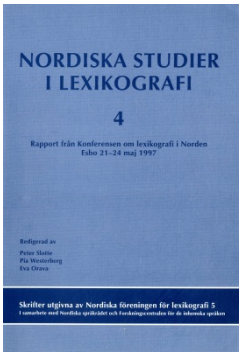
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: Dansk standard for indholds- og strukturbeskrivelse af leksikalske datasamlinger - Eksempler på anvendelser i leksikografisk arbejde

Forfatter: Anna Braasch

Kilde: Nordiska Studier i Lexikografi 4, 1997, s. 17-26
Rapport från Konferens om lexikografi i Norden, Esbo 21.-24. maj 1997

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

Dansk standard for indholds- og strukturbeskrivelse af leksikalske datasamlinger – Eksempler på anvendelser i leksikografisk arbejde

The first version of a *Proposal for Danish Standard: Lexical data collections - Description of data categories and data structure - Part 1: Taxonomy for the classification of information types* has been sent out for comments by Dansk Standard (The Danish Standard Association). This contribution deals with two main applications of the taxonomy: description of the information content of an existing lexical data collection for exchange/reuse purposes and planning of a new lexical data collection with a view to reusability. The Centre for Language Technology had the opportunity to work experimentally with both applications of the taxonomy within the framework of different projects. We provide here a brief description of the tasks where we applied the taxonomy. We conclude with some relevant aspects of our experience with the application of the taxonomy.

1 Indledning

STANLEX-gruppen[1], der er en tværinstitutionel arbejdsgruppe under Dansk Standard, har fået til opgave at udarbejde et forslag til dansk standard for indholds- og strukturbeskrivelse af leksikalske datasamlinger.

Projektets baggrund og målsætning, samt arbejdsprocessens første fase indtil forsommeren 1995 er beskrevet i Braasch (1995).

Standarden skal omfatte to hovedafsnit. *Del 1: Taksonomi til klassifikation af oplysninger* (herefter STANLEX-taksonomien) blev i en første version udsendt til høring i oktober 1996. Svarfristen var 1. januar 1997 og der er kommet knap 40 tilbagemeldinger. Alle indkomne kommentarer behandles i skrivende stund og relvante forslag til ændringer indarbejdes i den endelige version. *Del 2*, der vil behandle strukturen i leksikalske datasamlinger, er også under udarbejdelse.

På forsiden af det udsendte dokument står der "Forslaget skal yderligere bearbejdes og kan derfor ikke anvendes som dansk standard". Det betyder at både taksonomien og de dertil knyttede definitioner stadig er under revision (nu ikke mindst på grundlag af de indkomne høringssvar) og derfor kan det her præsenterede afvige fra forslagets endelige version.

---

[1] Gruppens medlemmer er: Anna Braasch, Dorthe Duncker, Gert Engel, Claus Bo Jørgensen, Hanne Jensen, Margrethe H. Møller, Bodil Nistrup Madsen, Ole Norling-Christensen, Hanne Ruus

2 Forudsætninger og behov for genbrug af leksikalske datasamlinger

I dag anvendes edb overalt i fremstillingen af materiale til trykte ordbøger: der er mange muligheder, såsom tekstbehandlingsprogrammer, systemer til struktureret ordbogsredigering, orddatabaser osv. man taler således om datamatstøttet leksikografi. Det leksikografiske materiale lagres på et elektronisk medium (disk eller diskette, bånd, CD-ROM) dvs. at det er maskinlæsbart. Dermed åbner der sig nye, spændende perspektiver for det leksikografiske arbejde.

Informationsstrømmen i samfundet vokser med de nye teknologiske muligheder, og samtidig vokser også behovet for hurtig og effektiv elektronisk behandling af naturligt sprog. Sprogindustrien udvikler og producerer forskellige typer af sprogteknologiske værktøjer bl.a. til maskinoversættelse, informationssøgning, automatiske svarsystemer. Disse værktøjer har i de fleste tilfælde også en indbygget ordbog, et såkaldt **leksikonmodul**. Da det er meget tidkrævende og derfor ganske dyrt at udarbejde leksikografisk materiale fra bunden f.eks. til en ny trykt ordbog eller til et leksikonmodul er der et stort behov for at genbruge det samme leksikografiske eller udnytte det til flere forskellige formål. Udnyttelse af maskinlæsbare leksikalske data er naturligvis højt prioriteret i udvikling af sprogteknologiske værktøjer.

Den tekniske udvikling gør det muligt at håndtere store mængder data meget effektivt; samtidig stiller den maskinelle behandling af ordbogsdata store krav til systematisk, entydig og præcis beskrivelse af datasamlingens indhold (oplysningstyper) og struktur. Leksikografisk materiale – ordbogsdata – kan bestå af mange forskellige slags oplysninger og de kan være struktureret på forskellig vis, alt efter hvilken ordbogstype, hvilket forlag eller hvilket slags projekt der er tale om. Den måde oplysningstyperne er udvalgt og organiseret på i en given leksikalsk datasamling (som kan være en trykt eller elektronisk ordbog, en ord- eller termbase, hhv. et leksikonmodul) sammenfattes i en **beskrivelsesmodel**. Det er en nødvendig forudsætning at kende en given datasamlings beskrivelsesmodel for at kunne arbejde med og bygge videre på de foreliggende data til en ny anvendelse. Ligeledes er det nødvendigt at udarbejde en beskrivelsesmodel der tager videst mulig hensyn til de påtænkte anvendelser når man vil producere nye leksikalske data som skal kunne udnyttes til forskellige formål.

Beskrivelsesmodellen, der svarer til en meget detaljeret og eksplicit redaktionsvejledning for en trykt ordbog mht. valg af oplysninger, deres detaljeringsgrad hhv. indbyrdes rækkefølge osv. produceres enten i forbindelse med ordbogens planlægning og udarbejdelse, eller den kan produceres vha. tilbundsgående analyse (f.eks. ved automatisk parsing) af allerede eksisterende ordbogsdata. Optimalt er det naturligvis hvis en datasamling bliver beskrevet parallelt med at den udarbejdes, og at beskrivelsesmodellen kan blive testet i det leksikografiske arbejde.

For at sikre at en leksikalsk datasamlings beskrivelse er velegnet som grundlag for planlægning af dataudveksling eller af genbrug af data i nye anvendelser, er det hensigtsmæssigt at beskrivelsen bygger på et gennemarbejdet, alment kendt – og accepteret og brugt! – klassifikationssystem. Formålet med standarden er at opfylde dette behov.

3 Om standardens anvendelsesmuligheder

Forslaget til standarden omtaler to væsentlige anvendelsesmuligheder for taksonomien, jf. Forslag til dansk standard (1966:10):

- beskrivelse af en datasamlings oplysninger med henblik på udveksling af leksikalske data
- planlægning af indholdet i en leksikalsk datasamling.

Brugen af en standard for lagring og udveksling af data anbefales bl.a. ved udarbejdelse af trykte og elektroniske ordbøger for mennesker samt ved planlægning af ord- hhv. termbaser for sprogteknologiske anvendelser. Også på internationalt plan findes der en række initiativer og forslag til harmonisering hhv. standardisering af indhold og struktur af leksikalske datasamlinger, og interessen for et fælles udvekslingsformat er voksende.

4 Praktisk anvendelse af STANLEX-taksonomien

Center for Sprogteknologi (CST) deltager aktivt i udarbejdelsen af standarden og det var derfor oplagt at undersøge dens anvendelsesmuligheder inden for et par af Centrets aktuelle sprogteknologiske projekter.

Da forslaget til standarden og hermed også STANLEX-taksonomien dengang endnu var under udarbejdelse, var vi indstillet på at anvende denne foreløbige version. Fremgangsmåden vil i det store og hele blive den samme, uanset om der senere vedtages nogle ændringer i taksonomiens hierarkiske struktur, detaljeringsgrad eller af de anvendte kategorinavne, og uanset eventuelle tilføjelser. Samtidig er vi opmærksomme på at sådanne ændringer muligvis vil kunne medføre en nødvendig revision af de første undersøgelsesresultater og de deraf følgende beslutninger.

Da CST i 1996 netop stod for at skulle arbejde med et par projekter som også omfatter opgaver, der falder inden for standardens erklærede anvendelsesområder, var det nærliggende afprøve STANLEX-taksonomien som redskab i arbejdet med disse opgaver. Det drejede sig først og fremmest om et internationalt udviklingsprojekt, OTELO[2]. Senere påbegyndte vi planlægningen af en stor dansk sprogteknologisk ordbog (STO) som CST har taget initiativ til.

I begge projekter er der brug for systematisk beskrivelse af leksikalske oplysninger, men da projekterne er ganske forskellige med hensyn til bl.a. udgangspunkt, metode og mål, har vi grebet sagen an på forskellig vis. Her vil vi koncentrere os om OTELO-projektet fordi STO-projektet befinder sig endnu i pilotfase.

5 OTELO-projektet

Center for Sprogteknologi deltager sammen med en række repræsentanter for sprogindustrien (jf. fodnote 2) i dette projekt. Målsætningen er at konstruere et omfattende, kommercielt sprogteknologisk system, fortrinsvis for opgaver inden for maskinoversættelse. Det omfatter bl.a. en

[2] OTELO er akronym for EU-projektet *O*pen *T*ranslation *E*nvironment for *LO*calisation (LE-2703) Deltagere: Logos GmbH., SAP AG., GMS GmbH. med Lotus Development som koordinator.

central enhed hvorigennem brugerne (f.eks. oversættere af fagtekster) skal kunne trække på alle tilknyttede ressourcer. Systemets kerne skal være en central leksikalsk database hvori der skal kunne integreres forskellige leksikalske datasamlinger, eksempelvis SAPterm som er en stor, flersproget termbase for mange fagområder, eller PaTrans[3] engelsk-danske petrokemiske termbase som er meget specialiseret og væsentlig mindre.

Projektet arbejdede i den her beskrevne arbejdsfase med fire forskellige oversættelsessystemers måde til at beskrive ord- og termer (Logos, GMS/METAL, SAP og PaTrans). Disse systemer har hvert sit leksikonmodul – dvs. et sæt af ord- og termbaser – som er ganske forskellige, både hvad størrelse, indhold og struktur angår, og de er derfor ikke kompatible. OTELO's centrale database skal kunne håndtere (ind- og udlæse, udveksle, vedligeholde, opdatere, overføre til andre systemer mm.) leksikalske data. Der skal desuden også kunne udtrækkes delmængder af oplysninger fra fællesbasen til nye, brugerbestemte anvendelser.

Målet er at gøre de leksikalske data kompatible vha. et overordnet, standardiseret fællesformat, OTELO-formatet, og via den centrale OTELO-database. I dette format fastlægges den fælles struktur som oplysningerne om det leksikalske materiale (dvs. indholdet af de enkelte datasamlinger) skal kunne organiseres i; formatet skal kunne opdateres efter behov.

Dertil udvikles der en beskrivelsesmetode således at de fire, til dels meget komplekse, leksikalske datasamlinger kan harmoniseres, sammenlignes og i sidste ende også kombineres på kryds og tværs gennem den centrale enhed.

For at opnå målet skulle der løses tre delopgaver:

- hvert systems leksikalske materiale skulle beskrives af firmaet der har udviklet det
- de fire datasæts beskrivelser skulle gøres sammenlignelige vha. et fælles system til klassifikation af oplysningerne
- den centrale databases oplysningstyper og beskrivelsesstruktur skulle fastlægges.

5.1 Beskrivelse af en datasamlings oplysninger: PaTrans

Det første trin bestod hovedsageligt i at klassificere de enkelte datasamlingers oplysninger entydigt. CST deltog i dette arbejde med beskrivelsen af de leksikalske oplysningstyper der er brugt i maskinoversættelsessystemet PaTrans' leksikonmodul. Modulet omfatter en engelsk og en dansk almensproglig ordbase samt en engelsk-dansk termbase. Den dertil hørede detaljerede kodningsmanual dokumenterer systemets leksikografiske beskrivelsesmodel.

Oplysningerne i PaTrans systemets ordbaser blev klassificeret ved hjælp af STANLEX-taksonomien, jf. figur 1. Der skal knyttes to bemærkninger til den her afbildede taksonomi. For det første er denne en nyere version (juli 1997) end den vi anvendte (oktober 1996). For det andet er taksonomien opstillet på basis af mange forskellige typer af leksikografiske datasamlinger (beskrevet i Braasch, 1995), herunder ordbøger der f.eks. også indeholder etymologiske og/eller fonetiske oplysninger. Taksonomien er dermed langt mere generel end det kræves af de sprogteknologiske anvendelser.

---

[3] PaTrans-systemet er udviklet af CST i samarbejde med firmaet Lingtech til oversættelse af petrokemiske patenttekster fra engelsk til dansk. I kommerciel anvendelse siden 1995.

Bemærkning:
Oplysningstyper som er anvendt i PaTrans-systemets leksikonmodul er fremhævet med fedt.

Hovedkategorier	**Kategorier**	**Subkategorier** [1]
administrative oplysninger	▫ intern henvisning	
	▫ ekstern henvisning	▫ litteraturhenvisning ▫ kildehenvisning
	▫ **oplysning om indsamling og bearbejdning af data** ▫ **tekniske oplysninger**	
etymologiske oplysninger	▫ oprindelse ▫ parallel	
fonetiske oplysninger	▫ prosodiske træk ▫ segmentale træk	
grafiske oplysninger	▫ ortografiske oplysninger	▫ **stavning** ▫ orddeling
	▫ grafisk symbol	
grammatiske oplysninger	▫ **ordklasse** ▫ **køn**	
	▫ **bøjningsoplysninger**	▫ **stamme** ▫ **bøjningsparadigme** ▫ **bøjningsform**
	▫ **orddannelsesoplysninger**	
	▫ **syntaktiske oplysninger**	▫ **syntaktisk ramme** ▫ **specifikation af syntaktisk ramme** ▫ **specifikation af hjælpeverbum** ▫ **syntaktisk funktion**
		fortsættes ...

Hovedkategorier	**Kategorier**	**Subkategorier** [1]
semantiske oplysninger	▫ emneklassificerende oplysninger	▫ klassifikationssystem ▫ standardiseret emneklassifikation ▫ ikke-standardiseret emneklassifikation
	▫ oplysninger om semantiske relationer	▫ begrebssystem ▫ position i begrebssystem ▫ generisk over-/under- ordningsrelation ▫ partitiv relation ▫ successiv relation ▫ kausal relation ▫ associativ relation ▫ antonymi ▫ metonymi ▫ ækvivalensrelation inden for ét sprog ▫ **ækvivalensrelation mellem to** eller flere **sprog** ▫ ækvivalensbegrænsning
	▫ indholdsspecificerende oplysninger	▫ leksikalsk parafrase ▫ analytisk definition ▫ denotativ definition ▫ ostensiv definition ▫ udvidelse af definition ▫ faglig forklaring ▫ **indholdsspecificerende træk** ▫ overført betydning
sprog		
sprogbrugs-oplysninger	▫ brugseksempler	▫ tekstudsnit ▫ ordforbindelse
	▫ brugsoplysninger	▫ tidslig dimension ▫ rumlig dimension ▫ kommunikativ dimension ▫ frekvens
	▫ evalueringsoplysninger	
struktur-oplysninger	▫ homografmarkering ▫ afsnitsmarkering	▫ lineær afsnitsmarkering ▫ hierarkisk afsnitsmarkering

[1] Blanke felter på subkategoriniveau betyder ikke, at der ikke kan dannes subkategorier til den pågældende kategori, men blot at nærværende standard ikke ønsker at foreskrive en endelig liste af subkategorier til kategorien.

Figur 1: *STANLEX-taksonomien (foreløbig version, juli 1997)*

Klassifikationen gjaldt i første omgang systemets danske og engelske almensproglige ordbase. Derefter har vi også inddraget den engelsk-danske termkomponent. På denne måde blev der arbejdet med to typer af leksikalske datasæt idet systemet håndterer det almensproglige ordstof anderledes (vha. to adskilte monolingvale ordbogskomponenter + en oversættelseskomponent) end termer (vha. en samlet bilingvalt komponent). Disse to typer afviger til dels fra hinanden mht. de anvendte oplysningstyper. De almensproglige etsprogskomponenter indeholder fyldige oplysninger til morfosyntaktisk identifikation af opslagsord og af ordets betydning, hvorimod opslagsordet i termbasen er koblet direkte til sin målsproglige ækvivalent, og derfor er der ikke brug for samme mængde oplysninger til identifikationen.

Oplysningerne er ikke hierarkisk strukturerede eller på nogen måde systematiserede i PaTrans – de er lineært formulerede; mnemotekniske navne bidrager til at leksikografen kan gennemskue og huske sammenhængen mellem dem. Databasen har faciliteter til at opretholde konsistens mellem de anførte oplysninger (det er f.eks. ikke muligt at angive køn ved et verbum i databasen). Klassificering af PaTrans' oplysningstyper ved brug af taksonomien var en *forholdsvis* enkel opgave, da systemet har en gennemkommenteret såkaldt trækspecifikation som anfører hver anvendt oplysningstype og for hver enkel type dens tilladte værdier. Eksempelvis beskriver linjen

'cat= v, n, adj, adv, ..., conj, p, pron.'

hvilke ordklasseangivelser der accepteres af systemet.

Denne metode svarer til brugen af tabellagte værdier i redaktionsvejledninger for trykte ordbøger, f.eks. i forkortelseslisten, som bør indeholde alle forkortelser der er brugt i ordbogen og kun dem. Den tidligere nævnte kodningsmanual benyttes i ordbogsarbejdet for at sikre korrekt og konsistent leksikografisk beskrivelse af ord. Manualen svarer omtrent til trykte ordbøgers meget detaljerede og komplette redaktionsvejledninger.

5.1.1 Sammenfatning af foreløbige erfaringer

Erfaringene fra denne proces kan sammenfattes således:

- Dokumentationen sammen med et grundigt kendskab til systemet var afhgørende for konsekvent tilordning af hver enkel oplysning til een og kun een af taksonomiens (hoved/sub)kategorier.
- Der var dog en vis usikkerhed mht. entydig klassificering i nogle tilfælde, især ved et par oplysningstyper inden for den grammatiske hovedkategori da PaTrans arbejder med detaljeret beskrivelse af ords syntaktiske forbindelighed der omfatter flere typer af valensoplysninger. I denne forbindelse opstod spørgsmålet om behov for et fjerde beskrivelsesniveau i taksonomien ('subsubkategori') som er siden blevet afklaret, jf. nedenfor (konklusionen).

  I STANLEX-taksonomien (figur 1) er oplysningstyperne der er anvendt i PaTrans-systemets leksikonmodul fremhævet med fedt.
- Resultatet var – trods den nævnte usikkerhed – en klar opdeling af oplysningerne, som derved ville være sammenlignelige med andre datasamlingers parallelle (STANLEX-baserede) beskrivelser.

5.2 Sammenligning af fire forskellige beskrivelser

Det næste skridt var at sammenligne beskrivelserne som er blevet udarbejdet af de fire systemers udviklere. Formålet var at opstille en fuldstændig liste der skulle dække alle oplysningstyper der er anført i et eller flere af de fire beskrivelse (dvs. en liste over **foreningsmængden)** samt at finde frem til de oplysningstyper der er repræsenterede samtlige fire beskrivelse (den såkaldte **fællesmængde**). Den sidstnævnte sammenstilling vil kunne danne den organisationsmæssige kerne i en fælles database. De oplysningstyper som er specifikke for de enkelte systemer, f.eks. dem som er nødvendige for vedligeholdelse af systemets database, er blevet udskilt; de bibeholdes, men uden for den egentlige fælles kerne. Derefter kunne vi påbegynde arbejdet med fastlæggelsen af den centrale databases oplysningstyper og beskrivelsesstruktur.

De enkelte (og oprindelig ganske uens strukturerede) beskrivelser er først blevet samordnet og er derefter, for overblikkets skyld, struktureret og sammenfattet i tabeller; nedenfor ses et udsnit af tabellen for oplysningstyper for verber.

Feature Type	*OTELO Feature/Data Type*	*Logos LEF*	*Metal MLIF*	*PaTrans*	*SAPterm*
• *Inflection:*					
	Inflection type / *integer*		Inflection type: <IN>		
Inflection Form:	**Person / *character, set***		Person: <PS>		
	Number / *string, set*	Number	Number: <NU>		
	Tense / *string, set*		Tense: <TN>	Tense	
	Aspect / *string, set*			Aspect	
	Mood / *string, set*		Mood: <MD>		
	Voice / *string, set*			Voice	
	Passive marker / *string, set*		Passive voice: <PV>	Passive form	
	Morphological verb form / *string*			Verb form	
• *Syntactic Information:*					
	Syntactic type / *integer*	Verb codes (SAL)	Kind of verb: <KVB>	Reflexive: <refl>	
Transitivity:	**Transitivity type / *string, set***	Transitivity	Transitivity type: <TT>	(Included in syntactic frame)	
Syntactic frame:	**Syntactic frame / *string, set (??)***	Verb codes (SAL)	Arguments: <ARGS> Predicate form:	1) Syntactic frame 2) Slot #s & semantic roles of arguments 3) Prep governance	

Figur 2: *Udsnit af en sammenlignende tabel, OTELO-projektet. (Kilde: Task 3.4: Integration Requirements: Common Lexical Resource Format. Arbejdsdokument, 1996. Udarbejdet af Susan McCormick, SAP AG.)*

Tabellerne blev derefter gennemarbejdet i fællesskab, med henblik på at finde frem til repræsentationen af alle de forskellige slags oplysninger dvs. foreningsmængden.

Samordningen, der bl.a. omfattede en delvis harmonisering af betegnelserne for de enkelte oplysningstyper, danner grundlaget for udarbejdelsen af den såkaldte formelle trækdefinitionsfil. I en trækdefinition kan forskellige betegnelser for ens – eller næsten ens – oplysningstyper koordineres; dette gør det muligt i den centrale base at genkende og håndtere data fra forskellige systemer vha. konvertering til det fælles OTELO-format.

Sammenlignings- og harmoniseringsprocessen var temmelig indviklet fordi beskrivelserne er principielt forskellige både hvad metode og beskrivelsessprog angår.

Problemerne opstod fordi de enkelte systemers oprindelige beskrivelser

- ikke er beregnet til ekstern dokumentation og derfor er ikke de ikke detaljerede og eksplicitte nok
- ikke er struktureret på en sådan måde at de systemspecifikke og almene, lingvistiske oplysninger er tydeligt og konsekvent adskilte.

De fire systemer

- er designet til forskellige slags opgaver hhv. formål og er derfor meget forskellige mht. oversættelsesstrategi hvilket bl.a. påvirker mængden og arten af oplysninger som et systems leksikonmodul hhv. grammatikkomponent skal håndtere
- er forskellige med hensyn til lingvistisk tilgang i oversættelsesprocessen, hvilket påvirker oplysningsindholdets detaljeringsgrad (f.eks. ordklassers semantikbaserede underinddeling, valensbeskrivelse ved mere eller mindre finmaskede syntaktiske mønstre)
- håndterer forskellige sprog og sprogkombinationer (f.eks. kræver tysk mange træk til beskrivelsen af kasusflektion, engelsk ikke; for tysk er løs sammensætning af verber relevant, for dansk ikke) hvilket resulterer i forskellige krav til repræsentation hhv. detaljeringsgrad af de pågældende morfosyntaktiske oplysninger
- bruger forskellige datatyper (streng, bogstav, ciffer/tal eller boolsk værdi) til at udtrykke samme oplysningsindhold og er dermed ikke kompatible uden konvertering
- organiserer oplysningerne på forskellig vis (lineært hhv. hierarkisk eller relationelt struktureret, i forskellige slags tabeller der er kædet sammen på forskellig vis ...)

5.2.1 Erfaringer fra sammenligningsprocessen

Sammenfattende kan det siges, at der både leksikografisk og datamatisk er ganske store, principielle forskelle mellem de fire behandlede datasæt.

Erfaringene fra sammenligningsprocessen er i hovedtræk:

- Det var meget vanskeligt at sammenligne *beskrivelserne* af de forskellige datasæt mht. informationsindhold, da een og samme leksikalske oplysning(stype) var blevet klassificeret på forskellig vis af de enkelte beskrivelser. Vi havde ved sammenligningen ikke direkte adgang til leksikalske data. Det ville have hjulpet ganske væsentligt, hvis hvert system havde været beskrevet på samme måde og ved konsekvent brug af den samme, helst standardiserede taksonomi (ikke nødvendigvis STANLEX). Der er dog blevet gjort forsøg herpå bl.a. med anvendelsen af MARTIF (Machine-Readable Terminology Interchange Format), men det skete ikke fra starten og derfor er erfaringerne derfra ikke medtaget her.

- Systemerne bygger på forskellige lingvistiske principper og stiller derfor ikke ens krav mht. til hvilke oplysninstyper der er de centrale. Følgelig er detaljeringsgraden og repræsentationen inden for de enkelte (sub)kategorier meget varierende. Den anvendte klassificering (i tabellen, jf. figur 2, repræsenteret med kolonnen *Feature Category*) ligner meget STANLEX-taksonomiens opdeling mht. almene sproglige oplysninger, men den er samtidig mere sprogteknologisk, datamatisk eller 'maskinoversættelsesorienteret'. (Den vigtigste forskel er at der er flere kategorier – eller kasser – til håndering af administrative opysninger, tekniske hhv. systemspeci-

fikke oplysninger.) Først efter flere skridt var det muligt at gøre de fire systemers oplysningstyper sammenlignelige og oprette kolonnen *OTELO Feature/Data type*. Denne kolonne indeholder det aktuelle forslag til, hvordan den pågældene information skal betegnes og udtrykkes i den fælles trækdefinition. Tabellen relaterer dermed også hvert enkelt systems betegnelse til den fælles feltbetegnelse i databasen. (Den viste tabel repræsenterer et mellemstadium i processen.)

- Det fremgik af de færdige sammenlignende tabeller at der er en del oplysniner er umulige at tilordne de vedtagne (sub)kategorier, og derfor er der indført en oplysningstype med betegnelsen *Other* der er en slags samlekasse for uklassificerede oplysninger.

- Et yderligere problem opstod ved oplysninger der – ud fra forskellig lingvistisk observans – kan tilordnes forskellige kategorier, eksempelvis brugen af tematiske roller i valensbeskrivelse, som ligger i grænseområdet mellem syntaks og semantik.

Konklusion

STANLEX-taksonomien viste sig både brugbar, nyttig og effektiv i de praktiske anvendelser hvor CST har afprøvet den. Vi fortsætter med at anvende de her skitserede metoder ved udarbejdelsen af den sprogteknologiske ordbogs beskrivelsesmodel hvor udnyttelse af eksisterende data og produktion af nye data vil indgår side om side i det lekikografiske arbejde.

Spørgsmålet om hvorvidt der er brug for et yderligere specificerende niveau i taksonomien i form af 'subsub'-kategorier er siden konferencen blevet afklaret. STANLEX-gruppen vedtog ikke at indføre flere hierarkiske niveauer. Den standardiserede taksonomi bør holdes på et tilpas generelt niveau, men den kan af den enkelte bruger ved behov frit udbygges med de ønskede subsubkategorier.

Litteratur

Braasch, Anna 1995: Arbejdet med "Forslag om dansk standard for lagring og udveksling af leksikalske data". I: *Nordiske studier i leksikografi 3*. Red.: Ásta Svavarsdóttir, Guðrun Kvaran, Jón Hilmar Jónsson. Reykjavík, 69–81.

Forslag til dansk standard (1996). DSF 30564. Leksikalske datasamlinger. Indholds- og strukturbeskrivelse. Del 1: Taksonomi til klassifikation af oplysningstyper. (DS-udvalg: K-315. Projektleder: Klaus Søndergaard. Udarbejdet af STANLEX-gruppen.)

MARTIF: ISO DIS 12200.2. Terminology – Computer Applications – Machine-Readable Terminology Interchange Format (MARTIF). Udgivet af ISO (the International Organisation for Standardization)