes hun, Historien saa saaledes ud:

BISTER OG MUSEN

„Jeg anklager dig! Kom med, nu maa Retten
i sit Hus: afgøre Trætten!
som den traf Ingen Udflugter,
til en Mus, nej!
Bister sa' Jeg har ingenting for,
saa Dagen er vor."
Musen svared lidt spag:
„Jamen, bedste Hr. Hund —
mig forekommer
dog denne Sag
uden Nævning
og Dommer
som Tidspilde kun."
„Stik din Anke i Lommen
og vær uforfærden," sa' Bister
„Jeg sørger for Dommen

og
dømmer dig herved
til Døden!"

»Du hører jo ikke efter,« sagde Musen strengt til Alice. Hvad sidder du og tænker paa?«

»Undskyld...« sagde Alice meget høfligt. »Men var du ikke kommet til den femte Sproning?«

Fig. 1. Fra Alice i Æventyrland (Carroll 1946). Alice er optaget af at stirre på musens lange hale, mens musen fortæller denne "lange og bedrøvelige" historie om hunden Bister.

store tekstmængder, strømme af tekst, typisk avis- og telegrambureau-tekster.

Hvis allerede eksisterende, trykte, ordbøger skal lagres i databaser eller anden eksplicit struktureret form, til brug for revision og/eller alternativ præsentation, skal de ikke blot kunne læses af maskinerne; det læste skal også kunne strukturanalyseres, så de enkelte oplysningstyper og relationerne imellem dem klargøres. Hertil kræves i højere grad end for simple tekstundersøgelser en tolkning af oplysningernes grafiske fremtræden, eller rettere: en tolkning af de maskinlæsbare data som repræsenterer denne fremtræden.

Korpuser og konkordanser, samt automatisk eller blot maskinunderstøttet excerpering, kræver - ligesom enhver anden datalingvistisk undersøgelse - tekster i naturligt sprog, som kan læses af en datamat. Men indtil videre er de fleste korpuser fremstillet ved (gen)indtastning eller optisk læsning (OCR, Optical Character Reading) af trykt tekst - med efterfølgende korrekturlæsning - skønt mange af de pågældende tekster faktisk allerede fandtes i maskinlæsbar form.

Automatisk excerpering er behandlet af Robert Amsler (1986), som bl.a. indfører begrebet "NewsWire Lexicography". Fra en tekststrøm på flere hundrede tusinde ord pr. dag fra telegrambureauet Associated Press uddrager han, ved scanning efter ret enkle kriterier, leksikografisk interessante passager. Fx finder han implicitte definitioner af nye ord og udtryk ved at søge efter forekomster som "acronym for", "defined as", "usually called", "new name". Metoden foregribes af Pia Riber Petersen (1984:18-19) i hendes beskrivelse af, hvilke træk i et

belæg der tyder på, at et dansk ord er en leksikalsk nyhed; blandt de signaler, hun nævner, er citationstegn, samt ord som "såkaldt", "kaldes", "ordet", "dvs.". Lenders (1986) har brugt en tilsvarende fremgangsmåde til at finde centrale begreber hos Kant og afklare deres indhold. I det danske kulturministeriums regi undersøges det for tiden, om dagbladsartikler kan formidles til læsehandicappede via datatransmission / syntetisk tale. Hvis dette bliver en realitet, må også dansk leksikografi kunne drage nytte af denne tekststrøm.

Hovedtesen hos Amsler er, at maskinlæsbar tekst med fordel kan udforskes langt mere, end det sker i dag, med henblik på at skaffe leksikalske oplysninger. NewsWire Lexicography er kun ét eksempel blandt de mange muligheder han foreslår. Men desuden gennemgår han de vanskeligheder, som især data fra den grafiske industri, typisk fotosætterier, frembyder: Den typografiske kodnings eneste formål er at frembringe et bestemt visuelt mønster på papir, ikke at bevare informationsindholdet i en form, der kan bruges til andre formål. Også det modsatte gælder, anfører han: Der mangler ligeledes faciliteter til at omsætte databasers indhold til velformet trykt tekst; de eksisterende rapportgeneratorer er ikke tilstrækkelige.

Ifl. Amsler savnes der altså nye måder at repræsentere maskinlæsbar information på, metoder der både tjener det formål, som teksten oprindelig produceredes til, og nye formål, som kan være vidensbaserede edb-programmer henholdsvis grafiske produkter: "I have great interest in solving this problem, but no quick solutions to offer ...".

3. Tekstformidlingsprocessen

En del af meningen med denne artikel er at gøre opmærksom på, at sådanne nye repræsentationer faktisk er ved at blive fastlagt i form af en række internationale standarder. Men før jeg går nærmere ind på dem, vil det være nyttigt at analysere problemstillingen "læsning af maskinlæsbare tekster" lidt nøjere, prøve at fastslå problemernes art. Som udgangspunkt bruges et norsk forslag til niveaudeling af tekstformidlingsprocessen (fig 2, næste side); det er udarbejdet af Geir Andresen (1987) fra INGRAF, det norske institut for grafisk forskning.

Ideen er, at vejen fra forfatter over forlag eller redaktør til den færdige grafiske præsentation kan opdeles i mange små trin eller niveauer. På hvert trin træffes der - og kan der kun træffes - bestemte slags beslutninger: Underliggende niveauer skal ikke påvirke beslutninger som er truffet på niveauer højere oppe i skemaet, men understøtte disse. På den anden side må beslutninger truffet på ét niveau baseres på kendskabet til mulighederne på de underliggende niveauer, idet alt, som ligger over et niveau, skal kunne understøttes af det apparat, som er tilgængeligt på niveauet.

(1) På første (øverste) trin foreligger der "noget", der skal formidles som tekst, - tanker der endnu ikke er færdigt udformet i ord eller billeder.

Trinn ved tekstformidling

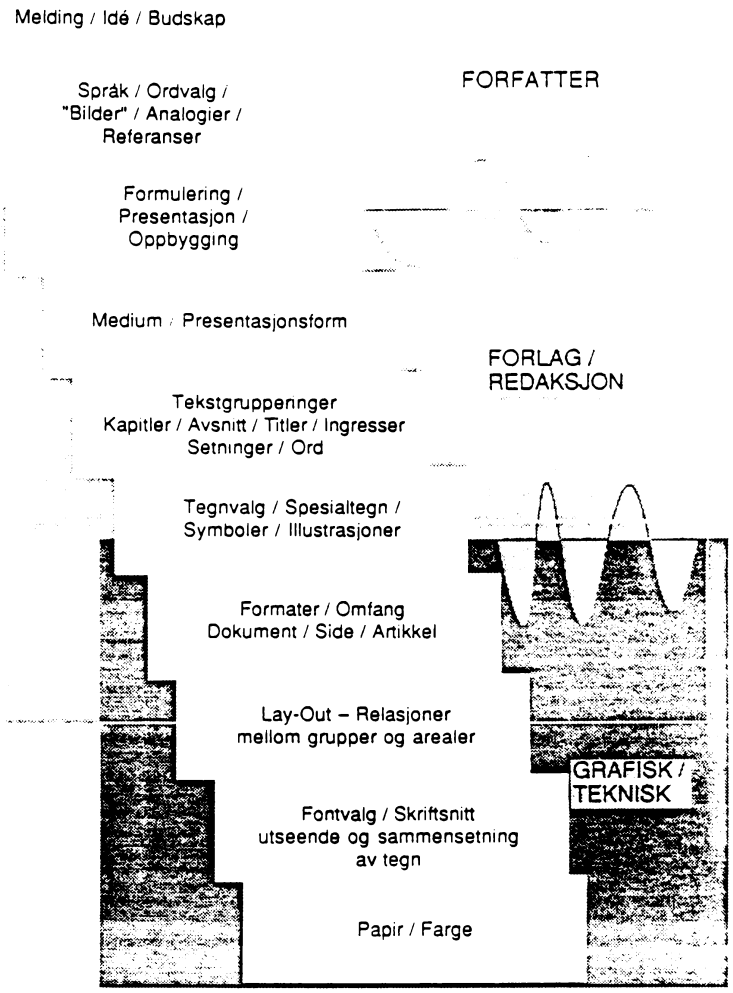


Fig. 2.
Efter Andresen (1987).

Underliggende nivåer skal ikke påvirke beslutninger som er gjort på overliggende nivåer, men underbygge disse. Overliggende beslutninger gjøres ut fra erfaringer fra underliggende nivåer.

(2) Dernæst sættes tankerne i ord; metaforer, henvisninger og sammenligninger udtænkes for bedst muligt at formidle budskabet.

(3) På tredje trin formuleres og disponeres teksten; den opdeles i indledning, forskellige afsnit og underafsnit, slutning.

(4) Valg af medie/præsentationsform er en vigtig beslutning, men svær at placere i skemaet. I den viste version er det anbragt som fjerde trin, i en anden version først som sjette. Her besluttes det, om teksten skal blive til en bog, en tidsskriftartikel, eller fx en avisnotits.

(5) Indholdet og dets rækkefølge er lagt fast. Tilbage står den detaljerede opdeling af teksten i overskrifter, brødtekst, lister, tabeller, fodnoter, billedtekster. Her træffes ikke længere beslutninger om indholdet, kun om formen. Men, i hvert fald i princippet, er der tale om generisk - artsmæssig - beskrivelse: Det markeres, at der er tale om forskellige slags

tekst; men der siges stadig intet om, hvilke skriftsnit og andre grafiske virkemidler der skal bruges til at udtrykke dem.

(6) Nu skal der tænkes på illustrationer, herunder specialtegn, lydskrifttegn, små "billeder i teksten", "ikoner", "sigler".

(7, 8) På trin 7 fastlægges de generelle typografiske regler for det pågældende medie, og på trin 8 indpasses den konkrete tekst heri.

(9, 10) Endelig handler de to nederste trin om selve sætningen og trykningen af teksten.

4. Fire slags maskinlæsbar tekst

Skemaet er - naturligvis - en abstraktion; i det virkelige liv vil hverken forfatter, forlag eller teknik opleve grænserne så skarpt. Afgrænsningen, rækkefølgen og antallet af forskellige trin kan diskuteres; som nævnt er Andresen selv i tvivl om placeringen af medie/præsentationsform. Den diskussion skal ikke tages her. Skemaet skal blot bruges til at klargøre, at der er forskellige slags maskinlæsbar tekst, og at forskellene bl.a. skyldes, at de hører til forskellige trin eller niveauer på tekstens vej fra tanke til tryk.

Omkring trin 2-3 er der tale om, hvad vi kunne kalde rå tekst, ord der er ordnet i sætninger, som måske igen er ordnet i afsnit, den slags tekst, som vil være nødvendig og tilstrækkelig for de fleste datalingvistiske undersøgelser, den slags tekst, som en traditionelt arbejdende ordbogsredaktør ville overføre udsnit af til sin excerptseddel.

Omkring trin 5 kommer den generiske mærkning ind, den der opdeler materialet i forskellige slags tekst (fx forord, kapitler, noter) uden at beskrive, hvordan forskellene skal vises i det færdige produkt. Også denne mærkning kan i visse tilfælde være nyttig for leksikografen eller datalingvisten.

Omkring trin 7-8 tilføjes layout mærkningen. Det kan i mange tilfælde gøres alene på basis af den generiske mærkning samt nogle generelle regler for det pågældende layout: nu skal det fx besluttes, om noten skal være en fodnote eller stå i et særligt noteafsnit, om en given fremhævelsestype skal vises med kursiv eller kapitæler. På dette niveau beriges teksten med koder for forskellige skriftsnit (kursiv, fed) og skriftstørrelser. Ofte vil der, især hvis der ud over teksten også er billeder, ikke kun blive taget hensyn til de forskellige tekststykkers art, men også æstetiske hensyn, fx til sidernes udseende. Under alle omstændigheder bliver den generiske mærkning hermed igen overflødig for så vidt angår det konkrete grafiske produkt. Derimod vil den bevare sin værdi for alle andre anvendelser af teksten: anden grafisk udformning, electronic publishing, datalingvistiske analyser. Ofte vil det være denne sidste slags maskinlæsbar tekst, et fotosætteri vil kunne bidrage med, og det kan være ganske svært at rekonstruere den rå tekst eller den generisk kodede tekst herfra.

Ved de to nederste og sidste trin er vi dybt inde i sættemaskinens og trykkemaskinens indre. Herfra vil vi næppe få brugbare

maskinlæsbare data; men vi vil få en trykt tekst, hvis indhold gerne skulle svare til det, vi fik ud af at maskinlæse tekster fra et af de tidligere niveauer. Dog kan det være et problem for traditionel filologisk anvendelse af de maskinlæsbare tekster, at ombrydning i linier (, spalter) og sider ofte først sker på disse sidste trin.

Alt efter, på hvilket tidspunkt af processen teksten gøres maskinlæsbar (eller gives fri til forskningsformål), kan vi altså skelne mellem rå tekst, generisk kodet tekst, typografisk kodet tekst og egentlige grafiske data (fx bit-map, en redegørelse for, hvor på siden der skal være sort, og hvor hvidt). Rå og generisk kodet tekst fås fx fra tekstbehandlingsanlæg; generisk og typografisk kodet tekst fra sætterier. De egentlige grafiske data hører til i og lige før sættemaskinen (jf. fig. 3).

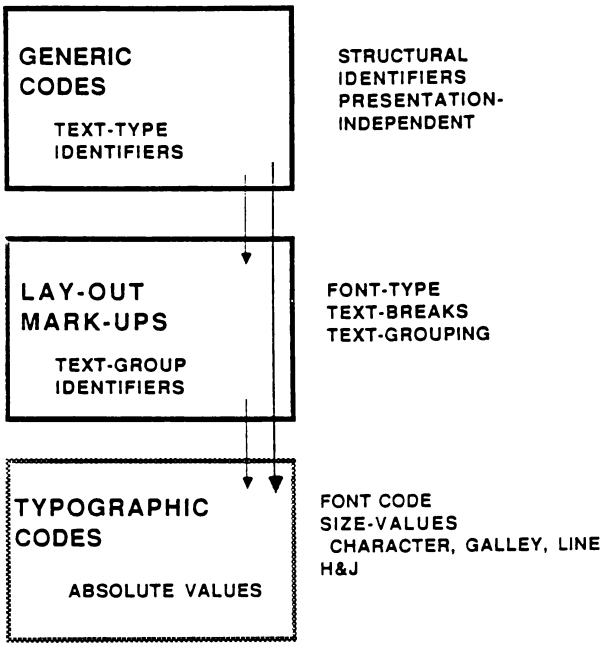


Fig. 3. De tre slags kodet maskinlæsbar tekst, som i artiklen kaldes generisk kodet, typografisk kodet, og egentlige grafiske data. Efter Andresen (1987).

Bortset måske fra de egentlige grafiske data vil den rå tekst i princippet stadig være til stede som koder, der repræsenterer enkelttegn (bogstaver, interpunktionstegn, osv.). Men jo længere vi kommer ned i skemaet, desto mere er den rå tekst gemt imellem alle mulige andre koder. Linier hørende til to forskellige spalter står nu måske sammen, og i værste fald foreligger teksten blot som en bit map.

Opgaven ved læsningen vil være at afdække den rå tekst samt så meget af den generiske kodning, som der er brug for til den påtænkte anvendelse; endvidere at kunne videregive resultatet i en form, som er entydig, og som er brugbar for anvenderen.

5. Genotype / fænotype

Ud fra synspunktet "læsning af maskinlæsbar tekst" kan et trin i Andresens model anskues således: Der foreligger en maskinlæsbar tekst fra det foregående trin; denne tekst undergår en transformation, hvis inddata er (1) uddata fra foregående trin, og (2) nogle yderligere regler, som hører til trinet. Som nævnt

ovenfor kan fx layout-mærkninger eventuelt foretages maskinelt (trin 7-8); i så fald vil de "yderligere regler" være eksplicit formulerede. Men reglerne kan også være en grafikers mere eller mindre intuitive viden. I begge tilfælde gælder det imidlertid, at trinets uddata er en entydig funktion af to variable, nemlig inddata og de nye regler. Derimod er uddata i almindelighed ikke således beskafne, at inddata entydigt kan rekonstrueres ud fra dem.

To eksempler, baseret på erfaringer fra ordbogsområdet, kan belyse dette:

(1) Om et skriftsnitmærke (fx ordinær-, kursiv-start) står før eller efter et blanktegn eller et punktum kan ikke ses i det færdige produkt; det er derfor ligegyldigt, og selv den bedste korrekturlæser vil ikke have haft anledning til at ændre derved: genotypen (den kodede repræsentation) er forskellig; men fænotypen (det synlige resultat) er det samme. Der kan også forekomme flere skriftsnitmærker efter hinanden, af hvilke kun det sidste har betydning for den færdige, trykte teksts udseende.

(2) Gennem flere trin i modellen kan det være underforstået, at startparentes med hensyn til skriftsnit følger den efterfølgende tekst, slutparentes den foregående. Dette giver normalt "pæne" resultater; men en i ordbogen hyppigt forekommende oplysningstype står i parentes, indledes med ordinær (opret) skrift, og afsluttes med kursiv (skrå skrift). I korrekturen konstateres, at dette giver et visuelt "grimt" resultat: kombinationen af ordinær startparentes med kursiv slutparentes giver ikke det rette indtryk af, at der er tale om én og samme parentes. Det besluttes derfor, at alle parenteser skal være ordinære. Dette indbygges i selve sættemaskinen, på trin 9. Man definerer blot, at en "kursiv" parentes har samme udseende som en "ordinær". De maskinlæsbare data er ikke blevet ændret; og igen resulterer forskellige genotyper i samme fænotype. Først når (dele af) det typografiske system udskiftes, viser problemet sig: det nye system "ved" ikke, at kursive parenteser skulle se ordinære ud.

Ved udarbejdelse af programmer, der skal tolke denne slags maskinlæsbar tekst, er det nødvendigt at tage hensyn til sådanne mulige flertydigheder.

Ordene genotype (anlægspræg) og fænotype (fremtoningspræg) er lånt fra genetikken. De er udmøntet af den danske plantefysiolog og genetiker W. L. Johansen (1857-1927). Analogien kunne trækkes så vidt, at eks. 1 ovenfor svarer til de tilfælde, hvor forskellige genotyper (pga. dominans) giver samme fænotype, mens eks. 2 er det tilfælde, hvor samme genotype (pga. forskelligt miljø) resulterer i forskellige fænotyper.

6. "Generisk" - hvad er det egentlig?

Med vilje har jeg i det foregående ikke givet nogen præcis definition af "generisk". Thi hvad der er artsmæssig og hvad der er typografisk mærkning af en tekst, afhænger i nogen grad af, fra hvilket synspunkt sagen anskues.

Fra ordbogsredaktørens synspunkt vil et større antal forskellige oplysningstyper inden trykningen skulle reduceres til fx tre skriftsnit: halvfed, kursiv, ordinær, i en vis udstrækning modificeret med forskellige slags parenteser og interpunktionstegn. Den generiske inddeling af oplysningerne reduceres altså til en - mindre detaljeret - typografisk. Men for typografen betyder en mærkning som "{k}" (jf fig. 8) ikke nødvendigvis "kursiv"; faktisk oversættes den i fotosætteriet til "brug format 2". Og "format 2" kan i sætteriet frit defineres. I praksis indebærer definitionen ikke blot kursiv skrift af bestemt snit og størrelse, men desuden at fx det franske orddelingsprogram skal benyttes. Ønskede man imidlertid en anden skrift end kursiv, kunne det lige så vel indkodes som en del af "format 2". Hvad ordbogsforfatteren må opfatte som en typografisk kodning, kan typografen altså med samme ret anse for en generisk.

7. Tekniske problemer

Maskinlæsbar tekst vil ofte blive leveret på magnetbånd eller diskette; men også direkte kommunikation fra én maskine til en anden forekommer. For alle tre medier gælder, at der findes flere forskellige standarder for, hvordan dataene lagres/transmitteres. Ikke mindst disketteområdet har været kaotisk: Der er 8", 5 1/4", 3 1/2" disketter. De kan være formatteret enkelt- eller dobbeltsidet og med forskellig tæthed, dvs. antal spor pr. tomme. Nogle er indspillet med konstant hastighed, andre med en hastighed, der varierer med læse/skrivehovedets afstand fra diskettens centrum. Filerne kan være organiseret under (MS/)DOS, CP/M eller et helt tredje operativsystem. Selv en 360 kB (MS/)DOS-diskette kan - hvis den er formatteret og/eller skrevet på et 1,2 MB drev - ikke altid læses på andre slags drev.

Dette er dog ikke det værste; som regel er det muligt at få en tekstfil overført til anden maskine, hvis man vil betale for det ("tekst" skal her forstås som den datatype der blot - set fra datamaten - er én vilkårligt lang, sekvens af tegn, uden anden datastruktur). Dermed er vanskelighederne imidlertid ikke forbi.

Alt efter hvilken konvention, der er brugt, vil tegnene tilhøre et alfabet med i alt 128 tegn (7-bit kode) eller 256 tegn (8-bit kode). Men betydningen af de enkelte tegn og tegnkombinationer afhænger bl.a. af, hvilket tekstbehandlingssystem der er anvendt. Selv ved tekster, der er skrevet med det samme tekstbehandlingsprogrammel, kan kodningen være forskellig, dels fordi visse tegn og koder kan defineres af brugeren, dels fordi ikke alle maskiner råder over samme tegnsæt; jf. 9.3 Tegnstandarder nedenfor.

Overførsel af tekstfiler fra et system til et andet er, som nævnt, mulig; bl.a. har UNI*C faciliteter hertil. Men dels er konverterings- eller kommunikationsudstyr dyrt, dels er en rent mekanisk overflytning kun sjældent tilstrækkelig. Den kan klare de fysiske forskelle på disketterne og flytningen fra et operativsystem til et andet. Generelle forskelle mellem de almindeligste tekstbehandlingssystemer kan naturligvis også klares med generelle konverteringsprogrammer; men så snart teksten rummer andre tegn end det engelske alfabet, kan det gå galt, med

mindre forlaget eller sætteriet for hvert enkelt manuskript lægger et stort arbejde i at lave individuelle konverteringstabeller. Rummer teksten fx skemaer og tabeller, eller kemiske og matematiske formler, er det næsten sikkert, at det går galt.

En undersøgelse (Møller 1987:8-9) viste da også, at kun 1/3 af de danske sætterier kunne modtage disketter, og at praktisk taget ingen af dem reklamerede for det. Indtil videre ser det altså ud til, at disketten kun er et realistisk alternativ til papirmanuskriptet for kunder med store mængder af tekster, der overholder en ensartet konvention.

Moderne tekstbehandlingsprogrammer får stadig flere faciliteter til at præsentere teksterne flot: fremhævelser, forskellige skrifter, tabulering, lige højremargen, forskellige spaltebredder, flerspaltet tekst, administration af fodnoter og registre. Desk-top Publishing er blevet et modeord. Men generelt kan det siges, at jo mere forfatteren har udnyttet sådanne grafiske faciliteter, desto vanskeligere er det at få et andet system til at læse og tolke teksten korrekt. Selv de firmaer, der lever af at sælge diskettekonverteringsudstyr, anbefaler derfor (Vail 1987), at man i stedet anvender generisk kodning. Denne kan aftales individuelt mellem sætteri og kunde. Men der findes også både typografisk orienterede standarder (INGRAF 1985; Cave 1986) og det helt generelle SGML, som omtales nedenfor.

Alt dette nævner jeg af to grunde. Dels fordi sprogligt materiale fra tekstbehandlingsanlæg også kan være af interesse for datamatstøttet leksikografi og anden datalingvistik, dels fordi problemerne med tekst fra avancerede tekstbehandlingssystemer ligner dem, man støder på, når man prøver at læse tekst fra fotosætterier, fx avis- og bogtekster. Hvert enkelt sætteri har sit eget kodesystem - også selv om apparaturet er det samme. Og selv det samme sætteri kan have ændret sine koder siden sidst; det giver visse vanskeligheder, når fx en ordbog skal revideres hvert tredje eller femte år. En fordel ved sætteridata frem for visse "avancerede tekstbehandlingsdata" er dog, at sætteriet normalt vil råde over et ikke-ombrudt (typografisk-generisk) arkivformat, som bevarer de typografiske fremhævelser, men alligevel er rimeligt tilgængeligt for genbrug.

8. Praktiske erfaringer

Med henblik på fotosætning af tekst fra ordbogsredigeringssystemet Compulexis (bl.a. ODSS 1987); overførsel af ordbogsdata fra forskellige typografiske systemer (jf. figg. 4-9) til et redaktørorienteret tekstbehandlingssystem og tilbage igen, samt strukturanalyse af dataene med henblik på databaselagring; fremstilling af en retrogradordbog på basis af RO (1986); m.fl. opgaver; har jeg udviklet en række analyse- og konverteringsprogrammer. De er skrevet i Pascal og kan stilles til rådighed for andre, som vil forsøge sig på området.

Min grundlæggende erfaring har været, at det altid er nødvendigt at gennemføre en total analyse af kodestrukturen i den pågældende tekst. Thi enten er en beskrivelse ikke til at få fat i, eller også viser det sig, at den er ukorrekt eller ufuldstændig.

Fremgangsmåden er i øvrigt følgende:

(1) Kig overfladisk på dataene. Det kan fx gøres med faciliteten VIEW i programmet QuickDos (Gazelle Systems, Provo, Utah, USA), eller ved at udskrive nogle sider under anvendelse af en rutine der omsætter ikke printbare (kontrol)tegn til deres talværdi. Programmer til "hex-dump" kan naturligvis også anvendes. Eksemplet i fig. 4 viste at den pågældende fil - som var kopieret til diskette fra magnetbånd - indledes med en "header" og afsluttes med en "trailer", som begge var irrelevante. Efter traileren lå der desuden nogle helt tilfældige datarester:

```
700000001645Cpiovol_Header(0)1 567075606 aJob_
Archive Mon Dec 21 10:00:06 1987(10)  (10)-ost /dev
/rmt/0yy CpiovolVersion3.1.1.1(10) (0) (0) (0) (0) (0) (0)
(0) (0) (0) (0) (0) (0) (0) (0) (0) (0) (0) (0) (0) (0) (0) (0)
00123-----00A(10)0707070004040212241007750001460000
144000002110373041631534210000160000070644600123--
---00A(0) (1)P5♥A(1)T(1)P4♥A(1)P1♥ Rei▒.(2)"(1)
P4♥A. (1)P2♥fk Academy; America; Associate; (1)P1♥
(i biografannonce) (1)P2♥(omtr.) (1)P1♥betinget ti
```

Fig. 4. Begyndelsen af Engelsk-dansk Ordbog (EDO 1988). Dele af headeren er oversprunget.

Det første kig kan også afsløre, at det foreliggende analyseprogrammel ikke uden videre kan anvendes. Russisk-dansk Ordbog, som blev lavet hos RECKU (nu UNI*C, København), foreligger således i en slags hulkortformat, hvor to "hulkort" / linier tilsammen bestemmer én linies tegn:

```
 I na præp.m.akk. 1 om retningen: (hen, ind, ned, om, op, over, ud) på;
+  ll ÜÜÜÜÜÜÜÜÜÜÜÜ l ÜÜÜÜÜÜÜÜÜÜÜÜ
  til; mod; i; v'ewat' ü st'enu hænge op på væggen; v'yglânut' ü dr'uga
+              kkkkkkkk+kkkkkkk                    kkkkkkkkkkk+kkkkkkk
  kaste et blik på sin ven; v'yjti ü 'ulicu gå udenfor; gå ud på gaden;
+                          kkkkkkk+kkkkkkk
  dor'oga ü Moskv'u vejen mod Moskva; 'exat' na S'ever rejse nordpå; leh'
+ kkkkkkkk+kkkkkkkk                  kkkkkkkkkkkkkkk                 kkkk
  na div'an lægge sig (hen, ned etc.) på sofaen; perevest'i ü d'atskij
+ kkkkkkkkk                                      kkkkkkkkkk+kkkkkkkkk
```

Fig. 5. Udsnit af Russisk-dansk Ordbog (RDO 1985). To linier, den første indledt med blanktegn, den næste med "+" definerer tilsammen èn linies tegn. I "plus-linien" betegner "l" fed og "k" kursiv kyrillisk skrift; ü markerer kursiv, blanktegn ordinær latinsk skrift; "+" under "ü" betegner tilden, der gentager opslagsordet.

(2) Næste trin er at køre programmet TGNTAL, der udskriver første forekomst af hvert af de (højst) 256 tegn. Desuden optæller det antallet af forekomster af hvert enkelt tegn. Programmet giver mulighed for at overspringe indtil 32.767 tegn i starten. Herved kan fx den irrelevante "header" overspringes. Det er også muligt at overspringe tegnene A..Z, a..z, hvis det første kig har vist, at de blot repræsenterer sig selv, dvs. følger ASCII konventionen.

(3) Resultatet af denne første undersøgelse gennemgås nøje, idet de udskrevne tekstprøver sammenholdes med den trykte tekst. Ét resultat har hver gang vist sig: antallet af højreparenteser i ordbogsdataene er en smule lavere end antallet af venstreparenteser; nogle parenteser er altså ikke afsluttede, hvilket sidenhen kan give vanskeligheder, når parenteser skal bruges som kriterium for forekomsten af bestemte oplysningstyper (jf. fig. 6, næste side).

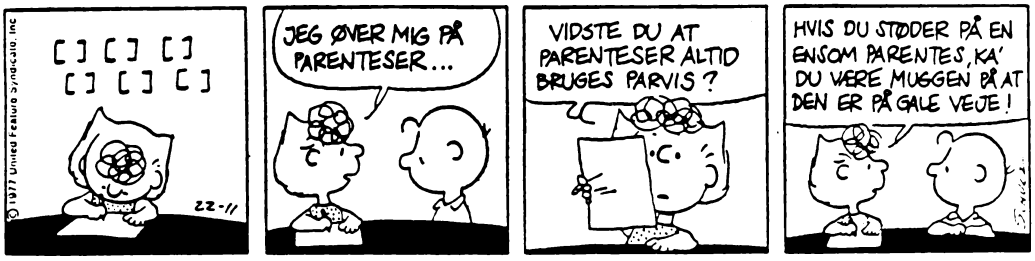


Fig. 6. Fra "Radiserne"; citeret efter dagbladet Politiken.

Gennemgangen afslører, hvordan hvert enkelt af de 128 eller 256 tegn er brugt, herunder fx hvordan æ, ø, å er repræsenteret. Desuden vil den vise, at visse tegn bruges til at indlede og/eller afslutte flertegnssymboler for grafiske tegn og andet, som ikke repræsenteres af enkelttegn:

```
93 ]    0.00%
        lse, sikkerhed (i optr(den), aplomb. (2)" (1)P4♥apoc
        alypse (1)P1♥R(17)fe(17)sjp(17)fdk(17)felips▒ (1)P
        2♥sb (1)P1♥)benbaring; (1)P2♥the A. (1)P1♥Johannes
        ' ]benbaring. (2)" (1)P4♥apocarp (1)P1♥R(17)sj(p
        (17)feka:p▒ (1)P2♥sb (bot) (1)P1♥ flerfoldsfrugt. (2)
        " (1)P4♥apocope (1)P1♥R(17)fe(17)sjp(17)fdk(17)fepi
        ▒ (1)P2♥sb (gram) (1)P
```

Fig. 7. For hvert konstateret tegn udskriver programmet ca. 127 tegn før og efter den første forekomst, som understreges. Desuden udskrives en tabel med det absolutte antal forekomster af hvert tegn. Her vises første forekomst af "Å" i (EDO 1988). Ved sammenligning med den trykte tekst konstateres det bl.a., at kontroltegnene <1>, <2> og <17> indleder flertegnssymboler.

(4) De næste to programmer bygger på den erfaring, at en "kode", dvs. et flertegnssymbol, ofte består enten af "kodestart" + et indhold af vilkårlig længde + "kodeslut" eller af et kodemærke + et fast antal tegn, som da hører med til koden.

Programmet STJRNTAL blev oprindelig lavet til at finde og optælle Gyldendals "stjernekoder", som både indledtes og afsluttedes med en "stjerne" (asterisk, ASCII-tegn nr. 42), heraf navnet; stjernerne er sidenhen afløst af { og } :

```
28e}    4.53%
        f2} {o}gå videre,; forlænge: {k}{f23} un mur; {f23
        } {26a}, de(13)(10){o}vedblive at; {f23} {k}q. dan
        s le m{2e}me emploi {o}lade {28e}n vedblive i(13)
        (10)samme virksomhed. {h}2. {o}vedvare, blive ved,
         fortsætte: {k}la pluie {f23}e;(13)(10){o}forlænge
        s, strække sig. {h}-it{2
```

Fig. 8. I Fransk-dansk Ordbog (FDO 1980) er venstre- og højre-"tuborg" tydeligt nok kodeindleder og -afslutter. Her vises første forekomst af koden for "è". Kontroltegnssekvensen <13><10> markerer postgrænser i filen ("ny linie"); de er irrelevante for ordbogsteksten og skal blot konverteres til blanktegn (ordmellemrum).

Programmet KODETAL kan i samme gennemløb af teksten behandle flere forskellige kodemærker, hvert med et individuelt defineret antal efterfølgende tegn; for nylig er det blevet udvidet, så det også kan "kigge bagud", idet en gruppe tekster viste sig at repræsentere accentbogstaver ved bogstavet efterfulgt af et symbol for accenten. Også dette program tæller antal forekomster af hver kode og udskriver den først fundne med kontekst:

```
<17>mg    4.95%
          (1)P2Vfit able-bodied (seaman); (am (1)P1Vform for
          ) (1)P2VB. A. (Bachelor of Arts). (1)P1V(2)" (1)P4Va
          back (1)P1VR(17)fe(17)s)b(k (1)P2Vadv: taken (17)
          mg (1)P1Vforbliffet. (2)" (1)P4Vabacus (1)P1VR(17)
          sj(b(17)fek(17)fes (1)P2Vsb (pl -es el. abaci (1)
          P1VR(17)sj(b(17)fesai) kugleramme, regnebr(t; (1)
          P2V(arkit)
```

Fig. 9. I (EDO 1988) repræsenterer kontroltegnet <17> efterfulgt af netop to andre tegn forskellige specialtegn, bl.a. fonetiske tegn. "<17>mg" er således tilden, der gentager opslagsordet.

(5) Konvertering af teksten til den ønskede form. Konverteringsprogrammet bør - for en sikkerheds skyld - udformes således, at det melder fejl og udskriver kontekst, hvis det møder tegn eller koder, der ikke udtrykkelig er defineret som tilladte.

9. Standarder

Et af problemerne ved at arbejde med maskinlæsbar tekst fra mange forskellige kilder er, som nævnt, at hvert system, hver leverandør, har sine egne koder. Yderligere kan man ikke regne med, at kodesystemet er fuldt dokumenteret.

Der er imidlertid håb om, at dette efterhånden vil ændre sig. Den internationale standardiseringsorganisation, ISO, har i de seneste år arbejdet intenst på at fastlægge standarder på området, og standardiseringsarbejdet støttes af både EF, USA og de nordiske lande. Faktisk har EF-landene - og dermed også Danmark - forpligtet sig til fra 1988 at stille krav om at standarderne overholdes ved alle offentlige indkøb af informationsteknologi og datakommunikation (Vejl. 1987).

9.1 Dokumentarkitektur

Blandt de mest ambitiøse af disse standarder er ISO/DIS 8613 (1987) Office Document Architecture / Interchange Format (ODA/ODIF). Den er endnu ikke vedtaget, men foreligger som udkast ("DIS" = Draft International Standard). Standardens formål er at muliggøre udveksling af kontordokumenter (rapporter, fakturaer, breve, notater osv.) ved hjælp af datakommunikation eller ved forsendelse af lagermedier (magnetbånd, disketter, etc.).

Skønt standarden specielt skal dække kontordokumenter, er den så generel, at den må kunne dække næsten alt andet også. Prisen for denne generalitet er, at standarden bliver så abstrakt og så omfattende, at næppe andre end specialister magter at sætte sig ind i den. Det er ikke en standard for kontorassistenter,

leksikografer eller datalingvister, men én som måske bliver indbygget i fremtidige tekstbehandlings-, informations- og kommunikationssystemer.

Ved "dokumentarkitektur" forstår ODA/ODIF "det sæt af regler, der angiver et udvekslet dokuments struktur"; og hovedideen er, at et dokument kan beskrives ved hjælp af to strukturer: en logisk struktur og en layout struktur. Den logiske struktur opdeler indholdet af et dokument i stadig mindre dele på grundlag af den måde, mennesker (logisk) opfatter indholdet på, dvs. i kapitler, afsnit, underafsnit, paragraffer og billeder. Layoutstrukturen opdeler indholdet i samlinger af sider, enkeltsider, arealer på siderne (fx spalter), linier. De to strukturer repræsenterer forskellige (og i princippet indbyrdes uafhængige), men komplementære syn på dokumentets indhold. Hver af dem beskrives som et hierarki af objekter; disse repræsenteres af såkaldte attributter, som definerer objekternes egenskaber og deres indbyrdes sammenhæng.

I Andresens skema (fig 2) indføjes den logiske struktur omkring trin 3-5, layoutstrukturen på trin 5-8.

Bortset fra filologers ønske om at kunne referere til en bestemt linie på en bestemt side i en trykt tekst, er layoutstrukturen mindre interessant i nærværende sammenhæng. Væsentligt er det blot, at den er adskilt fra den logiske struktur.

9.2 SGML, en standard for generisk mærkning

Den logiske struktur kan behandles i henhold til en anden standard, ISO 8879 (1986), Standard Generalized Markup Language (SGML), som er et system til mærkning af tekstdele efter deres art (den flere gange nævnte generiske mærkning). Selve standarden er tung læsning; men den ledsages af en række tillæg (Annex A - I), som rummer beskrivelser, der er mere tilgængelige for ikke-dataloger. Joan Smith (1986, 1987) har givet flere enkle introduktioner til emnet; som en god indføring kan desuden anbefales FORMEX (1985). Descriptive Tools (1987:139-44) giver en gennemgang, som især sigter mod leksikografiske anvendelser.

SGML er en generel formalisme, der beskriver et dokument som en hierarkisk struktur, et træ. Den kan virkeliggøres med vilkårlige symboler, af hvilke enkelte må reserveres til mærkningsformål. Syntaksen bygger på teorien for deterministiske finitte automater (tilstandsmaskiner) og muliggør derfor konstruktion af rimeligt enkle programmer, parsere, som er i stand til at behandle de strukturerede dokumenter. I gennemgangen nedenfor følges den SGML-konvention (FORMEX, 1985) som benyttes af EF's Kontor for Officielle Publikationer; de reserverede tegn er her "<" ("mindre end") og "&" ("og"), der benyttes på følgende måde:

En given kategori med navnet "xxx" indledes med koden "<xxx>" og afsluttes med "</xxx>". Mellem disse koder kan underordnede kategorier i vilkårligt mange niveauer indledes og afsluttes.

Længere udtryk, som hyppigt skal gentages, samt tegn der ikke er til rådighed i det anvendte alfabet, kan "stenografisk" noteres som "&yy;" - altså et navn, "yy" indledt af "&" og afsluttet af ";". Fx kunne "&SGML;" repræsentere "Standard Gene-

ralized Markup Language", og "&ag;" kunne repræsentere lille a med accent grave. Får man brug for de reserverede tegn, her "mindre end" og "og" tegnene, må de da symboliseres ved fx "&me;" og "&og;".

```
<entry>
  <hwsec>
    <hwlem id=1234567>Map</hwlem>
    <pron>(mæp)</pron>, <pos lit='sb.' hom=1>
    <vfl id=1234567>Also <vd>6–7 <vf>mappe</vf>, <vd>6–8
      <vf>mapp</vf>.
    <etym id=1234567>ad. L. <cf>mappa</cf>, in class.L. 'table-cloth,
      napkin', but in med.L. used <lab>transf.</lab> in the combination
      <cf>mappa mundi</cf> (see <xlem>Mappemonde</xlem>).
      <enote>Cf. the synonymous OF. <cf>mappe</cf> (rare: also in
        Rousseau <I>c<R>1770), Sp. <cf>mapa</cf>, Pg.
        <cf>mappa</cf>, G. <cf>mappe</cf> (obs.: the mod. sense
        'portfolio' is not directly connected).</etym>
  <signif s4=4>
    <sen4 par=1 lit='1.' s6=4>
    <sengp>
      <sen6>
        <stxt id=1234567>A representation of the earth's surface or a
          part of it, its physical and political features, etc., or of the
          heavens, delineated on a flat surface of paper or other
          material, each point in the drawing corresponding to a
          geographical or celestial position according to a definite scale
          or projection.
          <snote>A hydrographical map is now more usually called
            a <cf>chart</cf> (formerly † <cf>card</cf>).
        <qbank id=1234567>
          <quot id=1234567>
            <qdat>1527
            <srce><auth>R. Thorne</auth> in Hakluyt
              <wk>Voy.</wk> (1589) 257
            <qtxt>To make a bigger and a better mappe.
          <quot id=1234567>
            <qdat>1589
            <srce><auth>G. Harvey</auth> <wk>Pierce's
              Super.</wk> Wks. (Grosart) II. 130
            <qtxt>The great Mapp of Mercator.
          <quot id=1234567>
            <qdat>1601
            <srce><auth>Shaks.</auth> <wk>Twel. N.</wk>
              <SC>iii.</SC> <R>ii. 84
            <qtxt>He does smile his face into more lynes, then is in the new
              Mappe, with the augmentation of the Indies.
          <quot id=1234567>
            <qdat>1625
            <srce><auth>N. Carpenter</auth> <wk>Geog. Del.</wk>
              <SC>i.</SC> <R>vii. (1635) 166
            <qtxt>A Geographicall Mappe is a plaine Table, wherein the
              Lineaments of the Terrestriall Spheare are expressed.
          <quot id=1234567>
            <qdat>1760
            <srce><auth>Johnson</auth> <wk>Idler</wk> No. 97 [5
            <qtxt>A rivulet not marked in the maps.
          <quot id=1234567>
            <qdat>1867
```

Fig. 10. SGML-mærket tekst (korrektur) fra the New Oxford English Dictionary; den hierarkiske struktur er tydeliggjort ved indrykninger. Efter Benbow (1986).

Betydningen af kategorinavne og "stenogrammer" fastlægges i en dokumenttypedefinition (DTD), som nøje gør rede for kategorierne og de hierarkier, de kan indgå i, samt for de benyttede tegnsæt.

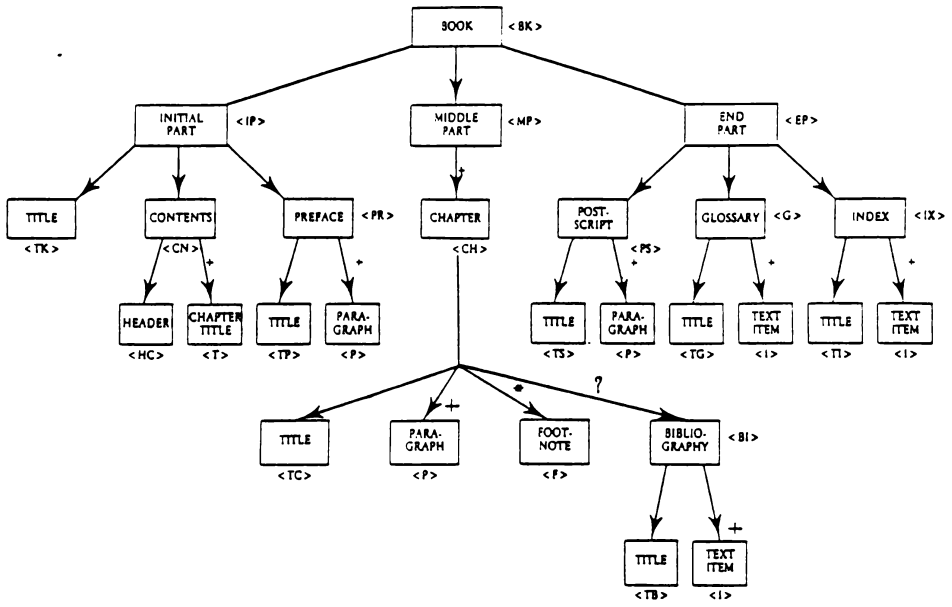


Figure 7

```
< !ELEMENT --          MIN  CONTENT --
  1 BK                 0 0  (IP, MP, EP)
  2 IP                 0 0  (TK, CN, PR)
  3 TK                 - 0  (#CDATA)
  4 CN                 0 0  (HC, T+)
  5 (HC|T)             - 0  (#CDATA)
  6 PR                 0 0  (TP, P+)
  7 (TP|P)             - 0  (#CDATA)
  8 MP                 0 0  (CH+)
  9 CH                 0 0  (TC, P+, F*, BI?)
 10 (TC|F)             - 0  (#CDATA)
 11 BI                 0 0  (TB, I+)
 12 (TB|I)             - 0  (#CDATA)
 13 EP                 0 0  (PS, G, IX)
 14 PS                 0 0  (TS, P+)
 15 TS                 - 0  (#CDATA)
 16 G                  0 0  (TG, I+)
 17 TG                 - 0  (#CDATA)
 18 IX                 0 0  (TI, I+)
 19 TI                 - 0  (#CDATA)
>
```

Figure 8

Fig. 11. En ret generel beskrivelse af dokumentstrukturen for en (fag)bog. Såvel i træet (øverst) som i dokumenttypedefinitionen (DTD) betegner "?" at det pågældende element forekommer ingen eller én gang, "*" at det forekommer ingen, én eller flere gange, og "+" at det forekommer én eller flere gange; umærkede elementer forekommer netop én gang. I DTD'en genskrives elementerne i venstre kolonne som vist i højre kolonne; "#CDATA" er en terminal kategori, nemlig tekst som ikke er yderligere opdelt, men kan rumme ethvert tegn fra et andetsteds defineret alfabet. Efter FORMEX (1985).

I dokumenttypedefinitionen kan det også fastlægges, at ikke alle kategorier behøver mærkning, idet bl.a. indledning af en sideordnet kategori, eller afslutning af en overordnet, vil være tilstrækkelig mærkning; kolonnen med "-" og "O" (= "omit") i DTD (fig. 11) angiver, i hvilken udstrækning dette skal være tilladt. Yderligere forenklinger kan opnås ved at definere forkortede mærkninger for hyppigt forekommende kategorinavne. Også de koder (kontroltegn m.fl.) som i tekstbehandlingsfiler markerer fx ny linie, nyt afsnit, tabulering og forskellige fremhævelser, kan defineres som forkortede SGML-mærkninger. Endvidere kan tekststrenge defineres som værende på én gang data og mærkninger. Med fuld udnyttelse af disse muligheder vil man næsten helt kunne undgå synlig mark-up selv i så kompliceret tekst som en ordbogsartikel (Erlandsen & Norling-Christensen 1988).

SGML-parseren er et program, som på basis af dokumenttypedefinitionen kontrollerer, at en given tekst er i overensstemmelse med den definerede struktur; desuden fuldstændiggør den mærkningen af tekster med reduceret / forenklet mærkning.

Mens jeg kan være i tvivl om, hvornår og i hvor høj grad ODA / ODIF slår igennem (den er som sagt meget ambitiøs), er det næsten sikkert, at SGML vil vinde frem. EF bruger det (FORMEX 1985); det er på vej ind i førende producenters tekstbehandlingssystemer; der er defineret dokumenttyper med bred anvendelighed, fx til sædvanlige videnskabelige afhandlinger (Smith 1987); og standarden er desuden taget i brug af store internationale forlag, især til videnskabelige tidsskrifter, som enten i deres helhed eller blot med abstracts og bibliografiske oplysninger skal indlægges i informationsbaser. Mærkningen betyder, at oplysninger om fx forfatternavn og -adresse, abstract, noter, bibliografi (og herunder de enkelte indførsler med forfatter, titel etc. identificeret), register m.v. kan genbruges i informationsbaser, selektive og generelle kataloger osv. Også til ordbogsarbejde er standarden taget i brug, først og fremmest af The New Oxford English Dictionary (Benbow 1986; Cowlishaw 1987; Descriptive Tools 1987:144-6); men flere SGML-baserede ordbogssystemer er på vej (Erlandsen & Norling-Christensen 1988). Gevinsten ved at anvende et system af denne art vil være, at ordbogsdataene kan bruges og genbruges, dels i forskellige medier, idet mange forskellige præsentationsformer kan afledes af den samme, medieuafhængige og veldefinerede, strukturbeskrivelse, dels i nye produkter som kombinerer data fra forskellige kilder.

9.3 Tegnstandarder

Hvorledes de 128 (7-bit kode) eller 256 (8-bit kode) forskellige bitkombinationer skal tolkes, afhænger af anvendelsen. De kan tolkes som heltal noteret i det binære talsystem (00000000 = 0; 01111111 = 127; 11111111 = 255), eller som bogstaver og andre grafiske tegn. Hvis en grafisk repræsentation ikke findes (visse kontroltegn), ikke er tilgængelig på en given printer eller skærm, eller det af andre grunde er ønskeligt entydigt at identificere en bitkombination, kan heltallet bruges, fx angivet som "<tal i titalssystemet>". Denne konvention er brugt i figg. 4, 7, 8 og 9, samt i det følgende.

ASCII (American Standard Code for Information Interchange) er den klassiske 7-bit standard; på et enkelt tegn nær ("$" =

<36>) er den identisk med den internationale referenceversion (fig. 12) af DS/ISO 646 (1974); denne standard er karakteristisk ved, at en række tegn ikke er internationalt definerede, idet det overlades til de nationale standardiseringsorganisationer at definere dem. I Danmark bruges de åbne pladser til æ, ø og å, i Tyskland til ä, ö, ü, ß , i Frankrig til ç og de øvrige accenterede bogstaver, osv. En sådan standard er anvendelig, så længe man holder sig til ét sprog, men klart uhensigtsmæssig, hvor flere sprog blandes.

#	Number sign	2/3
¤	Currency sign	2/4
@	Commercial at	4/0
[	Left square bracket	5/11
\	Reverse solidus	5/12
]	Right square bracket	5/13
{	Left curly bracket	7/11
|	Vertical line	7/12
}	Right curly bracket	7/13

Tegn (*Character*)	Position (Se ISO 646)	
	Kolonne (*Column*)	Række (*Row*)
Æ	5	11
Ø	5	12
Å	5	13
æ	7	11
ø	7	12
å	7	13

b7	b6	b5	b4	row \ column	0	1	2	3	4	5	6	7
0	0	0	0	0	NUL	TC. (DLE)	SP	0	@	P	`	p
0	0	0	1	1	TC (SOH)	DC	!	1	A	Q	a	q
0	0	1	0	2	TC, (STX)	DC,	"	2	B	R	b	r
0	0	1	1	3	TC, (ETX)	DC,	£(#)	3	C	S	c	s
0	1	0	0	4	TC. (EOT)	DC.	$(¤)	4	D	T	d	t
0	1	0	1	5	TC, (ENQ)	TC. (NAK)	%	5	E	U	e	u
0	1	1	0	6	TC. (ACK)	TC. (SYN)	&	6	F	V	f	v
0	1	1	1	7	BEL	TC. (ETB)	'	7	G	W	g	w
1	0	0	0	8	FE. (BS)	CAN	(	8	H	X	h	x
1	0	0	1	9	FE. (HT)	EM	)	9	I	Y	i	y
1	0	1	0	10	FE, (LF)	SUB	*	:	J	Z	j	z
1	0	1	1	11	FE, (VT)	ESC	+	;	K	@	k	@
1	1	0	0	12	FE, (FF)	IS. (FS)	,	<	L	@	l	@
1	1	0	1	13	FE, (CR)	IS, (GS)	-	=	M	@	m	@
1	1	1	0	14	SO	IS, (RS)	.	>	N	^ @@	n	- @@
1	1	1	1	15	SI	IS. (US)	/	?	O	_	o	DEL

Fig. 12. DS/ISO 646 (1974). Tegnene i kolonne 0 og 1 er de såkaldte kontroltegn; af de øvrige kan visse antage forskellige, sprogafhængige værdier. T.v. vises nederst de specielt danske værdier (DS 2089, 1974), øverst værdierne i standardens internationale referenceversion.

Dette tegnsæt kan udvides på to måder. For det første bestemmes det, at komma <44>, anførselstegn <34> m.fl. tegn, kombineret med <08> (BACKSPACE) repræsenterer cedille, trema, m.fl. diakritiske tegn: ligesom på en skrivemaskine "slår man to tegn oven i hinanden". Men desuden fastlægger en særlig standard, ISO 2022 (1986), hvorledes man kan bruge kontroltegnene <14> "Shift out", <15> "Shift in" og <27> "Escape" til at springe

til andre alfabet-definitioner, af hvilke der efterhånden findes adskillige.

Antallet af tilgængelige tegn er blevet udvidet gennem databehandlingens korte historie. Endnu for 10 år siden var 6-bit kode almindelig (64 tegn, kun versaler); i dag kan det meste udstyr håndtere 8-bitkoder og dermed op til 256 tegn ad gangen. Fælles for de ISO-standardiserede kodetabeller er det imidlertid, at grundstrukturen i ISO 646 bevares: hver tabel består af 32 kontroltegn og indtil 96 grafiske tegn. 8-bit standarderne kombinerer så blot to 7-bit tabeller, der lægges ved siden af hinanden. Herved bliver ikke blot kolonnerne 0 og 1, men også 8 og 9 reserveret til kontroltegn. Et eksempel herpå (fig. 13) er DS/ISO 8859-1 (1987), "Latinsk alfabet Nr. 1", som dækker de fleste nordiske og vesteuropæiske sprogs behov. ISO 8879 serien

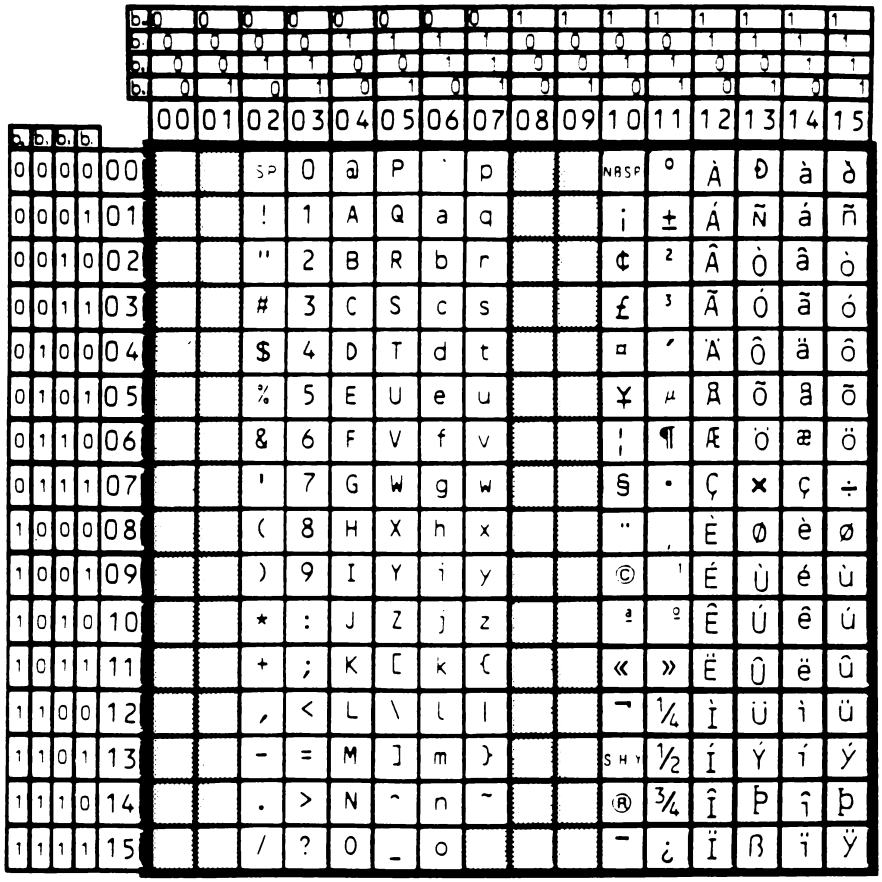


Fig. 13. DS/ISO 8859-1, Latinsk alfabet nr. 1. I denne som i den foregående figur findes et felts decimalværdi ved at gange kolonnens nummer med 16 og lægge liniens nummer til. "NBSP" (non-breaking space) er et blanktegn der ikke skal opfattes som ordmellemrum; "SHY" (soft hyphen) kan markere et potentielt orddelingspunkt. Venstre halvdel er ASCII alfabetet, højre halvdel et udvalg af andre latinske bogstaver m.m.

betragter "kombinerede bogstaver", dvs. ligaturer (fx dansk æ) og bogstaver med diakritiske tegn (fx ü), som enkelttegn. For at få plads til alle har en geografisk-sproglig opdeling i fire delvis ovelappende latinske alfabeter været nødvendig. Foruden nr. 1 rummer også nr. 4 (ISO/DIS 8859/4) de særlige danske tegn. Andre dele af denne serie og af en parallel serie, ISO 6937 (som bruger flertegnskombinationer for "kombinerede bogstaver"), omfatter desuden arabisk, græsk, hebraisk og kyrillisk. En del af de nævnte standarder er endnu ikke færdigbehandlede, men foreligger som forslag.

Standarder kan fastlægges på flere måder: af nationale eller internationale standardiseringsorganisationer, eller som "de facto industristandarder", hvor én tilstrækkelig indflydelsesrig producent (på dette område ofte IBM) sætter normen. De ovenfor omtalte, og specielt 8859-serien, understøttes - og er delvis udarbejdet - af ECMA (European Computer Manufactures Association); IBM's alfabeter afviger stærkt herfra. På større IBM-anlæg bruges EBCDIC (Extended Binary-Coded-Decimal Interchange Code), på PC'ere et alfabet, hvis USA-version ses anvendt på figg. 4-9. Med de netop introducerede operativsystemer (MS-)DOS 3.30 og OS/2 erstattes dette nu af en række nye "Code Pages", af hvilke nr. 850 (Multilingual; fig 14) nok vil dække de flestes behov. Den rummer stort set de samme tegn som DS/ISO 8859-1, men også her, som i det tidligere IBM-tegnsæt, er de er placeret helt andre steder i skemaet!

Hex Digits 1st → 2nd ↓	0-	1-	2-	3-	4-	5-	6-	7-	8-	9-	A-	B-	C-	D-	E-	F-
-0		►		0	@	P	`	p	Ç	É	á	▒	└	ð	Ó	
-1	☺	◄	!	1	A	Q	a	q	ü	æ	í	▒	┴	Ð	ß	±
-2	☻	↕	"	2	B	R	b	r	é	Æ	ó	▓	┬	Ê	Ô	‗
-3	♥	‼	#	3	C	S	c	s	â	ô	ú	│	├	Ë	Ò	¾
-4	♦	¶	$	4	D	T	d	t	ä	ö	ñ	┤	─	È	õ	¶
-5	♣	§	%	5	E	U	e	u	à	ò	Ñ	Á	┼	ı	Õ	§
-6	♠	▬	&	6	F	V	f	v	å	ù	ª	Â	ã	Í	µ	÷
-7	•	↨	'	7	G	W	g	w	ç	ù	º	À	Ã	Î	Þ	¸
-8	◘	↑	(	8	H	X	h	x	ê	ÿ	¿	©	╚	Ï	Þ	°
-9	○	↓	)	9	I	Y	i	y	ë	Ö	®	╣	╔	┘	Ú	¨
-A	◙	→	*	:	J	Z	j	z	è	Ü	¬	║	╩	┌	Û	·
-B	♂	←	+	;	K	[	k	{	ï	ø	½	╗	╦	▌	Ù	¹
-C	♀	∟	,	<	L	\	l	|	î	£	¼	╝	╠	▬	ý	³
-D	♪	↔	-	=	M	]	m	}	ì	Ø	¡	¢	═	¦	Ý	²
-E	♫	▲	.	>	N	^	n	~	Ä	×	«	¥	╬	Ì	´	■
-F	☼	▼	/	?	O	_	o	⌂	Å	ƒ	»	┐	¤	▀	¯	

Fig. 14. IBM's Code Page 850 (Multilingual); den minder om det tidligere PC-alfabet, men har flere accent-bogstaver; endvidere er yen- og centtegnene flyttet og har givet plads for Ø og ø.

10. Et dansk (eller nordisk?) udvekslingsformat?

Et veldefineret udvekslingsformat for maskinlæsbare tekster ville have den fordel, at der skulle foretages færre kodeanalyser og skrives færre konverteringsprogrammer (jf. 8. Praktiske erfaringer ovenfor), end hvis alt skal kunne konverteres til alt. Formatet kunne baseres på et udvalg af ISO-tegnsættene og på SGML. Jeg er i gang med, til intern brug på forlaget, at fastlægge et sådant format og hører derfor gerne, hvis nogen har kendskab til noget allerede eksisterende. Desuden indbydes andre interesserede til at deltage i fastlæggelsen, hvorved en bredere anvendelighed kunne sikres.

Tak for hjælp!

Geir Andresen, INGRAF, har venligst stillet upubliceret materiale til rådighed. Anna Braasch, EUROTRA-DK, samt Jane Rosenkilde Jacobsen og Hanne Ruus, begge Københavns Universitet, har gennemlæst manuskriptet og bidraget med værdifulde kommentarer.

Referencer

Amsler, Robert A. (1986), Deriving Lexical Knowledge Base Entries from Existing Machine-Readable Information Sources. Manuskript, dateret 16 May 1986, til The Workshop on Automating the Lexicon, Marina di Grosseto 19.-23. maj 1986.

Andresen, Geir (1987), Institutt for Grafisk Forskning, Oslo, Trinn ved Tekstformidling. Foredragsmanuskript til EPMARKUP, konference for The European Publishers Markup User Group, Amsterdam, 19.-20. februar 1987.

Benbow, Timothy (1986), The New Oxford English Dictionary Project: An Introduction, SGML Users' Group Bulletin vol.1, nr.2, 1986:65-74.

Carroll, Lewis (1946), Alice i Æventyrland og Bag Spejlet. Oversat af Kjeld Elfelt og Mogens Jermiin Nissen, København, Thorkild Becks Forlag, 1946.

Cave, Francis (1986), Typographic Markup Techniques. Fourth Working Draft. Udkast til British Standard. PIRA, Leatherhead, Surrey, England, 15.8.1986.

Cowlishaw, M.F. (1987), LEXX - A programmable structured editor, IBM Journal of Research and Development vol.31 nr.1, januar 1987:73-80.

Descriptive Tools (1987): The DANLEX Group (Ebba Hjorth, Jane Rosenkilde Jacobsen, Bodil Nistrup Madsen, Ole Norling-Christensen, Hanne Ruus), Studies in Computational Lexicography: Descriptive Tools for Electronic Processing of Dictionary Data., Lexicographica Series Maior 20. Tübingen, Max Niemeyer Verlag, 1987.

DS 2089 (1974): Dansk Standard. 7-bit kodet tegnsæt for databehandling. København, Dansk Standardiseringsråd 1974.

DS/ISO 646 (1974): Dansk Standard. 7-bit kodet tegnsæt for databehandling. København, Dansk Standardiseringsråd 1974.

DS/ISO 8859-1 (1974): Dansk Standard. Elektronisk informationsbehandling. 8-bit kodede grafiske tegnsæt. Del 1: Latinsk alfabet nr. 1 København, Dansk Standardiseringsråd 1.9.1987.

EDO (1988): Jens Axelsen, Engelsk-dansk Ordbog, 11. udgave, København, Gyldendal 1988.

Erlandsen, Jens, og Ole Norling-Christensen (1988), A SGML-Based Lexicographical Workstation. Foredrag til COLING 88, Budapest; endnu upubliceret.

FDO (1980): N.Chr.Sørensen, Fransk-dansk Ordbog, 8. udgave ved I.-L. Dalager, København, Gyldendal 1980.

FORMEX (1985): Formalized Exchange of Electronic Publications. Standard generalized mark-up language (SGML) as described in Appendix B of the FORMEX manual, Luxembourg, Office for Official Publications of the European Communities, 1985.

INGRAF (1985): INGRAFs anbefalte markeringssystem, Oslo, Institutt for Grafisk Forskning, september 1985.

ISO 2022 (1986): International Standard ISO 2022. Information processing - ISO 7-bit and 8-bit coded character sets - Code extension Techniques. 3. udgave, Schweiz, International Standard Organization maj 1986.

ISO 8879 (1986): International Standard ISO 8879. Information processing - Text and office systems - Standard Generalized Markup Language (SGML). Schweiz, International Standard Organization 15.10.1986.

ISO/DIS 8613 (1987): Forslag til International og Dansk Standard. ISO/DIS 8613 Elektronisk informationsbehandling. Teksthåndtering. Arkitektur og udvekslingsformat for kontordokumenter (ODA). København, Dansk Standardiseringsråd 1987.

Lenders, W. (1986), A Computer Aided Study in the Semantic Function of Verbs in Philosophical Texts, foredrag ved 13. International ALLC Conference, 1.-4. april 1986, Norwich, UK.

Møller, Gregers (1987), Kun et ud af tre sætterier, De Grafiske Fag 1/87 vol. 83:8-9.

ODSS (1987): Supplement til Ordbog over det danske Sprog. Prøvehæfte. Udgivet af Det danske Sprog- og Litteraturselskab, København, Gyldendal 1987.

Petersen, Pia Riber (1984), Nye Ord i Dansk 1955-75, København, Gyldendal 1984.

RDO (1985): Jørgen og Valentina Harrit: Russisk-dansk Ordbog, 2. udgave, København, Gyldendal 1985.

RO (1986): Retskrivningsordbogen. Udgivet af Dansk Sprognævn, København, Gyldendal 1986.

Smith, Joan M. (1986), Generic Markup of Documents the Standard Way, SGML Users' Group Bulletin vol.1 nr.1 1986:9-11. Heri også bibliografi (s. 8).

Smith, Joan M. (1987), The Standard Generalized Markup Language (SGML) for Humanities, Litterary and Linguistic Computing vol.2 nr.3, 1987:171-175.

Vail, Simon (1987), The Road to Conversion, Graphic Repro vol. 7, nr. 4, april 1987:30-35.

Vejl. (1987): Vejledning om valg af tekstbehandlingssystem. København, Administrationsdepartementet 1987.