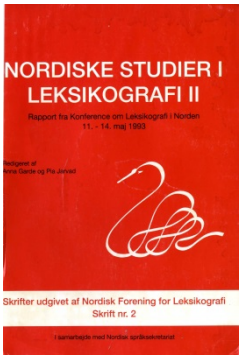
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Parsing af ikke-strukturerede ordbogsdata

Forfatter: Claus Bo Jørgensen

Kilde: Nordiske Studier i Leksikografi 2, 1993, s. 130-137
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.

Parsing af ikke-strukturerede ordbogsdata

Claus Bo Jørgensen

Abstract

Firmaet TEXTware har gennem de seneste 4 år konverteret traditionelle data til et stort antal ordbøger til struktureret tekst, dvs. tekst med eksplicit markering af en hierarkisk struktur, i overesstemmelse med den internationale SGML-standard. SGML-teksten er blevet anvendt i redigeringssystemet GestorLEX eller som et led i produktionen af elektroniske ordbøger eller til begge formål.

Erfaringerne fra dette arbejde peger på et antal typiske problemer i forbindelse med parsing, dvs. automatisk strukturering, af eksisterende, ikke-strukturerede ordbogsdata - typisk fotosætter- eller tekstbehandlingsfiler. Problemerne spænder fra banale tegnsætningsfejl over strukturer, der er problematiske for en maskine men ikke for en menneskelig læser, til principielt uløselige fortolkningsspørgsmål.

De enkelte problemtyper og mulige løsninger bliver præsenteret, og det diskuteres, hvordan og i hvilket omfang erfaringerne fra parsing kan udnyttes til at højne kvaliteten af såvel trykte som elektroniske ordbøger.

Indledning

Med ikke-strukturerede ordbogsdata menes i denne artikel f.eks. ordbogsdata skrevet med et almindeligt tekstbehandlingsprogram. Disse data er struktureret i den forstand, at der er mere eller mindre præcist formulerede redaktionsregler, der foreskriver i hvilken rækkefølge de forskellige typer af information skal gives, og hvordan de skal præsenteres (tegnsætning, skrifttype, skriftsnit). De er imidlertid ikke eksplicit struktureret, idet tekstbehandlingsfilen kun indeholder oplysninger om formateringen, dvs. den ønskede behandling, ikke om den logiske struktur.

Data af denne type har vi levet med i flere århundreder, og det betragtes ofte som helt problemfrit at læse og forstå sådanne data. Dette gælder i nogen grad, så længe læseren er et menneske. I det øjeblik man sætter et EDB-program til at fortolke sådanne data, f.eks. for bedst muligt at kunne præsentere dem og søge i dem i en elektronisk ordbog, afsløres der midlertid en hel række af problemer. En del af disse knytter sig selvfølgelig til EDB-programmets begrænsede "evner", mens mange også er reelle problemer for en menneskelig læser. EDB-programmet er blot med sin konsekvens og enfoldighed særlig velegnet til at afsløre dem.

De ideelle ordbogsdata til udnyttelse via et EDB-program indeholder eksplicit mar-

kering af, hvilken type af information de enkelte dele af en artikel giver, og hvordan delene hænger sammen. Den formalisme TEXTware har valgt til strukturbeskrivelsen er SGML, en international standard (se f.eks. Norling-Christensen, 1992), men formalismen i sig selv er ikke vigtig, hvis blot to væsentlige principper er overholdt:

1 det der markeres er den logiske struktur: hvilken type af information et stykke tekst giver - ikke hvordan teksten skal præsenteres (eller på anden måde behandles) til et givet formål.

2 der er tale om en hierarkisk struktur, en træstruktur, med etiketter klistret på de enkelte mindste enheder i artiklen og med non-terminale elementer, altså knuder der hver samler et antal mindre enheder.

En ordbogsartikel som struktureret tekst vil typisk indeholde en hel del mere information (en finere inddeling) end den tilsvarende traditionelle, typografisk markerede tekst, og skift mellem de større logiske enheder falder ikke nødvendigvis sammen med typografiske skift.

Spørgsmålet er nu, hvordan man automatisk kan komme fra typografisk markeret tekst til struktureret tekst.

Fra formateret til struktureret tekst

Svaret på dette spørgsmål er, at det kan man typisk ikke, hvis man tager ordet automatisk helt bogstaveligt. Der vil altid (efter vores erfaring) være artikler, der ikke umiddelbart passer ind i den struktur, man forsøger at genkende - og der vil være artikler der kan analyseres på flere forskellige måder.

Som basis for selve parsingen har man (i bedste fald) det sæt redaktionsregler der har ligget til grund for ordbogsskrivningen og som foreskriver:

- hvilke oplysninger der gives
- hvilken rækkefølge
- med hvilken skrifttype og -snit
- afgrænset med hvilke skilletegn
- evt. taget fra et lukket vokabular

Opgaven går så ud på at kombinere ovenstående regler med de faktisk foreliggende data, og sætte en parser til herudfra at finde ud af hvad leksikografen har tænkt og ment!

Processen kan deles op i to ikke helt uafhængige dele:

- genkendelse af informationstyper for den enkelte tegnstreng (afgrænsning af terminale elementer - svarende til tag-og-tekst (feltnavnindhold) data)
- sammenhægtning af informationstyper i træstruktur afgrænsning af non-terminale elementer - hvilke informationer hører sammen)

Genkendelse af informationstyper kan i en del tilfælde foretages rent lokalt, det vil sige uden at kigge på den kontekst en tegnstreng står i, ud fra selve tegnstrengen og dens formatering. Dette vil typisk være tilfældet ved f.eks. stil- og emneangivelser, der tages fra en lukket liste.

Det lader til at være en god idé at foretage så meget som muligt af denne type genkendelse, inden den egentlige strukturparsing starter.

Selve parsingen omfatter derefter genkendelse af de resterende (ikke entydige) informationstyper, og sammenhægtningen af disse i en træstruktur.

Afhængig af formålet med parsingen kan man vælge enten at få så mange som muligt af ordbogens artikler genkendt og struktureret med så få ændringer som muligt, eller at få artiklerne passet ind i en mere konsekvent struktur, med et større antal håndrettelser til følge.

Til parsingen anvendes en grammatik, der kan betragtes som en udvidelse af den SGML-grammatik, der beskriver den resulterende struktur. Udvidelserne er det, der sørger for at genkende kombinationer af skilletegn, formatskift og kontekstinformation som implicitte angivelser af den logiske struktur.

For at give et indtryk af arbejdets omfang kan det oplyses, at en SGML-grammatik på 1 1/2 til 2 A4-sider typisk vil modsvares af en parsegrammatik på 20-25 A4-sider.

Typer af problemer

Under selve parsingen (eller senere) viser problemerne sig på tre måder:

1 ikke-analyserede artikler
2 flertydige artikler: flere strukturer foreslået
3 kun én analyse, men forkert struktur foreslået

Inden for hver gruppe er der igen flere typiske årsager til problemet:

1) Ikke-analyserede artikler:

- fejl i tegnsætning

Tegnsætning er den sværeste del af en ordbog at læse korrektur på, og det er ofte let for et menneske at forstå meningen på trods af en formel fejl.

NB: en del fejl kan normalt fjernes ved at gøre parse-grammatikken mindre restriktiv (dvs. acceptere en afvigende tegnsætning som implicit markering af den samme struktur), men resultatet af en sådan ændring bør kontrolleres grundigt, da man risikerer at introducere fejl af typen entydig men forkert analyse!

- fejl i fontmarkering

Disse er lettere at fange ved korrekturlæsning. Til gengæld kan de meget let opstå når man tilføjer eller sletter data i et tekstbehandlingsprogram.

- uventet struktur

Redaktionsregler er åbne for fortolkning. Dette viser sig tydeligst i større ordbøger med skiftende redaktører eller redaktionsgrupper.

2) Flertydige artikler:
- flere fortolkninger af en enkelt tegnstreng

Ofte anvendes samme formatering til en del forskellige informationstyper, også med samme position i data. Det kan f.eks. være skellet mellem optionel del af kildesprogsfrase og ikke-tabellagt betydningsspecifikation: begge i kursiv med parentes omkring.

- flere fortolkninger af tegnsætning

Spørgsmålstegn og udråbstegn som rene indholdstegn (del af kildesprogsfrase eller oversættelsesækvivalent) eller som kombinerede indholds- og skilletegn (i så fald: svarende til adskillelse med komma eller semikolon?).

Komma som indholdstegn eller som skilletegn.

- flere mulige placeringer (niveauer) i strukturen

Samme type af information kan ofte placeres mere eller mindre højt i strukturen, dvs. gældende for en større eller mindre del af artiklen. En betydningsspecifikation kan således fortolkes som gældende for den nærmest følgende oversættelse alene, eller for alle de følgende oversættelser.

3) Forkert analyserede artikler:
- fejl i tegnsætning
- fejl i skrifttype eller skriftsnit
- ikke-genkendt flertydighed

Praktiske løsningsforslag

Problemerne med ikke-analyserede eller forkert analyserede artikler kan siges at være principielt uinteressante.

Hvis det er tegnsætningsfejl eller fejl i brugen af skrifttype/-snit der er årsagen, skal disse blot rettes og ariklerne parses igen.

Fejl i form af uventede strukturer kræver et valg mellem rettelse i data og udvidelse af grammatikken. Valget afhænger i meget høj grad af, hvad den strukturmarkerede tekst skal bruges til. Som oftest har opgaven for vores firma været: lav en elektronisk bog - billigst muligt - med alle de eksisterende data i, og det har givet meget lidt restriktive grammatikker. Hvis formålet derimod er at gå over til at redigere ordbogen på struktureret form, vil man ofte benytte lejligheden til at rydde lidt op i de eksisterende variationsmuligheder.

De interessante problemer er de flertydige konstruktioner. Også på dette område bærer TEXTware's løsningsforslag præg af, at ekstra arbejde (højere pris) skal kunne forsvares med en tilsvarende gevinst i udnyttelsen af data.

Desuden kræver de principielt flertydige konstruktioner en egentlig leksikografisk stillingtagen, og det er ikke vores job.

Der er derfor i praksis anvendt følgende "løsnings"-modeller:

Flertydig informationstype (f.eks. betydningsspecifikation vs. optionel del af kildesprogsfrase):

Eksempel:	
se sig godt for (før man handler)	[y] regarder à deux fois
Frase *se sig godt for* (før man handler)	Frase *se sig godt for*
Tyd1	Tyd1 TydSpec *(før man handler)*
Tyd2 [y] regarder à deux fois	Tyd2 [y] regarder à deux fois

I en passiv ordbog vil de to muligheder anvende hver sit sprog, og i praksis har det vist sig at en ret kort liste med hyppige ord (præpositioner, artikler, evt. pronominer og enkelte substantiver) er nok til at kunne afgøre, hvilket sprog der er anvendt.

Denne flertydighed har vi derfor normalt taget stilling til i hvert enkelt tilfælde.

I en aktiv ordbog er der ikke mulighed for at skelne mellem funktionerne ved hjælp af sproget. Den eneste mulighed ligger i en egentlig "forståelse" af teksten.

Da vi i de elektroniske ordbøger plejer at gøre både kildesprog og metasprog søgbare, er der ikke noget umiddelbart tab af funktionalitet ved ikke at skelne mellem disse, så en flertydighed hér er blevet opløst ved altid at vælge den ene fortolkningsmulighed (del af kildesprogsfrase).

Flertydig tegnsætning:

Eksempel:

Voulez-vous de la bière? -- Va pour la bière! Vil De have øl? -- ja, lad gå!

allez toujours! bliv bare ved! hæng bare i! lad Dem ikke forstyrre!

est-ce qu'on va au cinéma, dis, papa? skal vi så i biografen, hvad, far?

on dirait man skulle tro, man kunne fristes til at sige;
on dirait de loin une barque på afstand kunne man godt tro, at det er en båd; det kunne godt ligne en båd på afstand

allons donc! hvor vil de hen? hvad behager! nej hør nu! *(iron.)* javist! nå! nå skynd dig nu! *(beroligende)* nå nå!

allez! (sp.) start! *(i boksning)* time!

Ved spørgsmålstegn og udråbstegn er der, afhængig af hyppigheden, enten taget stilling

til hvert enkelt tilfælde, eller der er konsekvent valgt én af mulighederne (betydningsadskillelse som ved komma eller som ved semikolon - skellet mellem rent indholdstegn og kombineret skille- og indholdstegn vil ofte fremgå af konteksten).

Komma har vi normalt betragtet som skilletegn, men igen er der ikke nogen forskel med hensyn til funktionalitet i de elektroniske ordbøger, idet vi plejer at bruge semikolongrænsen som kriterium for linieombrydning. Det kan altså heller ikke ses i den elektroniske bog, om kommaet er skille- eller indholdstegn.

Flertydigt niveau i strukturen:

Eksempel:

I. **navle** *en -r*
1 *(anat)* navel; *(fagl.)* umbilicus; T *(lettere vulg.)* belly-button;
2 *(bot.)* hilum

Krop	Krop
Tyd1	Tyd1
| TydSpec *anat*	
|	Tyd2
| Tyd2	| TydSpec *anat*
| Ækvi navel	| Ækvi navel
| Tyd2	Tyd2
| TydSpec *fagl.*	TydSpec *fagl.*
| Ækvi umbilicus	Ækvi umbilicus
| Tyd2	Tyd2
| TydSpec *lettere vulg.*	TydSpec *lettere vulg.*
| Ækvi belly-button	Ækvi belly-button

Endnu en gang har de praktiske hensyn fået overtaget: der sker mindst skade ved en forkert fortolkning, hvis man vælger den højest mulige placering (det størst mulige virkefelt). I værste fald giver dette fejl, som er gennemskuelige for den menneskelige bruger (to modstridende TydSpecifikationer knyttet til samme oversættelsesækvivalent), mens valg af det mindst mulige virkefelt giver "usynlige" fejl i form af manglende information om de følgende ækvivalenter (i eksemplet ovenfor vil 'umbilicus' ikke kunne findes som anatomisk term).

I forbindelse med denne sidste type af flertydigheder kan det bemærkes (kommentar fra Ole Norling-Christensen), at endog den leksikograf, der har skrevet artiklen, kan være i tvivl om den rette fortolkning. Dette blot for at minde om i hvor høj grad de traditionelle værktøjer til ordbogsredigering har fokuseret på *præsentation* snarere end på *struktur*.

Problemernes hyppighed

Et lille antal stikprøver fra forskellige af TEXTwares parse-opgaver afslører, at de fleste problemer, der kræver manuel indgriben i processen (rettelse i data eller markering af én af de mulige analyser som den foretrukne), skyldes de "banale", dvs. principielt uinteressante, fejltyper. Typisk står tegnsætningsfejl for 30-40% af tilfældene, fejl i skrifttype/-snit for yderligere 20-30%, mens flertydigheder og egentlige strukturfejl (brud på redaktionsreglerne) hver giver 10-20%. Endelig er der ofte en mindre gruppe af egentlige kodningsfejl, dvs. fejlkodning af specialtegn eller lignende.

Fordelingen på de enkelte problemtyper skyldes i høj grad kontraktmæssig afgrænsning af TEXTwares ydelser over for forlagskunderne. Den "interessante" flertydighed med hensyn til en betydningsspecifikations niveauplacering er således typisk langt det hyppigste problem, men en fuldt tilfredsstillende løsning kræver som nævnt egentlig leksikografisk stillingtagen.

Konklusion

En automatisk behandling af traditionelle, formaterede ordbogsdata afslører altså en række problemer ved fortolkningen af disse. De banale problemer som fejl i tegnsætning og valg af skrifttype/-snit kan let løses, men understreger dog et rimeligt krav til ethvert ordbogsredigeringsværktøj: præsentationen af artiklerne skal ikke være en del af selve ordbogens data men alene afspejle den logiske struktur.

De interessante problemer, flertydighederne, kræver egentlig leksikografisk stillingtagen og er derfor dyre at løse. Værdien af en teoretisk tilfredsstillende løsning afhænger desuden i høj grad af den planlagte udnyttelse af de strukturerede data.

Strukturoplysningerne er afgørende, når data skal behandles automatisk, som f.eks. ved udnyttelse i forbindelse med maskinoversættelse (meget kritisk) eller blot i elektroniske bøger, hvor maskinens rolle er at formidle ordbogens data til en menneskelig læser.

Desuden må det konkluderes, at muligheden for at udnytte strukturen under præsentationen af data for en menneskelig bruger, er større ved elektronisk formidling end ved udgivelse i trykt form. For det første er der de traditionelle hensyn at tage til læseligheden (et roligt skriftbillede uden for mange forskellige skrifttyper og skriftsnit) og den trykte bogs omfang (omend tendensen går i retning af mindre komprimeret, mere "luftig" opsætning). For det andet er der den grundlæggende forskel, at den én gang trykte præsentation ikke kan ændres af læseren, hvis en anden synsvinkel på data ønskes! I den elektronisk formidlede form koster en luftig opsætning derimod ikke noget, og læseren kan selv (i nogen grad) vælge den præsentation af data der passer til det aktuelle formål.

Det er således ikke tilfældet, at en trykt udgave af en ordbog nogensinde vil kunne udnytte den strukturelle information i ordbogsdata i samme grad som en elektronisk udgave. Det er imidlertid stadigvæk en glimrende idé at det er de samme strukturerede data der ligger til grund for begge udgivelser.

Litteratur

Ole Norling-Christensen, 1992. *Struktureret redigering af ordbøger*, i: Nordiske studier i Leksikografi. Nordisk forening for leksikografi, Oslo, 1992.