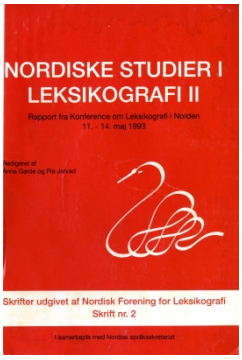
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Ordbogsredaktion med COMPULEXIS hos Munksgaard

Forfatter: Anders Geertsen

Kilde: Nordiske Studier i Leksikografi 2, 1993, s. 87-95
Rapport fra Konference om leksikografi i Norden, 11.-14. maj 1993

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



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.

Ordbogsredaktion med COMPULEXIS
hos Munksgaard

Anders Geertsen

Ethvert ordbogsprojekt starter med at forlag, hovedredaktør og forfatterne diskuterer redaktionsprincipperne for ordbogen. Disse redaktionsprincipper skrives ned og redigeres, så de mere og mere udgør en redaktionsvejledning, dvs et dokument, som kan blive en daglig manual for alle involverede i ordbogen, men mest af alt naturligvis en manual for forfatteren/forfatterne. En vigtig opgave for redaktionsvejledningen er at fastlægge principperne for hvordan de enkelte ordbogsartikler opbygges.

Feltopdelt data

Helt grundlæggende betragter vi en ordbogsartikel som en samling mindre elementer, hver med deres leksikografiske indhold. Hvis man arbejder med databaser ville man sige, at hver artikel er en "record", og at denne record kan bestå af en række "felter". En typisk ordbogsartikel kan fx se således ud i den trykte ordbog:

> **arm**[3] *vt & refl* **1.** bevæbne, opruste **2.** forsyne, udruste (*armed with the truth*) **3.** armere, klargøre (*fx bombe*) ■ ***in arms*** under våben, (op)rustet; ***lay down one's arms*** nedlægge våbnene; ***take up arms*** gribe til våben; ***under arms*** under våben; ***be up in arms*** (sædv. fulgt af *against, about*) øve væbnet modstand imod

Når forfatterne og forlaget arbejder med denne artikel vil vi samtidig betragte den som en sekvens af felter, hver med deres specielle type information. Hvert felt har i følgende eksempel en lille tre-bogstavskode:

HWD	arm
HOM	3
PSA	vt
PSA	refl
LV3	1
TSL	bevæbne
TSL	opruste
LV3	2
TSL	forsyne

TSL	udruste
EXA	armed with the truth
LV3	3
TSL	armere
TSL	klargøre
SEB	fx bombe
IDM	in arms
TSL	under våben
TSL	(op)rustet
IDM	lay down one's arms
TSL	nedlægge våbnene
IDM	take up arms
TSL	gribe til våben
IDM	under arms
TSL	under våben
IDM	be up in arms
GRA	sædv. fulgt af *against, about*
TSL	øve væbnet modstand imod

En ordbogs redaktionsvejledning skal helt overordnet fastlægge og beskrive hvilke typer elementer eller felter, der skal kunne indgå i en typisk artikel. Visse felter vil være til stede i alle artikler, mens andre, fx feltet IDM med eventuelle idiomatiske udtryk, kun vil optræde i nogle artikler. Samtidig beskriver redaktionsvejledningen hvordan indholdet af de forskellige felter vil komme til at se ud på tryk i den færdige ordbog.

Mens det er en forholdsvis enkel opgave at fastlægge hvilke felter, der skal anvendes for at kunne opbygge de enkelte artikler i ordbogen, er det en langt mere kompliceret opgave at opstille et fornuftigt og entydigt regelsæt for rækkefølgen af felter, iterationer af felter, hierarkier og autoriseret input i de enkelte felter. Lad mig give eksempler på disse fire problemstillinger:

Rækkefølgen af felterne. Eksempel: Et idiomatisk udtryk (IDM) vil ofte være efterfulgt af en grammatisk oplysning (GRA), og derefter af en oversættelse (TSL). Vi har derfor en typisk sekvens af tre felter: IDM + GRA + TSL:

...	
IDM	be up in arms
GRA	sædv. fulgt af *against, about*
TSL	øve væbnet modstand imod
...	

Hvis denne sekvens optræder hyppigt, og hvis vi bevidst ønsker netop denne rækkefølge, frem for en anden, må vi beskrive den i redaktionsvejledningen, og eventuelt sikre, at vores software kan genkende den og favorisere den frem for andre mønstre af felter.

Iterationer af mindre grupper af felter. Eksempel: En artikel vil i mange tilfælde blive opdelt i flere mindre afsnit, hver indledt af et arabertal. Man kan derfor betragte arabertallet med dets efterfølgende oversættelser som en fasttømret gruppe, som kan gentages efter behov:

```
...
LV3     1
TSL     bevæbne
TSL     opruste

LV3     2
TSL     forsyne
TSL     udruste
...
```

Når man opbygger nye ordbogsartikler, eller redigerer i bestående, skal man være klar over, at et enkelt element i en sådan fasttømret gruppe ikke kan slettes, flyttes eller forandres, uden at det har indflydelse på de andre elementer i gruppen. Til gengæld kan en sådan gruppe ofte flyttes til et andet sted i artiklen, blot man flytter alle felter i gruppen som en samlet blok. Igen skal redaktionsvejledning og software gøre forfatteren opmærksom på dette.

Hierarkiet af felter. Eksempel: Visse felter, eller grupper af felter, er underordnet andre felter, eller grupper af felter. Et brugseksempel tilknyttet en arabertalsgruppe er underordnet denne gruppe. Denne relation mellem overordnede, sideordnede og underordnede elementer kan måske bedst illustreres ved hjælp af indryk:

```
HWD     arm
HOM     3
PSA     vt
PSA     refl

        LV3  1
                    TSL     bevæbne
                    TSL     opruste

        LV3  2
                    TSL     forsyne
                    TSL     udruste

                            EXA  armed with the truth
...
```

Forholdet mellem over-, side- eller underordning af elementer er centralt for den typo-

grafiske manifestation af ordbogsartiklen. Populær sagt vil underordnede elementer ofte arve typografiske egenskaber fra overordnede elementer.

Autoriseret input i visse felter. Eksempel: Oplysninger om opslagsordets ordklasse hører hjemme i feltet "PSA" (Part of Speech A-language):

```
...
PSA     vt
PSA     refl
...
```

En opgave for redaktionsvejledning og software er naturligt nok at fastlægge hvilke forkortelser der anvendes for hvilke ordklasser, dvs fastlægge autoriseret input i PSA felterne.

Redaktionsvejledningen

Artiklernes opbygning, anvendte felter, typiske sekvenser og iterationer af felter, hierarkier m.v. forklares og beskrives grundigt og ved hjælp af mange eksempler i redaktionsvejledningen. Vi gør et stort arbejde for at skrive så grundige og detaljerede redaktionsvejledninger som muligt. Brugen af hvert enkelt felt forklares indgående, og vi giver eksempler på typiske sekvenser af felter, iterationer, mønstre m.v.. Det følgende viser et udsnit fra en redaktionsvejledning, som beskriver hvordan opslagsord, som er forkortelser, behandles ved hjælp af felterne FFA:

"Feltet Full Form anvendes når et opslagsord eller en oversættelse er en forkortelse. Både for opslagsord, som er forkortelser, og for oversættelser, som er forkortelser, kan det være ønskeligt at kunne angive den fulde form.

Til den fulde form af A-sprogs ord (dvs opslagsord) anvendes feltet FFA. Til den fulde form af B-sprogs ord (dvs oversættelser) anvendes feltet FFB:

I CX-Basic:

```
  HWD     EEC
  PSA     forkort
→FFA      European Economic Community
  TSL     EØF
→FFB      Det Europæiske Økonomiske Fællesskab
```

I den trykte ordbog: **EEC** *forkort = European Economic Community* EØF = *Det Europæiske Økonomiske Fællesskab*"

Ensartet feltstruktur for alle vores ordbøger

Selv om ordbøger er - og skal være - forskellige, har de mange iboende lighedspunkter. Enhver bilingval ordbog opererer med et A- og et B-sprog, med oversættelser (ækvivalenter), brugseksempler og idiomatiske udtryk. Ordklasseoplysninger og semantiske etiketter vil også være at finde i de fleste tosprogsordbøger. Vi kan derfor i meget høj grad ensrette brugen af felterne på tværs af alle vores ordbøger. Konkret betyder det, at et idiom altid står i et IDM felt, og at en ordklasseoplysning for A-sprogsord altid står i et PSA felt osv. Denne enighed gælder ikke blot feltnavnene, men også hele regelsættet for typiske iterationer, sekvenser af felter og hierarkiske relationer mellem felter. Og enigheden gør det muligt at lagrer alle vores ordbøger i én fælles konsensus-database. Sammenligninger af lemma-udvalg eller idiomatiske udtryk, vending af ordbøger m.v. kan foretages ud fra denne konsensus-database.

Redaktionsværktøjer

Alle Munksgaards ordbøger bliver redigeret og lagret i et specielt ordbogssystem, COMPULEXIS Dictionary System. Dette system, der er udviklet af COMPULEXIS i Oxford, har to hovedelementer: **CX-Basic**, der bygger på WordPerfect 5.1 og som bruges til indtastning/redigering af data, og **CX-Master**, som bygger på en ORACLE SQL database, og som bruges til at samle alle ordbogsdata, vedligeholde konsensus-databasen, generere korrekturprint og levere data til sats.

CX-Basic

CX-Basic er indtastnings- og redigeringsdelen af COMPULEXIS. Forfatterne vil normalt arbejde i CX-Basic. Kernen i CX-Basic er tekstbehandlingsprogrammet WordPerfect, version 5.1. Ved hjælp af en række makrofunktioner er denne WordPerfect-udgave specielt velegnet til at redigere feltopdelte data. For en forfatter ser en ordbogsartikel således ud i CX-Basic:

HWD	arm
HOM	3
PSA	vt
PSA	refl
LV3	1
TSL	bevæbne
TSL	opruste
LV3	2
TSL	forsyne
TSL	udruste
EXA	armed with the truth
...	

CX-Basic kan installeres og afvikles fra enhver IBM-kompatibel pc med DOS styresy-

stem. CX-Basic gør det muligt dels at redigere i eksisterende ordbogsartikler, dels at indtaste nye ordbogsartikler. Da CX-Basic er baseret på WordPerfect 5.1 indeholder CX-Basic ikke nogen egentlig database. CX-Basic producerer almindelige WordPerfect tekstfiler med feltopdelt data, som er beregnet til at blive importeret i CX-Master databasen, som er installeret hos Munksgaard.

CX-Master

CX-Master er den del af COMPULEXIS, hvor alle ordbogsdata lagres, alfabetiseres og systematiseres, og hvorfra forlaget kan generere satsfiler til de trykte ordbøger. Samtidig byder CX-Master på en række søgemuligheder, som gør det let at rette fejl i data, checke krydsreferencer m.v.

I CX-Master gemmes alle ordbogsdata i en SQL database. Alle de ordbogsdata, som redigeres af forfatterne i CX-Basic, vil blive importeret i CX-Master, og her gemt i databasen. Samtidig kan CX-Master eksportere data til CX-Basic i form af WordPerfect filer.

CX-Master kan vise - og printe - hvordan de ordbogsfiler, som forlaget har modtaget fra forfatterne i CX-Basic format, vil se ud på tryk i den færdige ordbog. Denne funktion anvendes bl.a. til at generere korrekturprint til forfatterne i takt med at arbejdet skrider frem. En anden mulighed er at lade forfatterne selve gererere disse korrekturudskrifter direkte fra CX-Basic.

Manifestationslogik

Både i CX-Basic og i CX-Master indtastes al tekst i ordinær, uden angivelse af fx fed, kursiv m.v. Tilsvarende skal forfatterne normalt ikke indsætte nogen form for tegnsætning: punktum, komma, semikolon, parenteser og lignende. Såvel skriftsnit som tegnsætning genereres automatisk af CX-Master eller CX-basic når filerne printes. Skriftsnit og tegnsætning skabes på basis af en analyse af:

- Feltkoden.
- Feltets placering blandt andre felter, dvs en analyse af iterationer, sekvens og hierarki.

Da hver leksikografisk oplysningstype har sit eget felt, kan man til enhver tid ændre skriftsnittet for en bestemt type oplysning. Eksempelvis sætte alle ordklasseangivelser i parentes - eller fjerne parentesen og sætte dem i petit.

Typisk arbejdsproces

Redaktionen af en ordbog består typisk af følgende etaper:

a) Eksport af data fra CX-Master hos Munksgaard

Munksgaard opbygger og vedligeholder i CX-Master databaser med bl.a. ordbøgerne i serien Munksgaards Ordbøger. I mange tilfælde vil dele af disse data skulle indgå i revisionen af en ordbog. Eller som halvfabrikata i en ny ordbog. Ordbogsarbejdet vil derfor ofte starte med eksport af ordbogsdata fra CX-Master.

Data bliver eksporteret i det WordPerfect format, som direkte kan anvendes i CX-Basic. Samtidig bliver data delt op i en række mindre filer, typisk à 100 Kb, så behandling af filerne i CX-Basic hos forfatterne bliver hurtig og bekvem.

b) Import af data i CX-Basic

Forfatterne vil modtage ordbogsartiklerne som CX-Basic filer (WordPerfect format) på disketter fra Munksgaard. Med en simpel DOS kommando lægges disse filer ind på harddisken på forfatterens pc.

c) Redaktion af data i CX-Basic

Forfatterne redigerer de enkelte artikler i CX-Basic, på egen pc. CX-Basic indeholder redigeringsfaciliteter, som bl.a. giver mulighed for:

- Redaktion af eksisterende ordbogsartikler og redaktion af enkelte felter i artikler.
- Sletning af artikler eller enkelte felter.
- Oprettelse af nye felter og nye artikler.
- Blokfunktioner for kopiering og flytning af felter eller hierarkiske blokke af felter, inden for én artikel, eller mellem artikler.
- Automatisk check af korrekt input i visse felter.

Forfatterne behøver ikke sikre en korrekt alfabetisk orden i artiklerne i CX-Basic. Når CX-Basic filerne bliver importeret i CX-Master alfabetiseres alle artikler på ny. Tilsvarende skal forfatterne ikke på nogen måde angive tegnsætning eller skriftsnit i CX-Basic filen.

d) Korrektur fra CX-Master, hos Munksgaard

Så snart en forfatter har afsluttet redaktionen af en fil i CX-Basic, sendes denne fil på diskette til Munksgaard, hvor der udskrives et korrekturprint på papir. I andre tilfælde vil vi lade forfatteren selv generere dette korrekturprint direkte fra CX-Basic. På dette print vil de enkelte ordbogsartikler være opstillet med korrekt skriftsnit og tegnsætning. Feltkoderne vil ikke være at se.

e) Rettelse af fejl i CX-Basic

Når forfatteren har korrekturprintet i hånden vil han kunne indtaste eventuelle rettelser og tilføjelser direkte i CX-Basic.

f) Satsfiler til den trykte ordbog

Når alle ordbogsartikler er redigeret færdig genererer Munksgaard de endelige satsfiler ud fra CX-Master basen.

- Og så til de spændende spørgsmål

Når man diskuterer feltopdelte data, redaktionsværktøjer og stringent kodning af leksikografiske data dukker der ofte de samme spørgsmål op igen og igen. Jeg vil her til slut prøve at svare helt punktuelt på et par af de mest hyppige spørgsmål:

I hvor høj grad sikres konsistente data i input-leddet?

Data indtastes og redigeres i CX-Basic hos forfatteren. I udformningen af CX-Basic er der sket en afvejning mellem forskellige hensyn. Det ene hensyn er til konsistente data. Her ønsker vi naturligvis at artiklerne i så høj grad som muligt skal leve op til redaktionsvejledningens anbefalede mønstre af felter. Vi sikrer dette ved at indbygge en række automatiske check i CX-Basic. Disse check omfatter især:

- Check af om feltnavnet er korrekt.
- Check af indhold i felter, hvor kun et kontrolleret vokabular er accepteret.
- Check af om visse centrale felter står korrekt i sekvensen af felter.
- Check af nummereringen af bogstavsektioner, arabertals- og romertalsafsnit.
- Check af hierarkier af felter.
- Check af om grupper af felter, som naturligt hører sammen, flyttes samlet ved "cut & paste"-operationer.

Disse check er dog ikke så omfattende, at der er tale om en fuldstændig "live parsing" af artiklens struktur. Og her kommer vi til det andet hensyn. Det handler om at vi ønsker os enkle standardværktøjer, og høj grad af frihed til selv at kunne konfigurere disse værktøjer. Vi har udtrykkeligt ønsket at forfatterne skulle arbejde i et kendt stykke software, - og her er WordPerfect ideelt. Desuden kan det afvikles på selv meget billige - og lidt ældre - 286-maskiner. Skulle CX-Basic foretage en komplet live parsing ville det stille væsentligt større krav til maskinen, og det ville i det hele taget ikke kunne være baseret på et stykke standardsoftware som fx WordPerfect.

Endelig er der et andet aspekt af sagen: Det er vores erfaring at man får det bedste resultat ved at gøre en stor indsats for at skrive en god redaktionsvejledning, og ved at få forfatterne til at følge den, samtidig med at man giver forfatterne et værktøj med en høj grad af frihed og fleksibilitet. Resultatet er, at vi nu og da får artikler fra CX-Basic, som ved import til CX-Master-basen viser sig ikke at overholde den struktur, som redaktionsvejledningen foreskriver. Disse afvigelser bliver da enten rettet under selve importen til CX-Master, eller senere i korrekturfasen, idet en forkert kombination af felter vil få manifestationslogik-modulet i CX-Master til at generere en forkert grafisk opsætning af artiklen.

Kan forfatterne se den færdige artikel på skærmen?

Nej, forfatterne kan ikke på skærmen se hvordan en artikel vil se ud i den trykte bog. I CX-Basic ses alle artikler som et WordPerfect-dokument, bestående af sekvenser af felter. Til gengæld kan vi udstyre forfatterne med en print-facilitet, som gør det muligt for dem at printe de færdige ordbogsartikler direkte fra CX-Basic.

Dette er et punkt, hvor vi er noget afventende: På den ene side ønsker de fleste nye forfattere at få lov til at se den færdige artikel på skærmen. Fra andre sammenhænge er de jo vant til WYSIWYG-faciliteter. Men på den anden side ved vi erfaringsmæssigt at vi får de bedste ordbogsartikler ved at tvinge forfatterne til at tænke i elementer med leksikografisk indhold fremfor i typografisk fremtoning.

Er det SGML?

Både ja og nej. Den feltopdelte og generiske kodning af alle typer information i en ordbogsartikel er i princippet en typisk SGML-kodning. Men i dag vil man intet sted i CX-Basic eller CX-Master finde en egentlig DTD - Document Type Definition - som følger syntaksen fra ISO-standarden. Ikke desto mindre findes der noget, som ganske modsvarer en DTD. Først og fremmest den meget detaljerede redaktionsvejledning, som beskriver anvendelsen af hvert enkelt felt, samt typiske sekvenser, hierarkier og iterationer af felter. Dernæst indeholder CX-Master et modul, som omformer sekvensen af felter i enhver artikel til et grafisk skriftbillede. Dette kan kun lade sig gøre, fordi CX-Master indeholder en formaliseret beskrivelse af alle felterne og deres indbyrdes relationer. Denne beskrivelse svarer ganske til en DTD, også selv om den ikke er formuleret i den af ISO vedtagne DTD-syntaks.