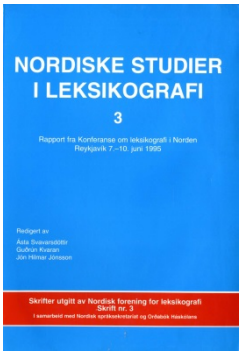
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Arbejdet med "Forslag om dansk standard for lagring og udveksling af leksikalske data"

Forfatter: Anna Braasch

Kilde: Nordiske Studier i Leksikografi 3, 1995, s. 69-81
Rapport fra Konferanse om leksikografi i Norden, Reykjavík 7.-10. juni 1995

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



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

Anna Braasch

Arbejdet med „Forslag om dansk standard for lagring og udveksling af leksikalske data“

This paper gives a brief report on the work done so far within the ongoing STANLEX project which is aiming at a proposal for standardization of storage and exchange of lexical data. The STANLEX group is affiliated to the Danish Standard Association. The current goal is to develop a clearly defined format (comprising both structure and content aspects) to support efficient sharing of machine readable dictionary data. Also the reusability and multifunctionality aspects of these resources will be strengthened by means of a general taxonomy covering both printed dictionaries and current requirements of lexicon modules integrated with natural language processing systems.

1 Indledning

Sprog i almindelighed — forståelse og håndtering af information spiller en stadig større rolle i det internationaliserede samfund, vi lever i. Der er stadig større behov for hurtig og præcis tekstbehandling, herunder produktion af tekster, oversættelse osv. „I den situation kan sprogteknologiske hjælpemidler få stor betydning“, fastslås det i en rapport om „Sprogteknologi“ fra Teknologinævnet. Computeren har for længst vundet indpas i al slags arbejde med tekster.

Sprogteknologiske produkter er bl.a. elektroniske ordbøger, termbaser, systemer for stave-, grammatik- og stilkontrol, samt systemer for maskinstøttet oversættelse og maskinel oversættelse. Disse produkter bruges, når man arbejder med sprog på elektronisk medium: fx skriver sine tekster på computer, anvender orddeling, stavekontrol eller slår op i en elektronisk ordbog, når man skriver på fremmedsprog.

Dansk er et lille sprogområde, og det betyder i mange henseender, at udvikling af sprogteknologiske produkter vil halte bagefter de store sprogs udvikling, om ikke andet så af økonomiske grunde. Det er dyrt at udvikle sådanne hjælpemidler, derfor er det vigtigt at økonomisere med de ressourcer, landet har.

2 Teknologinævnet

Den før nævnte pjece er udarbejdet af Teknologinævnet, som er en offentlig dansk organisation under Forskningsministeriet, og som iværksætter helhedsvurderinger af den teknologiske udviklings muligheder og konsekvenser.

Teknologinævnet har i 1993 nedsat et tværfagligt udvalg bestående af eksperter og repræsentanter for erhvervslivet (dvs. producenter og nuværende samt potentielle brugere af sprogteknologiske produkter) til at undersøge sprogteknologiens rolle i erhvervsmæssig tekstproduktion og oversættelsesarbejde.

Nordiske studier i leksikografi III, 69–81

Undersøgelsens resultater er beskrevet i en situationsrapport om anvendelsen af eksisterende sprogteknologiske værktøjer for dansk og vurderer, hvilke strategier der gør det muligt at komme på forkant med udviklingen med henblik på at opfylde fremtidige behov.

3 Teknologinævnets rapport

I rapporten *Dansk sprogteknologi — status, perspektiver og handlemuligheder* (Bech m.fl. 1994) fremsættes anbefalinger vedr. en koordineret dansk handlingsplan for fremtidige aktioner inden for det sprogteknologiske område, og der peges på fire særlige satsningsområder:

- Sprogteknologiske demonstrationsprojekter (afprøvning og systematisk evaluering)
- Etablering af en fælles standard for lagring og udveksling af leksikalske data (ved arbejdsgruppen *STANLEX*[1])
- Oprettelse af en stor dansk maskinbrugbar ordbog
- Oprettelse af netværk af danske termbanker

Det er åbenlyst, at især punkterne 2 og 3 kan være af interesse for leksikografer. Dette indlæg fokuserer på punkt 2.

Arbejdet med etableringen af en fælles standard for lagring og udveksling af leksikalske data er startet i september 1994 og forventes afsluttet ved omkring årsskiftet 1995/96. Vi er i skrivende stund (sommeren 1995) endnu ikke nået frem til et konkret, afrundet forslag om en sådan standard, derfor kan der her kun gives en redegørelse for selve målsætningen og den proces, arbejdsgruppen STANLEX er i gang med.

Alle leksikografer ved, hvor arbejdskrævende det er at lave ordbøger af høj kvalitet, og at materialet fra gode ordbøger bliver brugt igen og igen som grundlag eller kilde for nye slags udgaver, reviderede udgaver, som model for andre ordbøger osv. Dette er en form for genbrug af det arbejde, den viden og erfaring, der er opsamlet i en enkelt, stor ordbog, som måske er blevet til gennem indsats fra flere generationer af leksikografer.

I den efterfølgende fremstilling nævnes forskellige slags ordbøger. Nedenfor beskrives ganske kort, hvad de enkelte betegnelser står for.

Humanordbøger:
Fokus på *bruger*: ordbøger beregnet for mennesker
Medium: traditionelt tryk; i dag også elektronisk udgave

Maskinlæsbare ordbøger:
Fokus på *medium*: Ordbogsdata fremstillet og/eller lagret på elektronisk medium

[1] *STANLEX gruppens medlemmer (i alfabetisk orden):* Anna Braasch (CST; Terminologigruppen), Ib Elfving (Info:Partner A/S), Gert Engel (HHS, Institut for Erhvervsforskning), Hanne Jensen (IBM; efter d. 1.4.95 Munksgaards Forlag), Claus Bo Jørgensen (Textware), Bente Maegaard (CST), Bodil Nistrup Madsen (Kontaktperson; HHK, Institut for Datalingvistik; Terminologigruppen), Ole Norling-Christensen (Den Danske Ordbog), Hanne Ruus (KUA, Institut for Nordisk Filologi), Ragnar Heldt Nielsen (Teknologinævnet), Klaus Søndergaard (Dansk Standard), eksterne medlemmer.

Maskinbrugbare ordbøger:

Fokus på *bruger*: leksikalske datasamlinger beregnet til anvendelse i sprogteknologiske produkter

Termsamlinger:

Fokus på *indhold*: leksikografisk og faglig beskrivelse af termer

Medium: i denne sammenhæng fortrinsvis terminologi i databaseform

Bruger: udformet principielt til både mennesker og maskiner

Denne opstilling kan på ingen måde betragtes som typologi, den er snarere en ad hoc liste over de eksempler på leksikografiske produkter, som indgår i STANLEX-gruppens arbejdsgrundlag.

Sprogteknologiske produkter (fx et oversættelsessystem eller en såkaldt forfatterhjælp) inkluderer altid en eller anden form for maskinbrugbar ordbog, som et af systemets centrale moduler.

Sprogets hastige udvikling — måske især på det terminologiske område — kræver, at ordbogsarbejdet følger med. I fundamentet for maskinel sprogbehandling indgår også store basisordbøger for moderne almensprog, og da det er meget tidskrævende at udarbejde sådanne ordbøger, er det ønskeligt at kunne drage nytte af de eksisterende traditionelle humanordbøger (Braasch 1994a).

Den krævende systematik, der kendetegner maskinbrugbare ordbøger kan på den anden side virke gavnligt — eller bevidstgørende — ved udarbejdelse af ordbøger for mennesker (Braasch 1994b). Der åbnes nye specielle perspektiver i leksikografien, som også støttes af den moderne teknologi.

Der stilles store krav til den moderne leksikografi, både hvad oplysningsmængde og arbejdets kvalitet og hurtighed angår. Det er vigtigt, at kravene til det færdige produkt beskrives på et så tidligt tidspunkt i et leksikografisk projekt som overhovedet muligt. En dansk standard for lagring og udveksling af leksikalske data vil også kunne bidrage til hensigtsmæssig planlægning af nye typer ordbogsopgaver.

4 STANLEX arbejdsgruppen

På Teknologinævnets initiativ er der blevet nedsat en arbejdsgruppe, der fik navnet **STANLEX**, dels fordi gruppen skal beskæftige sig med **stan**dardisering af **lex**ikalske oplysninger, dels også fordi gruppen opfatter sig som en slags fortsættelse af DANLEX-gruppen, som udførte pionerarbejde inden for analyse og klassificering af ordbogsoplysninger til elektronisk behandling af ordbogsdata.

Arbejdsgruppen omfatter repræsentanter for institutioner, herunder også Center for Sprogteknologi (CST), virksomheder og organisationer, der har viden om og interesse i sprogteknologi, datamatisk leksikografi og terminografi samt standardisering. Arbejdsgruppen er således meget bredt sammensat. Projektet er placeret hos Dansk Standard, under standardiseringsudvalget for informationsteknologi, under udvalget for teksthåndtering.

5 Målsætningen

Hovedopgaven er at udarbejde „et forslag til indhold i en dansk standard for indholds- og strukturbeskrivelse af leksikalske oplysninger". Formålet med en fælles standard er „at lette genbrug og udveksling af leksikalske oplysninger" (citater fra intern målsætningserklæring). Da ordbogs- og terminologiarbejde er meget ressourcekrævende, kunne der opnås store besparelser ved at benytte ensartede metoder ved indholds- og strukturbeskrivelse af leksikalske data, der fx kan tænkes at indgå i sprogteknologiske værktøjer.

I denne sammenhæng er det måske på sin plads at give en kort definition af, hvad der her betegnes som **leksikalske data**. Vi har tidligere nævnt de ordbogstyper, som STANLEX-gruppen primært arbejder med. Leksikalske data er ganske enkelt elementer eller helheder i forskellige slags beskrivelser af „ord" — eller opslagsenheder — indeholdt i disse ordbogstyper eller leksikalske datasamlinger.

I det efterfølgende redegøres for de udførte arbejdsopgaver og delresultater der er opnået inden den 1. maj 1995, og som kan anses for at være væsentlige led i udarbejdelsen af et forslag om standardisering af leksikalske data.

Vi har i projektet indtil nu koncentreret os om at indsamle, sammenligne og beskrive eksempler på tilgængelige maskinlæsbare ordbogsdata ud fra indholds- og strukturmæssige kriterier, primært på grundlag af **DANLEX-taksonomi** (DANLEX 1987). Taksonomien er præsenteret i tabelform i tabel 1.

6 Arbejdsproces og status

6.1 Arbejdsprocessen

En grundlæggende erfaring er, at datalagring i forskellige systemer besværliggør dataudvekslingen; med andre ord for at kunne sikre problemfri dataudveksling er det nødvendigt at nå til enighed om formater, data er beskrevet i, og måder, de bliver lagret på. Dette gælder naturligvis også leksikalske datasamlinger. På dette grundlag arbejder vi på et forslag til harmonisering af beskrivelses- og lagringsmetoder.

Arbejdsprocessen er planlagt til at bestå af to hovedfaser:

- I den første fase behandles projektdeltagernes leksikalske datasamlinger enkeltvis og i sammenhæng: der foretages indholdsanalyse og -beskrivelse med henblik på at nå frem til en fælles klassifikation, der kan dække alle oplysningtyper der forekommer i datasamlingerne.

- I den anden fase arbejdes med en SGML-baseret (Standard Generalized Markup Language) strukturbeskrivelse, der skal føre frem til forslag til standardformat(er) til lagring og udveksling af data.

Målet er at producere en teknisk rapport med vejledning om muligheder og fordele ved udveksling og genbrug af data, samt et egentligt forslag til standardiserede lagrings- og udvekslingsformater.

Hovedkategorier	Kategorier	Subkategorier
etymologiske oplysninger	• parallel • oprindelse • datering	
fonetiske oplysninger	• prosodiske træk • segmentale træk	
grafiske oplysninger	• ortografiske opl.	• stavning • orddeling
	• grafisk symbol	
grammatiske oplysninger	• ordklasse	
	• bøjningsopl.	• paradigmeopl. • bøjningsform
	• orddannelsesopl.	
	• syntaktiske opl.	• valens • syntaktisk funktion
pragmatiske oplysninger	• tekstlige opl.	• citat • mulig kontekst
	• brugsopl.	• tidslig dimension • rumlig dimension • social dimension • frekvens
	• ekstern henvisning	• litteraturhenvisning • kildehenvisning
	• evalueringsopl.	
	• administrative opl.	• opl. om indsamling og bearbejdning af data • intern henvisning • homografnummer • tekniske oplysninger
semantiske oplysninger	• emneklassificerende opl.	
	• oplysninger om semantiske relationer	• generisk over/underordningsrelation • partitiv relation • successiv relation • kausalrelation • antonymi
	• indholdsspecificerende oplysninger	• leksikalsk parafrase • analytisk/syntetisk definition • denotativ definition • ostensiv definition • faglig forklaring • semantiske træk
	• ækvivalensopl.	• ækvivalensrelation inden for ét sprog • ækvivalensrelation mellem to eller flere sprog

Tabel 1: *DANLEX-taksonomien, generelle oplysningstyper.*

6.2 Udgangspunkt

Vi tager udgangspunkt i den tidligere nævnte DANLEX-taksonomi (tabel 1), der er opstillet af en gruppe forskere og leksikografer. Den grundlæggende undersøgelse omfattede en bred vifte af traditionelle leksikografiske opslagsværker, men maskinbrugbare ordbøger blev — naturligt nok — kun perifert nævnt. Siden dette arbejde er afsluttet (1986), har udviklingen af sprogteknologiske værktøjer medført et stadig stigende behov for maskinbrugbare ordbøger.

Derfor indgår nu også denne type ordbøger i STANLEX-gruppens arbejdsgrundlag, og det medfører, at den oprindelige taksonomi skal opdateres på forskellige punkter.

Den grundlæggende beskrivelse generaliserer oplysningskategorierne og indordner dem i en hierarkisk klassifikation, denne er DANLEX-taksonomien. Betegnelserne for opstillingens hovedkategorier er baserede på lingvistikkens deldiscipliner, såsom etymologi, grammatik, fonetik, semantik, pragmatik osv.

Denne taksonomi er derefter — stadig inden for rammerne af DANLEX-projektet — blevet afprøvet som beskrivelsesværktøj på forskellige ordbogstyper, der inddrager det danske sprog, dvs. tosprogs- og etsprogsordbøger; ordbøger ordnet efter onomasiologiske hhv. semasiologiske principper; almensproglige og fagsproglige, synkrone og diakrone, præskriptive og deskriptive typer osv. På denne måde har man undersøgt, hvor bredt et spektrum af ordbøger den opstillede taksonomi kan siges at dække.

Undersøgelsen førte til den konklusion, at det var muligt og hensigtsmæssigt at opstille en sådan generel taksonomi, der viste sig også at kunne dække de andre germanske sprogs og i stor udstrækning de romanske sprogs leksikografiske behov. Desuden antog man, at taksonomien også var anvendelig til beskrivelse af leksikalske datasamlinger „til brug i informationsteknologiske systemer" (DANLEX 1987). Da DANLEX-gruppen imidlertid afsluttede sin undersøgelse af ordbøger i 1985/86, har man følgelig ikke haft adgang til større sådanne datasamlinger for dansk. STANLEX-gruppen derimod, som i 1994, næsten 10 år senere er gået i gang med en videreførende undersøgelse, har adgang til et bredt udvalg af leksikalsk materiale udarbejdet til sprogteknologiske formål.

6.3 STANLEX materialet

Den største del af STANLEX-gruppens materiale stammer fra leksikografiske datasamlinger som deltagerne arbejder med til daglig. Hvert sæt data er forsynet med to slags beskrivelser.

For det første en *overordnet beskrivelse* af den pågældende datasamling, omfattende

- kort redegørelse for dens karakter (indhold, struktur, størrelse, status, medium, evt. grafisk fremtrædelsesform...)
- oversigt over datamatiske, lingvistiske og leksikografiske principper, der danner grundlaget for datasamlingen

Denne overordnede beskrivelse er udarbejdet for at sikre fælles teoretisk basis til den videre behandling af materialet.

For det andet en mere individuel, *detaljeret beskrivelse* af det pågældende leksikografiske materiale, som de enkelte medlemmer råder over, der omfatter

- fortegnelser over oplysningstyperne til sammenligning med DANLEX-taksonomiens kategorier hhv. subkategorier
- træk/værdiliste-erklæringer (dog ikke helt ned til de mindste detaljer)
- færdige ordbogsartikler (også kodet som databaseposter eller leksikalske regler) til illustration af brugen af træk i de forskellige oplysningstyper

Denne del af materialet er det centrale i arbejdsprocessen, idet DANLEX-taksonomien ønskes videre afprøvet og udbygget ved at inddrage sådanne ordbogstyper eller projekter i undersøgelsen, der endnu ikke var tilgængelige i den første halvdel af firserne — herunder især de sprogteknologiske.

Den nyeste viden stammer i høj grad fra arbejdet med datamatstøttet leksikografi og korpusundersøgelser samt maskinel sprogbehandling og andre grene af sprogteknologi. Det er naturligt, at der blandt STANLEX-gruppens medlemmer er mange repræsentanter for anvendt sprogteknologi, især maskinel og maskinstøttet oversættelse.

- Oversættelsessystemerne der indgår i undersøgelsen er:

  PATRANS (Center for Sprogteknologi (CST) og Lingtech): maskinoversættelsessystem til patenttekster

  METAL (Siemens, Handelshøjskole Syd): maskinoversættelsessystem til tekniske tekster

  WINGER/Info:Partner: støtteværktøjer til maskinoversættelse

  LMT (Logic-based Machine Translation) og TM2 (Translation Manager): maskinoversættelsesprogrammel fra IBM.

- Andre sprogteknologiske produkter, såsom elektroniske ordbøger, tekstsamlinger, leksikografisk software osv. er også repræsenterede (TextWare A/S)

- Desuden har gruppen adgang til terminologi i databaseform, nemlig til

  DANTERM-basen (Handelshøjskolen i København), som har til hensigt at tilgodese både menneske og maskine som bruger.

På den anden side deltager nogle relevante projekter inden for den sproglige dimension med betydelig vægt på det datamatiske aspekt i det leksikografiske arbejde:

- Moderne dansk sprog
  Den Danske Ordbog (under udarbejdelse), der bygger på et artikelformat, som beskrives som en SGML-struktur. Det benyttede datamatiske redigeringsværktøj (GestorLex) er et avanceret sprogteknologisk produkt.

- Historisk sprog
  er repræsenteret ved *Folkeviseprojektet*; her arbejdes med elektronisk tekstbase og ordbase. Projektet rejser specielle lingvistiske og datamatiske problemer (fx flere ortografiske sideformer og dertil hørende komplekse søgerutiner), som også indgår i STANLEX-gruppens overvejelser.

Fra forlagssiden deltager Munskgaards Forlag.

Hovedkategorier	**Kategorier**	**Subkategorier**
etymologiske oplysninger	• parallel • oprindelse • datering	**+ (under udarbejdelse)**
fonetiske oplysninger	• prosodiske træk • segmentale træk	
grafiske oplysninger	• ortografiske opl.	• stavning • orddeling
	• grafisk symbol	
grammatiske oplysninger	• ordklasse **+ køn**	
	• bøjningsopl.	• **bøjningsparadigme** • bøjningsform
	• orddannelsesopl.	
	• syntaktiske opl.	• **syntaktisk ramme** • syntaktisk funktion
	+ grammatisk specifikation	
pragmatiske oplysninger	• tekstlige opl.	• citat • mulig kontekst **+ kontekstbegrænsning**
	• brugsopl.	• tidslig dimension • rumlig dimension • **kommunikativ** dimension • frekvens **+ teksttypologisk opl.**
	• ekstern henvisning	• litteraturhenvisning • kildehenvisning
	• evalueringsopl.	**+ (under udarbejdelse)**
	• administrative opl.	• opl. om indsamling og bearbejdning af data • intern henvisning • homografnummer • tekniske oplysninger
semantiske oplysninger	• emneklassificerende opl.	**+ (under udarbejdelse)**
	• oplysninger om semantiske relationer	• generisk over/underordningsrelation • partitiv relation • successiv relation • kausalrelation **+ associativ relation** • antonymi **+ metonymi**
	• indholdsspecificerende oplysninger	• leksikalsk parafrase • **definition** • denotativ definition • ostensiv definition **+ udvidelse af definition** • faglig forklaring • **indholdsspecif.** træk
	• ækvivalensopl.	• ækvivalensrelation inden for ét sprog • ækvivalensrelation mellem to eller flere sprog **+ ækvivalensbegrænsning**

Tabel 2: *STANLEX-taksonomien, generelle oplysningstyper, status: sommeren 1995. Tilføjelser i forhold til DANLEX-taksonomien er markeret med + og fremhævet med fed. Andre ændringer er fremhævet med fed.*

6.4 Sammenligning af STANLEX-materialet med DANLEX-taksonomien

Arbejdsfase 1, dvs. sammenligning af STANLEX-materialet med DANLEX-taksonomien er nu gennemført, og resultaterne er nedfældet i tabelform. Derefter arbejder vi på at opstille den nye STANLEX-taksonomi. En foreløbig version af denne taksonomi er præsenteret i tabel 2. (Ændringer i forhold til DANLEX-taksonomien er fremhævet med fed.)

Især analysen af maskinbrugbare ordbøger bragte nye aspekter ind i opstillingen af den taksonomiske hierarki, derfor fokuserer fremstillingen nedenfor på denne ordbogstype. Et af de centrale spørgsmål var følgende: På hvilke punkter skal DANLEX-taksonomien udbygges for også at dække maskinbrugbare ordbøger?

De grundlæggende krav til formuleringen af oplysninger i maskinbrugbare ordbøger er, at den leksikografiske beskrivelse skal være

- eksplicit
- entydig
- udtømmende
- konsistent
- formaliseret

To vigtige aspekter bør fremhæves i forhold til traditionelle (humane) ordbøgers måde at bringe oplysninger på, nemlig ekspliciterings- og formaliseringsgraden. Disse to aspekter er specifikke for oplysninger i leksikografiske datasamlinger, der skal kunne bruges af maskiner, dvs. udnyttes i sprogteknologiske værktøjer. Maskinen skal have direkte adgang til hver bid af relevant information, som ellers kan gives implicit eller vha. eksempel, omskrivning eller forklaring i ordbøger for mennesker (jf. Braasch 1994) samtidig med at informationen skal gives i en fuldstændig fastlagt form i overensstemmelse med systemets formaliseringskrav.

Arbejdsforløbet omfattede blandt andet en detaljeret sammenligning mellem oplysningstyperne i de undersøgte datasamlinger og den oprindelige taksonomis kategorier. Det første spørgsmål var:

- Hvorvidt og hvordan kan oplysningstyper, der forekommer generelt i maskinbrugbare ordbøger, tilordnes DANLEX-taksonomiens kategorier hhv. subkategorier?

I dette skridt blev overensstemmelser og afvigelser mellem den oprindelige taksonomi og det nye sprogteknologiske materiale registreret.

Vi opstillede derefter sammenlignende tabellariske oversigter ud fra to synsvinkler:

1. Oversigter over DANLEX-taksonomiens kategorier/subkategorier og deres repræsentation i hver individuel datasamling, der indgår i STANLEX-materialet
2. Oversigter over den enkelte datasamlings oplysningstyper grupperet efter DANLEX-taksonomiens kategorier/subkategorier

Vi analyserede derefter nogle typiske ordbogsartikler fra hver datasamling og forsøgte at forsyne deres oplysningstyper med kategori/subkategori-betegnelser fra taksonomien. Det viste sig i en række tilfælde, at det ikke var nogen triviel opgave at tilordne en oplysning til en bestemt type og dermed en bestemt taksonomisk subkategori. Andre gange manglede der i taksonomien en dækkende subkategori. Især oplysninger, der hører til under hovedkategorien *Grammatiske oplysninger* viste sig at være vanskelige at håndtere.

Maskinbrugbare ordbøger fx til oversættelsessystemer er særlig detaljerede og eksplicitte inden for hovedkategorien *Grammatiske oplysninger*. De indeholder en række oplysninger, fx oplysning om ledfunktionen af valensbundne argumenter ved verber, der håndteres ud fra forskellige lingvistiske principper i de enkelte maskinbrugbare ordbøger. De kan derfor enten tilordnes den taksonomiske kategori *Syntaktiske oplysninger* (derunder subkategorien *Syntaktisk funktion*), eller kategorien *Grammatisk specifikation.*

Vi arbejder stadig med en udvidet liste af (sub)kategorier, der foreløbig omfatter oplysningstyper under kategorierne

- Syntaktisk ramme
- Syntaktisk funktion
- Grammatisk specifikation

under hovedkategorien *Grammatiske oplysninger.*

Disse oplysningstyper er velegnede til formalisering, fordi de forholdsvis enkelt kan systematiseres, fx udfra nogle lingvistiske eller datamatiske principper, og derefter kan de repræsenteres i en enkel formel kodeform i maskinbrugbare ordbøger.

De andre ordbogstyper gav anledning til at gå i dybden med andre taksonomiske kategorier:

- Humanordbøger bidrager primært med oplysningstyper inden for hovedkategorierne *Etymologiske oplysninger, Semantiske oplysninger* og *Pragmatiske oplysninger*. Under den sidst nævnte er en ny subkategori indført i taksonomien, nemlig *Teksttypologisk oplysning*. Arbejdet på *Den Danske Ordbog* er banebrydende på dette punkt, idet hvert tekststykke i korpuset er forsynet med 'Headers' — en slags overskrifter — som fastholder tekstkildens relevante egenskaber. Denne oplysning er systematisk opført i den leksikografiske beskrivelse (af eksempelmaterialet). Desuden er fx normative oplysninger også kun repræsenteret i humanordbøger, hvor de til gengæld spiller en vigtig rolle. Dette har vi endnu ikke arbejdet med i detaljer.

- Terminologibaser stiller særlige krav til detaljeringsgraden i kategorien *Indholdsspecificerende oplysninger* (især til subkategorierne for definitionstyper). De hører hjemme under hovedkategorien *Semantiske oplysninger*. Fagsystematisk beskrivelse hører til under den anden subkategori, nemlig *Oplysninger om semantiske relationer.* Vi har endnu ikke beskæftiget os mere dybtgående med disse oplysningstyper, men som det kan ses ved en sammenligning af den oprindelige (DANLEX) taksonomi og den opdaterede version (STANLEX), er der også her kommet nogle nye subkategorier til.

Det turde fremgå af fremstillingen ovenfor, at de forskellige ordbogstyper bidrager til opdateringen af taksonomien under hver sin(e) hovedkategori(er); man kan tale om en synergi-effekt.

På grundlag af de indtil nu udførte analyser og sammenligninger kan vi konkludere, at DANLEX-taksonomiens *hovedkategorier* og (med en enkelt undtagelse også) *kategorier* har vist sig at være dækkende for STANLEX-materialet. I kolonnen for *subkategorier* (dvs. på tredje niveau i hierarkiet) har vi fået brug for en del tilføjelser. For især de grammatiske oplysninger gælder det, at selvom kriterierne for maskinbrugbarhed af leksikalske datasamlinger i høj grad bygger på kravet om eksplicit og udtømmende formulering, er det ikke hensigtsmæssigt at udbygge taksonomien med et fjerde hierarkisk niveau, svarende til subkategorier af anden grad.

Grunden til denne beslutning er for det første, at taksonomien bliver tungere at arbejde med (især i de sammenhænge, hvor man ikke har brug for denne yderligere detaljeringsgrad); for det andet viste det sig, at oplysningstyperne på dette niveau i høj grad er system- og teoriafhængige, og derfor vil man ikke kunne opretholde princippet om generaliserbarhed.

7 Hvad skal/kan en sådan standard bruges til og hvordan?

Den her følgende redegørelse er baseret på en foreløbig skitse over STANLEX-gruppens opgaver nedfældet i notatet *Forslag til indhold i en dansk standard for indholds- og strukturbeskrivelse af leksikalske oplysninger*. Standarden vil omfatte

- klassifikation af oplysninger (i hovedkategorier, kategorier og subkategorier), dvs. en taksonomi
- opstilling af beskrivelsesmodeller for ordbogsstrukturer (makro- og mikrostruktur)

STANLEX-gruppen har foreløbig defineret to primære anvendelsesområder for en sådan standard. Det første er:

- *Standarden skal bruges ved udveksling af leksikalske data. . .*

En sådan situation foreligger fx, når man ønsker at udvide ordbogen i et maskinoversættelsessystem (det såkaldte *lexicon*) med ordbogsindgange fra et andet system eller fra en offentligt tilgængelig termbase. Der kan også være tale om at flette to maskinlæsbare ordbøgers materiale sammen til én enkelt, ny og større ordbog med det formål at udgive den i trykt eller elektronisk form. Hurtig udveksling af store mængder data er ikke længere et teknisk problem, men det forbliver en omfattende opgave at analysere og harmonisere datasamlinger som er meget forskellige, hvad indhold og struktur, men også hvad beskrivelsesmåde, dvs. præsentation af oplysninger, angår.

Inden sammenlægning af to eller flere leksikalske datasamlinger er det altså nødvendigt at analysere materialet og dertil har man brug for taksonomien. Analysen viser

- på hvilke punkter datasamlingernes oplysninger er kompatible
- på hvilke punkter der er indholdsmæssige afvigelser (fx i detaljeringsgrad af samme oplysning eller grundet forskelle i lingvistisk tilgang osv.)

- hvilke oplysninger der mangler i det samlede materiale i forhold til den planlagte beskrivelsesmodel

Datasamlinger bør i øvrigt forsynes med udførlig (og præcis) dokumentation og kan med fordel blive struktureret og beskrevet iht. den givne taksonomi.

Det andet anvendelsesområde skitseres på følgende vis:

- *Standarden kan . . . også bruges ved planlægning, udvikling og lagring af leksikalske data.*

Det er klart, at når man indser fordelene ved at have adgang til flere (indholds- eller størrelsesmæssigt — eller på anden måde) forskellige men kompatible leksikalske datasamlinger, vil man bestræbe sig på at udforme fremtidige ordbøger sådan, at de kan kombineres indbyrdes og udnyttes i flere sammenhænge. Resultatet vil blive at materialet kan bruges og genbruges i forskelligartede leksikografiske produkter.

Inden for standarden er det hensigten, at den opdaterede taksonomiske opstilling over leksikalske oplysningstyper skal fungere som generelt beskrivelsesværktøj og skal være udbygget med

- definitioner for alle kategorier og subkategorier
- eksempler på klassificering af faktiske oplysninger
- kommentarer, når der synes at være brug for uddybende forklaring

For at øge standardens anvendelighed i praktisk planlægning og udarbejdelse af leksikalske datasamlinger gives der supplerende eksempelmateriale i et anneks. Dette materiale vil efter de nuværende planer omfatte

- typeeksempler på beskrivelsesstrukturer i leksikografiske datasamlinger
- sammensætning af standardelementer til forskellige leksikografiske formål
- eksempler på hensigtsmæssige lagrings- og udvekslingsformater bygger på standardelementer

STANLEX-gruppen arbejder i indeværende, anden arbejsfase på at løse ovennævnte opgaver.

8 Afsluttende bemærkninger

Da STANLEX-projektet befinder sig midt i en arbejdsproces er der endnu ikke så meget andet at konkludere på nuværende tidspunkt end det jeg har kunnet sige sammenfattende om de delresultater, der er nået indtil nu.

Det danske initiativ står naturligvis ikke alene i den faglige omverden. Der er på internationalt plan forskellige typer projekter i gang inden for leksikografi og standardisering; nedenstående nævnes et par eksempler på denne aktivitet.

TEI: (Text Encoding Initiative, under organisationerne ACH, ACL, ALLC) har efterhånden løbet gennem en del år. Der udgives med mellemrum en ny (opdateret) version af publikationen „Guidelines For the Encoding and Interchange of Machine-Readable Texts" (TEI 1994). Foruden andre emner, der er relevante for forskellige typer af dokumenter har rapporten også et større afsnit om „Print Dictionaries". Afsnittet giver gode retningslinjer til at arbejde med makro- og mikrostruktur, hierarkisk strukturering af artikler og oplysningstyper og meget andet.

En anden type projekt er **EAGLES** (Expert Advisory Group on Language Engineering Standards), et komplekst EU-støttet projekt. En af dets undergrupper beskæftiger sig med at sammenligne hvordan human- og maskinbrugbare ordbøger koder morfosyntaktiske oplysninger. Alle EU-sprog er blevet inddraget. Formålet med dette delprojekt er at udarbejde et forslag til international standard til repræsentation af morfosyntaks i leksikalske datasamlinger.

Litteratur

Bech, Annelise/I. Elfving/G. Engel/J. Lund/B. Maegaard/B. Nistrup-Madsen/A. Melchiors/A. Møller/F. Svanholm/K. Aakjær/R. Heldt Nielsen 1994: Dansk sprogteknologi — status, perspektiver, handlemuligheder. *TeknologiNævnets rapporter* 1994/1. København.

Braasch, Anna 1994a: How Far Do Printed Dictionaries and MT-Lexicons Share Information? I: Paolo Alberto/Paul Bennett (ed.): *Lexical Issues in Machine Translation. Studies in Machine Translation and Natural Language Processing.* Vol. 8. Luxembourg: CEC, 117–133.

Braasch, Anna 1994b: Når maskinen tager én på ordet — ordbogsarbejde for maskinoversættelse. I: Anna Garde/Pia Jarvad (udg.): *Nordiske studier i leksikografi II. Skrifter undgivet af Nordisk Forening for Leksikografi.* Skrift nr. 2. København: LEDA, 29–38.

The DANLEX-GROUP (Ebba Hjorth/B. Nistrup-Madsen/O. Norling-Christensen/Rosenkilde Jacobsen/H. Ruus) 1987: *Descriptive Tools for Electronic Processing of Dictionary Data.* Studies in Computational Lexicography. Lexicographica Series Maior 20. Tübingen: Max Niemeyer Verlag.

TEI 1994 = C. M. Sperberg-McQueen/Lou Burnard (ed.): *Guidelines For the Encoding and Interchange of Machine-Readable Texts.* Chicago/Oxford.