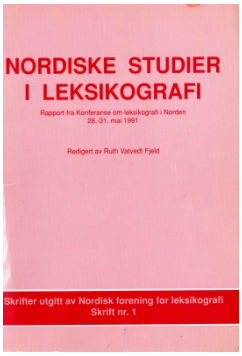
# NORDISKE STUDIER I LEKSIKOGRAFI


Titel: Struktureret redigering af ordbøger

Forfatter: Ole Norling-Christensen

Kilde: Nordiske Studier i Leksikografi 1, 1992, s. 447-454
Rapport fra Konferanse om leksikografi i Norden, 28.-31. mai 1991

URL: http://ojs.statsbiblioteket.dk/index.php/nsil/issue/archive



© Nordisk forening for leksikografi

**Betingelser for brug af denne artikel**
Denne artikel er omfattet af ophavsretsloven, og der må citeres fra den. Følgende betingelser skal dog være opfyldt:

- Citatet skal være i overensstemmelse med „god skik“
- Der må kun citeres „i det omfang, som betinges af formålet“
- Ophavsmanden til teksten skal krediteres, og kilden skal angives, jf. ovenstående bibliografiske oplysninger.


**Søgbarhed**
Artiklerne i de ældre Nordiske studier i leksikografi (1-5) er skannet og OCR-behandlet. OCR står for ’optical character recognition’ og kan ved tegngenkendelse konvertere et billede til tekst. Dermed kan man søge i teksten. Imidlertid kan der opstå fejl i tegngenkendelsen, og når man søger på fx navne, skal man være forberedt på at søgningen ikke er 100 % pålidelig.

Ole Norling-Christensen

# Struktureret redigering af ordbøger

Det standardiserede strukturbeskrivelsessprog SGML gør det muligt at beskrive vilkårligt komplicerede teksttrukturer og at opmærke teksten i overensstemmelse med beskrivelsen. Herved adskilles præsentation og indhold, og teksten bliver velegnet til alle slags datamatisk behandling, herunder syntakskontrol, komplicerede søgninger, automatiske ændringer; genbrug af definerede dele af en ordbogs oplysninger, præsentation af samme tekst med forskelligt lay-out. Teknikken og nogle praktiske anvendelser beskrives med en dansk-engelsk ordbogsartikel som gennemgående eksempel.

En del af meningen med de synspunkter og den teknik, som her skal introduceres, er at flytte fokus for leksikografens opmærksomhed fra formen henimod indholdet. Under det daglige ordbogsarbejde skal der ikke længere tænkes i typografiske kategorier, men i indholdsmæssige. Én gang for alle (omend med mulighed for senere ændringer) har leksikografen beskrevet væsentlige dele af sine redaktionsregler i et formelt sprog, gjort rede for, hvilke oplysningstyper ordbogen skal indeholde og for deres indbyrdes sammenhæng. Da de indholdmæssige kategorier således på forhånd er fastlagt, kan leksikografen herefter koncentrere sig om, hvilke af kategorierne der skal tages i brug i hvert enkelt tilfælde, og hvad der skal puttes i dem. Denne målsætning kan opfyldes på lige så mange måder, som der findes databasesytemer; men hvis eksisterende ordbøgers (makro- og) mikrostruktur skal indpasses i et databasesystem, når man oftest til så komplicerede modeller, at det nærmer sig det prohibitive. Lettere er det, når man starter forfra med et helt nyt ordbogsprojekt.

Vilkårligt komplicerede strukturer kan imidlertid beskrives i SGML, Standard Generalized Markup Language, og SGML kan bruges til en eksplicit strukturmarkering af ordbogsteksten. En ny tekst kan indskrives, og en eksisterende tekst rettes til, med en særlig SGML-editor, som understøtter strukturmarkeringen. Et syntakskontrolprogram, SGML-parseren, kan kontrollere om markeringerne i teksten er i overensstemmelse med strukturbeskrivelsen. Den strukturmarkerede tekst er velegnet til alle slags datamatisk behandling, herunder: indlæggelse i databaser; komplicerede søgninger; automatiske ændringer; udtrækning af definerede dele af ordbogens oplysninger til brug fx i andre ordbøger; præsentation ikke blot i trykt form, men også fx på dataskærm. På basis af DANLEX-gruppens undersøgelser og anbefalinger (DANLEX 1987) har softwarehuset TEXTware A/S og forlaget Gyldendal i fællesskab udviklet et redigeringssystem, GestorLEX, til ordbøger. Det kombinerer SGML-editor, -parser, og en lang række andre faciliteter.

Strukturbeskrivelsessproget SGML er en international standard (ISO 8879, 1986) beregnet til at beskrive, og styre behandlingen af, alle slags "dokumenter", fx forretningsbreve, EF-cirkulærer, ganske almindelige bøger, og altså også ordbøger. Dokumenterne beskrives som en træstruktur af *elementer*; et element kan bestå af andre elementer og/eller stumper af egentlig tekst, som er træets blade. Dokumentet er selv et element, træets rod.

## Indhold, struktur, præsentation

En grundlæggende tanke bag SGML er at adskille præsentation og indhold. Man beskriver, hvilken *slags* oplysninger dokumentet består af, og hvordan de hænger sammen, altså en *generisk* (artsmæssig) beskrivelse og kodning. Man beskriver derimod ikke, hvordan disse oplysninger skal præsenteres på tryk eller fx på en skærm. Præsentationsoplysningerne (skrifter, skriftstørrelser, interpunktionstegn og andre separatorer, spring til ny side eller nyt afsnit, osv.) kan føjes til bagefter som en funktion af strukturen. Til forskellig brug af samme dokument kan dettes strukturoplysninger meget vel oversættes til forskellige sæt af præsentationsoplysninger, uden at der overhovedet skal ændres i selve dokumentet. Et eksempel herpå er de danske Gyldendals Elektroniske Ordbøger og de tilsvarende norske fra Kunnskapsforlaget og svenske fra Almqvist & Wiksell (tidl.: Esselte Studium): det til grund liggende dokument er det samme; men skærmbilledet, fx Axelsen (1990), og bogsiden, fx Axelsen (1984), ser meget forskellige ud.

## Dokumenttypedefinitionen, DTD

Et fuldstændigt SGML-dokument består af en dokumenttypedefinition, DTD, som angiver reglerne for dokumentets struktur, samt selve dokumentet, hvori start og slut på hvert enkelt element er markeret. Et simpelt eksempel på en DTD er den generelle beskrivelse i *figur 1* af en (fag)bog; den er lånt fra (FORMEX 1985), men ændret en smule.

```
<!ELEMENT  BK          (IP, MP, EP) >
<!ELEMENT  IP          (TK, CN, PR) >
<!ELEMENT  TK          (#PCDATA) >
<!ELEMENT  CN          (HC, T+) >
<!ELEMENT  (HC, T)     (#PCDATA) >
<!ELEMENT  PR          (TP, P+) >
<!ELEMENT  (TP, P)     (#PCDATA) >
<!ELEMENT  MP          (CH+) >
<!ELEMENT  CH          (TC, P+, F*, BI?) >
<!ELEMENT  (TC, F)     (#PCDATA) >
<!ELEMENT  BI          (TB, I+) >
<!ELEMENT  (TB, I)     (#PCDATA) >
<!ELEMENT  EP          (PS, G, IX) >
<!ELEMENT  PS          (TS, P+) >
<!ELEMENT  TS          (#PCDATA) >
<!ELEMENT  G           (TG, I+) >
<!ELEMENT  TG          (#PCDATA) >
<!ELEMENT  IX          (TI, I+) >
<!ELEMENT  TI          (#PCDATA) >
```

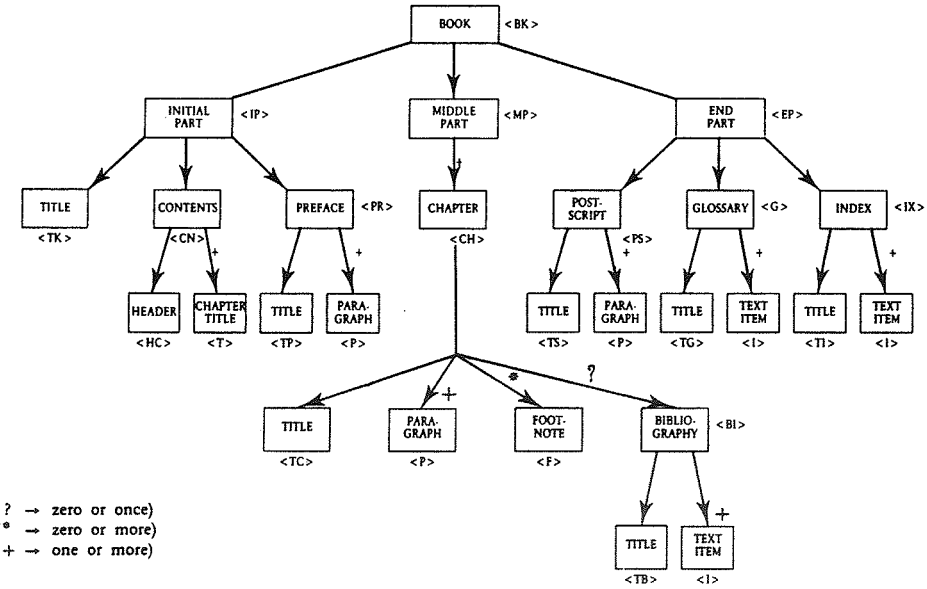
*Figur 1.* Dokumenttypedefinition (DTD) for en fagbog, jf. figur 2. (Efter FORMEX 1985)

DTD'en i figur 1 skal forstås således, at hvert element (venstre kolonne) består af ("genskrives som") de elementer, der står i parentesen i højre kolonne, i den anførte rækkefølge. #PCDATA symboliserer den egentlige tekst. Efter et elementnavn på højre side kan der stå et plus (+), en stjerne (*) eller et spørgmålstegn (?); de har følgende betydning:

+ Elementet skal forekomme mindst én gang, men kan forekomme flere gange efter hinanden;
* Elementet kan udelades, eller det kan forekomme én eller flere gange efter hinanden;
? Elementet kan udelades, eller det kan forekomme højst én gang.

Et umarkeret element skal forekomme netop én gang på den angivne plads.

De samme symboler er anvendt i den grafiske afbildning (figur 2) af (fag)bogens struktur. Fx består et kapitel <CH> af én titel <TC>, flere (+) afsnit <P>, måske nogle (*) fodnoter <F>, og måske én (?) bibliografi <BI>.




*Figur 2.* Struktur af en fagbog, jf. figur 1. (Fra FORMEX 1985)

## SGML og ordbøger

Som gennemgående eksempel i det følgende anvendes artiklen *afgjort* (figur 3) fra en dansk-engelsk ordbog (Axelsen 1984), som jeg har konverteret fra fotosatsdata til SGML-

strukturerede data, som dernæst er importeret til GestorLEX redigeringssystemet med henblik på udarbejdelsen af en ny udgave.

**afgjort** *adj (som er gået i orden)* settled; *(udpræget)* definite, F decided *(fx* advantage, improvement); *(om person)* decided *(fx* she was very decided); *adv* definitely, F decidedly *(fx* better); unquestionably *(fx* he is unquestionably the best man); *en ~ sag* a settled thing; *det er (så godt som) ~ at* it is (as good as) settled that.

*Figur 3.* Artiklen *afgjort* fra Jens Axelsen: Dansk-engelsk Ordbog (Axelsen 1984)

Ordbogen flyttes altså fra sætteriets tekstbehandlingssystem, hvis typografiske kodning (figur 4) skal omsættes til en generisk. Samtidig har redaktionen udtrykt ønsker om, at præsentationen af ordbogens oplysninger bliver mere eksplicit.

| | | | |
|---|---|---|---|
| FED | **afgjort** | | (fx |
| KURSIV | *adj* | ORDINÆR | she was very decided |
| | ( | | ); |
| KURSIV | *som er gået i orden* | KURSIV | *adv* |
| | ) | ORDINÆR | definitely |
| ORDINÆR | settled | | , |
| | ; ( | GROTESK | F |
| KURSIV | *udpræget* | ORDINÆR | decidedly |
| | ) | | (fx |
| ORDINÆR | definite | ORDINÆR | better |
| | , | | ); |
| GROTESK | F | ORDINÆR | unquestionably |
| ORDINÆR | decided | | (fx |
| | (fx | ORDINÆR | he is unquestionably the best man |
| ORDINÆR | advantage | | ); |
| | , | KURSIV | *en ~ sag* |
| ORDINÆR | improvement | ORDINÆR | a settled thing |
| | ); ( | | ; |
| KURSIV | *om person* | KURSIV | *det er (så godt som) ~ at* |
| | ) | ORDINÆR | it is (as good as) settled that |
| ORDINÆR | decided | | . |

*Figur 4.* Ordbogsartiklen opdelt i strukturmarkører og egentlig tekst.

Figur 5 viser en analyse (blandt mange mulige) af ordbogsartiklens opbygning. Grundlæggende består en ordbogsartikel af et **Hoved** med oplysninger om selve opslagsordet (her: ordklasse; men det kunne også være fx bøjning og udtale); en **Krop** med oversættelser af selve opslagsordet; og en **Fod** (andre ordbogsprojekter kalder dette afsnit for **Hale**) med mere eller mindre faste ordforbindelser, herunder egentlige idiomatiske udtryk. I det her viste eksempel er der to **Kroppe** og et indskudt "**Hoved1**". Det skyldes et redaktionelt ønske om, at ny ordklasse skal udløse ny artikel. På et senere stadium i konverteringsprocessen vil der fremkomme to adskilte artikler hver med sit homografnummer.

I den oprindelige artikel (figur 3) er de forskellige oversættelser af "afgjort" adskilt af komma eller semikolon; ofte er de desuden mærket med en **TydSpecifikation** (kursiv tekst i parentes) eller en **Sigel**, en versal i særlig skrift (grotesk). I eksemplet forekommer **F** = formelt sprog; i øvrigt bruger ordbogen bl.a. **S** = slang og **T** = talesprog. Der er grund til at antage, at variationerne (komma el. semikolon; TydSpec/Sigel el. ingenting) antyder et semantisk hierarki: de forskellige **Tyd** (en term, som skyldes Otto Jespersen) har betydningsmæssigt mere eller mindre med hinanden at gøre.

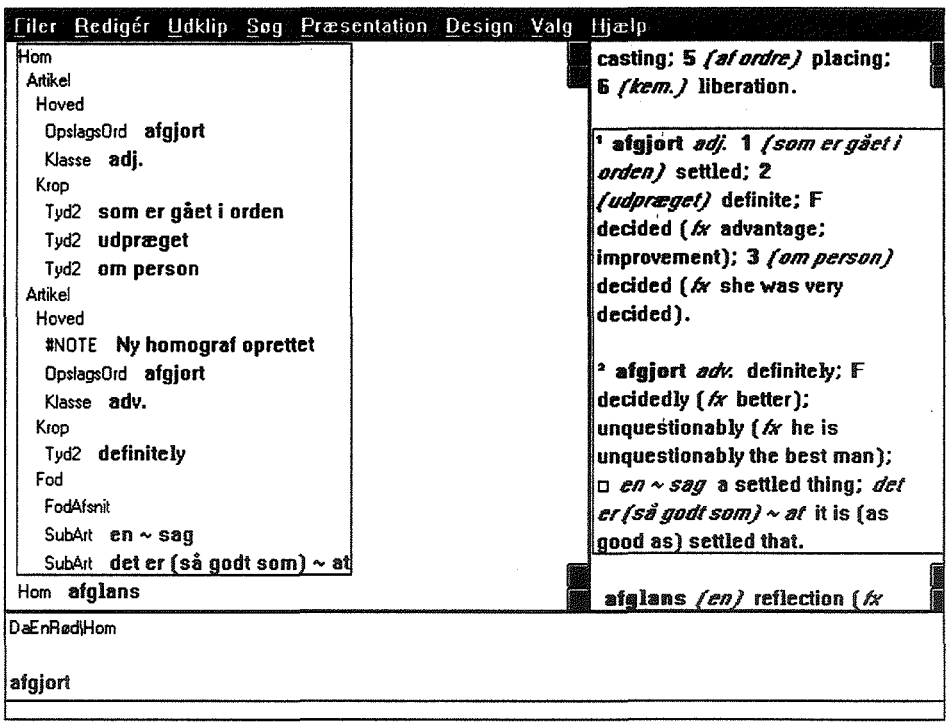
| | | | | | |
|---|---|---|---|---|---|
| Hoved | OpslagsOrd | | | | afgjort |
| | Klasse | | | | adj |
| Krop | Tyd2 | TydSpec | | | som er gået i orden |
| | | Tyd0 | Ækvival | | settled |
| | Tyd2 | TydSpec | | | udpræget |
| | | Tyd0 | Ækvival | | definite |
| | | Tyd0 | Sigel | | F |
| | | | Ækvival | | decided |
| | | | Eksempel | | advantage |
| | | | Eksempel | | improvement |
| | Tyd2 | TydSpec | | | om person |
| | | Tyd0 | Ækvival | | decided |
| | | | Eksempel | | she was very decided |
| Hoved1 | Klasse | | | | adv |
| Krop | Tyd2 | Tyd0 | Ækvival | | definitely |
| | | Tyd0 | Sigel | | F |
| | | | Ækvival | | decidedly |
| | | | Eksempel | | better |
| | | Tyd0 | Ækvival | | unquestionably |
| | | | Eksempel | | he is unquestionably the best man |
| Fod | SubArt | Udtryk | | | en ~ sag |
| | | Tyd0 | Ækvival | | a settled thing |
| | SubArt | Udtryk | | | det er (så godt som) ~ at |
| | | Tyd0 | Ækvival | | it is (as good as) settled that |

*Figur 5.* Strukturanalyse af artiklen *afgjort*

Efter aftale med ordbogens redaktør er det besluttet at strukturere tydene således: De enkelte **Tyd2** i en artikel har ikke meget til fælles; i den kommende ordbogsudgave skal flere **Tyd2** efter hinanden adskilles med semikolon og nummereres med arabertal. Én **Tyd2** kan omfatte en eller flere **Tyd0**, som semantisk er nærmere beslægtede; er der flere **Tyd0** under samme **Tyd2** adskilles de blot med semikolon. Det flertydige komma afskaffes altså som separator

mellem forskellige tyd. Den opmærksomme læser vil savne en **Tyd1**. Den er foreløbig ikke taget i brug, men er reserveret til en eventuel underinddeling af **Tyd2**, typografisk fx markeret med a), b), c) ...

Processen som overfører de typografiske data til SGML-formaterede data indebærer bl.a. en klargøring af de typografiske data, jf. Norling-Christensen (1988), samt en parsing, som er analog med maskinel syntaks-analyse af naturligt sprog. Den beskrives nøjere i Norling-Christensen (1992); her skal blot antydes, at hovedreglen for skift til ny **Tyd2** er, at der både forekommer et semikolon som separator og en **TydSpecifikation** eller en **Sigel**. Hvis der på **TydSpecifikationens** plads optræder en ny ordklasseangivelse, oprettes der derimod en helt ny artikel; for at fuldstændiggøre denne kopieres opslagordet dertil. **Hoved1** i figur 4 erstattes altså med et normalt **Hoved** bestående af opslagsordet "afgjort" og ordklasse-betegnelsen "adv".




*Figur 6.* Skærmbillede fra redigeringssystemet GestorLEX. I strukturvinduet t.v. ses artiklerne *afgjort* delvis udfoldet (til og med niveau 4). Tekstvinduet t.h. viser ordbogsteksten i et typografisk format, som automatisk genereres ud fra strukturen og et sæt præsentationsregler.

Processen omfatter også en konvertering af forkortelser etc. til en vedtagen standard; den indebærer bl.a. at alle forkortelser nu afsluttes af forkortelsespunktum. Resultatet af den automatiske proces ses i figur 6, som samtidig illustrerer, at det kun er forsvarligt at gennemføre automatiske ændringer af ordbøger, hvis der efterfølgende foretages en menneskelig kontrol. Bemærk nemlig, at **Foden** udelukkende rummer prøver på adjektivets brug; den skal altså flyttes op, hvor den hører til. I GestorLEX vil dette dog kun kræve nogle få tastetryk. For at sikre, at redaktøren er opmærksom på, hvor mekanikken har ændret hans tekst, sættes der noter ind, som han kan søge efter og tage stilling til.

### Ordbogsartiklen som SGML-struktur

Efter opdelingen i to artikler med samme opslagsord (homografer) kan artikelstrukturen beskrives med den DTD, som vises i figur 7. Det overordnede element **Hom** består af én eller flere **Artikel**; hvis der er flere, nummererer systemet dem med højtstående arabertal (homografnumre). I øvrigt er elementnavnene de samme som på figur 5.

```
<!ELEMENT Hom (Artikel+) >
<!ELEMENT Artikel (Hoved, Krop, Fod?) >
<!ELEMENT Hoved (OpslagsOrd, Klasse) >
<!ELEMENT (Opslagsord, Klasse) (#PCDATA) >
<!ELEMENT Krop (Tyd2+) >
<!ELEMENT Tyd2 (TydSpec?, Tyd0) >
<!ELEMENT Tyd0 ( (Sigel | TydSpec)?, Ækvival, Eksempel*) >
<!ELEMENT (TydSpec, Sigel, Ækvival, Eksempel) (#PCDATA) >
<!ELEMENT Fod (SubArt+) >
<!ELEMENT SubArt (Udtryk, Tyd0+) >
<!ELEMENT Udtryk (#PCDATA) >
```

*Figur 7.* En (forenklet) dokumenttypedefinition (DTD) for Dansk-engelsk Ordbog. Den lodrette streg i "(Sigel | TydSpec)" betegner "enten/eller". Flere elementnavne på venstre side i en genskrivningsregel, fx "(Opslagsord, Klasse)", er blot en forkortet skrivemåde for, at disse elementer genskrives på samme måde.

SGML er som nævnt en international standard. Det indebærer, at ordbogsdata, som er lagret i SGML-format, ikke blot kan importeres til, benyttes i, og eksporteres fra GestorLEX-systemet. De kan også udveksles mellem forskellige ordbogsprojekter, og de er lette at bearbejde datamatisk, også uden for GestorLEX. En nøjere beskrivelse af GestorLEX falder iøvrigt uden for denne artikels rammer. Her skal blot nævnes, at systemet er detaljeret beskrevet i TEXTware (1991), og at det opfylder praktisk taget alle de krav som opstilles i DANLEX (1987:239-251).

**Litteratur**

Axelsen, Jens. 1984. *Dansk-engelsk Ordbog.* 9. udgave (7. oplag 1991). København.

Axelsen, Jens. 1990. *Dansk-engelsk Ordbog.* 1. elektroniske udgave. København.

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

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 Europaen Communities.

ISO 8879. 1986. *International Standard ISO 8879. Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML).* Genève.

Norling-Christensen, Ole. 1988. Læsning af maskinlæsbare tekster. I: *Nordiske Datalingvistdage og Symposium for datamatstøttet leksikografi og terminologi 1987. Proceedings.* Institut for Datalingvistik, Handelshøjskolen i København.

Norling-Christensen, Ole. 1992. *Parsing Dictionary Data.* Et oplæg til Workshop on Tools and Methods for Practical Lexicography, Fifth EURALEX International Congress 4.-9. august, Tampere, Finland. (Under udarbejdelse).

TEXTware. 1991. *LexWrite Brugermanual.* Version 2: April 24, 1991. København. Og *LexDesign Brugermanual.* Version 2: April 29, 1991. København.