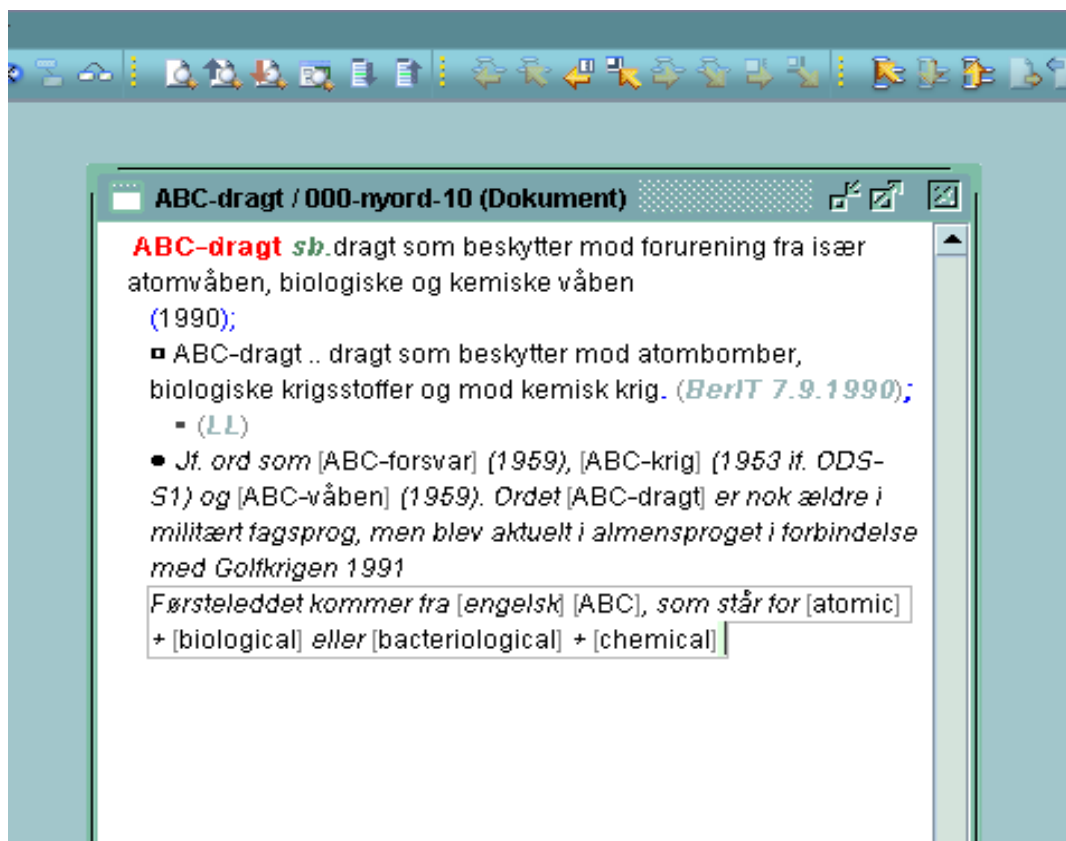


Harmonisering af de nordiske sprognævns databaser



De nordiske sprognævns arbejdsgruppe for sprogrøgt og sprogteknologi

Ida Elisabeth Mørch og Jakob Halskov, Dansk Sprognævn

September 2008

Indholdsfortegnelse

1.0 Indledning.....	3
1.1 Udredningens opdragsgiver og deltagere.....	4
1.2 Udredningens formål.....	5
1.3 Hovedspørgsmål.....	6
1.4 Rapportens struktur.....	6
1.5 Sammenfatning og anbefaling.....	6
1.5.1 Anbefalinger.....	6
1.5.2 Fordele og ulemper.....	7
1.5.3 Beslutning.....	8
1.5.4 Hvem skal gøre hvad?.....	10
1.5.5 Hvad er allerede gjort?.....	10
2.0 Svarbaser.....	10
2.1 Analyse af eksisterende strukturer.....	11
2.1.1 Danmark.....	11
2.1.2 Sverige.....	12
2.1.3 Norge.....	12
2.1.4 Finland.....	13
2.1.5 Island.....	14
2.1.6 Sammenligning af oplysningstyper i nordiske svarbaser.....	14
2.2 Udkast til fælles svarbasestruktur.....	15
2.3 Sammenligning af nordiske emnetaksonomier.....	17
2.3.1 Parring af ækvivalente kategorier i en metaemnetaksonomi.....	19
2.3.2 Sverige kontra Danmark.....	20
2.3.3 Norge kontra Danmark.....	21
2.3.4 Island kontra Danmark.....	22
2.3.5 Finland kontra Danmark.....	22
2.4 Internationale standarder for universelle emneklassifikationer.....	23
2.4.1 DDC (Dewey Decimal Classification).....	23
2.4.2 UDC (Universal Decimal Classification).....	23
2.5 Internationale standarder for lingvistiske emner.....	24
2.5.1 OLAC (Open Language Archives Community).....	24
2.5.2 GOLD (General Ontology for Linguistic Description).....	24
2.5.3 CLARIN (Common Language Resources and Technology Infrastructure).....	24
2.6 Internationale standarder for relevante datakategorier.....	25
2.6.1 Dublin Core.....	25
2.6.2 ISO 12620 (Data Category Registry).....	25
2.7 Udkast til fælles emnetaksonomi.....	27
2.8 Metaemnetaksonomiens konsekvenser.....	28
2.9 Konklusion og diskussion.....	29
3.0 Ordbaser.....	30
3.1 Analyse af eksisterende strukturer.....	30
3.1.1 Danmark.....	30
3.1.1.1 Ordsamlingsbasen.....	30

3.1.1.2 Retskrivningsordbogen.....	31
3.1.1.3 Nyordsordbogen.....	32
3.1.2 Sverige.....	33
3.1.3 Norge.....	34
3.1.4 Finland.....	34
3.1.5 Island.....	35
3.2 Sammenligning af ordbasestrukturer.....	35
3.3 Internationale standarder for ordbaser.....	35
3.3.1 Text Encoding Initiative (TEI).....	35
3.3.1.1 Nordisk netordbog.....	36
3.3.2 Lexical Markup Framework (LMF).....	37
3.3.3 Open Lexicon Interchange Format (OLIF).....	39
3.3.4 Multi Dictionary Formatter (MDF).....	39
3.3.5 Andre standarder for ordbaser.....	40
3.3.5.1 ISO-1951.....	40
3.3.5.2 DANLEX og STANLEX.....	41
3.4 Udkast/anbefalinger til fælles ordbasestruktur.....	43
3.4.1 Eksempel på ordbogsartikel i LMF.....	43
3.4.2 Eksempel på ordbogsartikel i det reducerede TEI-format.....	44
3.5 Konklusion.....	45
4.0 Software.....	45
4.1 Databaseprogrammel.....	45
4.1.1 Markedsoverblik.....	45
4.2 Native XML-databaser kontra konventionelle databaser.....	47
4.2.1 To eksempler: eXist og iLEX.....	48
4.3 Konsekvenser af overgang til XML-database.....	48
4.4 Terminologisk programmel.....	49
4.5 Konklusion.....	50
5.0 Overordnet konklusion og anbefalinger.....	51
Referencer.....	51
Bilag.....	52
1. Den danske emnetaksonomi.....	52
2. Den svenske emnetaksonomi.....	57
3. Den norske emnetaksonomi.....	59
4. Den islandske emnetaksonomi.....	61
5. Ækvivalensnøgle for nordiske emnetaksonomier.....	61
6. ISO-1951: Eksempel på monolingval ordbogsartikel i XML.....	65
7. Skema for den nordiske netordbog.....	67
8. Den finske emnetaksonomi.....	67
9. Den danske svarbases struktur.....	71
10. Udkast til fællesnordisk svarbasestruktur.....	72
11. Den danske ordbases struktur.....	73
12. Den danske Retskrivningsordbogs struktur.....	74
13. Den danske nyordsordbogs struktur.....	75
14. Sammenligning af oplysningstyper i de nordiske ordbaser.....	76
15. To eksempler på native XML-databasesystemer.....	78
16. En OLIF-fils struktur.....	79
17. Eksempel på bearbejdet nyordsexcerpt i LMF.....	80
18. Eksempel på bilingval ordbogsartikel i det nordiske netordbogsformat.....	81
.....	81

Fortegnelse over rapportens tabeller	
Tabel 1: Oplysningstyper i de nordiske svarbaser	s. 16
Tabel 2: Sammenligning af overkategorier i nordiske emnetaksonomier	s. 18-19
Tabel 3: Unikke kategorier i svensk og dansk emnetaksonomi	s. 22
Tabel 4A: Hovedkategorier i ISO 12620	s. 27
Tabel 4b: Ækvivalente kategorier i ISO 12620 og den danske emnetaksonomi?	s. 27-28
Tabel 5: Overlap mellem danske, svenske og norske emnekategorier	s. 28
Tabel 6: Modulerne i Text Encoding Initiative (TEI)	s. 37
Tabel 7: Standardisering af informationstyper i leksikalske data	s. 42
Tabel 8: Databaseprogrammel i de nordiske sprognævns svar- og ordbaser	s. 46
Tabel 9: Overblik over native XML-databasesystemer	s. 47-48

Fortegnelse over rapportens figurer	
Figur 1: En metaemnetaksonomi for alle nordiske emneklassifikationer	s. 10
Figur 2: Parring af nordiske kategorier i fælles metaemnetaksonomi	s. 20
Figur 3: Kernemodulet i Lexical Markup Framework	s. 38
Figur 4: Udvidelsesmoduler til LMF	s. 39
Figur 5: Hvordan anvendes LMF i et konkret projekt?	s. 39
Figur 6: Terminologisk arbejdsværktøj: iTERM	s. 50
Figur 7: Terminologisk arbejdsværktøj: TemaTres	s. 51

1.0 Indledning

Databaser til håndtering af ordsamlinger, ordbøger og svarsamlinger bliver nu anvendt i flere af de nordiske sprognævn. Med de nye generationer af databaseværktøjer og især med nye opmærkningsmuligheder i opmærknings sproget XML og beslægtede teknologier (se faktaboks 1 og 2) er der opstået nye muligheder for at beskrive og udveksle data fleksibelt og effektivt og dermed for at trække flere oplysninger ud af sprognævnenes samlinger end man tidligere troede muligt.

I løbet af 2007 har de nordiske sprognævn derfor sat fokus på indhold og struktur i deres sproglige databaser. I foråret 2007 blev der afholdt et temamøde om svarbaser i Stockholm og i efteråret (6. september 2007) på det nordiske netværksmøde blev diskussionen videreført. Derudover har der været en række bilaterale drøftelser af organiseringen af svar- og ordbaser mellem de enkelte nævn.

Det viste sig at stort set alle sprognævn har en eller flere ord- og/eller svarsamlinger, men at de er organiseret meget forskelligt, og at mange af dem ligger i forældede databasesystemer. Nogle sprognævn har for nylig skiftet til et nyt databasesystem, andre er i færd med at skifte, og atter andre overvejer stadig hvilket system de skal bruge. Tidspunktet er derfor gunstigt for at overveje en

fælles løsning, både hvad angår den tekniske og strukturelle problemstilling, inden nævnene igen får lagt sig fast på hver sin individuelle model, og der var derfor på det nordiske netværksmøde stor opbakning fra alle tilstedeværende til at undersøge mulighederne for en harmonisering.

Faktaboks 1: Hvad er XML?	Faktaboks 2: Hvorfor XML?
<p>XML står for <i>Extensible Markup Language</i> og er dermed beslægtet med andre opmærkningsprog som fx SGML¹ og det velkendte HTML².</p> <p><u>Fleksibilitet</u> Til forskel fra HTML kan man med XML helt selv navngive og definere alle elementer uden at spekulere på hvordan indholdet af elementerne skal præsenteres, fx</p> <pre data-bbox="167 824 662 1041"><CD> <titel>Kind of Blue</titel> <kunstner>Miles Davis</kunstner> <genre>jazz</genre> <årstal>1997</årstal> </CD></pre> <p><u>Uniformitet</u> Med XML kan man sikre at alle dokumenter er gyldige i forhold til et XSD skema³, fx</p> <p>at alle CD-dokumenter skal have 1 titel, 1 kunstner og 1 årstal og kan have 1 genre. Samt at indholdet af genre fx skal være enten "jazz", "pop", "rock" eller "klassisk".</p>	<ol style="list-style-type: none"> 1. Fordi man kan 100% adskille præsentation og indhold (til forskel fra HTML) og dermed kun behøver opbevare/opdatere indhold ét sted 2. Fordi XML er en åben standard 3. Fordi XML giver fleksibilitet + uniformitet 4. Fordi XML tilhører en hel familie af internationale standarder for indholdsredigering: <ul style="list-style-type: none"> ● XSD: skemaer som definerer hvordan et gyldigt XML-dokument skal se ud ● XSLT: transformationssprog som anvendes til at "oversætte" XML-dokumenter til fx anderledes XML-dokumenter eller HTML-dokumenter ● XQuery/XPath: søgesprog som anvendes til at fremfinde bestemte XML-dokumenter (fx alle CD'er af Miles Davis) ● XForms: specifikation af bl.a. brugergrænseflader til XML data (fx formularer på nettet)

1.1 Udredningens opdragsgiver og deltagere

Udredningens opdragsgiver er Nordens Sprogråd som i december 2007 bevilligede midler til projektet. Udredningens deltagere er organiseret i tre enheder: en styregruppe, en arbejdsgruppe og en kontaktgruppe.

Arbejdet ledes og koordineres af de nordiske sprognævns arbejdsgruppe for sprogrøgt og sprogteknologi som fungerer som styregruppe. Denne styregruppe har de følgende medlemmer: Torbjørg Breivik (Norge), Rickard Domeij (Sverige), Sabine Kirchmeier-Andersen (Danmark), Mikael Reuter (Finland), Eiríkur Rögnvaldsson (Island), Risten Turi (Norge).

Styregruppen nedsatte ultimo 2007 en arbejdsgruppe som har foretaget selve udredningen og udfærdiget nærværende rapport. Denne arbejdsgruppe har de følgende medlemmer: Jakob Halskov

1 Standard Generalized Markup Language

2 HyperText Markup Language

3 dvs. overholder den struktur der er defineret i skemaet

(Danmark) og Ida Elisabeth Mørch (Danmark).

Endelig har en kontaktgruppe med nøglepersoner fra hvert af de nordiske lande evalueret og kommenteret arbejdet undervejs. Denne gruppe består af: Birgitta Lindgren (Sverige), Sabine Rosenhart (Norge), Jóhannes B. Sigtryggsson (Island), Risto Widenius (Finland).

Desuden har en tidligere version af rapporten været sendt i høring hos følgende eksperter inden for terminologi og standardisering Håvard Hjulstad (Norge), Viggo Kann (Sverige), Bodil Nistrup Madsen (Danmark), Henrik Nielsson (Sverige). Vi takker alle for værdifulde input og kommentarer som er forsøgt indarbejdet i rapporten. Rapportens forfattere bærer naturligvis ansvaret for eventuelle misforståelser eller fejlagtige informationer.

1.2 Udredningens formål

Der er store funktionelle og prismæssige forskelle på de kommercielle databasesystemer, og det er ikke nemt at gennemskue hvilke systemer der egner sig til de særlige behov som sprognævnene har. Sprognævnene kan se klare fordele i at foretage en fælles undersøgelse af forskellige typer af databaser frem for at hvert sprognavn bruger tid og kræfter på at finde frem til en individuel løsning.

Uanset om en sådan undersøgelse fører frem til en beslutning om et fælles databasesystem, har sprognævnene stor interesse i at samarbejde om en fælles databasestruktur i form af en XML-standard der beskriver oplysningernes art og rækkefølge på en ensartet måde. En sådan harmonisering af sprognævnernes resurser (både internt på sprognævnene og tværgående mellem dem) har flere fordele. Det vil blive lettere:

1) for sprognævnernes medarbejdere (især nye medarbejdere) at orientere sig i andre sprognævns baser i forbindelse med tværnordiske projekter

fx anvende (fælles) emneord i stedet for (nationale) nøgleord

2) at sammenkoble orddatabaserne og fx danne ordbøger mellem de nordiske sprog

3) at koble svarbaserne sammen og fx at søge på tværs af de nordiske sprog i svarbaserne

4) at sammenkoble sprognævnernes databaser med andre sproglige resurser

En fælles klassifikation af sproglige fænomener som typisk forekommer i svarsamlinger, muliggør således ikke blot en hurtig og smidig søgning på bestemte spørgsmålstyper og beslægtede fænomener for det enkelte navn, men også muligheden for at forske systematisk kontrastivt i de nordiske sprognævns samlinger på en hidtil ukendt måde.

Blandt fordelene ved at bruge en XML-baseret standard er at specifikationerne er organiseret hierarkisk. Det betyder at man kan have en fælles kerne af beskrivelser som så kan specificeres mere detaljeret afhængigt af hvad der er nødvendigt for et givet sprog. Fx vil alle ordbaser indeholde visse typer af morfologiske oplysninger, som for eksempelvis finsk og grønlandsk vedkommende vil kunne specificeres yderligere.

Endvidere vil man kunne arbejde med forskellige grader af finkornethed i fx en emneklassifikation af sprognævnernes svar, hvor nogle sprognavn kan vælge en mere generel fælles kategori som fx *ortografi*, mens andre kan være mere specifikke og yderligere opdele spørgsmål om ortografi i hhv.

orddannelsestegn og bogstaver, og måske yderligere opdele spørgsmål vedrørende bogstaver i *bogstav-lyd forhold, translitteration, fremmede bogstaver* osv. (se bilag 1 for den danske emnetaksonomi).

1.3 Hovedspørgsmål

Sprognævnene ønsker hurtigst muligt i begyndelsen af 2008 at foretage en udredning af følgende spørgsmål

- 1) Hvordan kan sprognævnenes databaser strukturelt og indholdsmæssigt harmoniseres med hinanden og andre videnskabelige databaser?
- 2) Hvilke databasetyper er bedst egnet til at imødekomme sprognævnenes særlige behov?

I udredningen indgår overvejelser om internationale opmærkningsstandarder som TEI og diverse ISO-standarder, samt de standardiseringsprocesser som er i gang i EU-samarbejdet CLARIN og på nationalt niveau. Endvidere overvejes det om det klassifikationssystem for svar som er udviklet i Danmark, og som man også forventer at bruge i Norge, vil kunne anvendes i de andre nordiske lande.

1.4 Rapportens struktur

Rapporten er struktureret i tre hovedafsnit, nemlig et om svarbaser, et om ordbaser og et om software. Fokus er især rettet mod afsnittet om svarbaser fordi det er her der virkelig mangler standarder, og fordi klart strukturerede og standardiserede svarbaser vil lette sprognævnenes daglige rådgivningsarbejde.

1.5 Sammenfatning og anbefaling

Dette afsnit indeholder et slags ”executive summary” af harmoniseringsrapporten, nemlig dens anbefalinger (1.5.1), fordele og ulemper (1.5.2), beslutningsoplæg (1.5.3), opgavefordeling (1.5.4) og nuværende projektstatus (1.5.5).

1.5.1 Anbefalinger

På basis af analyserne i afsnit 2 (om svarbaser), afsnit 3 (om ordbaser) og afsnit 4 (om software) er rapportens overordnede anbefalinger følgende:

1. **Primært svarbaser:** Oprettelse og ibrugtagelse af en fællesnordisk metaemnetaksonomi (se faktaboks 3 og figur 1)
 - Opgave: de enkelte sprognævn beholder deres nuværende emnetaksonomier, men disse kobles til den fælles taksonomi (se ækvivalenstabellen I bilag 5 for et første forsøg på en kobling)
 - Opgave: definitioner og eksempler udarbejdes i det omfang det er nødvendigt
2. **Alle baser:** Det anbefales at anvende **XML** og beslægtede teknologier i så vid udstrækning som muligt.
 - Vigtigst for indholds- og strukturbeskrivelser

- Ikke afgørende om selve data ligger opmærket i XML eller fx i en relationel database
3. **Alle baser:** Oprettelse af et fællesnordisk DCR for centrale datakategorier
 - NB: Mange er allerede defineret i fx ISO 12620
 4. **Ordbaser:**
 - Struktur: Følg ISO-standarden Lexical Markup Framework (LMF) og anvend den i fremtidige baser.
 - Indhold: Brug det fællesnordiske DCR (punkt 3)
 5. **Svarbaser:** Der anbefales en fælles grundstruktur for svarbaseartikler, men det er mindre afgørende end punkt 1.
 6. **Alle baser:** Der kan være stordriftsfordele ved at sprognævnene anvender fælles redigerings- og publiceringsværktøjer, men det er sandsynligvis urealistisk.
 - Opgave: Der bør udvikles en fælles søgeportal på internettet som kan tilgå alle de nordiske databaser (=> databaseskemaer og bagdøre skal foreligge).

1.5.2 Fordele og ulemper

Et af formålene med at harmonisere struktur og emnetaksonomier for sprognævnenes databaser er som sagt at gøre det lettere at søge på tværs af baserne. Dette vil gavne forskere og sprognævnsansatte når de ønsker at søge på tværs af de nordiske sprog, men det vil også gavne de ansatte (og eksterne forskere) som ønsker at søge på tværs af egne baser. På lidt længere sigt vil harmoniseringen imidlertid også kunne komme slutbrugere (borgerne) til gode.

Konkrete eksempler på fordelene ved databaseharmoniseringen er således at:

- **sprognævnenes ansatte:**
 - **kan søge systematisk i egne** baser og udtrække svar som vedrører mere eller mindre specifikke emner
 - kan søge systematisk i **andres** baser uden først at skulle undersøge de enkelte sprognævns emnetaksonomier og deres respektive strukturer
 - **nye medarbejdere** kan hurtigt fremfinde relevant information fra svar- og ordbaser, som gamle medarbejdere måske tager for givet
 - **svartjenesten bliver mere effektiv** og kan koncentrere sig om de vanskelige spørgsmål
- **eksterne sprogforskere:**
 - får også mulighed for at søge på tværs af alle baser - også selvom de ikke forstår alle de nordiske sprog
- **slutbrugere (borgere):**
 - vil kunne søge i alle baser og få direkte eller indirekte hjælp af metaemnetaksonomien. De vil kunne anvende emneord i stedet for nøgleord. De vil fx kunne få (pædagogiske) definitioner på sprogvidenskabelige emner samt forslag til yderligere søgninger, fx på samme (eller beslægtede) emne(r) i andre nordiske baser.

Vi ser ikke som sådan nogen ulemper ved at harmonisere de nordiske sprognævns databaser, men man skal ikke være blind for at denne overgang har en række konsekvenser der involverer tid og penge. Det vil nemlig kræve:

- **Løsning #1: harmonisering "light"**
 - **kobling** af alle nationale emne kategorier til den fællesnordiske emnetaksonomi
 - **udvikling** af en fælles søgeportal på internettet
 - åbning af bagdøre til alle relevante nordiske databaser (kræver lokale systemadministratorer og adgang til fx databaseskemaer)
 - udvikling af søgeportalens grænseflade
 - udvikling af søgeportalens funktionalitet (fx søgescripts, opdateringsscripts)
- **Løsning #2: gennemgribende harmonisering**
 - **Alt fra løsning #1 samt:**
 - **konvertering**
 - af eksisterende emne kategori angivelser til den fælles standard
 - af eksisterende databasestrukturer til fælles standarder
 - af eksisterende databaser til XML
 - fra SQL (fx Access), SGML og lignende: relativt enkelt
 - fra regneark (fx Excel): lidt mere tidskrævende
 - **omprogrammering** af eksisterende, integrerede databaseløsninger
 - fx automatisk registrering af e-post i SQL-baseret svarbase (Sverige)
 - fx automatiske webløsninger der interagerer med en SQL-baseret svarbase (fx Frågelådan i Sverige)
 - **ibrugtagelse** af nyt redigerings- og publiceringssoftware (XML-baseret).

Både løsning #1 og løsning #2 vil kræve mindst én it-specialist ved hvert sprognavn (til åbning af databasebagdøre med videre) og en terminologisk kyndig koordinator, men løsning #2 vil desuden kræve omfattende **efteruddannelse** af hele det videnskabelige personale (det tager tid at vænne sig til en ny grænseflade til redigering/indtastning) og **yderligere it-eksptise** til konverteringen af databaseformat og omprogrammeringen af eksisterende webløsninger.

Vi anbefaler at man vælger løsning #1 (harmonisering "light"), da dette er det mest realistiske projekt og i praksis vil give stort set lige så store fordele som løsning #2. På længere sigt kan man naturligvis arbejde sig frem mod en mere gennemgribende harmonisering, fx i forbindelse med allerede planlagte overgange til nyt software.

1.5.3 Beslutning

Sprognævnene skal beslutte om de vil gå videre med udarbejdelsen af en metaemnetaksonomi på basis af denne rapport. Der er kun taget et første skridt i retning af en sådan metaemnetaksonomi (se

ækvivalensnøglen i bilag 5). Beslutter sprognævnene at gå videre, vil hvert enkelt sprognavn skulle bidrage aktivt med præcise definitioner af deres emnekategorier (herunder en validering af forslaget i bilag 5) og desuden tilgængeliggøre, dvs. åbne bagdøre til, deres relevante databaser så en fælles søgegrænseflade kan udvikles.

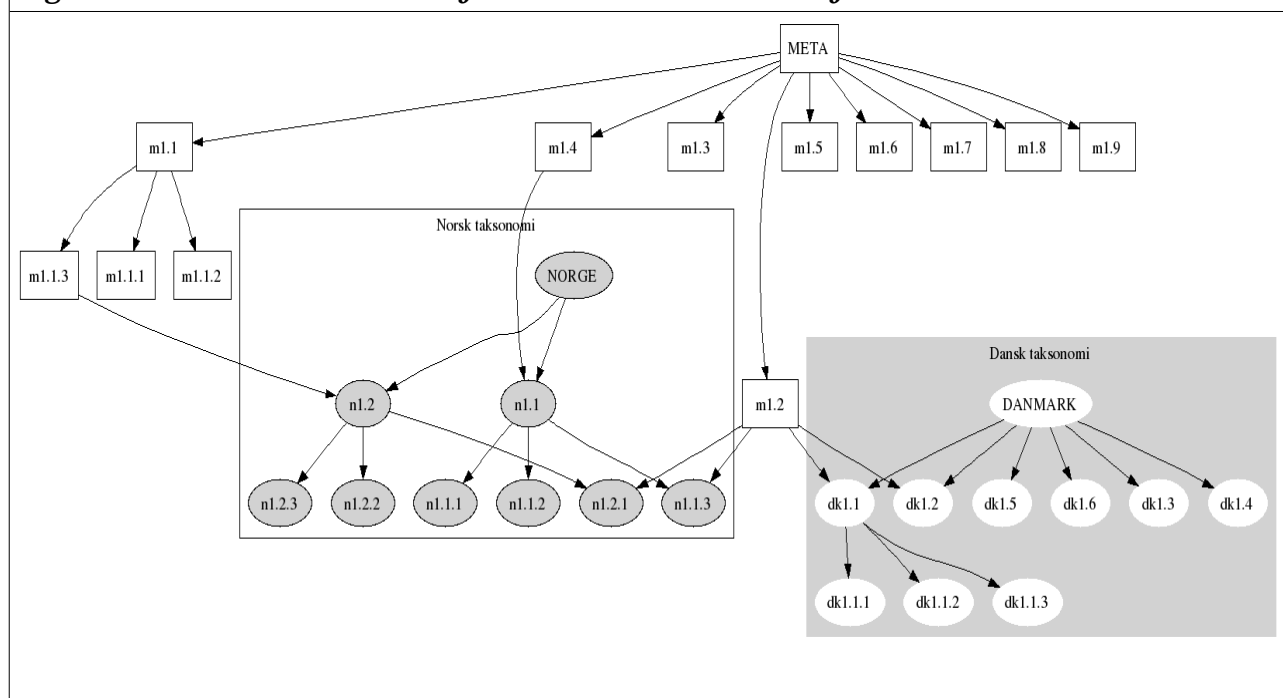
Faktaboks 3: Hvad er en taksonomi?

En mængde kontrollerede termer som repræsenterer kategorier der er organiseret hierarkisk og anvendes til klassifikation (og efterfølgende fremfinding) af data.

En metaemnetaksonomi er en emnetaksonomi over emnetaksonomier. Med andre ord en samordning af emnekategorier fra flere forskellige taksonomier i en enkelt, fælles taksonomi.

Figur 1 illustrerer således hvordan de opdigtede emnekategorier n1.2.1 og n1.1.3 fra den norske emnetaksonomi og emnerne dk1.1 og dk1.2 fra den danske emnetaksonomi vurderes at være synonyme og kan repræsenteres med en enkelt kategori i metaemnetaksonomien, nemlig kategorien m1.2.

Figur 1: En metaemnetaksonomi for alle nordiske emneklassifikationer



Den konkrete udformning af en sådan fællesnordisk emnetaksonomi vil uvægerligt indebære en hel del begrebsafklaring. **Ideelt set** burde man opbygge en komplet ontologi for domænet lingvistik med klare definitioner af samtlige begreber og udlede den fælles emnetaksonomi på denne basis. I praksis vil man sandsynligvis, grundet tidspres, kunne nøjes med at koble alle emnekategorierne fra de eksisterende taksonomier, anvende eksisterende definitioner og kun udarbejde nye definitioner i tvivlstilfælde (se eksemplet i afsnit 2.7). Den fælles emnetaksonomi kan med fordel ligge online i XML-format således at den kun behøver vedligeholdes ét sted, og således at alle baser gradvist kan gå over til at referere direkte til de fællesnordiske emnebetegnelser.

Virkeliggørelsen af denne idé kræver imidlertid en koordineret og løbende arbejdsindsats primært i opbygningsfasen, men også i den løbende vedligeholdelse af metaemnetaksonomien.

1.5.4 Hvem skal gøre hvad?

Gennem den fælles proces med at udarbejde metaemnetaksonomien vil de enkelte sprognævn få gennemgået deres kategorier og evt. opdage om der er emner/kategorier de mangler i deres taksonomi. Og hvis formålet opfyldes, vil man derefter have lettere adgang til at søge efter det præcise emne i andre baser – også selvom man ikke kender den lokale emnetaksonomi eller kan finde et brugbart stikord/søgeord. Det er måske ikke ofte at man hidtil har haft behov for at søge i hinandens baser, men det kan være at de praktiske vanskeligheder som har været forbundet med det, har gjort at man slet ikke har forsøgt.

Som nævnt ovenfor er det en forudsætning for et vellykket resultat at de relevante personer fra de involverede nævn deltager med deres kendskab til de eksisterende kategorier og brugssituationer. Den manuelle del af arbejdet består i at udarbejde en kobling mellem ens egen emnetaksonomi og den metaemnetaksonomi som bygges op. I dette indgår bl.a. afklaring og definition (også på engelsk) af hvad de forskellige emnekategorier dækker over, herunder en validering af omsætningsnøglen i bilag 5. Når koblingen mellem den nationale taksonomi og metaemnetaksonomien foreligger i en gennemarbejdet omsætningsnøgle, kan man i de fleste systemer mere eller mindre automatisk tilføje de relevante kategorinumre til basen (men dette vil ikke være absolut nødvendigt).

1.5.5 Hvad er allerede gjort?

Der er allerede udarbejdet et udkast til en omsætningsnøgle (bilag 5) for alle nævnenes emnekategorier. Desuden foreligger der en oversigt over de strukturerende elementer i de forskellige svar- og ordbaser (se henholdsvis afsnit 2.1.6 og bilag 14). Endvidere indeholder rapporten et forslag til en fælles svarbasestruktur (se afsnit 2.2) samt en anbefaling om at følge ISO-arbejdet med *Lexical Markup Framework* (LMF) tæt, og eventuelt anvende det som grundlag for en fælles ordbasestruktur (se afsnit 3.3.2).

2.0 Svarbaser

Denne del af rapporten fokuserer på de nordiske sprognævns svarbaser, dvs. databaser som indeholder de sproglige spørgsmål sprognævnene gennem tiden har modtaget fra borgerne og de svar de har afgivet. Først analyseres og sammenlignes de eksisterende svarbasestrukturer. Dernæst sammenlignes de emnetaksonomier de fleste sprognævn anvender til at organisere deres svarbaser. Sidstnævnte sammenligning inddrager endvidere en række internationale standarder. Endelig diskuteres et udkast til en fælles svarbasestruktur og, hvad endnu vigtigere er, en fælles emnetaksonomi for svarbaser.

2.1 Analyse af eksisterende strukturer

2.1.1 Danmark

Den grundlæggende struktur i den danske svarbase (der i øjeblikket indeholder ca. 10.000 svar) er som følger (se bilag 9 for den formelle struktur):

- Artikel
 - Hoved
 - Opslagsord
 - Emneklassifikation+
 - Oprettelse
 - Sidst redigeret
 - Nøgleord?
 - Henvisning?
 - Asset?
 - Krop
 - Spørgsmål
 - Svar
 - Gammelt svar?
 - Oplysninger om svaret
 - Svarkilde
 - Svardato
 - Svarforfatter
 - Spørger?

Følgende elementer i strukturen i bilag 9 er obligatoriske: hoved, krop, oplysninger om svaret, opslagsord, emneklassifikation, oprettelse, sidst redigeret, spørgsmål, svar, svarkilde, svardato og svarforfatter. Elementet ”emneklassifikation” kan tilmed forekomme flere gange (markeret med ”+”). Desuden indeholder hver artikel de følgende valgfrie elementer (markeret med ”?”): nøgleord, henvisning, asset, gammelt svar og spørger. Endelig er der de følgende obligatoriske attributter på rodelementet: id, status, regel, frekvensoplysninger og svartype.

Elementet ”nøgleord” kan indeholde et eller flere ord som skulle gøre det lettere at fremfinde artiklen ved søgning i databasen. Disse ord er typisk yderligere eksempler på det sproglige fænomen spørgsmålet vedrører. Elementet ”asset” indeholder en reference (et link) til en indscannet, elektronisk udgave af ældre, skriftlige svar.

Det obligatoriske element ”emneklassifikation” (barn til elementet ”hoved”) er centralt, idet spørgsmålets emne (eller som ofte er tilfældet: emner) her angives i forhold til en detaljeret emnetaksonomi som vil blive nærmere diskuteret og som er gengivet i sin helhed i bilag 1. Denne taksonomi fungerer således som en nøgle med hvilken man kan udtrække delmængder af svar som vedrører specifikke sproglige emner.

2.1.2 Sverige

Den svenske svardatabase er en del af en større database ved navn *Språklådan*. Ud over spørgsmål og svar indeholder *Språklådan* desuden fire andre typer poster:

1. spørgsmål (og svar)

2. litteraturhenvisninger
3. nyordsexcerpter
4. ordkort
5. emnekategorier

En **delmængde** af posterne med spørgsmål og svar udvælges til publicering på internettet og disse poster udgør *Frågelådan*. *Språklådan* indeholder knap 6000 svarposter (juni 2008) hvoraf ca. 2500 (juni 2008) vises i *Frågelådan*. Begge baser vokser således løbende. Hvert svar kan beskrives (**internt** i basen) med de følgende oplysningstyper:

- Spørger
- Oprindeligt spørgsmål/svar
- Publicérbart spørgsmål/svar
- Svardato
- Publiceringsstatus
- Henvisning/nøgleord
- Emnekategori+
- Svarforfatter
- Kommentar

Alle svar (men også litteraturhenvisninger og excerpter) kan kobles enten til ét eller flere ordkort, til én eller flere emnekategorier eller til både ordkort og kategori(er).

Den emnetaksonomi som anvendes for kategorielementet er gengivet i bilag 2.

Ud over den ovenstående svarbase administrerer det svenske sprognævn desuden en sverigefinsk version af *Språklådan*.

2.1.3 Norge

Det norske sprognævns svarbase indeholder følgende informationer:

- Svarbase (e-post)
 - tekst
 - dato
 - emne
 - kategori
 - artikel
- Ofte stillede spørgsmål
 - kategori
 - spørgsmål

- svar
- overskrift
- dato
- henvisning (intern)
- henvisning (ekstern)
- målform for teksten (bokmål/nynorsk)
- relevans for målform (bokmål/nynorsk/begge)
- kommentar

Hver gang nogen ændrer noget i en af baserne, registreres de følgende oplysninger: den redigerede post, brugernavn og dato.

Datatypen ”kategori” består af de følgende to dele:

- Beskrivelse/titel
- Nærmeste overkategori (”mor”)

I øjeblikket anvender det norske sprognavn sin egen emnetaksonomi (gengivet i bilag 3), men det danske udkast er også lagt ind i basen så det kan anvendes.

2.1.4 Finland

Det finske sprognavn (Focis) administrerer to svarbaser: en finlandssvensk version af Språklådan og en ren finsk svarbase (KITI). Førstnævnte har følgende struktur:

- Spørgsmål (Q)
- Svar (A)
- Område (AREA)
- Kategori (KAT)
- Spørger (QER)
- Svarer (AER)
- Oprettelsesdato (DAG)
- Sidst ændret dato (NYDAG)
- Svarkilde (SRC)
- Løbenummer (POSTNR)

Sidstnævnte har følgende struktur:

- Søgeord (HAKUSANA:1 PHRASE)
- Kategori (ERIKOISALA:2 PHRASE)
- Kontekst/eksempel (KONTEKSTI:3 TEXT)
- Kilde (LÄHDE:4 PHRASE)
- Nøgleord (ASIASANA:5 PHRASE)

Der anvendes i øjeblikket ingen emne kategorier i nogen af de to svarbaser (selvom der er et felt til denne information i baserne). Det gør der til gengæld i det finske nyordskartotek (se afsnit 3.1.4), og disse emne kategorier er gengivet i bilag 8.

2.1.5 Island

Det islandske sprognævn har en kombineret ord- og svardatabase ved navn Málfarsbankinn (<http://www.ismal.hi.is/malfar/>). Basen blev oprettet i 1998-2000 og indeholder ca. 8000 artikler med en simpel struktur.

Strukturen er som følger:

- **Artikelnummer** (fx 032/03280)
- **Opslagsord+** (fx F: lina, F: linna)
- **Kategori** (A: tilpasning av låneord, B: bøjning, D: ord, E: retskrivning, F: udtale, J: etymologi, L: udtryksmåde, M: betydning, N: navne, O: ordforbindelse, S: syntaks, Y: orddannelse)
- **Status** (B: artiklen har været ændret, V: artiklen er ikke færdiggjort)

Artikler som er blevet ændret gemmes i en liste over ufærdige artikler. For at vise dem til offentligheden skal redaktøren eksplicit markere dem som offentlige.

2.1.6 Sammenligning af oplysningstyper i nordiske svarbaser

Den korte gennemgang af strukturerne i de nordiske svarbaser illustrerede at der er en kerne af fælles oplysningstyper, men samtidig en række oplysninger som er mere eller mindre unikke for de enkelte sprognævn. Nedenstående skema (tabel 1) er et forsøg på at identificere denne fælles kerne af ækvivalente oplysningstyper (data kategorier) som kan danne basis for en fælles svarbasestruktur.

Skemaet viser at den eneste oplysningstype alle svarbaserne har tilfælles er "emne kategori". Derudover har stort set alle baser de to oplysningstyper, "svar" og "spørgsmål", selvom den islandske base eksempelvis blot indeholder usegenteret brødtekst. Mange baser har også metaoplysninger om eksempelvis spørgerens og svarerens navn (Sverige, Finland og Danmark) samt oprettelses- og redigeringsdato (Finland, Norge, Danmark). Desuden har fire af baserne et henvisningselement (Finland, Norge, Danmark, Sverige), mens tre af baserne har en særlig statusmarkering af de enkelte svar (Island, Sverige og Danmark). Endelig har de fleste baser et par unikke oplysningstyper, eksempelvis vedrørende sprogformen (Norge).

Tabel 1: Oplysningstyper i de nordiske svarbaser						
	ISLAND	FINLAND (KITI)	FINSV	SVERIGE	NORGE	DANMARK
Opslagsord	X	X				X
Emnekategori	X	X	X	X	X	X
Status	X			X		X
Svartype						X
Frekvensoplysninger						X
Regel ja/nej					?	X
Artikelnummer	X		X			X
”Brødtekst”	X				X	
Nøgleord		X		X		X
Spørger			X	X		X
Svarer			X	X		X
Kontekst/eksempel		X				X?
Spørgsmål	(X)		X	X	X	X
Svar	(X)		X	X	X	X
Oprindeligt svar				X		X
Oprindeligt spørgsmål				X		
Svardato				X	(X)	X
Overskrift					X	X
Oprettelsesdato			X		(X)	X
Sidst ændret dato			X		(X)	X
Kilde/henvisning		X			X	X
Henvisning (intern)				X	X	X
Kommentar				X	X	X
Tekstens sprogform					X	
Sprogformsrelevans					X	

2.2 Udkast til fælles svarbasestruktur

Da de nordiske svarbaser har så relativt få oplysningstyper tilfælles, virker det ikke hensigtsmæssigt at foreslå nogen rigid struktur som alle baserne skal overholde. Derimod vil vi alene foreslå at alle artikler som minimum indeholder de tre elementer ”emnekategori”, ”svar” og ”spørgsmål”. Endvidere vil det, rent søgemæssigt, være en fordel hvis alle artikler indeholder et element med ”opslagsord” eller ”nøgleord”. Endelig kunne det, rent æstetisk, være en fordel at samle alle ikke-sproglige metaoplysninger under et enkelt element og angive svarets status⁴, eksempelvis som en obligatorisk attribut på artikelelementet.

Vi vil derfor foreslå at de enkelte artikler i de nordiske svarbaser så vidt muligt overholder den følgende struktur (se bilag 10 for en formel gengivelse af XML-skemaet):

- Artikel

⁴ Med angivelse af om det er et internt udkast eller om svaret er klar til offentliggørelse.

- Opslagsord?
- Nøgleord?
- Emnekategori+
- Spørgsmål
- Svar
- Henvisning?
- Metadata
 - Svardato
 - Sidst ændret?
 - Spørgernavn?
 - Svarernavn?
 - Demografiske oplysninger?
 - Sprogform?
 - Kommentar?
 - ...

Som nævnt i indledningen er der de følgende store fordele ved at berige alle baser med oplysningstypen ”emnekategori” og sikre sig at alle sprognævne anvender emnekategorier som er defineret i en fælles metaemnetaksonomi og organiseret konceptuelt:

1. **Nye medarbejdere** kan hurtigt fremfinde relevant information fra svar- og ordbaser, som gamle medarbejdere måske kan i søvne.
2. **Svartjenesten bliver mere effektiv** og kan koncentrere sig om de vanskelige spørgsmål.
3. Sprognævnenes ansatte kan søge systematisk i **egne** baser og udtrække svar som vedrører mere eller mindre specifikke emner.
4. Sprognævnenes ansatte kan søge systematisk i **andres** baser uden først at skulle undersøge de enkelte sprognævns emnetaksonomier og deres respektive strukturer.
5. **Eksterne sprogforskere** (og andre interesserede) får også mulighed for 3. og 4. også selvom de ikke forstår alle de nordiske sprog.

Fordelene er således mange og gælder både gamle medarbejdere, nye medarbejdere og eksterne forskere (på lidt længere sigt også almindelige borgere).

Anvendelsen af elementerne nøgleord/opslagsord vil naturligt nok afhænge af de enkelte sprognævns softwareløsninger. I den danske svarbase er der således integreret et ”ordhjul” som nødvendiggør tilstedeværelsen af et opslagsord i alle artikler og et nøgleordsfelt som øger chancen for at fremfinde relevante artikler ved søgning. I Norge anvendes derimod en ekstern søgemotor hvilket gør et nøgleordsfelt overflødigt. De ikke-sproglige oplysninger i elementet ”metadata” kan være meget forskelligartede og mere eller mindre detaljerede. Demografiske oplysninger om spørgeren kunne eksempelvis omfatte personens alder, køn, bopælskommune og så videre. Det anbefales at elementet ”svardato” i det mindste er obligatorisk, da denne oplysning har vist sig

vigtig i stort set alle sprognævne.

Med hensyn til navngivningen af de enkelte elementer vil vi foreslå at der oprettes et såkaldt *Data Category Registry* (DCR) hvori elementerne, dvs. datakategorierne, dokumenteres med id-nummer, definitioner, forklaringer og ikke mindst udtømmende lister over alle de skandinaviske oversættelser af elementnavnene. Dermed skulle det blive lettere at søge på tværs af alle de skandinaviske svarbaser efter eksempelvis svar der er afgivet i 1998 uden først at skulle undersøge hvad ”svardato” hedder på eksempelvis finsk.

2.3 Sammenligning af nordiske emnetaksonomier

Vi bevæger os nu fra spørgsmålet om svarbasernes struktur til spørgsmålet om indholdet af et enkelt element i svarbaserne. Det foregående afsnit dokumenterede nemlig at oplysningstypen, ”emnekategori”, var en vigtig fællesnævner i alle de nordiske svarbaser. I forhold til tværskandinaviske svarbasesøgninger er det derfor af afgørende betydning at indholdet af netop oplysningstypen ”emnekategori” standardiseres. I dette afsnit vil vi således forsøge at sammenligne de emne kategorier der allerede finder anvendelse i de skandinaviske sprognævn, undersøge i hvilken udstrækning disse kategorier er identiske og endelig foreslå hvordan en fælles emnetaksonomi kunne udarbejdes.

Tabel 2 indeholder de mest overordnede emne kategorier der i dag anvendes i de forskellige sprognævns svarbaser. De komplette emnetaksonomier er gengivet i henholdsvis bilag 1 (Danmark), 2 (Sverige), 3 (Norge), 4 (Island) og 8 (Finland). I bilag 5 har vi forsøgt at parre alle svenske, norske, islandske og finske emne kategorier med de danske kategorier som et første skridt på vejen mod en fællesnordisk emnetaksonomi.

Tabel 2: Sammenligning af overkategorier i nordiske emnetaksonomier

Danmark	Island	Sverige	Norge	Finland
1.1 Morfologi	B: Bøjning Y: Orddannelse O: Ordforbindelse	1.1 Hur_ska_ordet_se_ut 1.2 När_ska_en_viss_ordf orm_anvendas 1.6 Namn_och_tilltal	A2 Korleis skal ordet bøyast?	sananmuodostus (morfologi)
1.2 Leksis	D: Ord	1.8 Fack_och_ämnesspråk (1.4 Skrivregler)	B1 Spørsmål om enkeltord	sanasto (leksis)
1.3 Ortografi	E: Retskrivning L: Udtryksmåde		A1 Korleis skal ordet (namnet) skrivast? A4 Skriveregler	oik. (retskrivning)
1.4 Semantik	M: Betydning	1.7 Ord_och_frassamlingar		semantiikka
1.5 Interpunktion		1.4 Skrivregler		
1.6 Pragmatik				
1.7 Etymologi	J: Etymologi A: Tilpasning_av_låneo rd		C3 Språket gjennom årene	etymologia
1.8 Layout			C4 Tekst og stil	
1.9 Videnscenter- spm	N: Navne	1.10 Språkinläring 1.12 Språkvetenskapliga_o mråden		

1.10 Fonetik	F: Udtale		A3 Korleis skal ordet (namnet) uttalast?	fon.
1.11 Syntaks	S: Syntaks	1.3 Konstruktioner_och_m eningsbyggnad	B2 Spørsmål om to eller flere ord i sammenheng	
1.12 Sproglig_variation		1.9 Talspråk_och_språkvar ieteter 1.5 Text_och_stil	C2 Språkvarianter	
2 XX		1.11 Främmande_språk 1.13 Övrigt	C1 Mer grammatikk C5 Annet	

Tabel 2 illustrerer i hvert fald fire forskellige problemstillinger:

- Hvem er emnetaksonomiens **målgruppe**?
- Bør der, helt grundlæggende, skelnes mellem **råd og regler**?
- Hvor **hierarkisk** skal emnetaksonomien være?
- Hvor **tematisk** omfattende bør emnetaksonomien være? Bør den også indeholde ikke-sproglige emner, eksempelvis videnscenterspørgsmål?

På et helt overordnet plan er det interessante ved eksempelvis den norske og svenske kontra den danske emnetaksonomi blandt andet at den norske og svenske målgruppe lader til at være den lidt bredere offentlighed, mens den danske målgruppe tydeligvis er sprogforskere og sprognævnets ansatte. Island ligger et sted midt i mellem. Der arbejdes imidlertid, både i Danmark og Norge, med at udvikle en nøgle til oversættelse mellem den interne emnetaksonomi og en offentlig udgave.

I nærværende projekt er fokus i første omgang at gøre det lettere for sprognævnenes ansatte (og sprogforskere generelt) at søge på tværs af hinandens databaser. Den væsentligste udfordring består således i 1) at undersøge hvilke emnekategorier der kan anses for ækvivalente, 2) at oprette en fællesnordisk emnetaksonomi som indeholder samtlige emner og 3) at ansøre de enkelte sprognævn til at anvende denne taksonomi aktivt. Dog vil det sandsynligvis ikke være noget stort problem at indføre både fagsproglige og almensproglige emneangivelser (på alle de nordiske sprog) i den fælles taksonomi, således at både af lægfolk og fagfolk kan få glæde af den via en fælles søgeportal på internettet.

Den norske emnetaksonomi er desuden interessant på et andet punkt; nemlig at der helt grundlæggende skelnes mellem regler ("Kva er rett?") og råd ("Hvordan bør jeg ordlegge meg?"). Denne skelnen findes ikke (særligt entydigt i hvert fald) i de andre nordiske emnekategorier. Årsagen hertil kan blandt andet være at der kan være forskel på de enkelte sprognævns myndighedsområder, eksempelvis ser det ud til at Norge har flere beføjelser i forhold til fastlæggelsen af stavning af stednavne og personnavne end Dansk Sprognævn har. Omvendt har Dansk Sprognævn mere direkte beføjelser når det gælder retskrivning generelt.

Dette giver anledning til at understrege at det **ikke** er projektets formål at ensrette de nordiske emnetaksonomier. Dette vil hverken være muligt eller ønskværdigt. Til gengæld vil vi i de følgende afsnit kort gennemgå de forskellige taksonomier og illustrere hvordan man med fordel kunne samle al denne viden i en enkelt, fælles resurse. Desuden vil vi helt konkret tage de første skridt i retning af at identificere et antal ækvivalente emnekategorier.

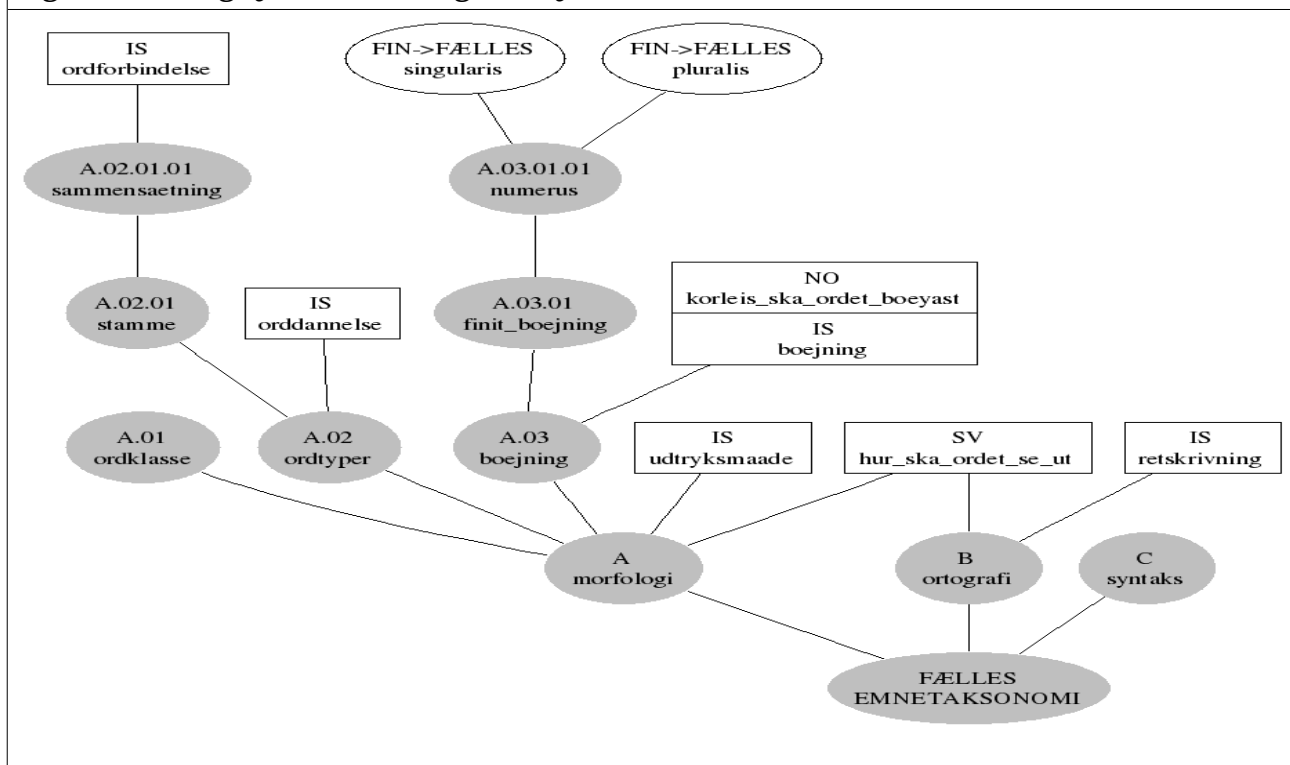
2.3.1 Parring af ækvivalente kategorier i en metaemnetaksonomi

Spørgsmålet om hvor hierarkisk og tematisk omfattende en sådan fælles emnetaksonomi bør være er af central betydning for harmoniseringsprojektets videre forløb. Vi har ladet os inspirere af et stort terminologisk projekt inden for biomedicin, nemlig Unified Medical Language System (UMLS)⁵, hvor man ved hjælp af en såkaldt metatesaurus har sammenkoblet en lang række forskellige tesaurusser og termlister i én stor begrebsdatabase. Formålet har blandt andet været at muliggøre effektiv fremfinding af specifikke begreber i lægevidenskabelige artikler uanset hvilke synonymer eller forkortelser der konkret er anvendt i teksterne (altså begrebsbaseret søgning). Da nærværende harmoniseringsprojekt til en vis grad har samme formål, virkede det oplagt at anvende en harmoniseringsmodel som minder om metatesaurusen i UMLS og altså sammenkøre de enkelte sprognævns emnetaksonomier i en overnational emnetaksonomi. Dog skal det understreges at en metatesaurus indeholder begreber og en metaemnetaksonomi indeholder emnekategorier.

Ligesom med UMLS kan den fællesnordiske emnetaksonomi i princippet være så hierarkisk, eller rettere polyhierarkisk⁶, som nødvendigt uden at de enkelte sprognavn behøver ændre på strukturen eller kategoriinventaret i deres nuværende taksonomier. Årsagen er at en metaemnetaksonomi består af alle emnekategorier fra de nationale emnetaksonomier som enten parres (hvis de er ækvivalente eller nær-ækvivalente) eller føjes til hierarkiet som selvstændige underkategorier (hvis de ikke har nogen ækvivalenter i de andre nationale taksonomier).

Alligevel vil det sandsynligvis være lettest at anvende den danske emnetaksonomi som grundlag, da den er den mest omfattende og mest hierarkiske. Den skal blot udvides med ekstra kategorier, skandinaviske synonymer, definitioner og eksempler. Man kan forestille sig at et lille fragment af en sådan parring kunne tage sig ud som gengivet i figur 2 hvor cirkler repræsenterer emnekategorier og firkanter repræsenterer synonymer (der altså hører til de pågældende kategorier).

Figur 2: Parring af nordiske kategorier i fælles metaemnetaksonomi



5 http://www.nlm.nih.gov/research/umls/about_umls.html

6 Et begrebssystem hvor de enkelte knuder kan have mere end én overkategori.

Figur 2 viser således at den svenske emnekategori "Hur ska ordet se ut?" anses for ækvivalent med to forskellige overordnede kategorier i metaemnetaksonomien, nemlig "ortografi" og "morfologi", mens en islandsk og en norsk kategori er ækvivalente med én og samme, mere underordnede kategori i metaemnetaksonomien, nemlig "morfologi bøjning". Som et andet eksempel har den danske emnetaksonomi begrebet "morfologi -> bøjning -> finit_bøjning -> numerus", men ikke de mere specialiserede begreber "singularis" og "pluralis" (som fx anvendes i Finland). I dette tilfælde vil det således være oplagt at tilføje de to underkategorier til kategorien numerus i metaemnetaksonomien (se figur 2).

2.3.2 Sverige kontra Danmark

Dette afsnit indeholder en detaljeret sammenligning af den svenske og den danske emnetaksonomi, samt et forslag til hvordan de svenske emnekategorier kan parres med de danske.

Et forsøg på at parre alle svenske emnekategorier til de tolv overordnede kategorier i den danske taksonomi gav følgende resultat:

1. Hur_ska_ordet_se_ut (12)	->	{ morfologi (6), ortografi (5), fonetik (1) }
2. Fack_och_ämnesspråk (11)	->	leksis (11)
3. Ord_och_frassamlingar (1)	->	semantik (1)/ syntaks (1)
4. Talspråk_och_språkvarieteter (4)	->	{ sproglig_variation (3), {Ø} (1) }
5. När_ska_en_viss_ordform_anvendas (6)	->	{ morfologi (5), {Ø} (1) }
6. Namn_och_tilltal (4)	->	morfologi (4), fonetik, ortografi, semantik
7. Främmande_språk (9)	->	{ {Ø} (6), semantik (1), etymologi (2) }
8. Konstruktioner_och_meningsbyggnad (9)	->	{ syntaks (5), {Ø} (2), semantik (1), morfologi (1) }
9. Text_och_stil (2)	->	{ sproglig_variation (1), {Ø} (1) }
10. Skrivregler (12)	->	{ layout (2), ortografi (3), interpunktion (2), morfologi / ortografi (2), {Ø} (2), morfologi (1) }
11. Språkinläring (3)	->	videnscenter-spm (3)
12. Språkvetenskapliga_områden (7)	->	{ {Ø} (4), videnscenter-spm (2), etymologi (1) }
13. Övrigt (3)	->	{ {Ø} (2), lexis (1) }

Tallene i parentes angiver hvor mange underkategorier de enkelte overkategorier indeholder (svensk side) eller hvor mange underkategorier der er blevet parret (dansk side). Den tomme mængde, {Ø}, betyder at der ikke kunne findes nogen dansk ækvivalent for den pågældende svenske

underkategori.

Ovenstående forsøg viser at man ikke bare entydigt kan parre de overordnede emnekategorier med hinanden, idet den overordnede svenske kategori *Hur_ska_ordet_se_ut* eksempelvis indeholder underordnede emner der tilhører hele tre forskellige overordnede kategorier i den danske taksonomi, nemlig morfologi, ortografi og fonetik. Endvidere er der en hel del 'huller' i den danske emnetaksonomi, idet seks af underkategorierne til kategorien *Främmande_språk* eksempelvis ikke har nogen modsvarende kategorier i den danske taksonomi. Som tidligere nævnt behøver dette imidlertid ikke være noget problem hvis man blot etablerer en altfavnende metaemnetaksonomi, identificerer ækvivalente begreber og parrer disse med denne metaemnetaksonomi.

Tabel 3 giver et mere detaljeret billede af de emneområder som er unikke i henholdsvis den svenske og den danske emnetaksonomi.

Unikt for svensk taksonomi		Unikt for dansk taksonomi	
<ul style="list-style-type: none"> • FACK-OCH_ÄMNESSPRÅK <ul style="list-style-type: none"> ○ alle underkategorier • FRÄMMANDE_SPRÅK <ul style="list-style-type: none"> ○ Språken_i_Norden ○ Latin_och_grekiska ○ Tyska ○ Teckenspråk ○ Minoritets-_och_invandrarsspråk • TALSPRÅK_OCH_SPRÅKVARIETETER <ul style="list-style-type: none"> ○ Finlandssvenska • SPRÅKVETENSKAPLIGA_OMRÅDEN <ul style="list-style-type: none"> ○ Retorik ○ Språk_och_kön ○ Språksociologi ○ Språkpsykologi ○ övriga_... 		<ul style="list-style-type: none"> • VIDENSCENTER-SPM <ul style="list-style-type: none"> ○ alle (undt. undervisning, sprogpolitik og sprogteknologi) • LEKSIS <ul style="list-style-type: none"> ○ alle • ETYMOLOGI <ul style="list-style-type: none"> ○ laan importmaade • PRAGMATIK <ul style="list-style-type: none"> ○ alle 	

Tabellen viser ret entydigt at den danske emnetaksonomi bør suppleres med eksempelvis en række fagområder, fremmedsprog og minoritetssprog for at kunne blive opgraderet til en altfavnende metaemnetaksonomi. Omvendt viser tabellen også at et område som pragmatik tilsyneladende spiller en mindre fremtrædende rolle i den svenske emnetaksonomi. Endelig illustrerer sammenligningen at man med fordel kan anvende flere forskellige emnekategorier fra en metaemnetaksonomi for at klassificere et givet svar; grænselandet mellem syntaks og semantik er et notorisk eksempel.

Med hensyn til fagområderne kunne man med fordel tage udgangspunkt i internationale standarder som DDC eller UDC (læs mere herom i afsnit 2.4).

2.3.3 Norge kontra Danmark

I den norske emnetaksonomi er der en helt overordnet skelnen mellem råd og regler som ikke ses i den danske. Desuden er der en fremtrædende distinktion mellem enkeltord og flerordsforbindelser som heller ikke findes i den danske emnetaksonomi. Endelig er de følgende kategorier unikke for

den norske taksonomi:

- Syntaks: ”determinativ”, ”og/å” (=og/at foran verbum)
- Layout: ”brevoppsett”, ”kursiv”, ”underskrift”
- Ortografi: ”mellomrom”⁷
- Etymologi: ”avløysarord”
- Leksis: ”dataord”, ”latin og gresk”, ”til nynorsk”
- Videnscenter-spm -> Sprogpolitik: ”samnorsk”
- Norrønt
- Pragmatik: ”helsing”

Ligesom da vi sammenlignede den svenske med den danske emnetaksonomi, er det også åbenbart at én kategori i den norske ofte modsvarer af flere kategorier i den danske og omvendt. I bilag 5 har vi forsøgt at parre samtlige norske kategorier med de danske. Unikke norske kategorier som ”brevoppsett” og ”kursiv” kunne eksempelvis tilføjes som nye underkategorier til ”1.8 Layout” i metaemnetaksonomien.

2.3.4 Island kontra Danmark

Den islandske emnetaksonomi har en meget flad struktur uden underkategorier. Til gengæld modsvarer de 12 overkategorier meget godt overkategorierne i den danske taksonomi. Eksempelvis må der antages at være direkte ækvivalens mellem ”M: Betydning” og ”1.4 Semantik”, F: Udtale” og ”1.10 Fonetik”, ”S: Syntaks” og ”1.11 Syntaks”, J: Etymologi” og ”1.7 Etymologi”. Desuden er der en delvis ækvivalens mellem ”B: Bøjning” og ”1.1 Morfologi”. Kategorien ”L: Udtryksmåde” kræver en nærmere definition, men er antageligvis delvist dækket af den danske kategori ”1.3 Ortografi”.

2.3.5 Finland kontra Danmark

Den finske emnetaksonomi minder om den islandske i og med at den kun har ét niveau. Til gengæld indeholder den et meget større antal emnekategorier. En stor del af disse er imidlertid forskellige fagområder (jf. bilag 8). Ud af i alt ca. 181 kategorier er 130 således fagområder, mens kun 51 er decideret sprogvidenskabelige emnekategorier (markeret med ”*” i bilag 8). De fleste af de 51 kategorier har ækvivalente kategorier i den danske emnetaksonomi (jf. bilag 5), men der er dog et mindre antal som ikke findes i den danske, nemlig:

- rekommendation
- namn på språk
- bildlig/metafor
- tidningsspråk, diktspråk, reklamspråk (dog 1.12.2.1 stilvarianter?)
- textforskning
- hälsingar
- konversationsanalys

⁷ Den danske emnetaksonomi har dog ”1.3.4.2 særskrivning”

Med hensyn til overkategorierne findes 5 af de 12 danske kategorier også i det finske inventar, nemlig morfologi (sananmuodostus), leksis (sanasto), semantik (semantiikka), etymologi (etymologia) og fonetik (fonetiikka) jf. tabel 2.

2.4 Internationale standarder for universelle emneklassifikationer

Gennemgangen af internationale standarder for emneklassifikationer er baseret dels på almene emnetaksonomier som Dewey Decimal Classification (DDC) og Universal Decimal Classification (UDC) og dels på lingvistiske emnetaksonomier så som OLAC og GOLD. Endelig vil vi kigge på to relevante taksonomier over datakategorier, nemlig Dublin Core og ISO 12620. Dog vil udgangspunktet i høj grad være den lingvistiske emnetaksonomi som allerede finder anvendelse i Dansk Sprognævn.

2.4.1 DDC (Dewey Decimal Classification)⁸

Dewey Decimal Classification (DDC) blev oprindeligt udviklet af Melvil Dewey i 1876 og har primært fundet anvendelse i biblioteksverdenen i ikke mindre end 22 forskellige reviderede udgaver.

DDC er et forsøg på at strukturere al viden i ti hovedkategorier som hver især har ti underkategorier der igen selv har yderligere ti underkategorier. Taksonomien indeholder altså med andre ord i alt 1000 kategorier.

Et eksempel på en emnekategori på øverste niveau er ”600 Technology (applied science)”. På næstøverste niveau kan denne kategori yderligere udspecificeres til eksempelvis ”630 Agriculture” og så fremdeles.

Fordelen ved DDC er at den er meget modulær; man kan med et simpelt decimaltal krydskategorisere ved at kombinere kategorier fra vidt forskellige dele af taksonomien.

En komplet liste over samtlige 1000 emnekategorier i DDC kan downloades her:

<http://www.oclc.org/dewey/resources/summaries/default.htm>.

DDC er fundamentet for en mere udtømmende, men samtidig også mere kompleks emnetaksonomi, nemlig Universal Decimal Classification (UDC), som kombinerer DDC-kategorier med forskellige andre tegn, fx komma, kolon, parenteser osv. Styrken ved DDC er imidlertid netop, at det er et enkelt system som ofte opdateres.

2.4.2 UDC (Universal Decimal Classification)⁹

Universal Decimal Classification (UDC) er en emnetaksonomi til biblioteksbrug udviklet af de belgiske bibliotekarer Paul Otlet og Henri la Fontaine i slutningen af det 19. århundrede. Systemet minder om DDC, men har desuden den følgende notation:

- +: **kombination**, fx 59+636 (zoologi og dyreavl)
- /: **udvidelse**, fx 592/599 (systematisk zoologi, alt fra 592 til og med 599)
- : **relation** (fx 17:7, relation mellem etik og kunst)
- []: **undergruppering** (fx 31:[622+669](485), statistik om minedrift og metallurgi i Sverige)
- =: **sprogangivelse** (fx =20 på engelsk; 59=20 zoologi på engelsk).

⁸ Kilde: www.wikipedia.org

⁹ Kilde: www.wikipedia.org

2.5 Internationale standarder for lingvistiske emner

DDC og UDC er universelle emnetaksonomier som dækker alt mellem himmel og jord. I forhold til de nordiske svarbaser er det imidlertid klart at et enkelt fagområde spiller en altdominerende rolle, nemlig lingvistikken. Dette afsnit vil således give et meget groft overblik over nogle internationale standarder med relevans for **lingvistiske** emnetaksonomier.

2.5.1 OLAC (Open Language Archives Community)

Open Language Archives Community (OLAC) er et internationalt samarbejde mellem forskningsinstitutioner og individer som ønsker at skabe et verdensomspændende virtuelt bibliotek over sprogresurser ved 1) at skabe konsensus om ideelle metoder til arkivering af sprogresurser og 2) at udvikle et netværk af kompatible arkiver og søgeværktøjer for disse resurser.

OLAC har blandt andet udgivet et "Linguistic Subject Vocabulary" (jf. <http://www.language-archives.org/REC/field.html>), men overordnet set er dets standarder ikke detaljerede nok til nærværende formål.

2.5.2 GOLD (General Ontology for Linguistic Description)

General Ontology for Linguistic Description (GOLD) blev oprindeligt lanceret af [Farrar and Langendoen \(2003\)](#). Visionen med GOLD var at løse problemet med heterogene og inkompatible opmærkningsstandarder for sproglige data, især for uddøende sprog. I dag er GOLD dog en meget mere generel standard som kan anvendes på ethvert sprog.

Will Lewis skabte den første version af ontologien ved at transformere og strukturere informationer fra "SIL International's [on-line glossary](#) of linguistic terms". En gruppe forskningsassistenter fra University of Arizona fordoblede derpå antallet af termer ved at granske den lingvistiske litteratur.

Siden november 2004 er et GOLD community blomstret op, og der er nu etableret en [Ontology Wiki](#) med det formål at anspore til en vedvarende udvidelse og løbende revision af GOLD.

GOLD er meget omfattende og indeholder en lang række definitioner af sprogvidenskabelige begreber som kunne være en stor hjælp i forbindelse med udarbejdelsen af en fællesnordisk emnetaksonomi.

2.5.3 CLARIN (Common Language Resources and Technology Infrastructure)

CLARIN er et igangværende projekt hvis målsætning er at tilgængeliggøre sprogresurser og sprogteknologi på europæisk plan og samtidig sikre kompatibiliteten mellem resurser på de forskellige europæiske sprog. En del af projektet går ud på at etablere såkaldte BLARKs (Basic Language Resource Kits) for de enkelte sprog og her er Dansk Sprognævn eksempelvis involveret i opbygningen af et stort almensprogligt tekstkorpus, men også et fagsprogligt korpus. En primær udfordring vil her være at udtænke og implementere en opmærkningsstandard (både for sproglige og metasproglige oplysningstyper) som følger fælles retningslinjer inden for CLARIN-projektet og dermed sikrer paneuropæisk kompatibilitet. Resultaterne af CLARIN-projektet vil således også være interessante for det nordiske databaseharmoniseringsprojekt (omend de ikke foreligger endnu).

2.6 Internationale standarder for relevante datakategorier

2.6.1 Dublin Core¹⁰

Dublin Core refererer til en mængde metadata-elementer udviklet med henblik på udveksling af informationskilder på tværs af forskellige fagområder. Standarden omfatter en række enkle konventioner for hvordan man kan beskrive ting online således at de bliver lettere at genfinde. Dublin Core bruges eksempelvis til at beskrive digitalt materiale som video, lyd, billeder, tekst og sammensatte medier som hjemmesider. De fleste implementeringer af Dublin Core gør brug af XML og RDF (Resource Description Framework).

Det forenkede Dublin Core Metadata Element Set (DCMES) består af de følgende 15 metadata-elementer:

1. Title
2. Creator
3. Subject
4. Description
5. Publisher
6. Contributor
7. Date
8. Type
9. Format
10. Identifier
11. Source
12. Language
13. Relation
14. Coverage
15. Rights

Hvert Dublin Core element er valgfrit og kan gentages. Selvom Dublin Core ikke indeholder nogen sprogvidenskabelige kategorier, kunne man sandsynligvis anvende DCMES som grundlag for udarbejdelsen af et antal fælles administrative (ikke-sproglige) oplysningstyper.

2.6.2 ISO 12620 (Data Category Registry)

ISO 12620, *Computer applications in Terminology – Data categories*, definerer og beskriver et større antal datakategorier som finder anvendelse i forbindelse med opmærkning af og søgning i terminologiske data.

Datakategorierne dækker både informationer som er relevante for svarbaser og ordbaser. De omfatter eksempelvis ikke blot sproglige informationer, men også en lang række rent administrative/ikke-sproglige oplysningstyper.

Dette afsnit undersøger hvilken delmængde af disse datakategorier kan være anvendelig i forhold til etableringen af en fællesnordisk emnetaksonomi. Afsnittet forsøger samtidig at identificere de emnekategorier som spiller en afgørende rolle i sprognævnenes svarbaser, men som ikke har nogen modsvarende datakategori i ISO 12620.

Nedenstående tabel (tabel 4A) indeholder alle hovedkategorierne i ISO 12620, mens den følgende tabel (tabel 4B) sammenholder samtlige kategorier i ISO 12620 med kategorierne i den danske emnetaksonomi og opremser de kategorier vi anser for ækvivalente.

¹⁰ Kilde: www.wikipedia.org

Overkategori	Underkategorier (eksempler)
A 1 term	
A 2 term related information	term type, grammar, usage, term formation, pronunciation, syllabification, hyphenation, morphology, term status, degree of synonymy
A 3 equivalence	degree of equivalence, false friend, directionality, reliability code, transfer comment
A 4 subject field	classification system, classification number
A 5 concept-related description	definition, explanation, context, example, nontextual illustrations, unit, range, characteristic
A 6 concept relation	generic, partitive, sequential, temporal, spatial, associative
A 7 conceptual structures	concept system, concept position
A 8 note	
A 9 documentary language	thesaurus name, thesaurus descriptor, nondescriptor, keyword, index heading
A 10 administrative information	terminology management, language symbol, concept identifier, source identifier, ...

Emnekategorier i dansk taksonomi	Datakategorier i ISO 12620
Morfologi ordklasse (1.1.1) Morfologi ordklasse substantiv [proprium] (1.1.1.7[.1]) Morfologi ordtyper stamme forkortning (1.1.2.1.3) Morfologi ordtyper stamme forkortning forkortelsestype initialforkortelse (1.1.2.1.3.4.2) Morfologi boejning finit_boejning numerus (1.1.3.2.7) Morfologi boejning finit_boejning genus (1.1.3.2.1)	part of speech (A 2.2.1) common/proper noun (A 2.2.5) abbreviated form (A 2.1.8) acronym (A 2.1.8.4)
Ortografi tegn orddannelsestegn bindestreg (1.3.1.2.1)	grammatical number (A 2.2.3) grammatical gender (A 2.2.2)
Syntaks kollokationer (1.11.1)	hyphenation (A 2.7)
Semantik fast_udtryk (1.4.5) Semantik synonym (1.4.2) Semantik betydning begrebsafklaring (1.4.4.1)	collocation (A 2.1.18.1)
Fonetik (1.10) Fonetik udtale stavelse (1.10.2.3)	phraseological unit (A 2.1.18) synonym (A 2.1.2) definition (A 5.1)
Etymologi (1.7)	pronunciation (A 2.5) syllabification (A 2.6)
	etymology/term provenance (A 2.4.1) neologism (A 2.4.1)

Tabel 4b: Ækvivalente kategorier i ISO 12620 og den danske emnetaksonomi?	
Leksis nyt_ord (1.2.10) Leksis varemaerke (1.2.9)	proprietary restriction (A 2.3.7)
Sproglig_variation lekter dialekt (1.12.2.3) Sproglig_variation stilvarianter slang/fagsproglig_stil (1.12.1.1/3)	geographical usage (A 2.3.2) slang register (A 2.3.3)

Tabel 4A og 4B viser at der om ikke andet så i hvert fald er et **delvist** overlap mellem datakategorierne i ISO 12620 og emnekategorierne i den danske emnetaksonomi. En lang række af datakategorierne i underkategorien A2 ("term related information") modsvarer kategorier på stort set alle niveauer i den danske svarbases emnetaksonomi, og deres definitioner kunne med fordel anvendes som dokumentation i en fællesnordisk emnetaksonomi. Dog er det vanskeligt at finde datakategorier i ISO 12620 som modsvarer underkategorier til "Videnscenter-spm", "Ortografi", "Syntaks", "Layout" og "Interpunktion". Omvendt indeholder ISO 12620 en lang række datakategorier vedrørende fagområder, semantiske relationer og andre konceptuelle oplysninger som, naturligt nok, er af knap så afgørende betydning for svarbaser som vedrører almensproglige spørgsmål og dermed i højere grad tager udgangspunkt i udtryksiden af sproglige tegn end indholdssiden.

2.7 Udkast til fælles emnetaksonomi

I bilag 5 har vi forsøgt at identificere ækvivalente kategorier i den norske, svenske, islandske, finske og danske emnetaksonomi. Det fremgår af henholdsvis bilag 1, 2, 3, 4 og 8 hvilke kategorier numrene i bilag 5 dækker over.

Tabel 5: Overlap mellem danske, svenske og norske emnekategorier				
		Danmark	Sverige	Norge
Antal kategorier i alt		317	93	126
Antal niveauer i taksonomien		7	2	4
Antal danske kategorier med ækvivalenter i mindst 2 taksonomier	99			
Antal danske kategorier med ækvivalenter i alle tre taksonomier	40			

Tabel 5 viser at den danske emnetaksonomi er den mest omfattende og dermed sandsynligvis også det bedste/letteste udgangspunkt for oprettelsen af en fællesnordisk metaemnetaksonomi. Tabellen viser imidlertid også at hvis vi kræver at en kategori skal have ækvivalenter i bare tre af de nordiske emnetaksonomier for at blive en del af en metaemnetaksonomi, så daler antallet markant.

Snarere end at foreslå en minimalistisk fællesnordisk emnetaksonomi baseret på **fællesmængden** af kategorier landene imellem, vil vi derfor anbefale at man gør den fælles taksonomi så inklusiv som muligt ved at tage **foreningsmængden** af alle de nordiske emnekategorier, identificere ækvivalente kategorier og dernæst gruppere og definere dem. At denne fælles emnetaksonomi så også kommer til at indeholde en hel del emnekategorier som enkelte sprognævn måske sjældent (og i nogle

tilfælde sandsynligvis aldrig) vil komme til at anvende vil næppe udgøre noget problem. Det afgørende vil være at kompatibiliteten mellem databaserne sikres, således at sproglige svar som omhandler sammenlignelige emner opmærkes så de efterfølgende kan fremfindes med identiske forespørgsler på tværs af alle databaserne.

I forhold til udarbejdelsen af fælles definitioner anbefaler vi at man i videst mulig udstrækning anvender eksisterende definitioner fra eksempelvis ISO 12620 og resurser som *Glossary of linguistic terms* (<http://www.sil.org/linguistics/glossaryoflinguisticterms/index.htm>). Endelig foreligger der et stort antal danske definitioner af sproglige begreber i eksempelvis Galberg Jacobsen (1996) som har været et af udgangspunkterne for den danske emnetaksonomi.

Hermed følger et eksempel på hvordan dette afklarings- og harmoniseringsarbejde kunne formaliseres (formatet er inspireret af ISO 12620).

1.8 *Layout*

DEFINITION (ENG): Layout is ...

DEFINITION (DA): Layout omfatter spørgsmål om regler og råd vedrørende placering og formatering af tekstlige og grafiske virkemidler på tryk.

EKSEMPEL (DA): Formatering af litteraturlister.

SYNONYMER:

NO1: Avsnitt, brevoppsett og punktopstilling

NO2: Brevoppsett/e-postoppsett

NO3: Helsing og underskrift

SV1: Skrivregler Stykke

SV2: Skrivregler Grafik

Fordelene ved en fællesnordisk emnetaksonomi er allerede beskrevet, men en enkelt ulempe er naturligvis at det vil kræve en smule resurser at holde den opdateret. Ideelt set burde taksonomien være tilgængelig som en online XML-kilde, således at man ved oprettelse af nye svarbaseartikler på de enkelte sprognævn helt automatisk anvender den til enhver tid nyeste version af emnetaksonomien. Desuden bør man på længere sigt overveje om man vil tage de fælles kategorinumre i brug i de nationale databaser. I første omgang vil det imidlertid ikke være absolut nødvendigt at ændre på de eksisterende svar i de forskellige nordiske svarbaser. Her mener vi at den udarbejdede ækvivalensnøgle i bilag 5 er et vigtigt første skridt, selvom de enkelte sprognævn naturligvis må verificere nøglen og melde tilbage med eventuelle rettelser og kommentarer.

2.8 *Metaemnetaksonomiens konsekvenser*

Selvom implementeringen af en nordisk metaemnetaksonomi har en lang række fordele, er det dog samtidig et initiativ som forpligter og involverer en vis mængde arbejde (se også afsnit 1.5.2) og koordination af samme.

- Der skal udarbejdes en **kobling** mellem ens eget nævns emnekategorier og metaemnetaksonomien
 - Herunder skal der, efter behov, udarbejdes definitioner samt

almensproglige/fagsproglige synonymer for mange emnekategorier

- Metaemnetaksonomien skal udvides efter behov
- Metaemnetaksonomien skal **tilgængeliggøres online**, fx som et XML namespace
- Der skal udpeges en **kontaktperson** fra hvert sprognavn hvoraf en **fuldtidskoordinator** bærer ansvaret for at sætte projektet i gang og udvælge emnekategorier, mens en **vedligeholdelsesansvarlig** løbende **opdaterer** metaemnetaksonomien
- Kontaktpersonerne skal huske at **indberette** enhver ændring/tilføjelse til deres nævns emnetaksonomi.

Nedenfor har vi skitseret en mulig tidsplan for projektet. Det vurderes at den største arbejdsbyrde ligger i at opnå konsensus om de udvalgte emnekategorier, med andre ord selve begrebsafklaringen herunder især udarbejdelsen af fælles definitioner (fase 1 og 2 i den skitserede plan).

2009				2010				2011	
jan.	apr.	jul.	okt.	jan.	apr.	jul.	okt.	jan.	apr.
1	2			3				4	

1	- Forberedelse af det praktiske samarbejde - Enighed om hvilke emnekategorier første fase omfatter (fx 3 niveauer) - Første identifikation af emnekategorier	3 mdr.	fuldtids-koordinator
2	- Fælles identifikation og definition af emnekategorier - Konsensusopnåelse - forskellige stadier - Offentliggørelse af de besluttede emnekategorier	ca. 9 mdr.	- hvert nævn ca. 1-1½ mdr. - koordinator på halv tid
3	- Implementering i de forskellige baser - Udvikling af søgegrænseflade (evt. eksternt udvikler)	6-12 mdr. 3 mdr.	- hvert nævn - koordinatoren
	I alt	1½-2 år	
4	- Evaluering og planer for fase 2 - Vedligeholdelse	fremover	- vedligeholdelsesansvarlig

2.9 Konklusion og diskussion

Vi vil anbefale at man anvender den udvidede danske emnetaksonomi som grundlag for etableringen af en metaemnetaksonomi. Som sagt vil det dog sandsynligvis kræve en del arbejde at opbygge en sådan, og det vil fordre en koordineret indsats som der ikke er afsat resurser til i dette projekt. Andre delkonklusioner og åbne spørgsmål på nuværende tidspunkt i projektet opridses i form af de følgende punkter:

- **Hvor hierarkisk skal den fælles emnetaksonomi være?**
 - Vi anbefaler 3 niveauer (gruppe -> kategori -> underkategori), da dette lader til at være den mest udbredte praksis.
 - Dog vil det altid være en balancegang at vurdere om en sammensat kategori som den norske "C1.4 Konsekvente former" bør optræde i emnetaksonomien som en selvstændig underkategori eller som en kombination af flere overordnede kategorier (artikel+genus+kongruens). Det giver større fleksibilitet at tillade en modulær angivelse af emne (sidstnævnte model) end at insistere på enkeltvalg fra en udtømmende, og dermed meget hierarkisk liste.

- **Hvem er emnetaksonomiens målgruppe?**
 - Er det sprogforskere og sprognævnansatte eller den almindelige befolkning?
 - Her mener vi man med fordel kan kombinere begge målgrupper ved eksplicit at tilføje både fagsproglige betegnelser (fx morfologi) og almensproglige betegnelser (fx hvordan skal ordet bøjes?) som synonymmer i den fælles emnetaksonomi. Og naturligvis markere hvilken målgruppe hvert synonym er rettet mod.
 - Alle overvejelser om målgrupper bør påvirke både udfærdigelsen af selve metaemnetaksonomien og udviklingen af brugergænsefladen til en fælles søgeportal.
- **Hvor tematisk omfattende skal emnetaksonomien være?**
 - Kun strengt sprogvidenskabelige kategorier eller også mere generelle emner?
 - Her anbefaler vi at man tager udgangspunkt i eksempelvis de tre øverste niveauer i den danske emnetaksonomi og dermed også inkluderer videnscenterspørgsmål, som ikke vedrører strengt lingvistiske emner, men emner der er vigtige for de borgere som retter henvendelse til sprognævne.
- **I hvilken udstrækning kan internationale standarder være en hjælp?**
 - Definitioner fra blandt andet ISO 12620, OLAC og GOLD kan med stor fordel genanvendes i arbejdet med etableringen af den fælles emnetaksonomi.
 - DDC eller UDC bør konsulteres for emner der vedrører specifikke fagsprog. Dette vil nemlig sikre at de nordiske baser ikke bare er kompatible med hinanden, men også er internationalt kompatible.

3.0 Ordbaser

Denne del af rapporten omhandler de nordiske sprognævns ordbaser, dvs. databaser som indeholder leksikografiske oplysninger af den ene eller anden art. På samme måde som i rapportens svarbaseafsnit begynder vi med at analysere og sammenligne de eksisterende ordbasestrukturer. Dernæst inddrages og analyseres en række internationale standarder for leksikografiske databaser, eksempelvis TEI, LMF og ikke mindst formatet bag den nordiske netordbog. Endelig diskuteres et udkast til en fælles ordbasestruktur.

3.1 Analyse af eksisterende strukturer

3.1.1 Danmark

Det danske sprognævn har tre ordbaser:

- En ordsamlingsbase (med excerpter)
- En Retskrivningsordbog (RO)
- En nyordsordbog (NOID)

3.1.1.1 Ordsamlingsbasen

Ordsamlingen indeholder et stort antal excerpter (ca. 1 million alt i alt) hvoraf de fleste stadig kun forefindes i fysisk form i et kartotek, mens ca. en fjerdedel forefindes i elektronisk form i en konventionel database (dvs. ikke "native XML"). Dansk Sprognævn er imidlertid ved at få indscannet samtlige kartotekskort og har netop fået konverteret den gamle database til en samling

XML-dokumenter der har den følgende struktur (se evt. det detaljerede XML-skema i bilag 11):

- Artikel
 - Opslagsord
 - Homograf+
 - Ordklasse
 - Ordets brug
 - Domæne
 - Emneklassifikation?
 - Retskrivningsordbogen (1955,1986,1996,2001)?
 - Belæg+
 - Asset
 - Dato
 - Belægstekst
 - Kildehenvisning
 - Koder
 - Excerpt
 - Kommentar?
 - Henvisning?

Hvert dokument i ordbasen indeholder altså som minimum et opslagsord og en eller flere homografer, der igen indeholder mindst ét belæg (dvs. excerpt) med en belægstekst samt en række metasproglige/administrative oplysninger. Endvidere er de tre elementer ”Ordklasse”, ”Ordets brug” og ”Domæne” også obligatoriske, idet deres indhold afgør om der skal oprettes en ny homograf i artiklen, eller om der blot skal oprettes et nyt belæg under en eksisterende homograf. Endelig kan hver homograf indeholde oplysninger om hvorvidt ordet står i én eller flere forskellige udgaver af Retskrivningsordbogen. Elementet ”asset” refererer til en billedfil med en indscannet udgave af belægget hvor man kan læse hele det oprindelige tekstudklip.

3.1.1.2 Retskrivningsordbogen

Den grundlæggende struktur i RO er som følger (se bilag 12 for en detaljeret beskrivelse):

- Homograf
 - Metadata?
 - {adj-artikel,adv-artikel,konj-artikel,pron-artikel,sb-artikel,vb-artikel,...}+
 - Hoved
 - Opslagsord
 - Delepunkter

- Opslagsord.altform?
- {bøjningsformer}+
- Hoved.altform?
- Krop?
 - Glosse+
 - Beskriver+
 - Stil
 - Forkortes+
 - Henvisning
 - Eksempel+
 - Kommentar

Som det fremgår af bilag 12 har Retskrivningsordbogen en meget hierarkisk struktur hvilket delvist skyldes at den er blevet halvautomatisk konverteret til XML fra et ældre SGML-baseret format der havde blandet præsentation og indhold.

3.1.1.3 Nyordsordbogen

Den grundlæggende struktur i nyordsordbogen er (se bilag 13 for en mere detaljeret gengivelse):

- Artikel
 - Opslagsord+
 - Udtale?
 - Ordklasse?
 - Bøjning?
 - Brug?
 - Betydning?
 - År?
 - Citater?
 - Henvisning?
 - Etymologi (hjemlig)?
 - Etymologi (udenlandsk)?
 - Kilde?
 - Administration

I modsætning til Retskrivningsordbogen har Nye Ord i Dansk en ret flad struktur og indeholder en del andre oplysningstyper, eksempelvis oplysninger om opslagsordets henholdsvis hjemlige (dvs.

danske) og udenlandske etymologi, herunder kildeangivelser. Elementet ”År” indeholder oplysninger om hvilket år opslagsordet blev observeret første gang.

3.1.2 Sverige

Det svenske sprognævn har tre typer ordbaser:

- Excerpter i Språklådan
- Ordkort i Språklådan
- Bilingvale og monolingvale ordbogsbaser (Lexin og nordisk netordbog)

Excerpter i Språklådan indeholder følgende oplysningstyper:

- Kilde
- Dato
- Tekst (kontekst)
- Kommentar
- Publiceringsstatus
- Henvisning/nøgleord
- Kategori
- Redaktør

Ordkort i Språklådan indeholder følgende oplysninger:

- Ord
- Orddannelse
- Bøjning
- Ordklasse
- Betydning
- Orddeling
- Udtale (billede/lyd)
- Redaktør
- Årstal
- Publiceringsstatus
- Kommentar

Lexin er oprindeligt et svensk projekt om produktion af ordbogsdata og leksika med særlig fokus på indvandrer målgruppen, men sidenhen har projektet udviklet sig til at omfatte alle de skandinaviske lande.

Lexins dataformat er ikke XML-baseret, og vi vil ikke kommentere det nærmere i denne udredning. Dog følger her et eksempel på hvilke oplysningstyper en bilingval ordbogsartikel typisk indeholder i

Lexin.

- Opslagsord (fx ”jätte”)
- Oversættelse (”giant”)
- Ordklasse (subst.)
- Ordklasse, oversættelse (noun)
- Betydning1 (sagofigur som är mycket större än en människa)
- Betydning2 (även bildligt om något stort i allmänhet)
- Oversættelse af betydning2 (also used figuratively of large objects)
- Sammensætning (atlantjätte)
- Oversættelse af sammensætning (big Atlantic liner)
- Løbenummer for sammensætning
- Bøjningsformer (jätten jättar)
- Udtale (²j'et:e)
- Løbenummer for opslagsord og varianter (98 7368, 99 8105)

Det nordiske netordbogsformat er til gengæld XML-baseret og nærmere beskrevet her: <http://www.csc.kth.se/tcs/projects/netordbog/format/>. Da dette format er en reduceret udgave af den internationale standard TEI (Text Encoding Initiative) vil det imidlertid først blive diskuteret i afsnittet ”Internationale standarder for ordbaser”.

3.1.3 Norge

Det norske sprognævn har medvirket til at udarbejde en lang række ordbøger, men de mest centrale er de følgende to ordbøger:

- Bokmålsordboka (65.000 opslagsord)
- Nynorskordboka (90.000 opslagsord)

Da den tekniske udformning og vedligeholdelse af disse ordbøger ikke varetages af selve det norske sprognævn, men af Universitetet i Oslo, indeholder rapporten ikke yderligere oplysninger om de norske ordbaser.

3.1.4 Finland

Det finske sprognævn har følgende ordbaser:

- En nyordsbase
- En navnebase

Begge baser er såkaldte TRIP-baser hostet på en unix-server. Ingen af dem er XML-baserede, og begge baser har en ret flad og enkel struktur.

Nyordsbasen indeholder eksempelvis de følgende oplysningstyper:

- Søgeord (HAKUSANA:1 PHRASE)
- Kategori (ERIKOISALA:2 PHRASE)
- Eksempel med kontekst (KONTEKSTI:3 1..1*TEXT ORIG)
- Kilde (LÄHDE:4 PHRASE)

Navnebasen indeholder disse oplysningstyper:

- Anbefalet form (SUOSITUS:1 TEXT)
- Bøjning (TAIVUTUS:2 TEXT)
- Kategori (ASIASANA:3 PHRASE)
- Geografisk placering (SIJAINI:4 TEXT)
- Henvisning (VIITE:5 TEXT ORIG)
- Baggrund (TAUSTAA:6 TEXT ORIG)
- Ikke-anbefalet form (HYLÄTTY:7 TEXT)
- Ændringer (MUUTOS:8 TEXT)
- Dato (PÄIVÄYS:9 TEXT)

3.1.5 Island

Det islandske sprognævn har en kombineret ord- og svardatabase ved navn Málfarsbankinn (<http://www.ismal.hi.is/malfar/>) som tidligere beskrevet.

3.2 Sammenligning af ordbasestrukturer

Bilag 14 indeholder en sammenligning af oplysningstyperne i de nordiske ordbaser. Udgangspunktet for sammenligningen er det nordiske netordbogsformat som anvendes i Sverige. Denne struktur er nemlig baseret på den internationale TEI-standard for opmærkning af ordbogsdata.

Bilag 14 viser at der er en del forskelle på oplysningstyperne i de 6 ordbøger/ordbaser. Dette er naturligt nok givet de forskellige formål ordbaserne tjener. Det er imidlertid muligt at identificere en kerne af oplysningstyper som går igen i langt de fleste baser. Elementet ”Opslagsord” optræder eksempelvis i samtlige baser, mens elementerne ”Definition/Betydning”, ”Ordklasse”, ”Eksempel/Citat” og ”Henvisning/Kilde” forekommer i 5 ud af de 6 baser. Endelig er der oplysninger om ”Udtale” i 4 ud af 6 baser.

Dog viser skemaet også at den forenkede TEI-standard mangler elementer for eksempel oplysninger om etymologi, emnekategori og publiceringsstatus.

3.3 Internationale standarder for ordbaser

Det ville være ideelt hvis alle de nordiske ordbaser var struktureret på samme måde efter et fælles skema. Hvis dette skema tilmed var en del af en international standard, ville det blive muligt for sprogforskere, også uden for Skandinaviens grænser, at søge i de nordiske ordbaser på problemfri facon. Dette afsnit undersøger således hvilke internationale standarder der findes for opmærkning af leksikografiske data og giver desuden en vurdering af hvilke af disse standarder der ville være de mest hensigtsmæssige at tage i anvendelse.

3.3.1 Text Encoding Initiative (TEI)

Som gengivet i tabel 6 består Text Encoding Initiative (TEI) i øjeblikket af i alt 22 forskellige XML skemaer (eller moduler).

Tabel 6: Modulerne i Text Encoding Initiative (TEI)

Figure 1. The TEI modules.

analysis	Simple analytic mechanisms
certainty	Certainty and uncertainty
core	Elements common to all TEI documents
corpus	Header extensions for corpus texts
declarefs	Feature system declarations
dictionaries	Printed dictionaries
drama	Performance texts
figures	Tables, formulae, and figures
gaiji	Character and glyph documentation
header	The TEI Header
iso-fs	Feature structures
linking	Linking, segmentation and alignment
msdescription	Manuscript Description
namesdates	Names and dates
nets	Graphs, networks and trees
spoken	Transcribed Speech
tagdocs	Documentation of TEI modules
tei	Declarations for datatypes, classes, and macros available to all TEI modules
textcrit	Text criticism
textstructure	Default text structure
transcr	Transcription of primary sources
verse	Verse structures

Det er alene ”core”, ”header” og ”textstructure” som er obligatoriske moduler. Derudover kan man tilvælge flere moduler efter behov, eksempelvis modulet, ”dictionaries”. Via hjemmesiden, <http://www.tei-c.org/Roma/>, er det endvidere enkelt at skræddersy ens egne dokumenttyper.

Selvom man alene anvender de tre obligatoriske moduler samt modulet ”dictionaries”, bliver det hurtigt tydeligt at TEI er en ekstremt omfattende standard, som til gengæld kun indeholder ganske få obligatoriske elementer. Ud over en header med forskellige administrative oplysninger, er elementerne ”text”, ”body” og ”div1” således den eneste struktur som skal være tilstede i en ordbogsartikel der følger TEI-standarden.

3.3.1.1 Nordisk netordbog

Nærværende projekt minder en del om det nordiske netordbogsprojekt som blev afsluttet i 2007 (jf. <http://www.csc.kth.se/tcs/projects/netordbog/>) og bl.a. med udgangspunkt i TEI's *dictionary base tag set* (jf. <http://www.tei-c.org/release/doc/tei-p4-doc/html/DI.html>) udviklede en fælles standard for opmærkning af nordiske ordbogsbaser (jf. <http://www.csc.kth.se/tcs/projects/netordbog/format/ordbog.dtd>), en fælles søgegrænseflade (Tvärslå, jf. <http://ordbok.nada.kth.se/>) og søgning i tekst på tværs af de skandinaviske sprog (Tvärsök).

Hovedresultaterne i Tvärslå-rapporten var som følger:

- Der er desværre intet standardformat for udveksling af ordbogsdata, og det nærmeste man kommer en fælles standard er TEI's standard for trykte ordbøger.
- Der er to problemer med denne standard
 - Den er for omfattende (7000 linjers DTD)
 - Den er for slap

- Man kan med fordel reducere og forenkle TEI-standarden kraftigt uden at skabe kompatibilitetsproblemer (bortset fra elementerne <def> og <index>)

Da DTD'er stort set er blevet erstattet af XML skemaer (XSD) har vi forsøgsvis konverteret ovennævnte DTD til XSD vha. et simpelt perl script (jf.

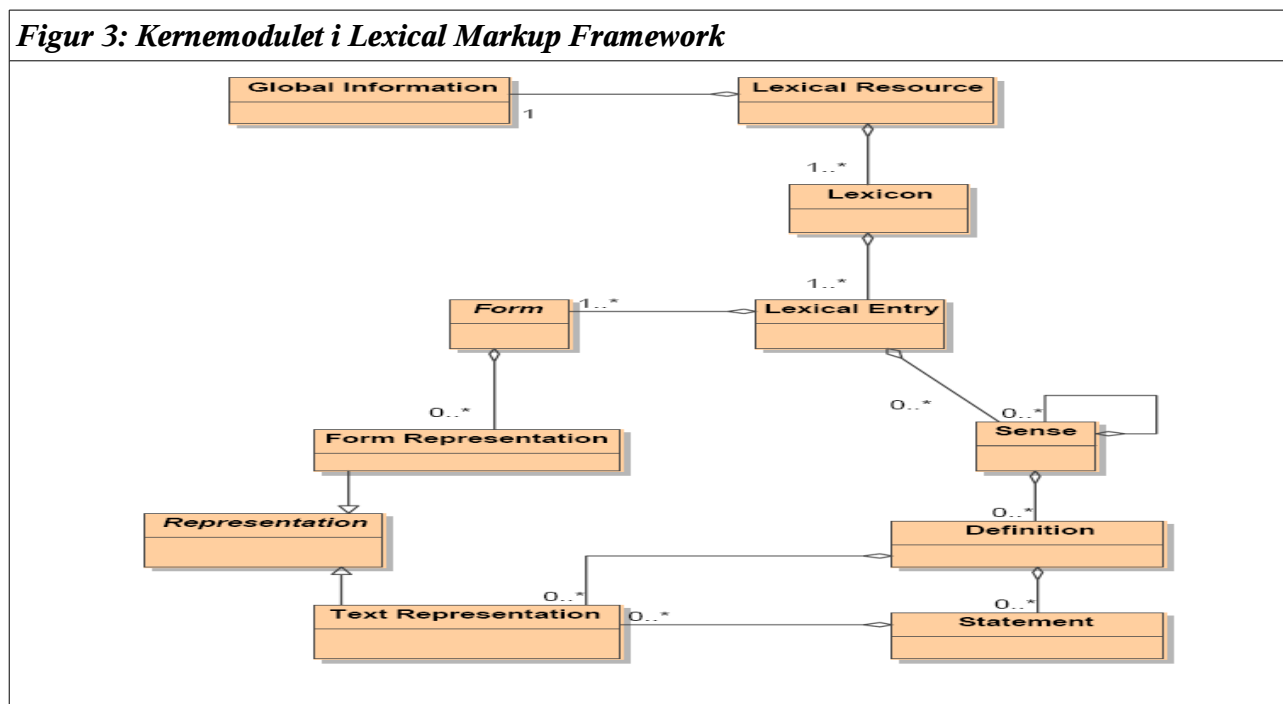
<http://www.mathling.com/xmlschema/dtd2xsd.pl>), og et fragment af dette skema er gengivet i bilag 7.

3.3.2 Lexical Markup Framework (LMF)

Selvom standarden vedrørende Lexical Markup Framework (LMF eller ISO-24613) stadig har DIS-status (Draft for International Standard), er det af flere årsager den mest oplagte standard at tage udgangspunkt i. Disse årsager vil blandt andet blive diskuteret i dette afsnit.

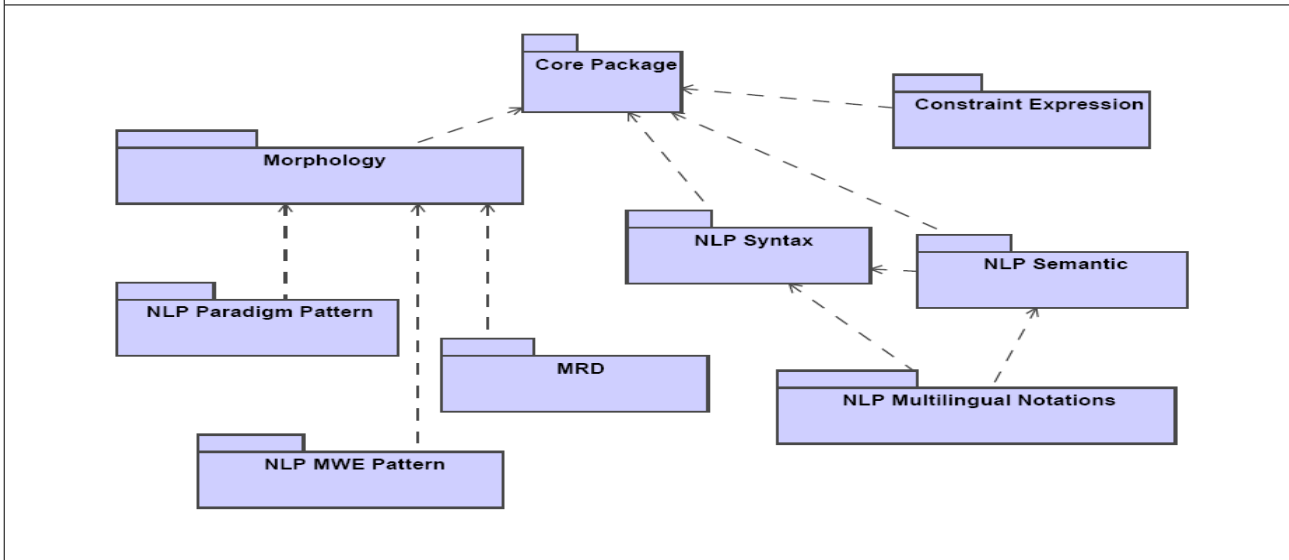
Kernemodulet i LMF har en struktur som er gengivet i figur 3. Mens notationen "1..*" betyder at elementet skal forekomme mindst én gang, betyder notationen "0..*" at det pågældende element er valgfrit, men kan forekomme et ubegrænset antal gange. Figuren fortæller således at en leksikalsk resurse skal indeholde ét eller flere leksika der hver især skal indeholde én eller flere ordbogsindgange ("Lexical Entry"), som skal indeholde mindst én "Form", men som dog ikke nødvendigvis behøver indeholde betydningsangivelser ("Sense") eller definitioner. Definitionerne på alle disse datakategorier er fastlagt i ISO DIS 24613:2007, og vi vil ikke i dette oversigtsafsnit komme nærmere ind på kategoriernes indhold og brug. Det vil vi derimod i afsnit 3.4.1 der byder på et konkret eksempel på en ordbogsindgang som følger LMF-standarden.

Figur 3: Kernemodulet i Lexical Markup Framework



Ud over dette kernemodul specificerer standarden en lang række udvidelsesmoduler hvoraf i hvert fald moduler som morfologi og semantik vil være nødvendige i de fleste ordbogsprojekter. Figur 4 gengiver en oversigt fra ISO 24613.

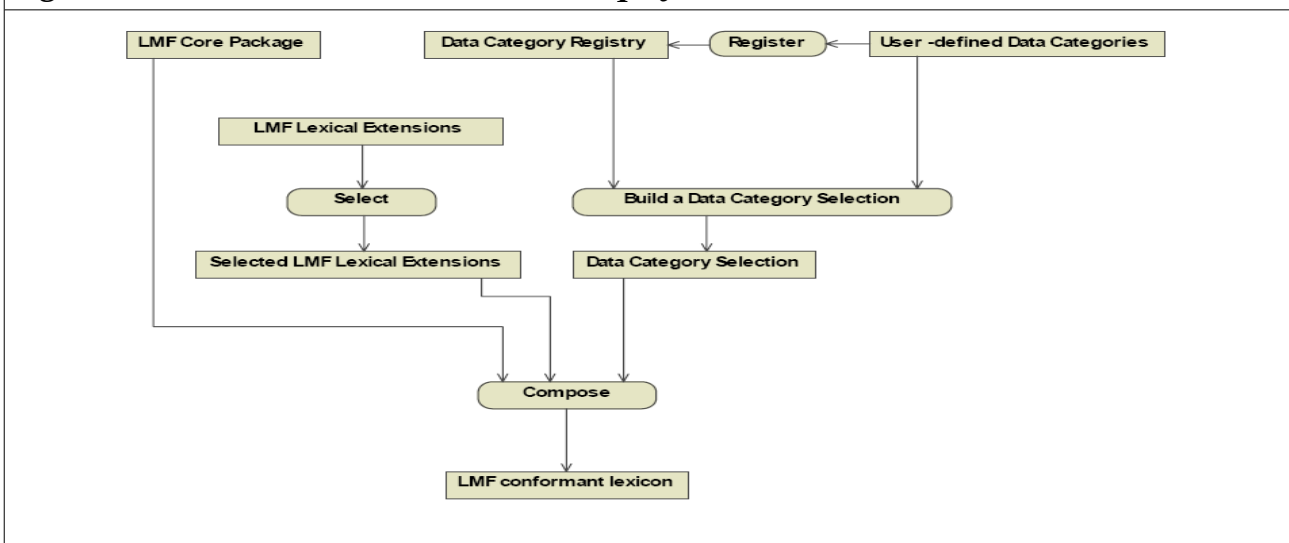
Figur 4: Udvidelsesmoduler til LMF



An LMF conformant lexicon is defined as the combination of an LMF core package, zero to many lexical extensions and a set of data categories. The combination of all these elements is described in the following UML activity diagram:

(ISO 24613:29007, p. 19)

Figur 5: Hvordan anvendes LMF i et konkret projekt?



Figur 5 illustrerer hvordan ordbogsartikler som følger LMF i virkeligheden har to kernebestanddele, nemlig

- **Strukturelle elementer** eller "high level specifications" (fx Lexical Resource, Global Information, Lexicon, Lexical Entry, Lemma, Wordform)
- **Indholdselementer** eller "low level specifications" (fx languageCoding, writtenForm, partOfSpeech m.fl.)

De strukturelle elementer kommer fra "LMF Core Package" samt et udvalg af LMF udvidelsesmoduler. Indholdselementerne er derimod **ikke** specificeret i LMF (dvs. ISO-24613), men stammer fra et såkaldt Data Category Registry (DCR). Et eksempel på et sådant DCR er ISO 12620 som også går under navnet "Computer applications in Terminology – Data categories" og blev analyseret i afsnit 2.6.2. Ofte vil man imidlertid have brug for en lang række datakategorier som ikke er defineret i noget eksisterende DCR og i sådanne tilfælde må man selv tilføje disse skræddersyede kategorier til standarden (jf. "User-defined Data Categories" i figur 5).

I afsnit 3.4.1 vil vi som sagt give et eksempel på en ordbogsartikel som er opmærket efter LMF-standardens og i den anledning nærmere forklare hvordan elementerne fra kernemodulet (figur 3) helt konkret spiller sammen med elementerne fra et udvalg af udvidelsesmodulerne (figur 4). Endelig vil vi diskutere fordele og ulemper ved denne standard kontra andre mulige standarder.

Dette afsnit vil vi imidlertid afslutte ved at henlede læserens opmærksomhed på at den seneste version af LMF kan ses på <http://lirics.loria.fr>, og online-redigering af ordbogsressurser som følger LMF-standardens kan foretages i betasoftwarepakken, Lexus (<http://www.lat-mpi.eu/>). Endelig kan de seneste versioner af datakategorierne udforskes her: <http://syntax.inist.fr/>.

3.3.3 Open Lexicon Interchange Format (OLIF)

Bilag 16 viser strukturen i en OLIF-fil¹¹. Denne standard følger et typisk design med en overordnet opdeling i et hoved og en krop, hvor hovedet indeholder forskellige administrative oplysninger om ophavsret (publicationStatement), tekstformat (contentInfo), workflow og så videre. Kroppen indeholder derimod et antal ordbogsindgange (entry) som hver især indeholder et obligatorisk element med monolingval information (mono) samt to valgfrie elementer med henholdsvis interne (crossRefer) og eksterne/bilingvale henvisninger (transfer). De monolingvale oplysninger skal omfatte en række nøgledata (keyDescription) som fx den kanoniske form (canForm), sprog, ordklasse (ptOfSpeech) og domæne (subjField).

De seneste oplysninger om standarden kan findes her: <http://www.olif.net>.

3.3.4 Multi Dictionary Formatter (MDF)

Multi Dictionary Formatter definerer omkring 100 datakategorier og fastlægger følgende struktur for en ordbogsindgang:

- lx (lexeme)
 - hm (homonym)
 - lc (citatform)
 - se (underindgang)
 - ph (fonetisk form)
 - ps (ordklasse, engelsk)
 - pn (ordklasse, national)
 - sn (betydning)
 - dv (definition, national)

¹¹ Kilde: Lieske (2001)

- de (definition, engelsk)
- rf (henvisning)
 - xv (eksempel)
- uv (brug)
- ev (encyklopædisk information)
- sy (synonym)
- an (antonym)
- mr (morfologi)
- bw (låneord)
- et (etymologi)
 - sg (ental), pl (flertal),

Læs mere om standarden på: <http://www.sil.org/computing/shoebox/MDF.html>.

3.3.5 Andre standarder for ordbaser

Ud over LMF (ISO-24613), ISO-12620, TEI, OLIF og MDF kan følgende standarder være af interesse for nærværende projekt.

1. ISO-16642 (Terminological Markup Framework - TMF)
2. ISO-24612 (Linguistic Annotation Framework – LAF) - draft
3. ISO-1951 (Presentation/representation of entries in dictionaries)
4. DS 2394-1 (STANLEX)

Heraf vil vi imidlertid alene kigge på ISO-1951 og DS 2394-1, idet ISO-24612 endnu ikke er godkendt og ISO-16642 i for høj grad er fokuseret på strukturering af terminologiske data snarere end leksikalske data i bredere forstand.

3.3.5.1 ISO-1951

ISO-1951 indeholder en formel beskrivelse af ordbogsartikler som eksemplificeret i bilag 6. Standarden fastlægger følgende struktur for eksempelvis en almensproglig, monolingval ordbogsindgang:

- DictionaryEntry
 - HeadwordCtn
 - Headword
 - PartOfSpeech
 - Pronunciation
 - Etymology

- DerivationBlock
- SenseGroup+
 - Definition
 - Example
 - CompositionalPhraseCtn
 - SynonymBlock
 - Antonym

De fleste informationstyper er velkendte, men standarden indeholder en del definitioner af de anvendte datakategorier som kunne vise sig relevante i forhold til opbygningen af et fællesnordisk DCR for ordbaser.

3.3.5.2 DANLEX og STANLEX

Hjorth (1987) og Madsen (1998) beskriver resultaterne af et stort udredningsarbejde vedrørende en standardisering af både indhold og struktur for leksikalske datasamlinger. Tabel 7 vedrører indholdsbeskrivelsen og sammenholder de overordnede informationstyper (inklusive eksempler) som henholdsvis DANLEX-gruppen og senere STANLEX-gruppen har anbefalet for opmærkningen af leksikalske data i eksempelvis ordbaser.

<i>Tabel 7: Standardisering af informationstyper i leksikalske data</i>	
DANLEX	STANLEX (DS 2394-1)
	Administrative oplysninger (fx intern/ekstern henvisning)
Etymologisk information	Etymologiske oplysninger
Fonetisk information (fx Prosodiske træk, Udtale)	Fonetiske oplysninger
Grafisk information	Grafiske oplysninger
Grammatisk information (fx Ordklasse, Syntaks, Bøjning)	Grammatiske oplysninger
Pragmatisk information (fx Kontekstuel information, Brug, Ekstern reference, Administrativ information)	
Semantisk information (fx Semantiske relationer, Ækvivalens, Emneklassifikation)	Semantiske oplysninger (fx Semantiske relationer, Indholdsspecificerende oplysn.)
	Sprog
	Sprogbrugsoplysninger (Brugseksempler, Brugsoplysninger)
	Strukturoplysninger

Både DANLEX-gruppens og STANLEX-gruppens arbejde vedbliver med at være en rigtig god rettesnor for ordbasers indholdsbeskrivelse. Selvom der er kommet flere overordnede informationstyper til i den endelige standard (DS 2394-1), svarer oplysningstyperne i DANLEX og STANLEX i store træk til hinanden. Der er også et tydeligt overlap mellem informationstyperne i tabel 7 og datakategorierne i ISO 12620 (se afsnit 2.6.2), og tillige de overordnede emnekategorier i fx den danske og den islandske emnetaksonomi (se tabel 2). Dog er informationstypen ”Semantiske relationer”, som jo især er af central betydning i terminologiarbejde, ikke eksplicit repræsenteret i nogen af de nordiske ordbaser (jf. bilag 14).

Hvad angår selve strukturbeskrivelsen følger her et eksempel på en ordbogsartikel som anvender STANLEX-taksonomiens oplysningstyper:

- artikel
 - EkstStruk
 - **Hoved**
 - Stavning+
 - GRAMM*
 - Ordklasse
 - Bøjningsparadigme
 - **Krop**
 - BETYDGRP+
 - IntStruk (betydningsnummer)
 - BETYDAFSN+
 - BETYDSPEC*
 - stand.emneklassifikation
 - kommunikativ dimension
 - indholdsspecifikation
 - eqv.rel.2 (oversættelse)
 - InternHenvisning
 - **Hale***
 - ORDFORBGRP+
 - ordforbindelse
 - overført betydning*
 - ordforb.eqv+ (oversættelse)

Dele af ovenstående struktur indeholder dele som minder om henholdsvis den danske Retskrivningsordbog, LMF og ISO-1951. Retskrivningsordbogen har således også en overordnet inddeling i et hoved med ortografiske og grammatiske oplysninger og en krop med betydningsmæssige oplysninger (bilag 12). I LMF håndteres denne opdeling med de to elementer ”Form” og ”Sense”, og i ISO-1951 anvendes elementerne ”HeadWord” og ”SenseGroup”.

Det overordnede element ”Hale” er imidlertid nyt i forhold til de andre ordbasestrukturer i denne rapport, men inddelingen i hoved-krop-hale er helt klassisk for ordbogsartikler, idet denne rækkefølge afspejler den typiske præsentation i papirordbøger med udtrykssiden først, dernæst indholdssiden og til sidst et afsnit med faste udtryk, idiomer og andre sproglige kuriositeter.

Endelig er elementet ”stand. emneklassifikation” interessant. Afsnit 2 dokumenterede hvor stor en betydning en standardiseret emneklassifikation har i forhold til sprognævnens svarbaser. Selvom de nordiske ordbaser kun i meget ringe udstrækning indeholder oplysninger om emneklassifikation i øjeblikket, kunne man med fordel tilføje en sådan oplysning og med mere terminologiske briller få opdelt indholdet af ordsamlingerne i forskellige fagområder eller domæner.

3.4 Udkast/anbefalinger til fælles ordbasestruktur

Da der som vist findes en række internationale standarder for ordbaser, vil vi ikke forsøge at foreslå en ny nordisk standard, men derimod give eksempler på ordbogsartikler som følger to af de mest lovende og/eller udbredte standarder, nemlig LMF og TEL.

3.4.1 Eksempel på ordbogsartikel i LMF

Bilag 17 viser et eksempel på hvordan et (bearbejdet) opslagsord fra den danske ordbase kunne tage sig ud i LMF. Hvis man analyserer beskrivelserne af de enkelte moduler i ISO DIS 24613:2007 (udkastet til standarden om LMF), kan man se hvordan store dele af strukturen i bilag 17 kommer fra standardens kernemodul, nemlig elementerne

- GlobalInformation
 - LexicalResource
 - Lexicon
 - LexicalEntry
 - FormRepresentation
 - Sense
 - Definition

mens mindre dele af strukturen kommer fra tilvalgsmodulet, ”Morphology”, nemlig elementerne

- Lemma
- WordForm

eller fra tilvalgsmodulet, ”NLP Semantics”, nemlig

- MonolingualExternalRef

eller tilvalgsmodulet, ”Machine Readable Dictionaries”, nemlig

- Context

Samtlige datakategorier er som sagt defineret i ISO DIS 24613:2007, men især elementerne "Form" (figur 3), "FormRepresentation" (figur 3 og 7), "Representation" (figur 3) og "TextRepresentation" (figur 3) kræver en nærmere forklaring.

Elementet "Form" er en abstrakt klasse som kan repræsentere et lemma, et leksem, en morfologisk variant af et leksem (en ordform) eller et morfem. I eksemplet i bilag 17 anvendes således to konkrete underkategorier til "Form", nemlig "Lemma" og "WordForm", i stedet for den abstrakte overkategori. Elementet "FormRepresentation" beskriver ortografiske varianter af en "Form". I eksemplet i bilag 17 har vi således forestillet os at der eksisterer en ortografisk variant af lemmaet "gyllebaron" som er geografisk afgrænset til øen Fyn. Elementet "Representation" er ligesom "Form" en abstrakt overkategori (i dette tilfælde til "FormRepresentation" og "TextRepresentation") som sjældent vil finde konkret anvendelse. Endelig anvendes elementet "TextRepresentation" til at beskrive særlige træk ved ortografien af definitioner (og dertilhørende statements). Det er således en oplysningstype der vedrører betydningen ("Sense") af et opslagsord snarere end dets udtryk ("Form").

En andet interessant aspekt af eksemplet i bilag 17 er at det afslører et helt tydeligt karakteristisk træk ved LMF-standarden. Nemlig at LMF følger en klassisk attribut-værdi-struktur hvilket gør den ret fleksibel og nem at tilpasse de opmærkningsbehov forskellige typer af ordbaser måtte gøre gældende. Eksempelvis kan vi til elementet "Lemma" tilføje attributten "partOfSpeech", "grammaticalGender" og "writtenForm" for at angive hvilken ordklasse og køn lemmaet tilhører samt hvorledes det udtrykkes rent sprogligt. Desuden kan vi nærmere beskrive elementet "FormRepresentation" ved at tilføje attributten "geographicalVariant" og angive at varianten er begrænset til Fyn.

Hvorimod attributter som "partOfSpeech", "grammaticalGender" og "writtenForm" stammer fra ISO 12620, stammer attributter som "genre", "kilde", "brug" med flere fra den danske ordbase og bør derfor defineres og tilføjes til et fælles DCR før de tages i anvendelse.

Afslutningsvis vil vi konkludere at styrken ved LMF især er dens meget **modulære opbygning**. Denne modularitet betyder nemlig at artikelstrukturen ikke bliver unødigt kompleks, idet der kun er et fåtal af obligatoriske elementer i kernemodulet, men til gengæld et hav af udvidelser som hver især indfører flere obligatoriske og valgfri elementer og naturligvis er gensidigt compatible. Desuden er standarden meget fleksibel, idet der er mulighed for nærmere at beskrive de enkelte elementer med en åben mængde attribut-værdi-par.

3.4.2 Eksempel på ordbogsartikel i det reducerede TEI-format

Bilag 18 indeholder et eksempel på en ordbogsartikel i det nordiske netordbogsformat, som altså er en reduceret udgave af TEI-modulet "dictionaries".

Oplysningstyperne i TEI og LMF minder en del om hinanden. Elementet "teiHeader" modsvarer således delvist elementerne "globalInformation" og de attributter der måtte være på elementet "Lexicon" i LMF. På samme vis modsvarer elementerne "entry" og "lexicalEntry" hinanden. Dog tillader det nordiske netordbogsformat, i modsætning til LMF, ikke en åben mængde attribut-værdi-par, men er udelukkende baseret på et inventar af præspecificerede elementer og attributter. Standarden er heller ikke så modulært opbygget, idet elementinventaret ikke er grupperet i forskellige moduler som kan stykkes sammen efter behov.

3.5 Konklusion

Vi vil anbefale at man overvejer at anvende, eller i hvert fald nærmere undersøge, den kommende ISO-standard LMF i forhold til opbygningen af **fremtidige** nordiske ordbaser. LMF har den fordel at det er den nyeste, mest fleksible (pga. de åbne attribut-værdi-par) og også enkleste standard (pga. dens modularitet). Den nordiske netordbog har den fordel allerede at være en etableret standard som har fundet anvendelse i flere forskellige ordbogsprojekter. Dog virker det enklere at udvide LMF end TEI med nye datakategorier (eksempelvis oplysninger om etymologi eller emnekategori) uden at dette kompromitterer formatets kompatibilitet.

4.0 Software

Dette afsnit indeholder en meget kursorisk gennemgang af det programmel de nordiske landes sprognævn i øjeblikket anvender til at administrere deres svar- og ordbaser (afsnit 4.1). Derpå følger et ganske overfladisk overblik over udvalgt databaseprogrammel på markedet. I afsnit 4.2 kommer en diskussion af de væsentligste forskelle mellem relationelle databaser og databaser som kan siges at være baseret på "native XML" (se definitionen i afsnit 4.1.1). I afsnit 4.3 opridses konsekvenserne af en overgang til en XML-database, og endelig gives der i afsnit 4.4 to eksempler på terminologisk programmel.

4.1 Databaseprogrammel

Tabel 8 indeholder en oversigt over databaseprogrammel som i øjeblikket finder anvendelse i de nordiske sprognævn med angivelse af om programmet er kommercielt eller frit tilgængeligt (open source) og af hvilken databaseteknologi der er tale om.

	Programmel	Type	Licens
Danmark	iLEX	Native XML	Kommercielt
Sverige	SQL Anywhere	Relationel db	Kommercielt
Norge	MySQL Access	Relationel db Relationel db	Open Source Kommercielt
Finland	SQL Anywhere TRIP	Relationel db Relationel db	Kommercielt ???
Island	Excel		Kommercielt

Tabellen viser med stor tydelighed at langt de fleste sprognævn anvender konventionelle databasesystemer, og kun Danmark anvender en native XML-database. Endvidere anvender de fleste sprognævn kommercielle platforme som ikke er open source.

4.1.1 Markedsoverblik

Som i tabel 8 kan man opdele databaseprogrammel efter licensforhold (den kommercielle industri kontra open source) eller efter den teknologiske platform (relationel database kontra native XML). I første omgang kan vi nævne fire store spillere i den kommercielle softwareindustri, nemlig

- Oracle
- Microsoft SQL Server
- DB2 fra IBM
- Sybase Adaptive Server eller SQL Anywhere

Deres største konkurrenter i det stadig mere populære og produktive open source-miljø er:

- MySQL
- Firebird
- PostgreSQL

Alle tre open source-alternativer er tilgængelige både på Windows og UNIX/Linux-platforme, så det bør ikke være sådanne overvejelser som forhindrer et sprognævn i at vælge et open source-databasesystem.

Vigtigere end dette licensbaserede skel er imidlertid spørgsmålet om databaseprogrammets teknologiske platform. Her tænkes på skellet mellem konventionelle (dvs. relationelle) databasesystemer og de nyere native XML-databasesystemer. Rationalet bag native XML-databaser er at data typisk opmærkes og transporteres i XML og transformeres til og fra XML. Derfor virker det nærliggende også at lagre data i XML-format. Selvom mange databasesystemer nu accepterer XML som input og genererer XML som output, betyder det ikke nødvendigvis at der er tale om en native XML-database. Sidstnævnte skal nemlig opfylde følgende krav¹²:

1. XML-dokumentet skal være den fundamentale (logiske) lagerenhed (ligesom en række i en tabel er lagerenheden i konventionelle databaser)
2. Databasen skal definere en logisk model (fx XPath eller XQuery) for et XML-dokument og anvende denne model i søgninger og lagringer.
3. Databasens fysiske lagermodel er underordnet (relationelle databasestrukturer eller indekserede filer er o.k.)

Fordelene ved at vælge et databasesystem som med rette kan kaldes ”native XML” er beskrevet i afsnit 4.2. Resten af dette afsnit vil således alene give et lille overblik over et antal udvalgte native XML-databasesystemer i form af den nedenstående tabel (tabel 9).

iLEX (http://www.emp.dk)	Kommerciel (anvendes mest til leksikografarbejde)
iTerm (http://www.i-term.dk/)	Kommerciel (skræddersyet til terminologiarbejde)

12 Kilde: www.wikipedia.org

eXist (http://exist.sourceforge.net)	Open Source (understøtter XQuery)
Tamino (http://www.softwareag.com/tamino)	Kommerciel
XIndice (http://xml.apache.org/xindice)	Open Source
Ozone (http://ozone-db.org)	Open Source (objektorienteret, javabaseret)
Sedna (http://modis.ispras.ru/sedna)	Open Source (understøtter XQuery, SQL-forbindelser)

Ud over et databasesystem som kan anvendes til lagring og effektiv søgning i samlinger af XML-dokumenter, er der imidlertid et antal andre behov de nordiske sprognævn kan have, nemlig:

- Dokumentredigering (både af XML-dokumenter og skemaer)
- Skemavalidering (XSD: Har et givet dokument den korrekte struktur?)
- Flere forskellige dokumentvisninger (transformationer med XSLT)
- Nem publikation af baseindhold på internettet (mere XSLT)
- Nem integration af multimediedata (fx lydoptagelser og scannede dokumenter)

Vi har ikke i denne rapport undersøgt i hvilken udstrækning det forskellige databaseprogrammer indeholder al den ovennævnte funktionalitet. Generelt kan man sige, at der måske er en tendens til at kommercielle produkter (fx iLEX og iTerm) integrerer et større antal funktioner og gør den daglige brug mere bekvem for brugerne, hvorimod man sandsynligvis bliver nødt til at vælge et antal forskellige open source-programmer hvis man vil have dækket alle sine behov uden at betale licens til nogen. Eksempelvis systemet eXist til lagring og søgning, Liquid XML¹³ til dokumentredigering, W3C's onlineservice¹⁴ til skemavalidering og så videre.

4.2 Native XML-databaser kontra konventionelle databaser

Som beskrevet i afsnit 4.1.1 er den væsentligste forskel mellem relationelle databasesystemer og native XML-databasesystemer at sidstnævnte grundlæggende er dokumentorienterede, mens førstnævnte er dataorienterede. Native XML-databaser er således udrustet med et dokumentorienteret søgesprog der som oftest følger en af de internationale standarder, dvs. enten XPath eller XQuery (som er en mere fleksibel og udvidet udgave af XPath).

Dokumentorienterede søgesprog som XPath og især XQuery passer naturligt nok bedre til dokumenter end et dataorienteret søgesprog som SQL der er optimeret til søgninger i tabeller og rækker. Et dokument er imidlertid en mere abstrakt størrelse som kan være vanskelig at presse ned i en tabelskabelon, især fordi en dokumentstruktur nemt bliver temmelig hierarkisk som illustreret ved de mange eksempler i bilagene. Det vil eksempelvis være vanskeligt med SQL at udtrække den delmængde af svar i den danske svarbase som indeholder præcis to kildehenvisninger og mindst tre

¹³ <http://www.liquid-technologies.com/>

¹⁴ <http://www.w3.org/2001/03/webdata/xsv>

materialiter-elementer på vilkårlige niveauer i dokumentet hvorimod det sagtens kan lade sig gøre med XPath/XQuery.

4.2.1 To eksempler: eXist og iLEX

Bilag 15 indeholder et eksempel på to Native XML-databaser. Den ene er open source og hedder eXist, og den anden er et kommercielt system ved navn iLEX som i øjeblikket anvendes af Dansk Sprognævn. De to skærmbilleder i bilaget illustrerer følgende udfordringer og nødvendige overvejelser i forhold til valg af databaseprogrammel:

1. Det kommercielle system er mere brugervenligt og byder på større funktionalitet.
2. Open source-løsningen kræver en *inhouse*-programmør og kan stadig være vanskelig for sprognævnets ansatte at anvende i det hele taget.
3. Open source-løsningen involverer ingen licensudgifter.
4. Open source-løsningen følger (i dette tilfælde) en international standard (XQuery) hvorimod den kommercielle løsning har defineret sit eget søgesprog.

Den væsentligste fordel ved det kommercielle system er helt givet dets brugervenlighed og den store vifte af integrerede funktioner. ILEX har således integreret redigering, visning, søgning, multimediedata (assets) og meget andet som gør at man ikke behøver anvende andet software i sit daglige arbejde med svar- og ordbaseadministration.

4.3 Konsekvenser af overgang til XML-database

Afsnit 4.2.1 illustrerede ganske kort de væsentligste konsekvenser ved at vælge et open source databasesystem frem for et kommercielt databasesystem. I dette afsnit vil vi mere overordnet diskutere hvilke konsekvenser det har at indføre en XML-database som sådan. Som tidligere opridset i sammenfatningen (afsnit 1.5) vurderer vi at det vil fordr:

1. Teknisk kyndigt personale, dvs. efteruddannelse (omfanget afhænger af softwarevalget)
 - Især tager det tid at vænne sig til en ny grænseflade til redigering/indtastning
2. En it-specialist, med fordel inhouse (til opsætning af software, transformation af data osv.)
3. Konvertering af eksisterende databaser
 - fra SGML eller SQL, fx Access (relativt enkelt)
 - fra regneark, fx Excel (lidt mere tidskrævende)
4. Omprogrammering af eksisterende, integrerede databaseløsninger
 - fx automatisk registrering af e-post i SQL-baseret svarbase (Sverige)
 - fx automatiske webløsninger der interagerer med en SQL-baseret svarbase (fx Frågelådan i Sverige).

Det kan således være en bekostelig og omfattende affære at skifte til en XML-database, idet eksisterende sprognævns personale ikke kan undgå at skulle afsætte en vis arbejdstid på

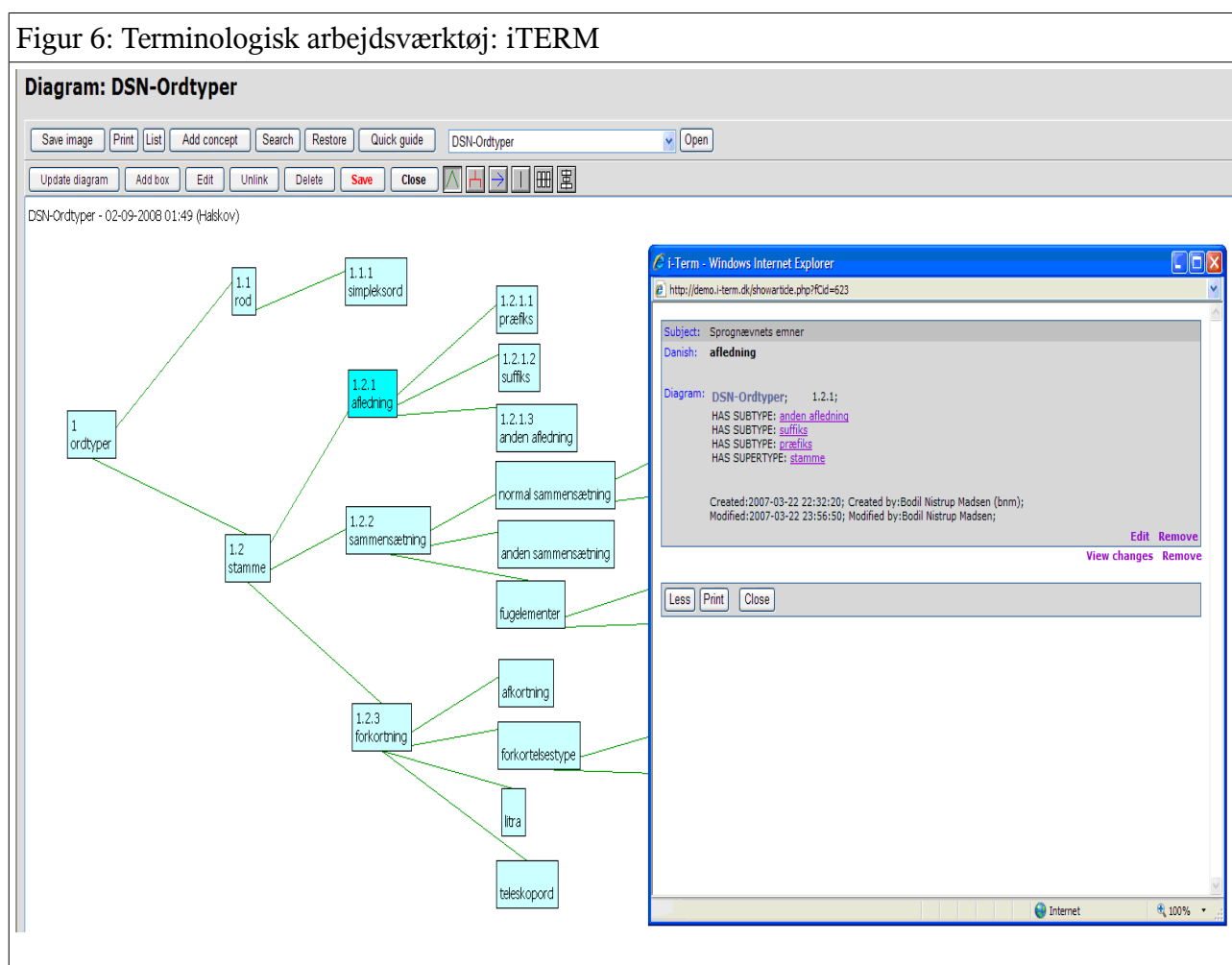
efteruddannelse, og idet nyt it-kyndigt personale skal hyres på kort sigt til at udføre både konvertering og omprogrammering, men sandsynligvis også på længere sigt til at ændre databasens opsætning, transformere data og så videre.

Derfor anbefaler vi som sagt at man i første omgang nøjes med at etablere en fælles emnetaksonomi, kobler denne til de eksisterende databaser og udvikler en fælles søgeportal (se model "light" i afsnit 1.5.2).

4.4 Terminologisk programmet

Arbejdet med etableringen af en fælles emnetaksonomi vil i høj grad involvere begrebsafklaring og andet terminologiarbejde. Her anbefaler vi at man også anvender relevant programmet hvormed man kan dele viden på tværs af landene, systematisere og effektivisere sit arbejde. I dette afsnit vil vi alene vise eksempler på to programmer, nemlig et kommercielt program (iTERM) og et open source program (TemaTres).

Figur 6: Terminologisk arbejdsværktøj: iTERM



Fordelen ved iTERM er at man både kan arbejde visuelt (til venstre i figur 6) og opbygge en hierarkisk struktur over relevante kategorier og underkategorier, men samtidig kan arbejde på listeform (til højre i figur 6) og eksempelvis tilføje definitioner for de enkelte kategorier.

Figur 7: Terminologisk arbejdsværktøj: TemaTres

The screenshot shows the TemaTres web interface. At the top, there is a search bar and the title 'TemaTres'. On the left side, there is a navigation menu with links for 'Hierarchical list', 'Alphabetic list', 'About...', 'Administration', 'ADMINISTRADOR, USUARIO', 'My account', and 'Logout'. Below the menu is a language dropdown set to 'english'. The main content area displays the title 'A.01 Morphology' with buttons for 'Add', 'Options', and 'Menu'. Below the title, there is a 'Scope note' section with a definition: 'Morphology is the field of linguistics which studies the internal structure of words.' There is also a 'Bibliographic note' section with a link to Wikipedia. Below these notes, there is a list of related terms with checkboxes and labels: '[x]NT A.01.01 Part_of_Speech', '[x]NT A.01.02 Inflection', '[x]EQ 1.1 morfologi (DA_emnetaksonomi / en)', '[x]EQ Hur_ska_ordet_se_ut? (SV_emnetaksonomi / en)', and '[x]EQ Korleis_ska_ordet_boeyast? (NO_emnetaksonomi / en)'. At the bottom of the page, there is a footer with the text 'MADS Zthes SKOS-Core XTM' and several small icons. The footer also includes the author's name 'Diego Ferreyra' and the URI 'http://localhost/tematres/'.

TemaTres har den fordel at det er gratis og frit tilgængeligt, men i modsætning til iTERM byder det på knapt så omfattende funktionalitet, eksempelvis er der ikke mulighed for at opbygge taksonomier visuelt. Desuden har TemaTres heller ikke lige så gode muligheder for at udskrive strukturbeskrivelser i XML som iTERM har.

4.5 Konklusion

ILEX er den mest brugervenlige og integrerede løsning, men samtidig også den dyreste – i hvert fald i forhold til licensudgifterne. Open source-programmel som eXist kan være udmærket, men vil ofte mangle brugervenlighed og funktionalitet, fx nem dokumentvisning, dokumentredigering og dokumenttransformation. I forhold til basal funktionalitet som lagring, indeksering og søgning kan man dog være mindst lige så godt tjent med en open source-løsning.

Uanset hvilket konkret stykke software man vælger at anvende, er det en fordel at vælge et native XML-databasesystem. Desuden skal man være opmærksom på, at sprognævnets personale skal **efteruddannes** uanset om man vælger det mest brugervenlige, kommercielle system eller ej. Endelig bør man være særlig påpasselig hvis man i øjeblikket anvender en meget integreret løsning, fx i forhold til automatisk registrering af svar eller generering af indhold til internettet, idet denne løsning sandsynligvis også vil skulle ændres hvis man skifter databaseplatform.

I forhold til terminologiarbejdet med opbygningen af den fælles emnetaksonomi kan man overordnet sige at der gælder de samme overvejelser ved valget mellem eksempelvis iTERM og TemaTres.

5.0 Overordnet konklusion og anbefalinger

Se afsnit 1.5.

Referencer

Bird, Steven; Gary Simons (2003) "Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources"

Farrar, Scott; Terry Langendoen (2003) "A Linguistic Ontology for the Semantic Web" I: *GLOT International* 7(3), pp. 97-100

Francopoulo, Gil; Thierry Declerck; Monica Monachini; Laurent Romary (2006) "The relevance of standards for research infrastructures" I: *Proceedings of LREC 2006*

Francopoulo, Gil et al. (2006) "Lexical Markup Framework (LMF) for NLP multilingual resources I: *Proceedings of LREC 2006*

Francopoulo, Gil et al. (2006) "Lexical markup Framework (LMF)" I: *Proceedings of LREC 2006*

Galberg Jacobsen, Henrik (1996) *Grammatisk talt – Anbefalede sproglige betegnelser*, Dansk Sprognævns skrifter 24, Dansk lærerforening 1996

Hjorth, Ebba et al. (1987) *Descriptive Tools for Electronic Processing of Dictionary Data, Studies in Computational Lexicography* (Lexicographica Series Maior 20), Tübingen, Niemeyer.

Huckstorf, Axel (2008) "The Multilingual European Thesaurus on International Relations and Area Studies" I: *Proceedings of TKE 2008*, Copenhagen Business School.

Kemps-Snijders, Marc; Menzo Windhpuwer; Peter Wittenburg; Sue Ellen Wright (2008) "A Revised Model for ISO Data Category Registry" I: *Proceedings of TKE 2008*, Copenhagen Business School.

Lieske, Christian (2001) et al. "The Open Lexicon Interchange Format (OLIF) Comes of Age" I: *Proceedings of the Machine Translation Summit VIII*, Galicia, Spanien

Madsen, Bodil Nistrup et al. (1998) *DS 2394-1 Leksikalske datasamligner. Indholds- og strukturbeskrivelse. Del 1: Taksonomi til klassifikation af oplysningstyper*. Dansk Standard.

Madsen, Bodil Nistrup; Hanne Erdman Thomsen "A Taxonomy of Lexical Metadata Categories" I: *Proceedings of LREC 2008*

Bilag

1. Den danske emnetaksonomi

- 1.1 morfologi
 - 1.1.1 morfologi ordklasse
 - 1.1.1.1 morfologi ordklasse adverbium
 - 1.1.1.2 morfologi ordklasse pronomen
 - 1.1.1.2.1 morfologi ordklasse pronomen indefinit_pronomen
 - 1.1.1.2.2 morfologi ordklasse pronomen personligt_pronomen
 - 1.1.1.2.3 morfologi ordklasse pronomen relativt_pronomen
 - 1.1.1.2.4 morfologi ordklasse pronomen tiltalepronomen
 - 1.1.1.2.5 morfologi ordklasse pronomen reflektivt_pronomen
 - 1.1.1.2.6 morfologi ordklasse pronomen possessivt_pronomen
 - 1.1.1.2.7 morfologi ordklasse pronomen demonstrativt_pronomen
 - 1.1.1.2.8 morfologi ordklasse pronomen interrogativt_pronomen
 - 1.1.1.2.9 morfologi ordklasse pronomen reciprokt_pronomen
 - 1.1.1.3 morfologi ordklasse adjektiv
 - 1.1.1.4 morfologi ordklasse praeposition
 - 1.1.1.4.1 morfologi ordklasse praeposition ikke_valensbaerende
 - 1.1.1.4.2 morfologi ordklasse praeposition valensbaerende
 - 1.1.1.5 morfologi ordklasse verbum
 - 1.1.1.5.1 morfologi ordklasse verbum hjaelpeverbum
 - 1.1.1.5.2 morfologi ordklasse verbum modalverbum
 - 1.1.1.5.3 morfologi ordklasse verbum sammensatte_verber
 - 1.1.1.6 morfologi ordklasse numerale
 - 1.1.1.6.1 morfologi ordklasse numerale kardinaltal
 - 1.1.1.6.2 morfologi ordklasse numerale ordinaltal
 - 1.1.1.7 morfologi ordklasse substantiv
 - 1.1.1.7.1 morfologi ordklasse substantiv proprium
 - 1.1.1.7.1.1 morfologi ordklasse substantiv proprium personnavn
 - 1.1.1.7.1.2 morfologi ordklasse substantiv proprium stednavn
 - 1.1.1.7.1.3 morfologi ordklasse substantiv proprium officielt_institutionsnavn
 - 1.1.1.7.2 morfologi ordklasse substantiv verbalsubstantiv
 - 1.1.1.7.3 morfologi ordklasse substantiv appellativ
 - 1.1.1.8 morfologi ordklasse artikel
 - 1.1.1.9 morfologi ordklasse konjunktion
 - 1.1.1.9.1 morfologi ordklasse konjunktion sideordningskonjunktion
 - 1.1.1.9.2 morfologi ordklasse konjunktion samordningskonjunktion
 - 1.1.1.9.3 morfologi ordklasse konjunktion underordningskonjunktion
 - 1.1.1.10 morfologi ordklasse interjektion
 - 1.1.2 morfologi ordtyper
 - 1.1.2.1 morfologi ordtyper stamme
 - 1.1.2.1.1 morfologi ordtyper stamme afledning
 - 1.1.2.1.1.1 morfologi ordtyper stamme afledning suffiks
 - 1.1.2.1.1.2 morfologi ordtyper stamme afledning praefiks
 - 1.1.2.1.1.3 morfologi ordtyper stamme afledning anden_afledning
 - 1.1.2.1.2 morfologi ordtyper stamme sammensaetning
 - 1.1.2.1.2.1 morfologi ordtyper stamme sammensaetning fugeelementer
 - 1.1.2.1.2.1.1 morfologi ordtyper stamme sammensaetning fugeelementer med_fugeelement
 - 1.1.2.1.2.1.2 morfologi ordtyper stamme sammensaetning fugeelementer uden_fugeelement
 - 1.1.2.1.2.2 morfologi ordtyper stamme sammensaetning normal_sammensaetning
 - 1.1.2.1.2.2.1 morfologi ordtyper stamme sammensaetning normal_sammensaetning andetled
 - 1.1.2.1.2.2.2 morfologi ordtyper stamme sammensaetning normal_sammensaetning foersteled
 - 1.1.2.1.2.3 morfologi ordtyper stamme sammensaetning anden_sammensaetning
 - 1.1.2.1.3 morfologi ordtyper stamme forkortning
 - 1.1.2.1.3.1 morfologi ordtyper stamme forkortning afkortning
 - 1.1.2.1.3.2 morfologi ordtyper stamme forkortning litra

- 1.1.2.1.3.3 morfologi ordtyper stamme forkortning teleskop
- 1.1.2.1.3.4 morfologi ordtyper stamme forkortning forkortelsestype
- 1.1.2.1.3.4.1 morfologi ordtyper stamme forkortning forkortelsestype initialforkortelse
- 1.1.2.1.3.4.2 morfologi ordtyper stamme forkortning forkortelsestype andre_forkortelser
- 1.1.2.2 morfologi ordtyper rod
- 1.1.2.2.1 morfologi ordtyper rod simpleksord
- 1.1.3 morfologi boejning
- 1.1.3.1 morfologi boejning infinit_form
- 1.1.3.2 morfologi boejning finit_boejning
- 1.1.3.2.1 morfologi boejning finit_boejning genus
- 1.1.3.2.2 morfologi boejning finit_boejning tempus
- 1.1.3.2.2.1 morfologi boejning finit_boejning tempus pluskvamperfektum
- 1.1.3.2.2.2 morfologi boejning finit_boejning tempus praeteritum
- 1.1.3.2.2.3 morfologi boejning finit_boejning tempus praesens
- 1.1.3.2.2.4 morfologi boejning finit_boejning tempus futurum
- 1.1.3.2.2.5 morfologi boejning finit_boejning tempus perfektum
- 1.1.3.2.3 morfologi boejning finit_boejning komparation
- 1.1.3.2.4 morfologi boejning finit_boejning kasus
- 1.1.3.2.4.1 morfologi boejning finit_boejning kasus dativ
- 1.1.3.2.4.2 morfologi boejning finit_boejning kasus nominativ
- 1.1.3.2.4.3 morfologi boejning finit_boejning kasus genitiv
- 1.1.3.2.4.4 morfologi boejning finit_boejning kasus akkusativ
- 1.1.3.2.4.5 morfologi boejning finit_boejning kasus oblik
- 1.1.3.2.5 morfologi boejning finit_boejning bestemthed
- 1.1.3.2.6 morfologi boejning finit_boejning diatese
- 1.1.3.2.6.1 morfologi boejning finit_boejning diatese aktiv
- 1.1.3.2.6.2 morfologi boejning finit_boejning diatese passiv
- 1.1.3.2.7 morfologi boejning finit_boejning numerus
- 1.1.3.2.8 morfologi boejning finit_boejning modus
- 1.2 leksis
- 1.2.1 leksis antal_laan
- 1.2.2 leksis indbyggerbetegnelse
- 1.2.3 leksis antal_ord
- 1.2.4 leksis fagterm
- 1.2.5 leksis produktnavn
- 1.2.6 leksis hvad_hedder_x
- 1.2.7 leksis erhvervsbetegnelse
- 1.2.8 leksis eksisterer_ordet
- 1.2.9 leksis varemaerke
- 1.2.10 leksis nyt_ord
- 1.3 ortografi
- 1.3.1 ortografi tegn
- 1.3.1.1 ortografi tegn andre_tegn
- 1.3.1.1.1 ortografi tegn andre_tegn symboler
- 1.3.1.1.2 ortografi tegn andre_tegn orddeling_linje
- 1.3.1.1.3 ortografi tegn andre_tegn parentes
- 1.3.1.1.4 ortografi tegn andre_tegn accenttegn
- 1.3.1.1.5 ortografi tegn andre_tegn prikker
- 1.3.1.1.6 ortografi tegn andre_tegn skraastreg
- 1.3.1.1.7 ortografi tegn andre_tegn anfoerelsestegn
- 1.3.1.2 ortografi tegn orddannelsestegn
- 1.3.1.2.1 ortografi tegn orddannelsestegn bindestreg
- 1.3.1.2.2 ortografi tegn orddannelsestegn bindestreg_fra_til
- 1.3.1.2.3 ortografi tegn orddannelsestegn bindestreg_usaedvanlige_sms
- 1.3.1.2.4 ortografi tegn orddannelsestegn apostrof
- 1.3.1.2.5 ortografi tegn orddannelsestegn bindestreg_faelles_orddel
- 1.3.2 ortografi store_smaa_bogstaver
- 1.3.2.1 ortografi store_smaa_bogstaver store_bogst_i_tekstbeg
- 1.3.2.2 ortografi store_smaa_bogstaver store_bogst_etter_tegn
- 1.3.2.3 ortografi store_smaa_bogstaver store_smaa_bogst

- 1.3.3 ortografi bogstaver
 - 1.3.3.1 ortografi bogstaver translitteration
 - 1.3.3.2 ortografi bogstaver bogstav_lyd_forhold
 - 1.3.3.3 ortografi bogstaver fremmede_bogstaver
 - 1.3.3.4 ortografi bogstaver alfabetisk_raekkefoelge
 - 1.3.3.5 ortografi bogstaver hjemlige_bogstaver
 - 1.3.3.6 ortografi bogstaver tal_bogstaver
- 1.3.4 ortografi et_to_ord
 - 1.3.4.1 ortografi et_to_ord sammenskrivning
 - 1.3.4.2 ortografi et_to_ord saerskrivning
- 1.3.5 ortografi saerlige_staveproblemer
 - 1.3.5.1 ortografi saerlige_staveproblemer stumme_bogstaver
 - 1.3.5.2 ortografi saerlige_staveproblemer stavning_af_fremmedord
 - 1.3.5.3 ortografi saerlige_staveproblemer dobbeltformer
 - 1.3.5.4 ortografi saerlige_staveproblemer en_eller_to_konsonanter
- 1.3.6 ortografi homonymer
 - 1.3.6.1 ortografi homonymer homofoner
 - 1.3.6.2 ortografi homonymer homografer

1.4 semantik

- 1.4.1 semantik semantiske_roller
 - 1.4.1.1 semantik semantiske_roller agens
 - 1.4.1.2 semantik semantiske_roller andre_semantiske_roller
 - 1.4.1.3 semantik semantiske_roller patiens
- 1.4.2 semantik synonym
- 1.4.3 semantik dansk_oversaettelse
- 1.4.4 semantik betydning
 - 1.4.4.1 semantik betydning begrebsafklaring
 - 1.4.4.2 semantik betydning konnotationer
 - 1.4.4.3 semantik betydning overfoert_betydning
 - 1.4.4.4 semantik betydning betydningsudvikling
- 1.4.5 semantik fast_udtryk
 - 1.4.5.1 semantik fast_udtryk ordsprog
 - 1.4.5.2 semantik fast_udtryk fast_vending
 - 1.4.5.3 semantik fast_udtryk kliche
 - 1.4.5.4 semantik fast_udtryk idiom
- 1.4.6 semantik antonym

1.5 interpunktion

- 1.5.1 interpunktion punktum
- 1.5.2 interpunktion kolon
- 1.5.3 interpunktion spoergsmaalstegn
- 1.5.4 interpunktion tankestreg
- 1.5.5 interpunktion semikolon
- 1.5.6 interpunktion komma
- 1.5.7 interpunktion udraabstegn

1.6 pragmatik

- 1.6.1 pragmatik deiksis
- 1.6.2 pragmatik meddelelsesstruktur
- 1.6.3 pragmatik udbredelse
- 1.6.4 pragmatik talehandling
- 1.6.5 pragmatik holdningstilkendegivelse
- 1.6.6 pragmatik implikatur
- 1.6.7 pragmatik praesupposition

1.7 etymologi

- 1.7.1 etymologi tidsudvikling
- 1.7.2 etymologi orddatering
- 1.7.3 etymologi laan
 - 1.7.3.1 etymologi laan importsprog
 - 1.7.3.1.1 etymologi laan importsprog fra_engelsk

- 1.7.3.1.2 etymologi laan importsprog fra_andre_sprog
- 1.7.3.2 etymologi laan importmaade
 - 1.7.3.2.1 etymologi laan importmaade indirekte_laan
 - 1.7.3.2.1.1 etymologi laan importmaade indirekte_laan fri_gengivelse
 - 1.7.3.2.1.2 etymologi laan importmaade indirekte_laan betydningslaan
 - 1.7.3.2.1.3 etymologi laan importmaade indirekte_laan oversaettelelseslaan
 - 1.7.3.2.1.4 etymologi laan importmaade indirekte_laan hybrid_laan
 - 1.7.3.2.1.5 etymologi laan importmaade indirekte_laan pseudolaan
 - 1.7.3.2.1.6 etymologi laan importmaade indirekte_laan paralleldannelse
 - 1.7.3.2.2 etymologi laan importmaade direkte_laan
 - 1.7.3.2.2.1 etymologi laan importmaade direkte_laan citatord
 - 1.7.3.2.2.2 etymologi laan importmaade direkte_laan tilpasset_laan
- 1.7.4 etymologi arveord

- 1.8 layout
 - 1.8.1 layout noter
 - 1.8.2 layout skrifttyper
 - 1.8.3 layout citater
 - 1.8.4 layout punkttopstillinger
 - 1.8.5 layout litteraturliste

- 1.9 videnscenter-spm
 - 1.9.1 videnscenter-spm sprogundervisning
 - 1.9.2 videnscenter-spm retskrivningshistorie
 - 1.9.3 videnscenter-spm om_opslagsvaerker
 - 1.9.4 videnscenter-spm navneforskning
 - 1.9.5 videnscenter-spm sprogteknologi
 - 1.9.6 videnscenter-spm litteraere_spm
 - 1.9.7 videnscenter-spm om_RO
 - 1.9.8 videnscenter-spm nordiske_talord
 - 1.9.9 videnscenter-spm sprogpolitik
 - 1.9.10 videnscenter-spm om_kommasystem
 - 1.9.11 videnscenter-spm om_Sproгнаevnet
 - 1.9.12 videnscenter-spm klarsprog

- 1.10 fonetik
 - 1.10.1 fonetik prosodi
 - 1.10.1.1 fonetik prosodi tryk
 - 1.10.1.2 fonetik prosodi intonation
 - 1.10.1.3 fonetik prosodi stoed
 - 1.10.2 fonetik udtale
 - 1.10.2.1 fonetik udtale enkeltlyde
 - 1.10.2.2 fonetik udtale lydudvikling
 - 1.10.2.3 fonetik udtale stavelse

- 1.11 syntaks
 - 1.11.1 syntaks kollokationer
 - 1.11.2 syntaks saetningsafgraensning
 - 1.11.2.1 syntaks saetningsafgraensning oversaetning
 - 1.11.2.2 syntaks saetningsafgraensning periode
 - 1.11.2.3 syntaks saetningsafgraensning helsaetningsstamme
 - 1.11.2.4 syntaks saetningsafgraensning helsaetning
 - 1.11.2.5 syntaks saetningsafgraensning ledsaetning
 - 1.11.2.5.1 syntaks saetningsafgraensning ledsaetning spoergeledsaetning
 - 1.11.2.5.2 syntaks saetningsafgraensning ledsaetning at_ledsaetning
 - 1.11.2.5.3 syntaks saetningsafgraensning ledsaetning indroemmelsesledsaetning
 - 1.11.2.5.4 syntaks saetningsafgraensning ledsaetning tidsledsaetning
 - 1.11.2.5.5 syntaks saetningsafgraensning ledsaetning betingelsesledsaetning
 - 1.11.2.5.6 syntaks saetningsafgraensning ledsaetning aarsagsledsaetning
 - 1.11.2.5.7 syntaks saetningsafgraensning ledsaetning hensigtsledsaetning
 - 1.11.2.5.8 syntaks saetningsafgraensning ledsaetning sammenligningsledsaetning
 - 1.11.2.5.9 syntaks saetningsafgraensning ledsaetning foelgeledsaetning

- 1.11.2.5.10 syntaks saetningsafgraensning ledsaetning relativsaetning
- 1.11.2.6 syntaks saetningsafgraensning andre_saetningstyper
 - 1.11.2.6.1 syntaks saetningsafgraensning andre_saetningstyper der_kloevning
 - 1.11.2.6.2 syntaks saetningsafgraensning andre_saetningstyper citeret_tale
 - 1.11.2.6.2.1 syntaks saetningsafgraensning andre_saetningstyper citeret_tale daekning
 - 1.11.2.6.2.2 syntaks saetningsafgraensning andre_saetningstyper citeret_tale indirekte_tale
 - 1.11.2.6.2.3 syntaks saetningsafgraensning andre_saetningstyper citeret_tale inkvit
 - 1.11.2.6.2.4 syntaks saetningsafgraensning andre_saetningstyper citeret_tale direkte_tale
 - 1.11.2.6.3 syntaks saetningsafgraensning andre_saetningstyper det_kloevning
 - 1.11.2.6.4 syntaks saetningsafgraensning andre_saetningstyper ufuldstaendig_saetning
 - 1.11.2.6.5 syntaks saetningsafgraensning andre_saetningstyper indskud
 - 1.11.2.6.6 syntaks saetningsafgraensning andre_saetningstyper skjult_saetning
 - 1.11.2.6.7 syntaks saetningsafgraensning andre_saetningstyper indlejret_saetning
 - 1.11.2.6.8 syntaks saetningsafgraensning andre_saetningstyper saetningsknude
- 1.11.3 syntaks saetningsled
 - 1.11.3.1 syntaks saetningsled adverbial
 - 1.11.3.1.1 syntaks saetningsled adverbial saetningsadverbial
 - 1.11.3.1.2 syntaks saetningsled adverbial holdningsadverbial
 - 1.11.3.1.3 syntaks saetningsled adverbial stedsadverbial
 - 1.11.3.1.4 syntaks saetningsled adverbial maadesadverbial
 - 1.11.3.1.5 syntaks saetningsled adverbial praepositionsforbindelse
 - 1.11.3.1.5.1 syntaks saetningsled adverbial praepositionsforbindelse styrelse
 - 1.11.3.1.6 syntaks saetningsled adverbial gradsadverbial
 - 1.11.3.1.7 syntaks saetningsled adverbial frit_adverbial
 - 1.11.3.1.8 syntaks saetningsled adverbial tidsadverbial
 - 1.11.3.1.9 syntaks saetningsled adverbial fast_adverbial
 - 1.11.3.1.10 syntaks saetningsled adverbial t_adverbial
 - 1.11.3.2 syntaks saetningsled praedikativ
 - 1.11.3.2.1 syntaks saetningsled praedikativ subjektspraedikativ
 - 1.11.3.2.2 syntaks saetningsled praedikativ frit_praedikativ
 - 1.11.3.2.3 syntaks saetningsled praedikativ objektspraedikativ
 - 1.11.3.3 syntaks saetningsled konjunkional
 - 1.11.3.4 syntaks saetningsled subjekt
 - 1.11.3.4.1 syntaks saetningsled subjekt formelt_subjekt
 - 1.11.3.4.2 syntaks saetningsled subjekt reelt_subjekt
 - 1.11.3.4.3 syntaks saetningsled subjekt indholdssubjekt
 - 1.11.3.5 syntaks saetningsled verbal
 - 1.11.3.5.1 syntaks saetningsled verbal valens
 - 1.11.3.5.1.1 syntaks saetningsled verbal valens transitiv
 - 1.11.3.5.1.2 syntaks saetningsled verbal valens intransitiv
 - 1.11.3.6 syntaks saetningsled apposition
 - 1.11.3.7 syntaks saetningsled objekt
 - 1.11.3.7.1 syntaks saetningsled objekt indirekte_objekt
 - 1.11.3.7.2 syntaks saetningsled objekt praepositionsobjekt
 - 1.11.3.7.3 syntaks saetningsled objekt direkte_objekt
- 1.11.4 syntaks kongruens
- 1.11.5 syntaks ledstilling
 - 1.11.5.1 syntaks ledstilling ekstraposition
 - 1.11.5.2 syntaks ledstilling slutfelt
 - 1.11.5.3 syntaks ledstilling centralfelt
 - 1.11.5.3.1 syntaks ledstilling centralfelt lette_led
 - 1.11.5.3.2 syntaks ledstilling centralfelt negerede_led
 - 1.11.5.4 syntaks ledstilling inversion
 - 1.11.5.5 syntaks ledstilling forfelt
- 1.11.6 syntaks nominal
 - 1.11.6.1 syntaks nominal adled
 - 1.11.6.1.1 syntaks nominal adled attributiv
 - 1.11.6.1.1.1 syntaks nominal adled attributiv attributivt_adjektiv
 - 1.11.6.1.1.2 syntaks nominal adled attributiv participiumsformer
 - 1.11.6.1.1.2.1 syntaks nominal adled attributiv participiumsformer praesens_participium
 - 1.11.6.1.1.2.2 syntaks nominal adled attributiv participiumsformer praeteritum_participium
 - 1.11.6.1.2 syntaks nominal adled andre_adled

- 1.11.6.2 syntaks nominal kerne
- 1.11.7 syntaks forbindelsesart
 - 1.11.7.1 syntaks forbindelsesart underordning
 - 1.11.7.2 syntaks forbindelsesart sideordning
 - 1.11.7.3 syntaks forbindelsesart samordning
- 1.11.8 syntaks saetningstype
 - 1.11.8.1 syntaks saetningstype spoergende
 - 1.11.8.2 syntaks saetningstype imperativ
 - 1.11.8.3 syntaks saetningstype fremsaettende
- 1.12 sproglig_variation
 - 1.12.1 sproglig_variation stilvarianter
 - 1.12.1.1 sproglig_variation stilvarianter slang
 - 1.12.1.2 sproglig_variation stilvarianter uformel_stil
 - 1.12.1.3 sproglig_variation stilvarianter fagsproglig_stil
 - 1.12.1.4 sproglig_variation stilvarianter formel_stil
 - 1.12.2 sproglig_variation lekter
 - 1.12.2.1 sproglig_variation lekter kronolekt
 - 1.12.2.2 sproglig_variation lekter standardsprog
 - 1.12.2.3 sproglig_variation lekter dialekt
 - 1.12.2.4 sproglig_variation lekter etnolekt
 - 1.12.2.5 sproglig_variation lekter sociolekt
 - 1.12.3 sproglig_variation medie
 - 1.12.3.1 sproglig_variation medie talesprog
 - 1.12.3.2 sproglig_variation medie skriftsprog
 - 1.12.3.3 sproglig_variation medie andre_medier

2 XX

2. Den svenske emnetaksonomi

- 1.1 hur_ska_ordet_se_ut?
 - 1.1.1 hur_ska_ordet_se_ut? -> Fraemmande_ordformer
 - 1.1.2 hur_ska_ordet_se_ut? -> Ett_ord_eller_flera
 - 1.1.3 hur_ska_ordet_se_ut? -> Uttal
 - 1.1.4 hur_ska_ordet_se_ut? -> Avstavning
 - 1.1.5 hur_ska_ordet_se_ut? -> Stavning
 - 1.1.6 hur_ska_ordet_se_ut? -> Transkription_och_translitterering
 - 1.1.7 hur_ska_ordet_se_ut? -> Genitiv
 - 1.1.8 hur_ska_ordet_se_ut? -> Komparation
 - 1.1.9 hur_ska_ordet_se_ut? -> Oevrig_boejning
 - 1.1.10 hur_ska_ordet_se_ut? -> Sammansaettningsfog
 - 1.1.11 hur_ska_ordet_se_ut? -> Foerledsanslutning_varm_korvgubbe
 - 1.1.12 hur_ska_ordet_se_ut? -> Oevriga_ordbildningsfraagor
- 1.2 naer_ska_en_viss_ordform_anvaendas?
 - 1.2.1 naer_ska_en_viss_ordform_anvaendas? -> Singular_plural
 - 1.2.2 naer_ska_en_viss_ordform_anvaendas? -> Bestaemd_obestaemd_form_och_artikel
 - 1.2.3 naer_ska_en_viss_ordform_anvaendas? -> Lilla_lille
 - 1.2.4 naer_ska_en_viss_ordform_anvaendas? -> De_dem_han_honom
 - 1.2.5 naer_ska_en_viss_ordform_anvaendas? -> Sin_hans
 - 1.2.6 naer_ska_en_viss_ordform_anvaendas? -> Tempus
- 1.3 konstruktioner_och_meningsbyggnaed
 - 1.3.1 konstruktioner_och_meningsbyggnaed -> Prepositionsbruk
 - 1.3.2 konstruktioner_och_meningsbyggnaed -> Partikelverb
 - 1.3.3 konstruktioner_och_meningsbyggnaed -> Att-strykning
 - 1.3.4 konstruktioner_och_meningsbyggnaed -> Annan_strykning
 - 1.3.5 konstruktioner_och_meningsbyggnaed -> Kongruens
 - 1.3.6 konstruktioner_och_meningsbyggnaed -> Oevriga_konstruktioner
 - 1.3.7 konstruktioner_och_meningsbyggnaed -> Subjektsregeln

- 1.3.8 konstruktioner_och_meningsbyggnad -> Ordfoeljd
- 1.3.9 konstruktioner_och_meningsbyggnad -> Oevrig_meningsbyggnad

- 1.4 skrivregler
 - 1.4.1 skrivregler -> Stor_liten_bokstav
 - 1.4.2 skrivregler -> Siffror_i_tidsuppgifter
 - 1.4.3 skrivregler -> Foerkortning
 - 1.4.4 skrivregler -> Siffror_i_ovrigt
 - 1.4.5 skrivregler -> Citattecken
 - 1.4.6 skrivregler -> Komma_och_kommatering
 - 1.4.7 skrivregler -> Andra_skiljetecken
 - 1.4.8 skrivregler -> Oevriga_tecken
 - 1.4.9 skrivregler -> Sortering_och_alfabetisering
 - 1.4.10 skrivregler -> Stycke
 - 1.4.11 skrivregler -> Grafik
 - 1.4.12 skrivregler -> Specialtext

- 1.5 text_och_stil
 - 1.5.1 text_och_stil -> Textbyggnad
 - 1.5.2 text_och_stil -> Stil_och_stilfigurer

- 1.6 namn_och_tilltal
 - 1.6.1 namn_och_tilltal -> Titlar_och_tilltal
 - 1.6.2 namn_och_tilltal -> Geografiska_namn
 - 1.6.3 namn_och_tilltal -> Personnamn
 - 1.6.4 namn_och_tilltal -> Oevriga_namn

- 1.7 ord_och_frassamlingar
 - 1.7.1 ord_och_frassamlingar -> Allmaenna_ord_och_frassamlingar

- 1.8 fack_og_aemnesspraak
 - 1.8.1 fack_og_aemnesspraak -> EU
 - 1.8.2 fack_og_aemnesspraak -> Juridiskt_spraak
 - 1.8.3 fack_og_aemnesspraak -> Mat_och_dryckspraak
 - 1.8.4 fack_og_aemnesspraak -> Medicinskt_spraak
 - 1.8.5 fack_og_aemnesspraak -> Mediespraaket
 - 1.8.6 fack_og_aemnesspraak -> Myndighetspraak_och_klarspraak
 - 1.8.7 fack_og_aemnesspraak -> Skoenlitteraert_och_religioest_spraak
 - 1.8.8 fack_og_aemnesspraak -> Sportsspraaket
 - 1.8.9 fack_og_aemnesspraak -> Teknik_och_dataspraak
 - 1.8.10 fack_og_aemnesspraak -> Fackspraak_ovrigt
 - 1.8.11 fack_og_aemnesspraak -> Fack_och_aemnesordsamlingar

- 1.9 talspraak_och_spraakvarieteter
 - 1.9.1 talspraak_och_spraakvarieteter -> Talspraaksdrag
 - 1.9.2 talspraak_och_spraakvarieteter -> Slang_svordomar_och_gruppspraak
 - 1.9.3 talspraak_och_spraakvarieteter -> Finlandssvenska
 - 1.9.4 talspraak_och_spraakvarieteter -> Dialekter_och_regionalt_spraak

- 1.10 spraakinlaerning
 - 1.10.1 spraakinlaerning -> Undervisning
 - 1.10.2 spraakinlaerning -> Svenska_foer_invandrare
 - 1.10.3 spraakinlaerning -> Spraakkunskaper

- 1.11 fraemmande_spraak
 - 1.11.1 fraemmande_spraak -> Fraemmande_paaverkan_paa_svenskan
 - 1.11.2 fraemmande_spraak -> Oeversaettning_och_tolkning
 - 1.11.3 fraemmande_spraak -> Teckenspraak
 - 1.11.4 fraemmande_spraak -> Minoritets_och_invandarspraak_i_Sverige
 - 1.11.5 fraemmande_spraak -> Spraaken_i_Norden
 - 1.11.6 fraemmande_spraak -> Latin_och_grekiska
 - 1.11.7 fraemmande_spraak -> Engelska

- 1.11.8 fraemmande_spraak -> Tyska
- 1.11.9 fraemmande_spraak -> Oevriga_fraemmande_spraak

- 1.12 spraakvetenskapliga_omraaden
 - 1.12.1 spraakvetenskapliga_omraaden -> IT_och_spraakteknologi
 - 1.12.2 spraakvetenskapliga_omraaden -> Retorik
 - 1.12.3 spraakvetenskapliga_omraaden -> Spraakhistoria_och_spraakfoeraendring
 - 1.12.4 spraakvetenskapliga_omraaden -> Spraak_och_koen
 - 1.12.5 spraakvetenskapliga_omraaden -> Spraaksociologi
 - 1.12.6 spraakvetenskapliga_omraaden -> Spraakpsykologi
 - 1.12.7 spraakvetenskapliga_omraaden -> Spraakvaard

- 1.13 oevrigt

3. Den norske emnetaksonomi

A Kva er rett? [Kva er rett skrivemåte, bøying, uttale eller teiknsetjing?

– Spørsmål om rettskriving og praktiske språkreglar der det ofte er fastsett kva som er norma.]

A1 Korleis skal ordet (namnet) skrivast?

- A1.1 Rett skrivemåte av enkeltord
- A1.2 Binde-s og binde-e
- A1.3 Namn
 - A1.3.1 Fornamn og etternamn
 - A1.3.2 Stadnamn
- A1.4 Transkripsjon

A2 Korleis skal ordet bøyast?

- A2.1 Substantiv – generelt om bøyinga
 - A2.1.1 Substantiv – enkeltord
 - A2.1.2 Engelske ord
 - A2.1.3 Forkortingar og forbokstavord
 - A2.1.4 Kjønn (genus)
 - A2.1.5 Latin og gresk
- A2.2 Verb
- A2.3 Adjektiv/determinativ

A3 Korleis skal ordet (namnet) uttalast?

- A3.1 Namn
 - A3.1.1 Fornamn og etternamn
 - A3.1.2 Stadnamn

A4 Skriveregler [Praktiske språkreglar for teiknsetjing, store og små bokstavar, særskriving og samskriving, orddeling, forkortingar o.l.]

- A4.1 Avsnitt, brevoppsett og punkttoppstilling
- A4.2 Forkortingar
- A4.3 Samanskriving, særskriving og orddeling
 - A4.3.1 Faste ordgrupper?
 - A4.3.2 Samansette ord i eitt
 - A4.3.3 Orddeling ved linjeskift
- A4.4 Samansetningar med spesielle ledd: forkorting, tal eller namn
 - A4.4.1 Enkle samansetningar
 - A4.4.2 Fleirledda samansetningar
- A4.5 Stor og liten bokstav
- A4.6 Andre namn
- A4.7 Tal
- A4.8 Teiknsetjing
 - A4.8.1 Komma
 - A4.8.2 Punktum, kolon og semikolon
 - A4.8.3 Spørjeteikn og ropeteikn

- A4.8.4 Bindestrøk og tankestrøk
- A4.8.5 Aksenteikn og apostrof
- A4.8.6 Hermeteikn, kursiv og sitat
- A4.8.7 Skråstrøk og parentes
- A4.8.8 Andre teikn
- A4.8.9 Mellomrom

B Hvordan bør jeg ordlegge meg (formulere meg)? Hvordan finner jeg den gode uttrykksmåten?

B1 Spørsmål om enkeltord (betydning, opprinnelse, valg av riktig ord, hva noe kalles, osv.)

- B1.1 Hva betyr det? [Saksopplysninger til den som er i tvil om ordet/ordene passer.]
 - B1.1.1 Betydning
 - B1.1.2 Lyder, tegn, former
 - B1.1.3 Ord
 - B1.1.4 Samanlikne to eller fleire
- B1.2 Hva kommer det av?
 - B1.2.1 Ordhistorie
 - B1.2.2 Ord (enkle og samansette)
 - B1.2.3 Samanlikne to eller fleire
- B1.3 Kan dette ordet brukes, og hvordan brukes det?
 - B1.3.1 Kan det brukast?
 - B1.3.2 Korleis brukast det?
 - B1.3.3 Val mellom to/fleire
- B1.4 Hva heter dette, hva kan det kalles?
 - B1.4.1 B. Kva heiter dette?
 - B1.4.2 Antonym
 - B1.4.3 Folk/ting frå ...
 - B1.4.4 Avløysarord/omsetting/nyord
 - B1.4.5 Frå andre språk
 - B1.4.6 Frå engelsk
 - B1.4.7 Til nynorsk
 - B1.4.8 Termer
 - B1.4.9 Dataord
 - B1.4.10 Andre termar

B2 Spørsmål om to eller flere ord i sammenheng (setningsbygning, samsvar, faste uttrykk, ordtak osv.)

- B2.1 Hva betyr det?
 - B2.1.1 Betydning
 - B2.1.2 Fleire ord saman
 - B2.1.3 Samanlikne to eller fleire
 - B2.1.4 Uttrykk/ordtak/sitat
- B2.2 Hva kommer det av?
 - B2.2.1 Ordhistorie
 - B2.2.2 Fleire ord saman
 - B2.2.3 Uttrykk/ordtak/sitat
- B2.3 Kan denne uttrykksmåten brukes, og hvordan brukes den?
 - B2.3.1 Kan det brukast?
 - B2.3.2 Korleis brukast det?
 - B2.3.3 Val mellom to/fleire
- B2.4 Hvilken uttrykksmåte bør jeg velge?
 - B2.4.1 et/ett
 - B2.4.2 og/å
 - B2.4.3 Genitiv (s og sin)
 - B2.4.4 -ing/-ning/-else
 - B2.4.5 Preposisjonar
 - B2.4.6 Pronomen
 - B2.4.7 Fleire ord
 - B2.4.8 Skrivemåte, fleirordsuttrykk
 - B2.4.9 Ordtak og faste uttrykk

C Diverse språktema

C1 Mer grammatikk

C1.1 Nye grammatiske betegnelser

C1.2 Ordklasser og former

C1.3 Setning, ledd og frase

C1.4 Konsekvente former

C1.5 Imperativ

C1.6 Preposisjonsbruk

C1.6.1 bokmål

C1.6.2 nynorsk

C2 Språkvarianter

C2.1 Fremmede språk

C2.2 Bokmål og nynorsk

C2.3 Mållover og -regler

C2.4 Dialekter

C2.5 Slang og sms (gruppespråk)

C2.6 Fagspråk og terminologi

C2.6.1 Ikt

C3 Språket gjennom årene

C3.1 Språkhistorie

C3.1.1 Norrønt

C3.2 Samnorsk

C3.3 Rettskrivningsendringer - reformer og former

C4 Tekst og stil

C4.1 Brevoppsett/e-postoppsett (maler?)

C4.2 Helsing og underskrift

C4.3 Stil og stilfigurer

C5 Annet

C5.1 Godkjenning av nye ord

C5.2 Engelsk/global påvirkning

C5.3 Generell korrektur/oversettelse

C5.3.1 bokmål og nynorsk

C5.4 Norsk i tall

C5.5 Alfabet og alfabetisering

C5.6 Diverse

4. Den islandske emnetaksonomi

A: tilpasning av låneord

B: bøjning

D: ord

E: rettskrivning

F: udtale

J: etymologi

L: udtryksmåde

M: betydning

N: navne

O: ordforbindelse

S: syntaks

Y: orddannelse

5. Ækvivalensnøgle for nordiske emnetaksonomier

Norsk->Dansk

A1 -> 1.3

Svensk->Dansk

1.1.1 -> 1.3.3.3

Finsk->Dansk

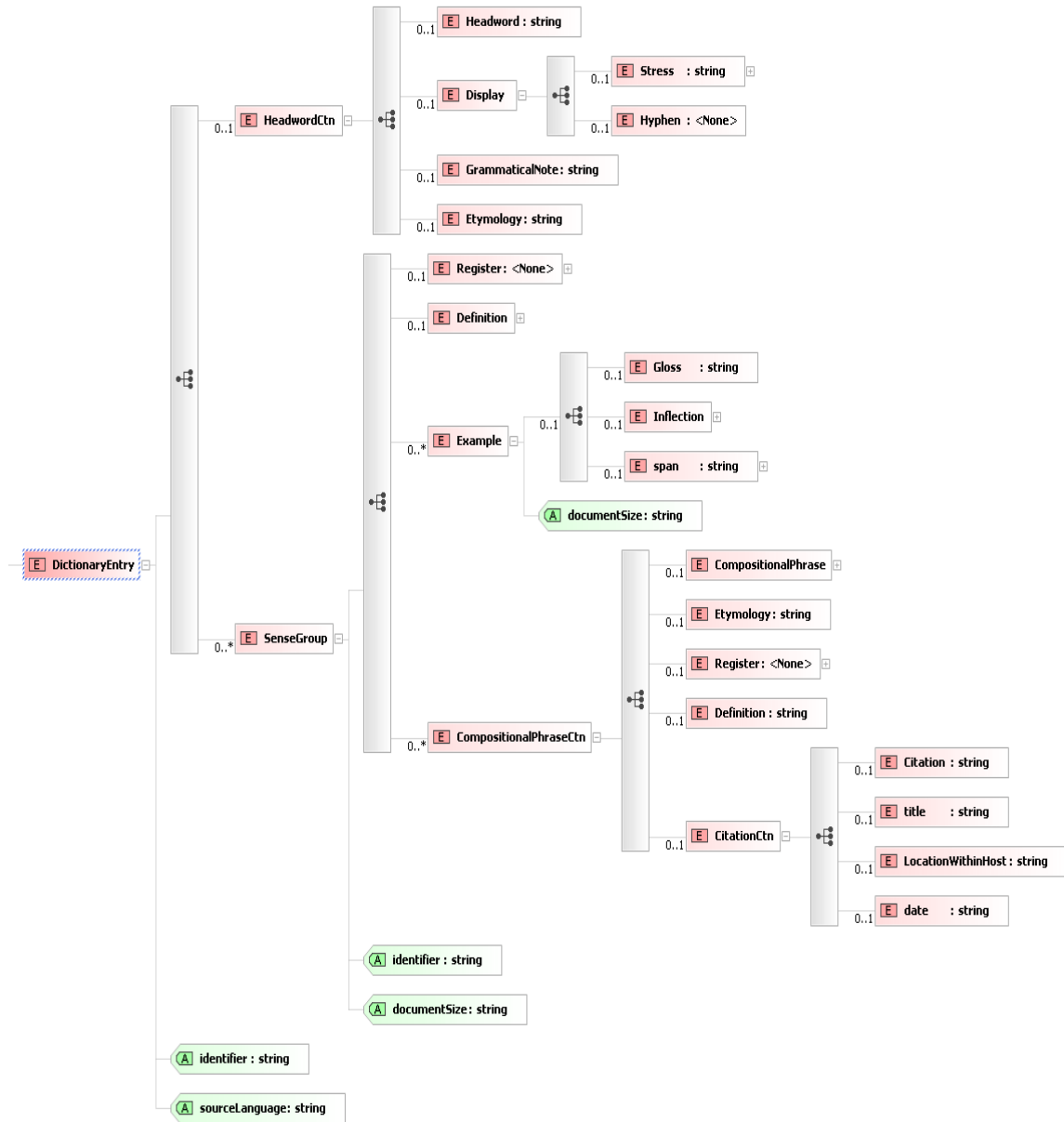
1 -> 1.1.1.3

A1.1 -> 1.3	1.1.2 -> 1.3.4	2 -> 1.1.1.1
A1.2 -> 1.1.2.1.2.1	1.1.3 -> 1.10	3 -> 1.12.1.1, 1.12.1.2
A1.3 -> 1.1.1.7.1	1.1.4 -> 1.3.1.1.2	4 -> 1.12.1.3, 1.2.4
A1.3.1 -> 1.1.1.7.1.1	1.1.5 -> 1.3	5 -> 1.2.7
A1.3.2 -> 1.1.1.7.1.2	1.1.6 -> 1.3.3.1	8 -> 1.12.1.2
A1.4 -> 1.3.3.1	1.1.7 -> 1.1.3.2.4.3	21 -> 1.12.1.3?
A2 -> 1.1.3	1.1.8 -> 1.1.3.2.3	22 -> 1.12.1.3, 1.2.4
A2.1 -> 1.1.3, 1.1.1.7	1.1.9 -> 1.1.3	23 -> 1.1.1.7.1.1, 1.1.1.7.1.2
A2.1.1 -> 1.1.1.7	1.1.10 -> 1.1.2.1.2.1	24 -> 1.1.1.7.1.1
A2.1.2 -> 1.1.3, 1.7.3.1.1	1.1.11 -> 1.1	25 -> 1.7
A2.1.3 -> 1.1.3, 1.1.2.1.3	1.1.12 -> 1.1	30 -> 1.10
A2.1.4 -> 1.1.3, 1.1.3.2.1	1.2.1 -> 1.1.3.2.7	31 -> 1.4.5, 1.11.1
A2.1.5 -> {Ø}	1.2.2 -> 1.1.1.8, 1.1.3.2.5	32 -> 1.4.5, 1.11.1
A2.2 -> 1.1.3, 1.1.1.5	1.2.3 -> 1.11.4	33 -> 1.4.5, 1.11.1
A2.3 -> 1.1.3, 1.1.1.3, 1.1.3.2.5	1.2.4 -> 1.1.3.2.4.4	63 -> {Ø}
A3 -> 1.10	1.2.5 -> 1.1.1.2.5	64 -> {Ø}, ALT!
A3.1 -> 1.10, 1.1.1.7.1	1.2.6 -> 1.1.3.2.2	65 -> 1.9.9
A3.1.1 -> 1.10, 1.1.1.7.1.1	1.3.1 -> 1.1.1.4, 1.11.3.1.5, 1.11.1	66 -> {Ø} ??
A3.1.2 -> 1.10, 1.1.1.7.1.2	1.3.2 -> 1.1.1.5.3	67 -> {Ø}, 1.2 ??
A4 -> 1.3, 1.8	1.3.3 -> 1.11.2.5.2	77 -> {Ø} ??
A4.1 -> 1.8	1.3.4 -> 1.11.2.6.4	80 -> 1.4.3
A4.2 -> 1.1.2.1.3	1.3.5 -> 1.11.4	81 -> 1.1.1.7.1.3
A4.3 -> 1.3.4, 1.3.1.1.2	1.3.6 -> 1.11	83 -> 1.12.2.1
A4.3.1 -> 1.4.5, 1.11.1	1.3.7 -> 1.11.3.4	84 -> {Ø}, 1.12.1?
A4.3.2 -> 1.1.2.1.2	1.3.8 -> 1.11.5	88 -> 1.1.2.1.3
A4.3.3 -> 1.3.1.1.2	1.3.9 -> 1.4.4	89 -> {Ø}, 1.12.1?
A4.4 -> 1.1.2.1.2.3	1.4.1 -> 1.3.2	95 -> {Ø}, 1.12.1?
A4.4.1 -> {Ø}	1.4.2 -> 1.1.2.1.3	100 -> {Ø}
A4.4.2 -> {Ø}	1.4.3 -> 1.3.3.6, 1.1.1.6	104 -> 1.1.3.2.7
A4.5 -> 1.3.2	1.4.4 -> 1.3.3.6, 1.1.1.6	105 -> 1.12.3.1 el. 1.12.2.3 (dialekt)?
A4.6 -> {Ø}	1.4.5 -> 1.8.3	109 -> 1.3
A4.7 -> 1.1.1.6, 1.3	1.4.6 -> 1.5.6	131 -> 1.1
A4.8 -> 1.5, 1.3.1	1.4.7 -> 1.5	132 -> 1.2
A4.8.1 -> 1.5.6	1.4.8 -> 1.3.1	133 -> 1.4
A4.8.2 -> 1.5.1, 1.5.2, 1.5.5	1.4.9 -> 1.3.3.4	134 -> 1.12.1.1
A4.8.3 -> 1.5.3, 1.5.7	1.4.10 -> 1.8	135 -> 1.12.1.1
A4.8.4 -> 1.5.4, 1.3.1.2.1	1.4.11 -> 1.8	139 -> 1.1.1.7.1.1
A4.8.5 -> 1.3.1.1.4, 1.3.1.2.4	1.5.1 -> 1.6.2	141 -> {Ø}
A4.8.6 -> 1.3.1.1.7, 1.8.3	1.5.2 -> 1.12.1	150 -> {Ø}, 1.9 ??
A4.8.7 -> 1.3.1.1.6, 1.3.1.1.3	1.6.1 -> 1.1.1.7.1	154 -> {Ø} ??
A4.8.8 -> 1.3.1.1	1.6.2 -> 1.1.1.7.1.2	166 -> 1.1.1.5
A4.8.9 -> 1.3.4	1.6.3 -> 1.1.1.7.1.1	171 -> 1.7.3
B1.1 -> 1.4	1.6.4 -> 1.1.1.7.1	172 -> 1.1.2.1.2, 1.1.1.3
B1.1.1 -> 1.4.4	1.7.1 -> 1.11.1, 1.4.5	173 -> 1.1.2.1.2, 1.1.1.1
B1.1.2 -> 1.4, 1.3.1	1.8.1 -> 1.2.4	175 -> 1.1.3.2.7

B1.1.3 -> {Ø}	1.8.2 -> 1.2.4	176 -> {Ø} (default i DK)
B1.1.4 -> 1.4.4.1	1.8.3 -> 1.2.4	177 -> 1.12.1.4
B1.2 -> 1.7	1.8.4 -> 1.2.4	179 -> 1.1.2.1.2
B1.2.1 -> 1.7	1.8.5 -> 1.2.4	180 -> 1.1.1.1.5.3
B1.2.2 -> {Ø}	1.8.6 -> 1.2.4	181 -> 1.10.2
B1.2.3 -> {Ø}	1.8.7 -> 1.2.4	
B1.3 -> 1.6	1.8.8 -> 1.2.4	
B1.3.1 -> 1.2.8	1.8.9 -> 1.2.4	
B1.3.2 -> 1.11, 1.1	1.8.10 -> 1.2.4	
B1.3.3 -> 1.6, 1.4.4.1	1.8.11 -> 1.2.4	
B1.4 -> 1.2.6	1.9.1 -> 1.12.3.1	
B1.4.1 -> 1.2.6	1.9.2 -> 1.12.1.1, 1.12.2.5	
B1.4.2 -> 1.4.6	1.9.3 -> {Ø}	
B1.4.3 -> 1.2.2	1.9.4 -> 1.12.2.3	
B1.4.4 -> 1.7.3.2.1.3, 1.2.10	1.10.1 -> 1.9.1	
B1.4.5 -> 1.7.3.1.2	1.10.2 -> 1.9.1	
B1.4.6 -> 1.7.3.1.1	1.10.3 -> 1.9.1	
B1.4.7 -> {Ø}	1.11.1 -> 1.7.3.1.2, 1.9.9	
B1.4.8 -> 1.2.4	1.11.2 -> 1.4.3	
B1.4.9 -> 1.2.4	1.11.3 -> {Ø}	
B1.4.10 -> 1.2.4	1.11.4 -> {Ø}	
B2.1 -> 1.4	1.11.5 -> {Ø}	
B2.1.1 -> 1.4	1.11.6 -> {Ø}	
B2.2 -> 1.7	1.11.7 -> 1.7.3.1.1	
B2.2.1 -> 1.7	1.11.8 -> {Ø}	
B2.3 -> 1.6	1.11.9 -> {Ø}	
B2.4 -> 1.1, 1.3	1.12.1 -> 1.9.5	
B2.4.1 -> 1.1.1.8	1.12.2 -> {Ø}	
B2.4.2 -> 1.1.1.5, 1.12.3.1	1.12.3 -> 1.7	
B2.4.3 -> 1.1.3.2.4.3, 1.1.1.2.6	1.12.4 -> {Ø}	
B2.4.4 -> 1.1.1.7.2, 1.1.2.1.1	1.12.5 -> {Ø}	
B2.4.5 -> 1.1.1.4, 1.11.1	1.12.6 -> {Ø}	
B2.4.6 -> 1.1.1.2	1.12.7 -> 1.9.9	
B2.4.7 -> {Ø}		
B2.4.8 -> 1.3.1.2.1		
B2.4.9 -> 1.4.5		
C1 -> 1.1, 1.11	Islandsk->Dansk	
C1.1 -> 1.2.4	A -> 1.7.3.2.2.2	
C1.2 -> 1.1.1, 1.1.2	B -> 1.1	
C1.3 -> 1.11	D -> 1.2	
C1.4 -> 1.11.4, 1.1.3.2.1, 1.1.1.8	E -> 1.3	
C1.5 -> 1.11.8.2	F -> 1.10	
C1.6 -> 1.1.1.4, 1.11.3.1.5, 1.11.1	J -> 1.7	
C1.6.1 -> {Ø}	L -> 1.3?	
C1.6.2 -> {Ø}	M -> 1.4	
C2 -> 1.12	N -> 1.9	
C2.1 -> {Ø}	O -> 1.1.2.1.2	
C2.2 -> {Ø}	S -> 1.11	
C2.3 -> 1.9.9	Y -> 1.1.2	
C2.4 -> 1.12.2.3		

C2.5 -> 1.12.1.1
C2.6 -> 1.2.4, 1.4.4
C2.6.1 -> 1.2.4
C3 -> 1.7, 1.9
C3.1 -> 1.7
C3.1.1 -> { \emptyset }
C3.2 -> { \emptyset }
C3.3 -> 1.9.2
C4 -> 1.8
C4.1 -> 1.8
C4.2 -> 1.8
C4.3 -> 1.12.1
C5 -> { \emptyset }
C5.1 -> { \emptyset }
C5.2 -> 1.9.9
C5.3 -> 1.4.3
C5.3.1 -> { \emptyset }
C5.4 -> 1.2.1, 1.2.3
C5.5 -> 1.3.3.4, 1.3.3
C5.6 -> { \emptyset }

6. ISO-1951: Eksempel på monolingval ordbogsartikel i XML



1. <DictionaryEntry identifier='A050672' sourceLanguage='de'>

2. <HeadwordCtn>

3. <Headword>Fliege</Headword>

4. <Display>Fl<Stress type='long'>ie</Stress><Hyphen/>ge</Display>

5. <GrammaticalNote>die; -, -n </GrammaticalNote>

6. <Etymology>1- mhd. vliege, ahd. fliege, eigtl. = die Fliegende; 3- für frz. mouche</Etymology>

7. </HeadwordCtn>

8. <SenseGroup identifier='A050672-s1'>

9. <Definition><Optional>in zahlreichen Arten vorkommendes</Optional> gedrungenes, kleines Insekt mit zwei Flügeln u. kurzen Fühlern.</Definition>

10. <Example>eine dicke, zudringliche, lästige F.;</Example>

11. <Example>die <Inflection type='plural'><RepeatSymbol></RepeatSymbol>n</Inflection> summen, schwirren, setzen sich auf das Fleisch;</Example>

12. <Example>eine F. fangen, verjagen,

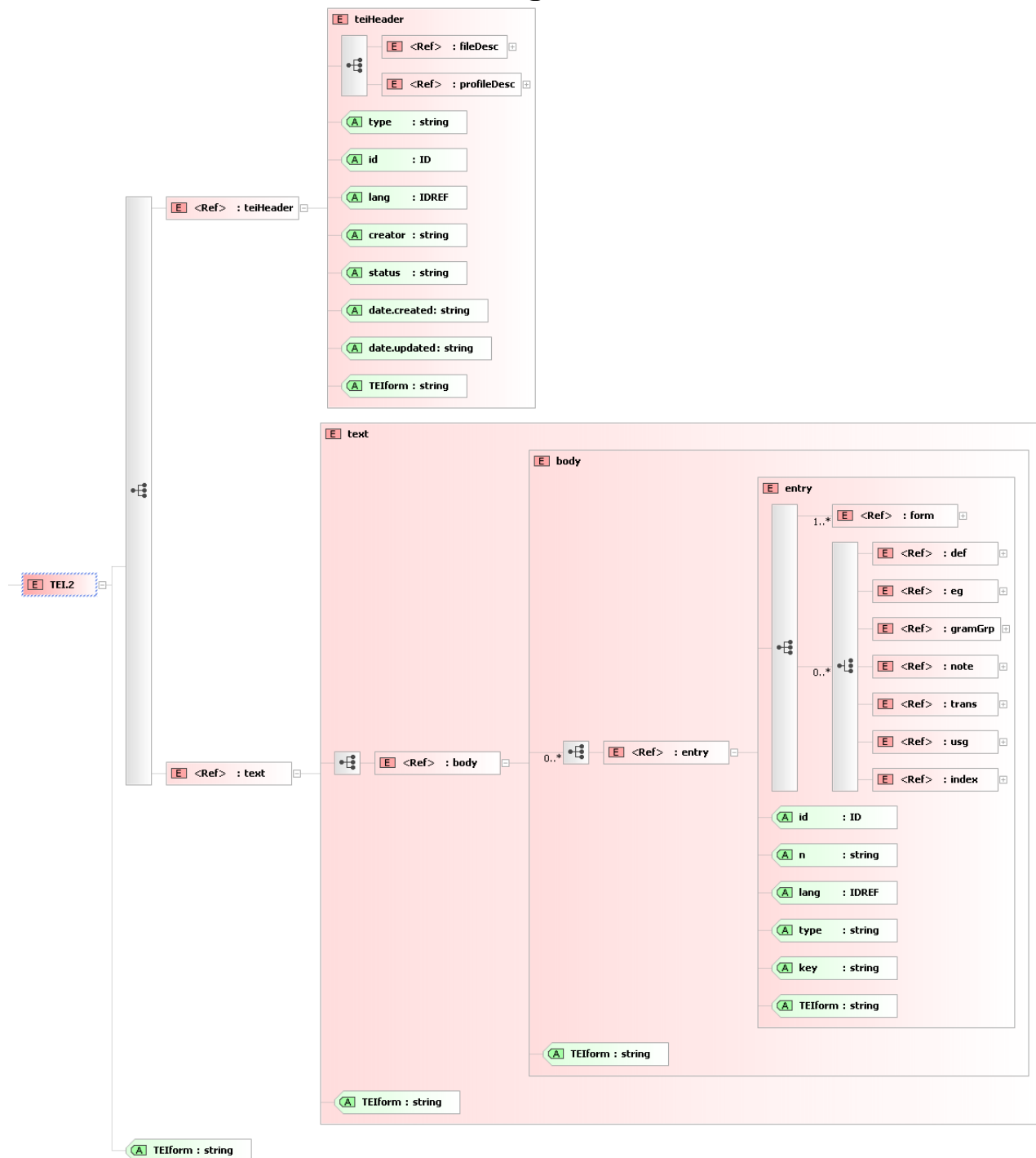
- totschlagen;</Example>
13. <Example>mit der [künstlichen]F. <Gloss>einer Nachbildung der Fliege)</Gloss>
angeln</Example>
14. <CompositionalPhraseCtn>
15. <CompositionalPhrase><Inflection type='plural'>zwei
<RepeatSymbol></RepeatSymbol>n</Inflection> mit einer Klappe
schlagen</CompositionalPhrase>
16. <Register freeValue='ugs.'/>
17. <Definition>einen doppelten Zweck auf einmal erreichen</Definition>
18. </CompositionalPhraseCtn>
19. <CompositionalPhraseCtn>
20. <CompositionalPhrase>eine, die F. machen </CompositionalPhrase>
21. <Etymology>nach dem raschen Davonfliegen der Fliegen</Etymology>
22. <Register freeValue='salopp'/>
23. <Definition>[schnell]weggehen;</Definition>
24. <CitationCtn>
25. <Citation>Da wird ... ein Hochschullehrer in jeder Vorlesung oder
Übung ... unter
Druck gesetzt und 'madig' gemacht - 'bis er ... an der Uni 'ne F.
macht'</Citation>
26. <title>Spiegel</title>
27. <LocationWithinHost>43, 226</LocationWithinHost>
28. <date>1977</date>
29. </CitationCtn>
30. </CompositionalPhraseCtn>
31. <CompositionalPhraseCtn>
32. <CompositionalPhrase>sich über die F. an der Wand
ärgern</CompositionalPhrase>
33. <Definition>(sich über jede Kleinigkeit ärgern);</Definition>
34. </CompositionalPhraseCtn>
35. <CompositionalPhraseCtn>
36. <CompositionalPhrase>jmdn. stört die F. an der Wand</CompositionalPhrase>
37. <Definition>(jmdn. stört jede Kleinigkeit);</Definition>
38. </CompositionalPhraseCtn>
39. <CompositionalPhraseCtn>
40. <CompositionalPhrase>umfallen wie die -n</CompositionalPhrase>
41. <Register freeValue='ugs.'/>
42. <Definition>in grosser Zahl sterben;</Definition>
43. </CompositionalPhraseCtn>
44. <CompositionalPhraseCtn>
45. <CompositionalPhrase>matt sein wie eine F. </CompositionalPhrase>
46. <Register freeValue='ugs.'/>
47. <Definition>sehr erschöpft sein</Definition>
48. </CompositionalPhraseCtn>
49. <CompositionalPhraseCtn>
50. <CompositionalPhrase>keiner F. etw. zuleide tun [kön-
nen]</CompositionalPhrase>
51. <Register freeValue='ugs.'/>
52. <Definition>sehr gutmütig sein u. niemandem etwas zuleide tun [kön-
nen]</Definition>
53. </CompositionalPhraseCtn>
54. </SenseGroup>
55. <SenseGroup identifier='A050672-s2'>
56. <Definition>als Querschleife gebundene Krawatte- </Definition>
57. <Example documentSize='5'>er trägt gern karierte od. gestreifte -n;</Example>
58. <Example>eine F. umbinden.</Example>
59. </SenseGroup>
60. <SenseGroup identifier='A050672-s3'>
61. <Definition>schmales, gestutztes Bärtchen auf der Oberlippe od. zwis-chen Unterlippe u.
Kinn.</Definition>
62. </SenseGroup>
63. <SenseGroup identifier='A050672-s4' documentSize='5'>
64. <Register freeValue='Schneiderei'/>

65. <Definition>gesticktes Dreieck zur Befestigung von Falten, Nähten od. Tascheneinschnitten.</Definition>

66. </SenseGroup>

67. </DictionaryEntry>

7. Skema for den nordiske netordbog



8. Den finske emnetaksonomi

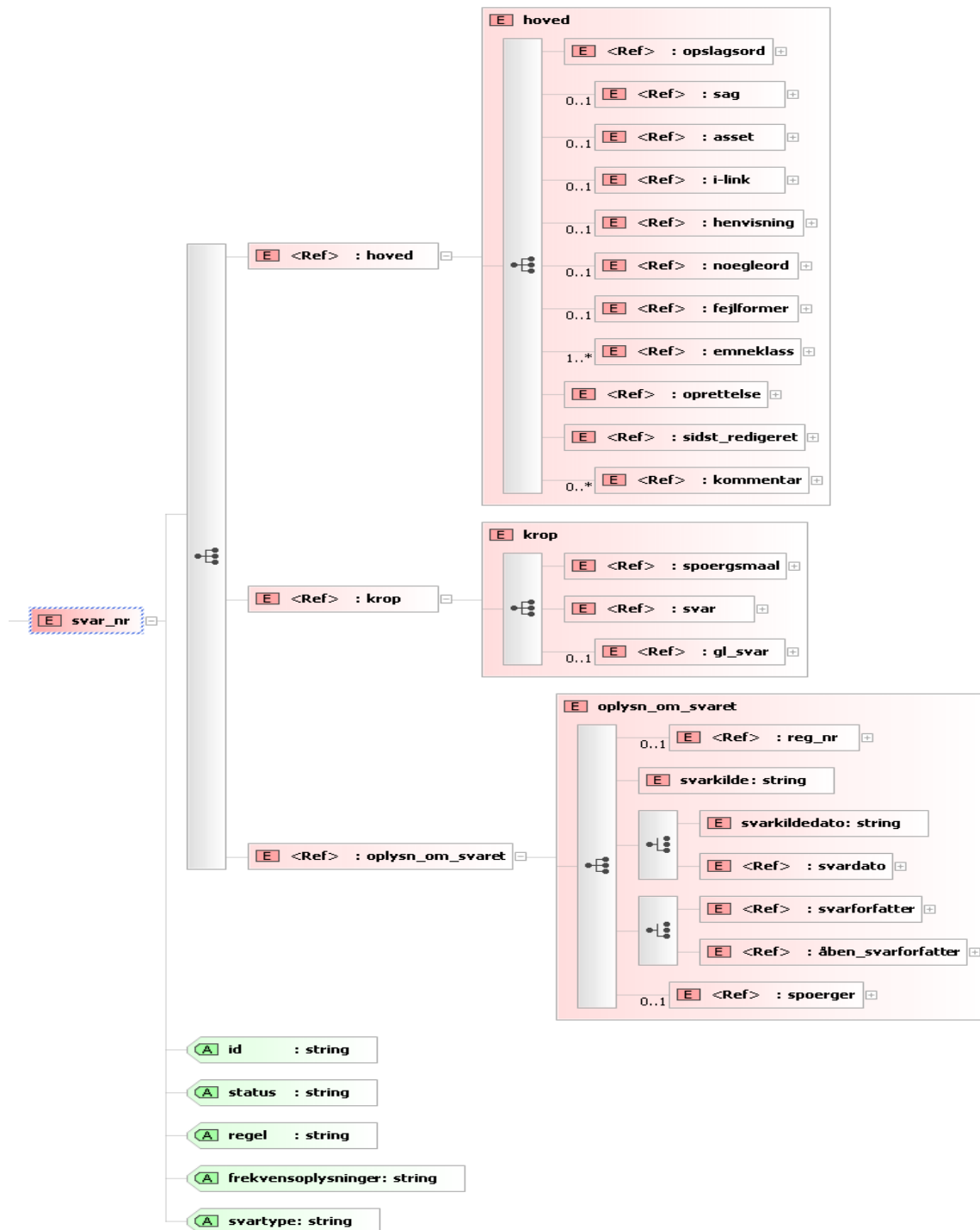
*1 adj. adjektiv

*2	adv.	adverb		
*3	alat.	alatyylissä, alatyylisesti	vulgär	
*4	amm.	ammattinharjoittajanimi, ammattinimike		facklig
*5	ammattinimekkeet	yrkesnamn		
6	anat.	anatomia, anatominen	anatomi	
7	antrop.	Antropologia	antropologi	
*8	ark.	arkikielessä, arkikielinen	vardaglig	
9	arkeol.	arkeologia	arkeologi	
10	arkkit.	arkitektur		
11	astrol.	astrologia, astrologinen	astrologi	
12	atk	tietokoneala, automaattinen tietojenkäsittely		informationsteknologi
13	biokem.	biokemia, biokemiallinen	biokemi	
14	biol.	biologia, biologinen	biologi	
15	diagnostiikka	diagnostik		
16	draama	(drama)		
17	el.	eläintiede	zoologi	
18	elekt.	elektroniikka	elektronik	
19	elok.	film		
20	eläinlääk.	eläinlääketiede	djurmedicin	
*21	erik.	i syfte som avviker från allmän språkbruk	subst.	substantiv
*22	erikoiskielet	fackspråk		
*23	erisn.	Erisnimi	person- eller platsnamn	
*24	etunimet	förmamn		
*25	etymologia	etymologi		
26	EU	Euroopan unioniin liittyvää sanastoa (ei sanakirjalyhenne)		EU
27	farm.	farmakologia, farmasia	farmakologi	
28	fil.	filosofia, filosofinen	filosofi	
29	filat.	filatelia	filateli	
*30	fon.	fonetiikka, foneettinen	fonetik	
*31	fraasi	fras		
*32	fraseologia	fraseologi		
*33	fras.	fraseologinen ilmaus (ei sanakirjalyhenne)	fras	
34	fys.	fysiikka, fysikaalinen	fysik	
35	fysiol.	fysiologia, fysiologinen	fysiologi	
36	geofys.	geofysiikka	geofysik	
37	geol.	geologia, geologinen	geologi	
38	geom.	geometria, geometrinen	geometri	
39	hall.	offentliga sammanhang		
40	halv.	halventava, halventavasti	förolämpande	
41	hammaslääketiede	tandvård		
42	*heng.	käytä lyhennettä usk.; vanhoissa sanalipuissa: ...		andlig
43	hist.	historia, historian tutkimus, historiallinen		histori
44	hoit.	hoitoala	vårdfack	
45	ilm.	ilmailu	aviation	
46	ilmat.	Ilmatiede	meteorologi	
47	iron.	ironinen	ironisk	
48	jalok.	jalokiviala	smycken (juvel)	
49	jap.	japansk		
50	juhlapäivien nimet	namn för högtidsdagar		
51	kaavanimet	plannamn		
52	kal.	kalastus	fiske	
53	kalevala	forddikter		
54	kansanperinne	folktradition		
55	kansanrunous	folkdikt		
56	kansankielinen	folklig		
57	kansat.	kansatiede	etnologi	
58	kasvatustiede/kasvatust.	pedagogik		
59	kasv.	kasvitiede	botanik	
60	kauppa	handel		
61	kat.	katolisk		
62	kem.	kemia, kemiallinen	kemi	
*63	keskusteluanalyysi	konversationsanalys		

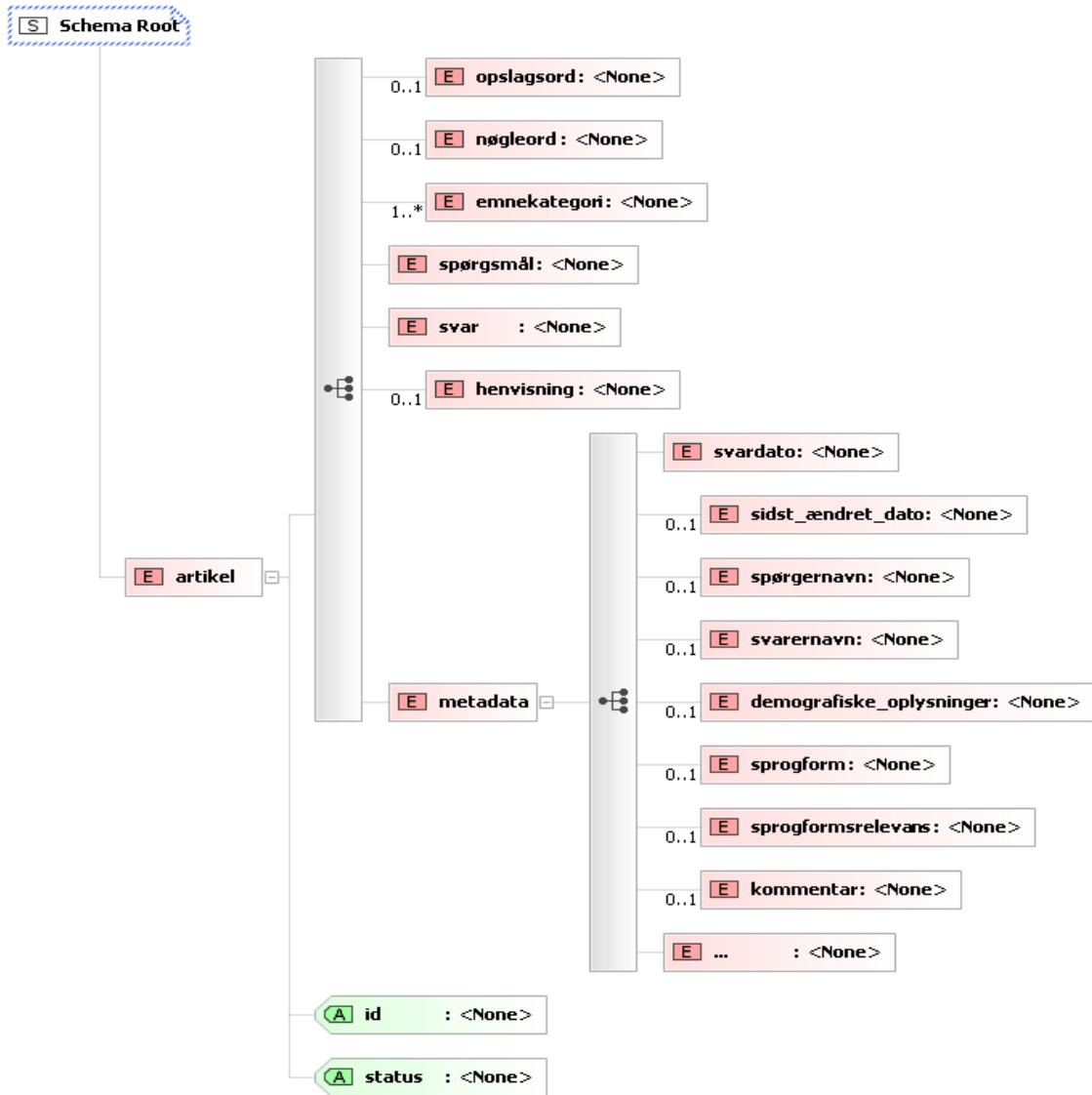
*64	kiel.	kielitiede, -oppi	språkvetenskap	
*65	kielipolitiikka	språkpolitik		
*66	kielitoimiston suosituks	rekommendation		
*67	kieltennimet	namn på språk		
68	kirj.	litteratur		
69	kirjall. tiede	litteraturvetenskap		
70	kirjap.	kirjapainoala	tryckeri	
71	korttip.	korttipelit	kortspel	
72	kosmet.	kosmetiikka	kosmetik	
73	kotal.	hushållning		
74	koul.	skolelevspråk		
75	koulu	skolväsen		
76	kultt.	kultur		
*77	kuv.	kuvakielessä, kuvallisesti	bildlig	
78	kuvat.	kuvataiteet	konst	
79	käs.	käsityöt	handarbete (textil)	
*80	kääntäminen	översättning		
*81	laitosten nimet	namn på instituter		
82	lakikieli	lagspråk		
*83	last.	lastenkielessä, lapsille puhuttavassa kielessä	barnspråk	
*84	lehtikieli	tidningsspråk		
85	leik.	leikillinen, leikillisesti	skojande	
86	liik.	käytä lyhennettä tal.; vanhoissa sanalipuissa: liikeala...	ekonomi	
87	log.	logiikka	logik	
*88	lyh.	lyhenne	förkortning	
*89	lyriikka	diktspråk		
90	lääk.	lääketeide	medicin	
91	maal.	maalausala	måleri	
92	maanm.	maanmittaus	geodetisk	
93	maant.	maantiede	geografi	
94	maat.	maatalous	jordbruk	
*95	mainoskieli	reklamspråk		
96	mat.	matematiikka	matematik	
97	matkaviestinala	mobilmäsen		
98	mer.	merenkulku	nautica	
99	met.	metalliala	metallurgi	
*100	metafora	metafor		
101	mets.	metsästys	jakt	
102	metsät.	metsätalous, metsätiede	forstbruk	
103	miner.	mineralogia	mineralogi	
*104	mon.	monikko, monikossa	pluralis	
*105	murt.	murteissa, murteellinen	folkmål	
106	mus.	musiikki	musik	
107	myt.	mytologia	mytologi	
108	nimistö	namnskick		
*109	oik.	rättskrivning		
110	oik.	oikeustiede, oikeusala	juridik	
111	opet.	opetusalan sanastoa (ei sanakirjalyhenne)	pedagogik	
112	opt.	optiikka	optik	
113	ort.	ortodox		
114	paikannimet	platsnamn		
115	paleont.	paleontologia	paleontologi	
116	palo- ja pelastustoimi	brand- och räddningsväsen		
117	pap.	paperi- ja selluteollisuus	pappersindustri	
118	pol.	politiikka, poliittinen	politik	
119	psyk.	psykologia, psykologinen	psykologi	
120	purjehdus	segling		
121	puut.	puuteollisuus	skogsindustri	
122	raam.	Raamatussa, Raamatun kielessä, raamatullinen	biblick	
123	rak.	rakennusala, rakennustaide	byggfack	
124	rautat./rautatie	rautatieala	järnväg	
125	rav.	näringsämnen		

126	retkeilysanasto	vandring		
127	run.	runousopissa	aristotelisk	estetik
128	runok.	runokielessä	diktlig	
129	ruok.	ruokatalous, ruoka-ala		matterminologi
130	ruotsinsuomi	sverigefinska		
*131	sanamuodostus	morfologi		
*132	sanasto	vokabulär		
*133	semantiikka	semantik		
*134	slangi	slang		
*135	slg.	slangissa	slang	
136	sos.	sosiaali- (ei käyttöalalyhenteenä sanakirjassa)		socialfack
137	sosiol.	sosiologia, sosiologinen	sociologi	
138	sot.	sotilaskielessä, sotilasala	militär	
*139	sukunimet	efternamn		
140	sukututkimus	genealogi		
*141	suositukset	rekommendation		
142	sähkötekn./sähk.	sähköala, sähkötekniikka		elteknik
143	taide	konst		
144	taidehistoria/taidehist.	taidehistoria		konsthistori
145	tal.	taloustiede, -elämä	ekonomi	
146	tav.	tavallinen, tavallisesti	vanligen	
147	tekn.	tekniikka, tekninen	teknik	
148	tekst.	tekstiiliala (sic)	textilfack	
149	tekstiilisana	textilfack		
*150	tekstintutkimus	textforskning		
151	teletekn.	teletekniikka	teleteknik	
152	teol.	teologia	teologi	
153	terv.	folkhälsa		
*154	tervehdykset	hälsningar		
155	tied.	naturvetenskap		
156	tiet.	informationsteknologi		
157	tilap.	tillfällig		
158	tilastokieli/tilastot.	tilastotiede		statistik
159	tutkimus	forskning		
160	täht.	tähtitiede	astronomi	
161	urh.	urheilu	sport	
162	usk.	uskonto, uskonnollisessa kielenkäytössä		religion
163	vaat.	vaatetusalan sanastoa (ei sanakirjalyhenne)		kläder, mode
164	valok.	valokuvaus	fotografi	
165	vanh.	vanhentunut	föråldrad	
*166	v.	verb		
167	valuutta	valutanamn		
168	viestintä	kommunikation		
169	virakieli	byrokrati		
170	voim.	voimistelu	gymnastik	
*171	vs.	vierassana (ei sanakirjalyhenne)	lånord	
*172	ya.	yhdysadjektiivi (ei sanakirjalyhenne)		sammansatt adjektiv
*173	yadv.	yhdysadverbi (ei sanakirjalyhenne)		sammansatt adverb
174	yht.	samhälle		
*175	yks.	singularis		
*176	yleisk.	yleiskielessä	allmän språkbruk	
*177	ylät.	ylätyylissä, ylätyylisesti	högspråk	
178	ymp.	ympäristö(nsuojele)sanastoa (ei sanakirjalyhenne)		miljö, naturskydd
*179	ys.	sammansättning		
*180	yv.	yhdysverbi (ei sanakirjalyhenne)		sammansatt verb
*181	ääntäminen	uttal		

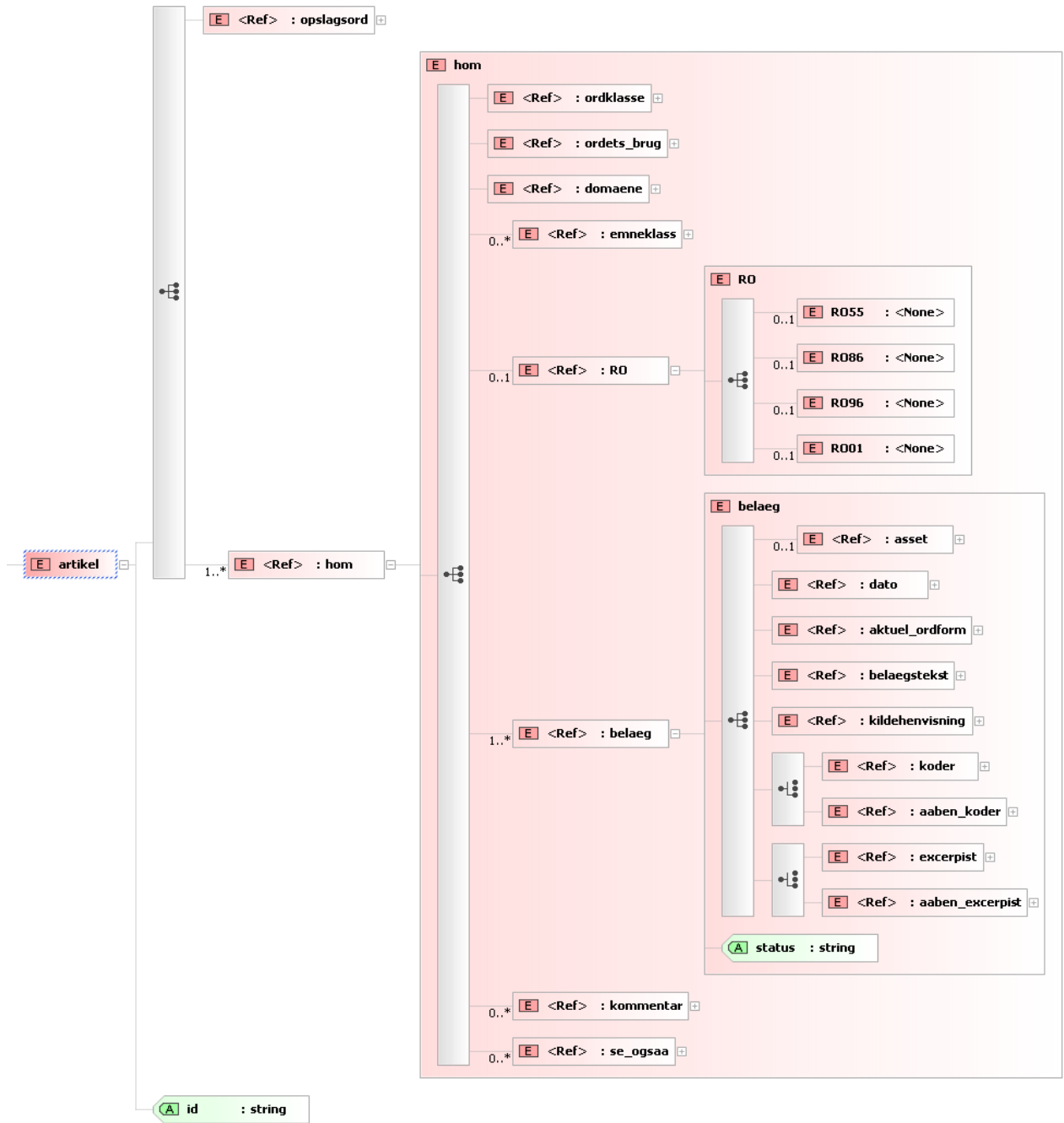
9. Den danske svarbases struktur



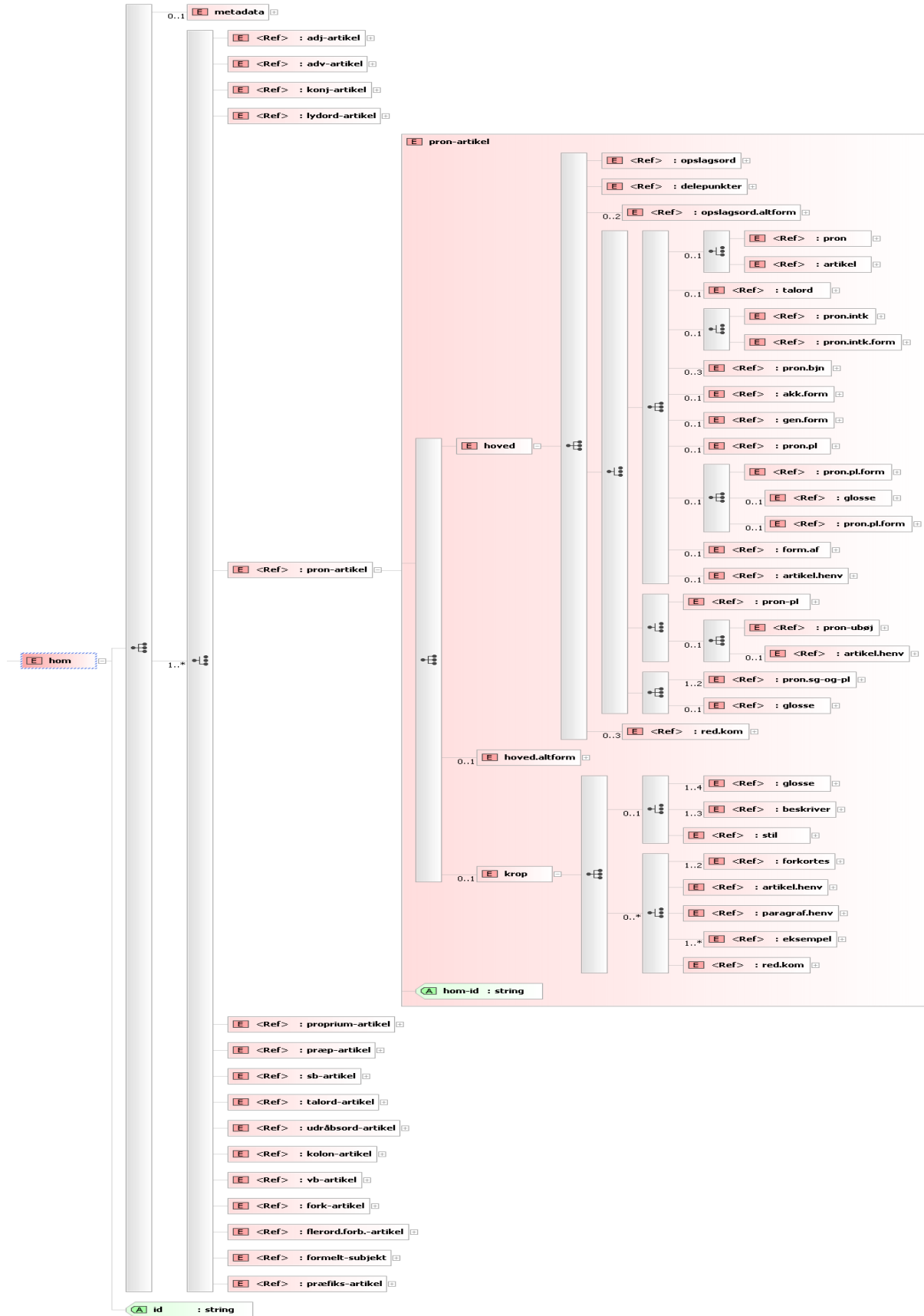
10. Udkast til fællesnordisk svarbasestruktur



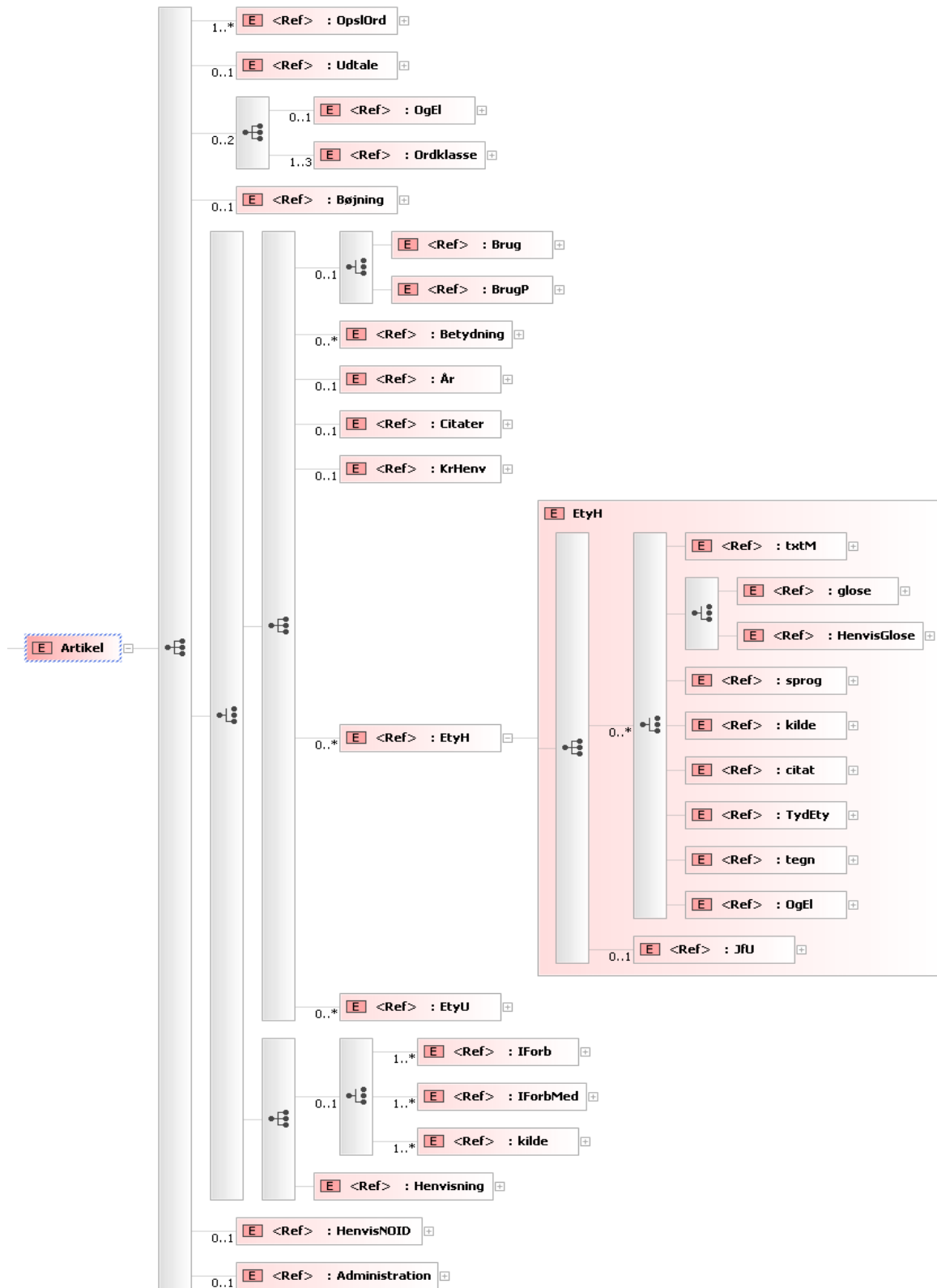
11. Den danske ordbases struktur



12. Den danske Retskrivningsordbogs struktur



13. Den danske nyordsordbogs struktur



14. Sammenligning af oplysningstyper i de nordiske ordbaser

	Sverige (Netord bogen)	Sverige (Lexin)	Danmark (RO)	Danmark (NOID+ ordbasen)	Sverige (Språklåda n)	Finland (nyordsbase)
teiHeader/fileDesc/titleStmt/title	X					
../author	X		X metadata/ medarb.	X koder/ excerpist	X inskrivare	
../principal	X					
teiHeader/fileDesc/publicationStmt/availability	X					
teiHeader/fileDesc/sourceDesc/bibl	X					
teiHeader/profileDesc/langUsage/language	X			X EtyU/sprog		
body/entry	X		X hom	X artikel/hom		
../form/ptr ../form/xptr	X		X refer ...	X Kilde KrHenv/Refer	X källa	X Kilde
../form/orth	X	X	X opslagsord	X OpslOrd	X ord/fras	X Søgeord
../form/hyph	X		X delepunkter		X avstavning	
../form/pron	X	X		X Udtale	X uttal	
../form/usg	X			X BrugP ordets_brug		
../gramGrp	X	X		X BjnForm	X ordbildning böjning	
../gramGrp/pos	X	X	X {sb,vb,adv ...}	X Ordklasse	X ordklass	

../gramGrp/gen	X		X {itk, itk.ds. ...}			
../gramGrp/mood	X		X {imp.for m ...}			
../gramGrp/number	X	X	X {ental.bes t ...}			
../gramGrp/tns	X		X {præs,pr æt ...}			
../gramGrp/per	X					
../gramGrp/subc	X					
../gramGrp/case	X					
../trans	X	X				
../def	X	X	X glosse	X Betydning/ Tyd	X betydelse	
../eg	X		X eksempel	X citat/txtD belegstekst	X text	X eksempel
../note	X					
../index	X					
Dato				X år/dato	X årtal/datum	
Publiceringsstatus					X	
Etymologi				X		
Kategori				X	X	
”Asset”				X	(X) uttalsfil	

15. To eksempler på native XML-databasesystemer

The screenshot shows the eXist Admin Client interface. The main window displays a table of resources, and a 'Query Dialog' window is open in the foreground. The query dialog shows a query input and its results.

Query Dialog

Query Input:

```
History: 1. for $o in collection("/db/svarbase")//ops
<opslagsord_i_begge_baser>.
{
  for $artikel in collection("/db/svarbase")//svar_nr.
  for $artikel2 in collection("/db/ordbase")//artikel.
  where $artikel//opslagsord = $artikel2//opslagsord and matches($artikel//opslagsord, '^t', 'i').
  return .
  <opslagsord>(data($artikel//opslagsord)) ((data($artikel/@id))</opslagsord>.
}
</opslagsord_i_begge_baser>.
```

Context: /db/ordbase

Results:

```
<opslagsord_i_begge_baser>.
<opslagsord>test (SV000000038)</opslagsord>.
<opslagsord>test (SV00001466)</opslagsord>.
</opslagsord_i_begge_baser>.
```

The screenshot shows the iLEX v 1.1 interface. The main window displays a document view, and a search panel is visible on the right side.

iLEX v 1.1 137s Licenseret til Dansk Sprogævn. Copyright (c) 2004-2008 EMP ApS

Projektpanel: svarbase

Opslag Design

"bedler" og "belder"

(at) personlige årsager/personlige årsager
(fælles) fodslag
&endash
101 er ude
10.-Klasse-Center
10 min. i halv tre/20 min. over 2
15 timer el. timers valgfag
1901'erne/enerne
½ gang mere
2-4000 kr, Bindestreg for fælles
24-7
24-7
3. a/3. A + genitiv
4% per gallon: kontamination
40 billister fanges der hver dag/4
600-MHz-Pentium III-processor,
60'ersound/60'er-sound
af/afaf
abbelat; appelat; abelat; ablat; o
abbelat/voblat
abekat

Opgavepanel

Søgepanel

Søg efter

<opslagsord>*

Resultat

Antal dokumenter: 740

10 min. i halv tre/20 min. over 2
15 timer el. timers valgfag
accenttegn til at angive tryk
afledning til silicium, silicificeret
Algier, Tunis
altan/terrasse
alternativ: om mere end to mulige
alternativ: om mere end to mulige

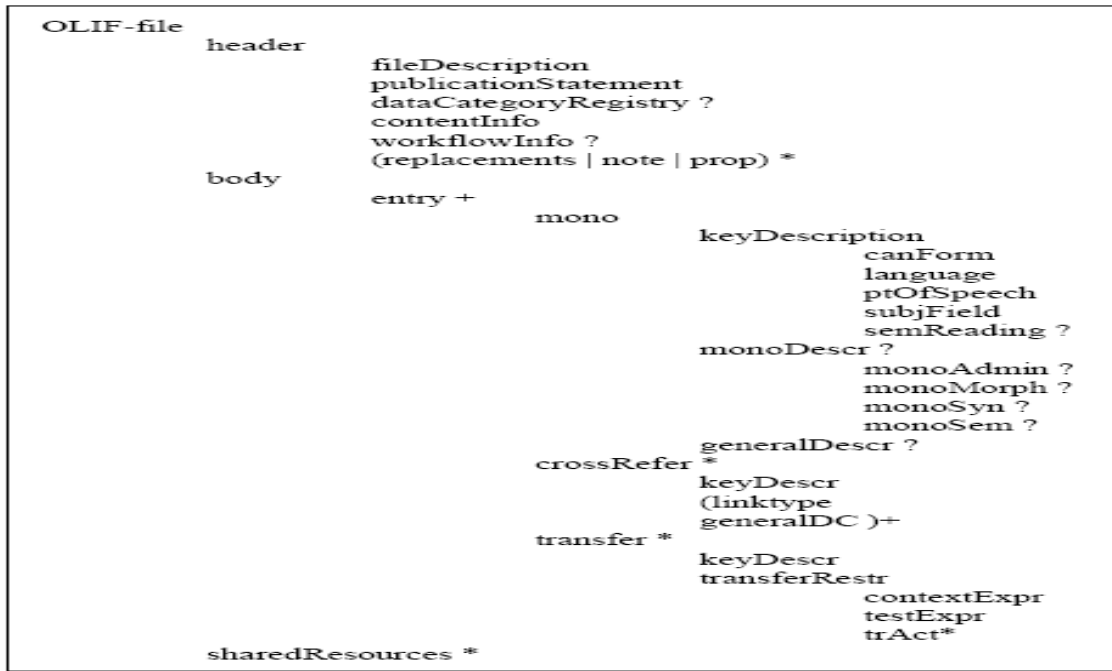
Document View:

"bedler" og "belder" / svarbase (Dokument)

```
<svar_nr frekvensoplysninger="nej" id="SV00001049" status="offentlig" svartype="NFS-svar">
<hoved>
<opslagsord>bedler" og "belder"
<noegleord>billede, billeder, måned, måneder
<emneklasse>1.10.2 fonetik udtale
<oprettelse>
<dato>27.04.2007
<redaktoer>EMTK
<sidst_redigeret>
<dato>06.06.2007
<redaktoer>IEM
<krop>
<spoergsmaal>
<afsnit>Hvorfors bytter så mange - især unge - danskere om på
<materialer>L og
```

ResId: SV00001049

16. En OLIF-fils struktur



17. Eksempel på bearbejdet nyordsexcerpt i LMF

```
<?xml version="1.0" encoding="UTF-8" ?>
- <LexicalResource dtdVersion="14" xmlns="http://www.w3.org/namespace/">
- <GlobalInformation>
  <feat att="approverResponsibility" val="JH" />
</GlobalInformation>
- <Lexicon>
  <feat att="language" val="da" />
- <LexicalEntry>
  <feat att="origination_Date" val="2008-04-17" />
  - <Lemma>
    <feat att="partOfSpeech" val="commonNoun" />
    <feat att="grammaticalGender" val="faelleskoen" />
    <feat att="writtenForm" val="gyllebaron" />
    - <FormRepresentation>
      <feat att="geographicalVariant" val="Fyn" />
      <feat att="writtenForm" val="blah blah" />
    </FormRepresentation>
  </Lemma>
  - <WordForm>
    <feat att="lexicalType" val="inflection" />
    <feat att="grammaticalNumber" val="plural" />
    <feat att="writtenForm" val="gyllebaroner" />
  </WordForm>
  <Sense id="gyllebaron1">
    <feat att="brug" val="ny" />
    <feat att="domaene" val="landbrug" />
    - <Context>
      <feat att="text" val="Jeg kan ikke forstå, at disse 'gyllebaroner' stadig kan opføre tusinder og atter tusinder af kvadratmeter, (...)" />
    </Context>
    - <Definition>
      <feat att="elementWorkingStatus" val="working" />
      <feat att="text" val="Nedsættende udtryk om en landmand med omfattende investeringer i svineindustrien" />
    </Definition>
    - <MonolingualExternalRef>
      <feat att="genre" val="Kultur_avis" />
      <feat att="kilde" val="Politiken" />
      <feat att="sektion" val="2" />
      <feat att="side" val="1" />
      <feat att="spalte" val="1" />
      <feat att="forfatter" val="Karsten R. S. Iversen" />
    </MonolingualExternalRef>
  </Sense>
</LexicalEntry>
</Lexicon>
</LexicalResource>
```

18. Exempel på bilingval ordbogsartikel i det nordiske netordbogsformat

```
SV_netordbog_exempel.xml
- <TEI.2>
- <teiHeader type="dictionary" date.created="2005-09-19" date.updated="2005-11-08">
- <fileDesc>
- <titleStmt>
  <title>Lexin svensk-engelsk ordbok</title>
  <author>Myndigheten för skolutveckling, Sverige</author>
  <principal>Viggo Kann</principal>
</titleStmt>
- <publicationStmt>
- <availability>
  <p>Muntligt tillstånd till användning</p>
</availability>
</publicationStmt>
- <sourceDesc>
- <bibl>http://lexin.nada.kth.se/sve-eng.html</bibl>
</sourceDesc>
</fileDesc>
- <profileDesc>
- <langUsage>
  <language id="sv" usage="source">Swedish</language>
  <language id="en" usage="target">English</language>
</langUsage>
</profileDesc>
</teiHeader>
- <text>
- <body>
- <entry>
- <form>
  <orth>jätte</orth>
</form>
- <gramGrp>
  <pos>noun</pos>
</gramGrp>
  <trans lang="en">giant</trans>
  <index>jätten</index>
  <index>jättar</index>
</entry>
- <entry>
- <form type="compound">
  <orth>atlantjätte</orth>
  <hyph>atlant|jätte</hyph>
</form>
- <gramGrp>
  <pos>noun</pos>
</gramGrp>
  <trans lang="en">big Atlantic liner</trans>
</entry>
```