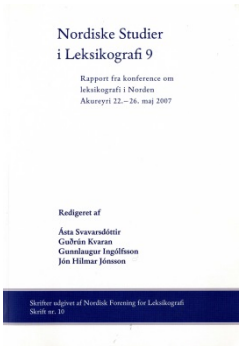
NORDISKE STUDIER I LEKSIKOGRAFI

Titel: DanNet: udvikling og anvendelse af det danske wordnet

Forfatter: Bolette Sandford Pedersen, Sanni Nimb og Lars Trap-Jensen

Kilde: Nordiska Studier i Lexikografi 9, 2008, s. 353-370
Rapport fra Konference om leksikografi i Norden, Akureyri 22.-26. maj 2007

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.

Bolette Sandford Pedersen, Sanni Nimb og Lars Trap-Jensen

DanNet: udvikling og anvendelse af det danske wordnet

A wordnet for Danish is under compilation as a joint project between the Centre for Language Technology at the University of Copenhagen and the Society for Danish Language and Literature. Both partners have recently been involved in the development of relevant lexical resources which are utilized as an important part of the current project, most importantly The Danish Dictionary, a corpus-based dictionary of modern Danish, and the international SIMPLE project. This article describes how the existing data are reused to create a much sought-after resource within Danish language technology. Typical examples and problems faced during the editing process are presented, with focus on polysemy, synonymy and semantic classification. Finally, we outline some perspectives for lexicographic products aimed at human users. Specifically, the development potential for the productive function of dictionaries is discussed, as are ways of improving the pedagogical function of learners' dictionaries.

1. Indledning

Det første wordnet, Princeton WordNet (Fellbaum 1998), blev udviklet i et psykolingvistisk forskningsmiljø i et forsøg på at afbilde det mentale leksikon, altså den orden som vi antager at ordene indtager i vores bevidsthed når vi forstår og producerer sprog. Men ud over at være en god platform for psykolingvistiske eksperimenter udgør wordnets også en interessant resurse for teknologisk forskning og udvikling inden for sprogteknologi og kunstig intelligens. I disse fagmiljøer er manglen på ordsemantik i stor målestok nemlig ofte den barriere der vanskeliggør opskalering af en prototype til et stort velkørende system. Også for leksikografien er wordnets interessante; de er i grunden blot ordbøger udformet ud fra en onomasiologisk tankegang med orddefinitionen som det væsentlige omdrejningspunkt, og derfor kan de give ny inspiration til hvordan især *digitale* ordbøger kan opbygges og anvendes i fremtiden.

I denne artikel giver vi først en introduktion til DanNet-projektet og beskriver i hvilken ramme det realiseres, nemlig som et samarbejdsprojekt mellem et sprogteknologisk miljø og et leksikografisk miljø (afsnit 2). Der er i høj grad tale om genbrug af eksisterende resurser, først og fremmest i form af genbrug af *Den Danske Ordbog*. Dernæst giver vi i afsnit 3 en række eksempler fra selve udviklingsarbejdet hvor polysemi, synonymi og semantisk klassifikation udgør nogle af de typiske problemstillinger der er i fokus når en traditionel ordbog skal genanvendes i form af et wordnet. Endelig kommer vi i afsnit 4 ind på an-

vendelser af wordnettet i ordbogsammenhæng. Vi ser blandt andet på hvordan et wordnet kan anvendes til at udvikle ordbøgers *sprogproduktive* funktion, samt på hvordan de kan bruges til at forbedre learnerordbøgers *pædagogiske* funktion.

2. DanNet – et leksikalsk-semantisk wordnet for dansk

2.1. Et samarbejdsprojekt

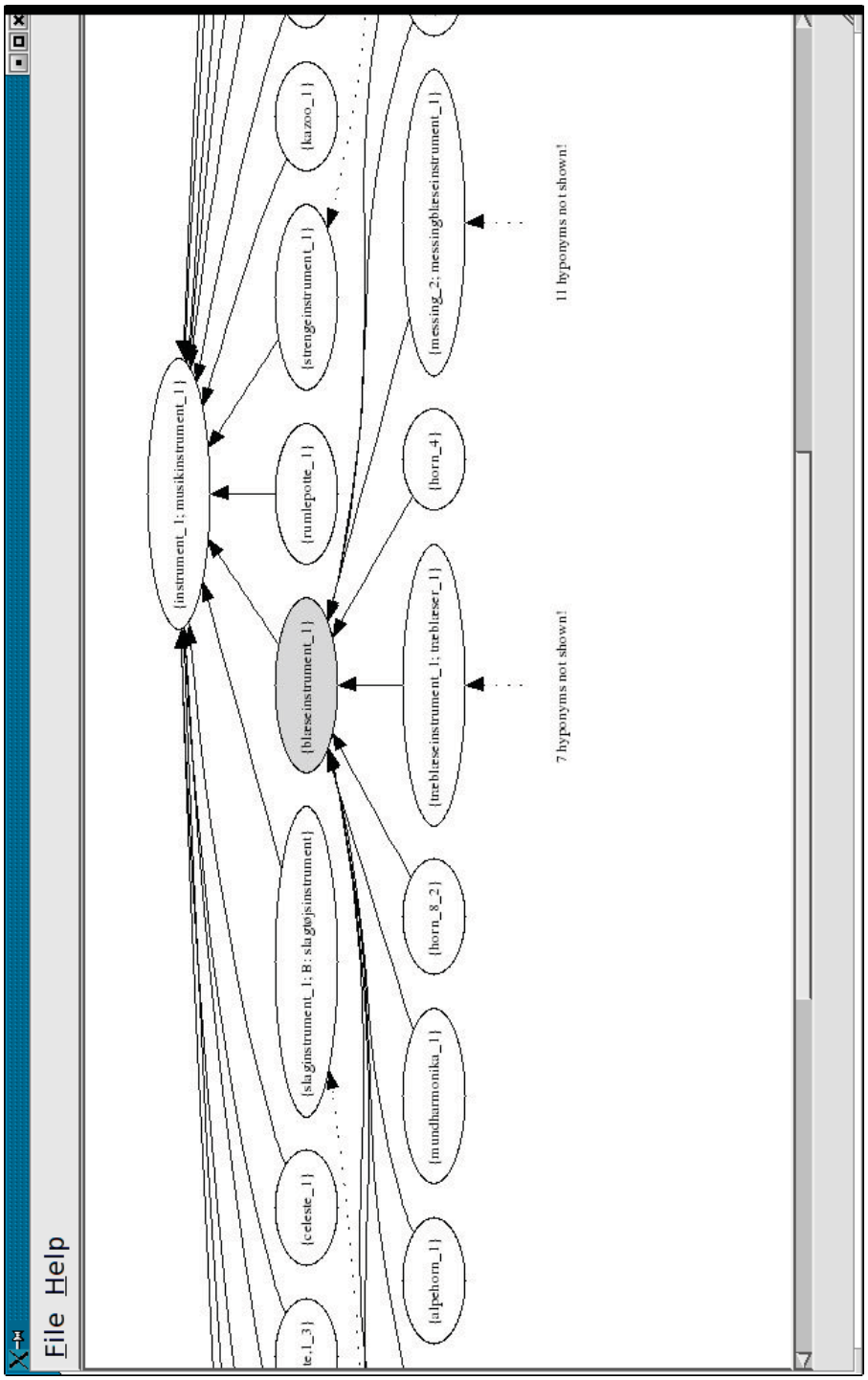
DanNet-projektet er et samarbejdsprojekt mellem Center for Sprogteknologi, Københavns Universitet (CST) og Det Danske Sprog- og Litteraturselskab (DSL). Hvert af disse to miljøer har inden for de senere år afsluttet et omfattende leksikalsk projekt hvis resultater nu tilsammen udgør et væsentligt udgangspunkt for udviklingen af det danske wordnet: *Den Danske Ordbog* (DDO) er en stor ordbog der på baggrund af korpusundersøgelser beskriver ords betydninger, primært ved hjælp af definitioner og brugseksempler (Lorentzen 2004). CST har deltaget i det internationale projekt SIMPLE (Semantic Information for Multifunctional, Plurilingual Lexica, jf. Lenci m.fl. 2000), hvor ontologiske modeller for formel betydningsbeskrivelse er blevet afprøvet og konsolideret på et bredt ordmateriale; en afprøvning på 10.000 begreber i dansk er udført i den danske del af projektet, SIMPLE-DK (Pedersen og Paggio 2004). Erfaringerne fra disse to projekter i henholdsvis det leksikografiske og det sprogteknologiske miljø har givet gode forudsætninger for at udarbejde et wordnet for dansk.

DanNet-projektets første fase er finansieret af Forskningsrådet for Kultur og Kommunikation og løber i perioden 2005–2008. Målet er i denne fase at nå op på en beskrivelse af 40.000 betydninger i dansk samt at få en delmængde af disse forbundet med det engelske Princeton WordNet via tværsproglige links. På længere sigt er det målet at nå op på en dækningsgrad der svarer til hele *Den Danske Ordbogs* ca. 100.000 betydninger. En række publikationer beskriver de forskellige stadier af projektet; jf. Pedersen & Asmussen (2006), Pedersen m.fl. (2006), Pedersen & Sørensen (2006) & Asmussen m.fl. (2007); se i øvrigt projektets hjemmeside www.wordnet.dk.

Et af målene med wordnettet er at det skal kunne anvendes i forskellige sprogteknologiske værktøjer; her tænkes især på begrebsbaseret søgning, men også på værktøjer der kan resumere og indeksere tekst på en mere intelligent måde end de gængse, mønsterbaserede værktøjer er i stand til.

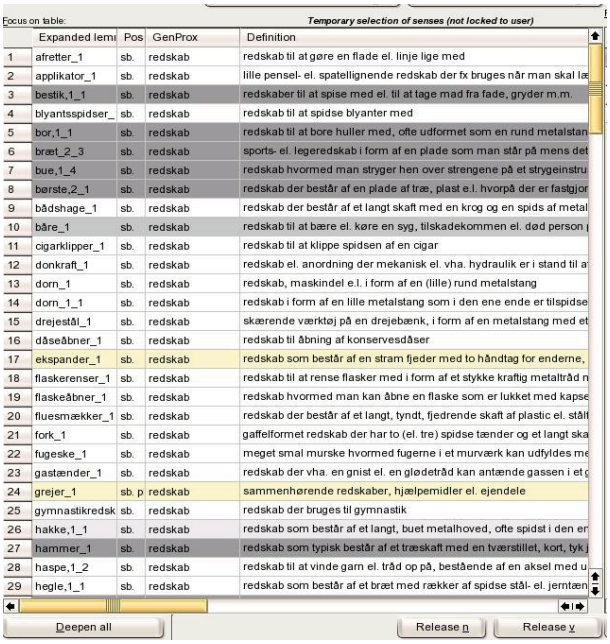
2.2. Basisbegreber der anvendes i DanNet

Som det er tilfældet i både Princeton WordNet og i EuroWordNet (Vossen 1999), anvender vi i DanNet såkaldte *synsets* som grundlæggende enheder i netværket. Et synset – eller synonymsæt – er det sæt af synonymer der tilsammen udgør et begreb. Fx kan man i figur 1 se at *slaginstrument* og *slagtøjsinstrument* optræder i samme synset og dermed opfattes som refererende til samme begreb.



Figur 1. Et wordnet er et netværk af relationer mellem begreber, eller såkaldte 'synsets'

Som udgangspunkt for den praktiske udvikling af wordnettet har vi udtrukket grupper af ord med samme overbegreb fra Den Danske Ordbog, som det fx ses i figur 2 for overbegrebet *redskab*. Herudfra bygges de over- og underbegrebshierarkier som danner wordnettets grundstruktur.



Figur 2. Genus proximum-oplysninger udtrukket fra Den Danske Ordbog

Ud over relationer mellem over- og underbegreber etableres der også en lang række andre relationer mellem de enkelte synsets. I figur 3 ses en oversigt over de relationer der tages i anvendelse. Vi har som udgangspunkt valgt at fokusere på de relationer der angives i den almindelige ordbogsdefinition; hvis det her er blevet skønnet relevant at angive hvilket materiale noget er lavet af, eller hvad noget bruges til, etableres disse relationer også i DanNet (via relationerne has_holo_madeof og for_purpose_of). Mange af relationerne nedarves desuden fra overbegreberne; for_purpose_of *spise* nedarves fx til alt spiseligt i hierarkiet.

Relationer i DanNet	**Eksempel**
concerns	*fodboldmål* concerns *sport*
for_purpose_of	*hammer* for_purpose_of *hamre*
fpo_object	*klipse* fpo_object *klips*
has_holo_madeof	*mel* has_holo_madeof *brød*
has_holo_member	*partimedlem* has_holo_member *parti*
has_holo_location	*oase* has_holo_location *ørken*
has_holo_part	*øje* has_holo_part *ansigt*
has_hyperonym	*birketræ* has_hyperonym *træ*
has_orthohyperonym	*vejtræ* has_orthohyperonym *træ*

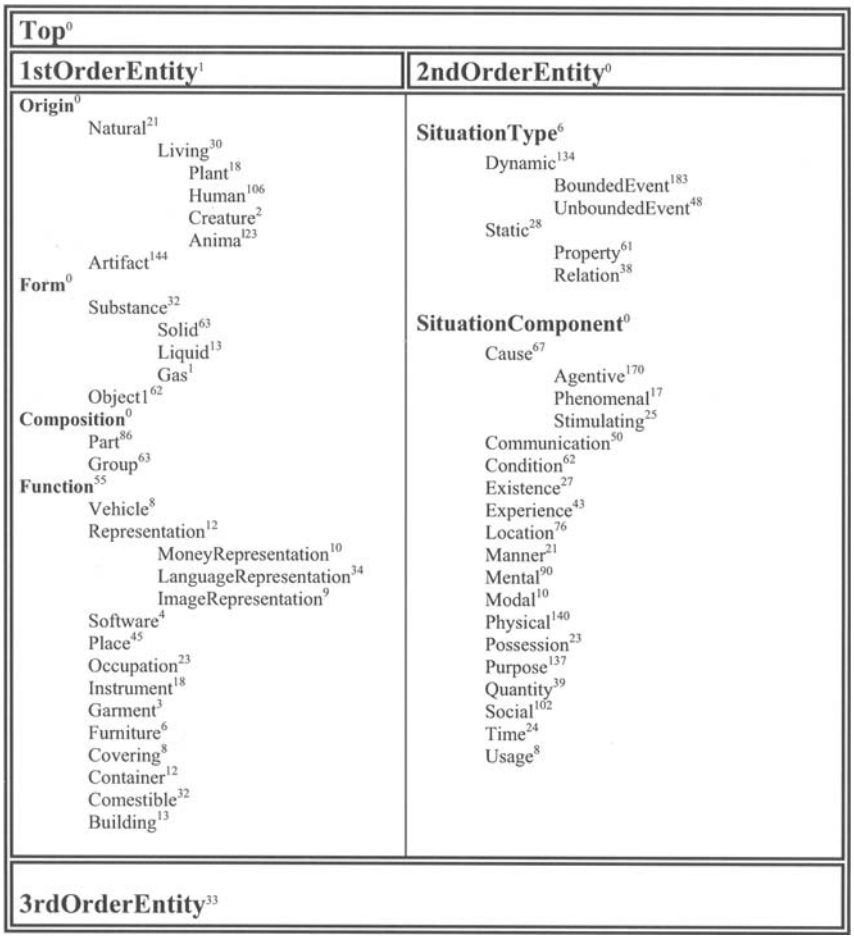
has_hyponym	*træ* has hyponym *birketræ*
has_mero_madeof	*brød* has_mero_madeof *mel*
has_mero_member	*parti* has_mero_member *partimedlem*
has_mero_part	*hånd* has_mero_part *finger*
has_mero_location	*ørken* has_mero_location *oase*
role_agent	*passager* role_agent *rejse*
role_patient	*modtager* role_patient *modtage*
made_by	*bagværk* made_by *bage*
near_synonym	*køkkentøj* near_synonym *kogegrej*
xpos_near_synonym	*behandle* xpos_near_synonym *behandling*

Figur 3. Relationer der anvendes i DanNet

Endelig forsynes alle begreber i DanNet med en top-ontologisk type. Her anvender vi EuroWordNets flerdimensionelle top-ontologi byggende på Lyons entiteter af 1., 2. og 3. grad, jf. figur 4. Fordelen ved at angive top-ontologiske typer er blandt andet at man kan identificere mængder af begreber der deler visse fælles egenskaber, men som ikke nødvendigvis er placeret under det samme danske overbegreb. Planter og dyr som mennesker spiser, er fx ikke placeret under det danske begreb *fødevarer*, men derimod under henholdsvis *planter* og *dyr*. Via den ontologiske type Comestible kan man imidlertid identificere dels tilberedte madvarer, dels de dyr og planter som mennesker typisk spiser. Man kan også have brug for at fremfinde en delmængde af de begreber der hører under et dansk overbegreb. Fx kan man udtrække alle jobfunktioner (som fx *maler* og *skolelærer*) via den ontologiske type Human+Object+Occupation. Disse er alle listet under det danske overbegreb *person*, men her finder man også *spanier*, *demokrat* og *fodgænger*.

3. Opbygningen af det danske wordnet med udgangspunkt i DDO

Selve opbygningen af det danske wordnet foregår som nævnt med udgangspunkt i DDO's betydningsinddeling og definitioner. I arbejdet udnytter vi dels at hver enkelt betydning i ordbogen kan udtrækkes som en entydigt nummereret enhed, dels at genus proximum for hver af disse betydninger kan udtrækkes direkte. Overbegrebet, der samtidig er et ord fra selve definitionen, er nemlig markeret i et selvstændigt felt i ordbogsstrukturen. I figur 5 ses et udsnit af ordbogsstrukturen for substantivet *flaskeåbner* med dets entydige betydningsnummer samt ordet *redskab* som genus proximum, markeret i feltet Genprox.



Figur 4: Top-ontologien der anvendes i DanNet (Vossen 2005)

```
flaskeåbner, sb.
<-Semdel>
   <-Semem>
     <-Denbet DanNetSemID="21020396">redskab hvormed man kan åbne en flaske som er
     lukket med kapsel el. prop
       <-Genprox>redskab
```

Figur 5. Udsnit af DDO's ordbogsstruktur med betydnings-id, definition og genus proximum

Wordnettet opbygges ved automatisk at udtrække alle de ord fra DDO der deler genus proximum, og derefter vurdere og evt. tilrette de data der fremfindes. I figur 2 ses fx et udsnit af udtrækket på overbegrebet *redskab*.

3.1. Efterredigering af DDO's genus proximum-oplysninger

Genus proximum-oplysningen fra DDO er dog langtfra altid entydig idet der i notationen hverken er skelnet mellem homografi eller et overbegrebs eventuelt flere betydninger. Derfor er der ofte tale om et større udredningsarbejde når man udtrækker et antal ord med samme overbegreb. I figur 6 ses et eksempel på hvordan samme overbegreb, i dette tilfælde *blad*, er brugt for ord af helt forskellig semantisk type, og hvor vi derfor i hvert enkelt tilfælde skal beslutte hvilken af *blad*s mange betydninger det drejer sig om.

	Expanded ler	Pos	Genf	Definition
4	basilikum_1_1	sb.	blad	friske el. tørrede blade fra denne
5	højblad_1	sb.	blad	blad som sidder højt oppe på stilk
6	billedblad_1	sb.	blad	blad med mange billeder og unde
7	koriander_1_2	sb.	blad	blade fra denne plante, brugt som
8	kirkeblad_1	sb.	blad	blad som regelmæssigt udgives til
9	klinge,1_1	sb.	blad	skarpt el. spidst blad på et (større)
10	koka_1_1	sb.	blad	blade fra denne busk der, ved tyg
11	pornoblad_1	sb.	blad	blad der især indeholder pornogra
12	bog,1_1	sb.	blad	trykte el. beskrevne blade af papi
13	timian_1_1	sb.	blad	friske el. tørrede blade fra denne
14	småblad_1	sb.	blad	lille, fint blad på en plante
15	bog,1_3	sb.	blad	indbundne el. sammenhæftede bl
16	småblad_2	sb.	blad	blad el. avis af et lille format
17	gratisblad_1	sb.	blad	blad der finansieres gennem annc

Figur 6. Ordene i venstre kolonne har alle samme genus proximum i DDO, nemlig det polyseme substantiv *blad*, og kræver derfor efterredigering

Ud over problemet med at udrede homografer og polyseme udtryk er det genus proximum der er angivet i DDO, ikke altid det systematisk set nærmeste overbegreb, idet DDO's primære formål er at formidle på den i hvert enkelt tilfælde mest hensigtsmæssige måde til ordbogsbrugeren snarere end at angive en stringent systematik inden for et semantisk felt. Fx har substantiverne *kagespade, ostehøvl, osteskærer, persillehakker, si* og *ske* det meget generelle overbegreb *redskab* i DDO, hvorimod substantiverne *dørslag, hjulpisker, hulske, hvidløgspresser, potageske, rivejern* og *æggedeler* alle har det systematisk set mere korrekte overbegreb *køkkenredskab*. I sådanne tilfælde vælger vi i DanNet at se bort fra DDO's overbegreb og markerer i stedet alle ordene som underbegreber til *køkkenredskab*, jf. figur 7.

Der er dog masser af tilfælde hvor det i høj grad kan diskuteres hvilket overbegreb der er mest passende. Fx har substantiverne *isspand, kedel, litermål, osteklokke, randform* og *saltkar* alle overbegrebet *beholder* i DDO, og substantiverne *foodprocessor, kogeapparat, sandwichrister* og *sodavandsmaskine* har overbegrebet

apparat. Men begge disse grupper af ord kunne lige så vel grupperes som køkkenredskaber på linje med substantiverne nævnt ovenfor. Det er altså meget ofte et spørgsmål om hvordan man vælger at opbygge wordnettet, og der er i rigtig mange tilfælde mere end én måde at inddele kategorierne af betydninger på. I de ovennævnte tilfælde har vi valgt at samle mange af betydningerne under det fælles overbegreb *køkkenredskab* og har derfor valgt at se bort fra DDO's genus proximum'er. I figur 7 ses nogle eksempler.

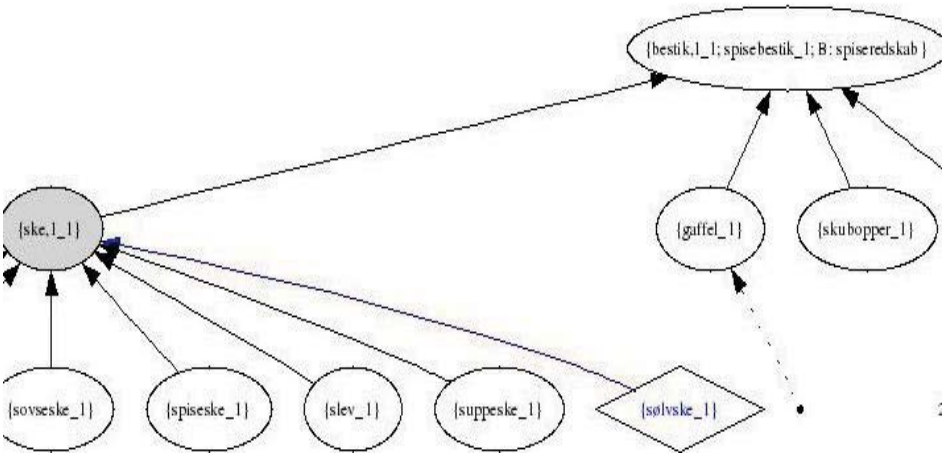
Figur 7. Udsnit af DanNet. Overbegrebet *kogegrej/køkkengrej/køkkenredskab/køkkentøj* med nogle af dets underbegreber efter færdigredigering og ændring af forskellige genus proximum'er fra DDO. Indramning viser hvor der under redigeringen af DanNet er ændret i genus proximum.

3.2. To typer af underbegreber

Et interessant problem som vi er stødt på i vores arbejde med at gruppere ordene i overbegreber og underbegreber, er de mange tilfælde hvor to ord ganske vist har samme overbegreb, men hvor de grupperer sig i forskellige slags underbegreber der er kompatible på tværs af grupperingerne, men udelukker hinanden inden for den enkelte gruppering (se også Pedersen & Sørensen 2006). Fx kan en stor gruppe af underbegreber have det tilfælles at de alle er et bestemt eksempel på overbegrebet med særlige fysiske kendetegn. Overbegrebet *træ* har således en masse træsorter som systematiske (taksonomiske) underbegreber, fx *bøgetræ*, *el* og *gran*. Et eksempel på et underbegreb til *træ* der går på tværs af denne store gruppe af taksonomiske underbegreber, er *vejtræ*. Et vejtræ er nemlig ikke en bestemt slags træ, men snarere et hvilket som helst træ der står langs en vej. Begrebet betegner i dette tilfælde funktionen af træet og ikke en specifik type træ

med særlige karakteristika, og et vejtræ kan sagtens også være et bøgetræ, hvorimod et grantræ ikke samtidig kan være et bøgetræ. Hvis vi betragter figur 8, ser vi et andet eksempel på det samme fænomen, her ved en række underbegreber til *ske*. De fleste underbegreber angiver i dette tilfælde i stedet hvilken **funktion** den pågældende ske har, men en af sketyperne skiller sig ud, nemlig *sølvske,* der i stedet betegner hvilket **materiale** skeen er lavet af. En sølvske kan godt også være en suppeske, hvorimod fx en dessertske ikke samtidig kan være en suppeske. I arbejdet med opbygningen af DanNet har vi valgt at skelne mellem den taksonomiske eller systematiske gruppe af underbegreber til et bestemt overbegreb og de underbegreber der går på tværs af den systematiske gruppe. De første er umarkerede; de sidste markerer vi med et særligt træk 'orthogonal' i databasen, visualiseret med en rombe i figur 9.

Figur 8. Forskellige typer af underbegreber til ske. Der skelnes mellem to typer overbegrebsrelationer, og *sølvske* markeres derfor efterfølgende med et særligt træk, jf. figur 9.



Figur 9. Sølvske markeret som en særlig type underbegreb til *ske*

For de konkrete substantivers vedkommende er denne skelnen mellem en systematisk gruppe af undergreber og en gruppe der adskiller sig fra denne, forholdsvis uproblematisk. Det er straks mere kompliceret for verbernes og verbalsubstantivernes vedkommende, og et af de fremtidige forskningsområder i DanNet er da også at undersøge om man kan opstille lignende skel mellem disse ordgruppers typer af forskellige underbegreber.

3.3. Synonymi i DanNet

En sidste meget vigtig ting der skal tages stilling til i den konkrete opbygning af wordnettet ud fra DDO's betydninger, er synonymi: Hvornår skal betydningerne fra DDO slås sammen i et enkelt synset (synonym-sæt)? Vi arbejder i wordnet med et mere udvidet synonymi-begreb end man gør i DDO, og der er derfor en lang række tilfælde hvor begreber er synonymer i DanNet, men ikke i DDO. Fx ser man i DanNet helt bort fra ordenes valør og opfatter ord som *bil* og *spand* som synonymer. Derudover har vi følgende retningslinjer: Hvis ordene er synonymer i DDO (markeret med SYN i den trykte ordbog), tilhører de samme synset i DanNet. Derudover kan de, hvis de har næsten enslydende definitioner i DDO og fx er angivet som nærsynonymer – i den trykte ordbog angivet vha. JF ('jævnfør') – også komme i samme synset i DanNet. Et eksempel på dette er substantiverne *køkkengrej*, *kogegrej*, *køkkenredskab* og *køkkentøj*, se figur 7 øverst. De har nemlig i DDO næsten ens betydningsbeskrivelser og er for nogles vedkommende angivet som nærsynonymer til hinanden:

> **køkkengrej**: redskaber der bruges ved madlavning JF husgeråd, køkkentøj
> **kogegrej**: redskaber til at lave mad med i et køkken
> **køkkenredskab**: redskab der bruges til madlavning JF køkkengrej
> **køkkentøj**: redskaber der bruges til madlavning og ved spisning SYN køkkenting, JF køkkengrej.

4. Anvendelser i ordbogssammenhæng

Wordnets bruges primært i sprogteknologisk sammenhæng og har bl.a. fundet anvendelse i forbindelse med forbedring af emnesøgning i søgemaskiner og til udvikling af internettets næste fase, det såkaldte Semantic Web. Men også inden for ordbøger til menneskelige brugere rummer det visse interessante muligheder. Vi vil her pege på tre områder hvor vi kan se et leksikografisk perspektiv:

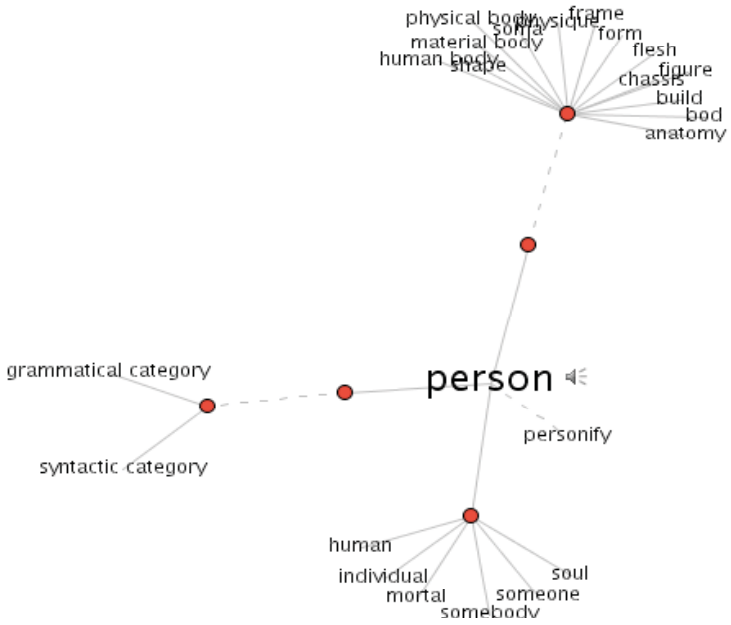
1. udvikling af ordbøgers sprogproduktive funktion
2. forbedring af learnerordbøgers pædagogiske funktioner
3. underholdningsfunktion, især krydsordshjælp

Man kunne hertil føje et fjerde område, viden om sproget, men fordi det er så generelt et formål at det næsten altid kan bringes i anvendelse, vil det ikke få selvstændig behandling.

4.1. Sprogproduktion: nuancer og variation

En leksikografisk resurse hvor ordforrådet ordnes efter dets begrebsindhold (onomasiologisk) frem for efter dets benævnelser (semasiologisk), er ikke nogen ny idé. Eksemplet par excellence er P.M. Rogets *Thesaurus of English Words and Phrases* fra 1852, og siden er der udsendt mange lignende begrebsordbøger, ikke mindst inden for fagsproglige områder. DanNet har meget tilfælles med begrebsordbøgerne: den systematiske ordning af ordforrådet, den onomasiologiske tilgang og vægtningen af relationer mellem begreberne i synonymer, antonymer, meronymer, hyperonymer og hyponymer. Ganske vist er kodningen i et wordnet foretaget med henblik på sprogteknologisk brug, dvs. med computere som "målgruppe", men med en vis redaktionel bearbejdning vil DanNet også kunne finde anvendelse som en selvstændig ordbog for menneskelige brugere. På internettet kan man også se eksempler på hvordan man har tilgængeliggjort wordnets for menneskelige brugere ved at knytte en særlig brugergrænseflade til resursen. Der er selvfølgelig en sammenhæng mellem graden af bearbejdning til menneskelige brugere, grænsefladen og kvaliteten af det leksikografiske produkt; eksempler kan ses på http://wordnet.princeton.edu/perl/webwn og www.online-thesaurus.net, mens en grænseflade med dansk hjælpetekst findes på www.onlineordbog.dk/wordnet/da. Sidstnævnte tilbyder også norsk og engelsk hjælpetekst.

En anden mulighed er at udnytte wordnettet ved at knytte det sammen med en anden ordbog. I Det Danske Sprog- og Litteraturselskabs projekt *ordnet.dk* vil DanNets oplysninger med tiden blive integreret for at kunne forbedre visse funktioner (jf. Lorentzen & Trap-Jensen 2006). Det er ikke mindst nærliggende fordi DanNet i forvejen henter sine oplysninger fra *Den Danske Ordbog*, som indgår i *ordnet.dk*. I *Den Danske Ordbog* bruges som nævnt – og bevidst – et snævert synonymibegreb, men et bredere udvalg af muligheder kan være en fordel for den der har brug for overblik over sprogets nuancer og variationsmuligheder i forbindelse med tekstproduktion (jf. Bergenholtz & Vrang 2005, Trap-Jensen 2005b). DanNets kodning af synonymer og nærsynonymer udgør netop et sådant bredere udvalg, og det kan man udnytte ved at gøre disse oplysninger tilgængelige i ordbogen, helst som en yderligere, klikbar mulighed ("se et større udvalg"), hvorved man kan bevare fordelen af både det snævre sæt synonymer (til receptions- og vidensformål, også for ikke-modersmålsbrugere) og det bredere sæt (til tekstproduktion, for brugere med modersmålslignende kompetence). På tilsvarende vis kan der ved en given betydning oplyses om alle øvrige relationer til betydninger der er registreret i DanNet.



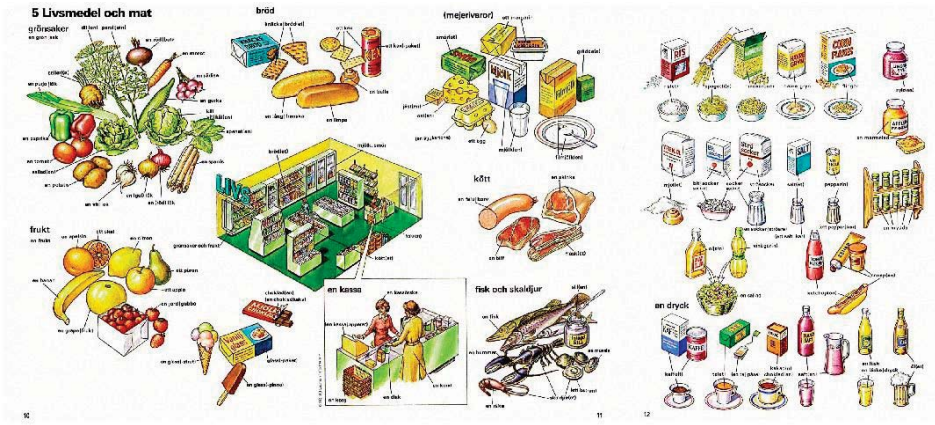
Figur 10. Visuel fremstilling af søgeordet *person* fra www.visualthesaurus.com

Brugen af DanNet som begrebsordbog kan også tilbydes som en selvstændig søgefacilitet inden for *ordnet.dk*. Muligheden for at foretage begrebsorienterede søgninger i DDO på nettet har lige fra begyndelsen været en vigtig anledning for DSL til at involvere sig i projektet (jf. Trap-Jensen 2005a). Det der kræves for at gøre det muligt, er dels at der oprettes et særligt felt til søgning på begreber, dels at der udvikles en grænseflade der kan vise resultatet på en hensigtsmæssig måde. En udfordring her er især hvordan man præsenterer meget store søgeresultater. En søgning som "vis alle ord der betegner en *person*" giver fx flere tusind resultater som det kan være svært at få overblik over. Det problem kan man forestille sig løst dels ved at folde hyponymgrupper sammen så de kun vises i deres helhed når man klikker på dem (fx 'mand', 'kvinde', 'idrætsudøver', 'kunstner'), dels ved at ordne de resterende forekomster i meningsfulde grupper efter karakteristiske fællestræk. En metode til automatisk gruppering efter et sådant princip (kaldet *feature detection method*) er beskrevet i Asmussen (2004).

Endelig kan man hjælpe brugeren til at få overblik over et betydningsområde ved at give en visuel fremstilling af det. Også det kan man finde eksempler på i eksisterende netpublikationer. I figur 10 ses et eksempel fra www.visualthesaurus.com.

4.2. Pædagogiske anvendelser

Mange elektroniske learnerordbøger har efterhånden ud over opslagsdelen også en afdeling med forskellige ordøvelser, grammatik og systematisk glosetræning. Hvis ordbogen publiceres som cd-rom, er der desuden mulighed for at lagre oplysninger lokalt på brugerens harddisk, og man kan derfor operere med en personlig tilpasning af ordbogen så brugeren kan afstemme den efter sit eget behov og niveau, markere bestemte ord som særlig vigtige, tilføje nye artikler osv. Fordi et wordnet til en vis grad også opdeler det sproglige felt i systematiske betydningsområder, kan det være interessant som leverandør af input til den slags betydningsoversigter. Det kræver naturligvis en videre forarbejdning så stoffet svarer til målgruppen. Man kan også peppe præsentationen op med billeder, lyd eller animationer. Et eksempel med billeder er vist i figur 11.



Figur 11. Oversigt over levnedsmidler fra svensk Lexin

Tilsvarende oversigter, men i mindre format, kan man give under de enkelte betydninger. I læringsmæssig sammenhæng er det vigtigt at man ikke nøjes med at tilegne sig det enkelte ord og dets betydning i isolation, men også er opmærksom på hvordan det adskiller sig fra andre betegnelser inden for samme betydningsområde. Det kan man gøre ved at tilbyde en mulighed for at se en samlet oversigt over ko-hyponymer, altså ord eller betydninger som har samme overbegreb, fx strengeinstrumenter, køkkenredskaber, ordklasser eller sygdomme.

4.3. Underholdning – krydsordshjælp

En hjemmeside der tilbyder en netordbog, er i konkurrence med alle andre hjemmesider på internettet og er derfor underlagt de samme betingelser som gælder

for nettet generelt. Hvis man vil fange og fastholde opmærksomheden hos netbrugere der blot kommer forbi siden mere eller mindre tilfældigt, må man benytte sig af de midler der egner sig til formålet. Dertil hører forskellige former for underholdning og spil. En oplagt anvendelsesmulighed er at bruge DanNet i kombination med *Den Danske Ordbog* som hjælp til krydsordsløsere. Man kan allerede finde forskellige former for krydsordshjælp på nettet, men ofte er de ikke særlig raffinerede. Som regel fungerer de ved at brugeren kan udfylde de dele af ordet han/hun kender i forvejen, og hjælpefaciliteten undersøger derefter hvilke ord i basen der matcher den indtastede streng. Søgefeltet kan være opdelt i felter som i krydsordsforlægget eller være et almindeligt søgefelt hvor forskellige jokertegn erstatter ét eller flere bogstaver.

Når denne hjælp ikke forekommer særlig avanceret, skyldes det dels at man ikke har mulighed for at angive om ordet optræder i bøjet form, dels at man ikke kan begrænse antallet af søgeresultater ved at oplyse hvilken type ledeord der er tale om. Enhver krydsordsløser ved at begge dele er helt afgørende når man løser krydsord. Det vil derfor være en stor forbedring hvis man kan supplere krydsordshjælpen med ordbogens oplysninger om ordenes bøjningsformer og wordnettets opdeling af ordforrådet i begrebsområder. Et enkelt eksempel kan illustrere hvordan.

Lad os antage at man er gået i stå i sin krydsord ved et ord på seks bogstaver. De eneste oplysninger man har, er at det slutter på *s*, og at ledeordet er "sport". Taster man de oplysninger ind i den krydsordshjælp der findes hos forskellige eksisterende udbydere, giver det fx hos Caplex 2244 resultater (se figur 12) og hos Nationalencyklopedins internettjänst 1600. Den Danske Online Ordbog oplyser ikke antallet af resultater, men viser den første side med 50 forekomster. Ved at bladre frem kan man finde ud af at der er i alt 14 sider resultater, dvs. ca. 700 forekomster. Det siger sig selv at det er alt for mange og alt for uoverskueligt et antal resultater. En søgning i DanNet viser derimod hurtigt at der kun er to sportsgrene der opfylder kriteriet: *isdans* og *tennis* (evt. *diskos* opfattet som en disciplin).

DanNet kan således bidrage væsentligt til forbedring af krydsordshjælp – hvad enten det er en integreret hjælp som i eksemplerne her eller ved udarbejdelse af et selvstændigt krydsordsleksikon, som jo traditionelt også indeholder systematisk ordnede oversigter med ordene fordelt efter antallet af bogstaver. Selvom der kan opnås en betydelig teknisk forbedring af krydsordshjælpen med resurser som DanNet og DDO, bør det dog retfærdigvis understreges at den underliggende base i andre henseender er mangelfuld. Det gælder ikke mindst med hensyn til navnestof og andet encyklopædisk materiale, som ofte udgør en stor del af krydsordenes ordforråd. Det er også grunden til at man især finder funktionen hos forlag der råder over både encyklopædi og ordbøger.

Figur 12. Et eksempel på krydsordshjælp fra www.caplex.no

5. Konklusion

I denne artikel har vi beskrevet hvordan vi udvikler det danske wordnet på baggrund af eksisterende ordbogsdata, og hvilke lingvistiske problemstillinger vi støder på når vi forsøger at opbygge så stringent en netværksstruktur over det danske ordforråd som muligt. Vi har foreslået anvendelser af wordnettet i en almenleksikografisk sammenhæng i et forsøg på at anvise nogle fremtidige, lidt alternative muligheder for brug af et dansk wordnet.

Den færdige DanNet-resurse vil forhåbentlig være med til at sikre at datalingvistiske og leksikografiske miljøer i Danmark kan udvikle bedre sprogteknologiske værktøjer. På længere sigt kan integrationen af wordnettet med andre allerede eksisterende danske ordbogsresurser og korpusser føre til en yderligere berigelse af wordnettet således at fx oplysninger om ordenes syntaktiske

egenskaber samt korpuseksempler knyttes til de enkelte synsets.

Litteratur

Ordbøger m.m.

Caplex: www.caplex.no. Oslo.

DanNet: www.wordnet.dk. Center for Sprogteknologi og Det Danske Sprog- og Literaturselskap 2005–2008. København.

DDO = *Den Danske Ordbog. Bind 1-6*. Udgivet af Det Danske Sprog- og Litteraturselskab. Hovedredaktører: Ebba Hjorth og Kjeld Kristensen. København: Gyldendal 2003-2005.

Den Danske Online Ordbog: www.ddoo.dk

Lexin: http://lexin.nada.kth.se/lexin.html. Stockholm.

Nationalencyklopedins internettjänst: www.ne.se. Göteborg.

SIMPLE: http://cst.dk/simple. København.

Anden litteratur:

Asmussen, Jørg 2004: Feature Detection – A Tool for Unifying Dictionary Definitions. I: Williams, Geoffrey and Vessier, Sandra: *Proceedings of the 11th EURALEX International Congress.* Lorient, 63–69.

Asmussen, Jørg, Bolette S. Pedersen & Lars Trap-Jensen 2007: DanNet: From Dictionary to WordNet. I: Kunze, Claudia, Lemnitzer, Lothar & Osswald, Rainer (eds.): *GLDV-2007 Workshop on Lexical-Semantic and Ontological Resources.* Universität Tübingen, 1–11.

Bergenholtz, Henning, & Vibeke Vrang 2005: Den Danske Ordbog bind 2 (E–H) og 3 (I–L) – en ordbog for folket eller for akademikere? I: *LexicoNordica 12,* 169–187.

Fellbaum, Christiane (red.) 1998: *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Lenci, Alessandro, Nuria Bel, Federica Busa, Nicoletta Calzolari, Elisabetta Gola, Monica Monachini, Antoine Ogonowski, Ivonne Peters, Wim Peters, Nilda Ruimy, Marta Villegas & Antonio Zampolli 2000: SIMPLE – A General Framework for the Development of Multilingual Lexicons. I: *International Journal of Lexicography 13,* 249–263.

Lorentzen, Henrik 2004: The Danish Dictionary at large: presentation, problems and perspectives. I: *Proceedings of the 11th EURALEX International Congress.* Lorient, 285–294.

Lorentzen, Henrik & Lars Trap-Jensen 2006: *ordnet.dk* – et nyt sprogligt opslagsværk på internettet. I: *Nordiske Studier i Leksikografi, NFL-skrift nr. 9*. København, 253–264.

Pedersen, Bolette S. & Jørg Asmussen 2006: DanNet – Fra ordbog til et leksikalsk-semantisk WordNet for dansk. I: *LEDA-Nyt 42,* 3–12.

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai H. Sørensen, Lars Trap-Jensen & Henrik Lorentzen 2006: DanNet – a WordNet for Danish. I: *Proceedings of the Third International WordNet Conference.* Jeju, 329–331.

Pedersen, Bolette S. & Patrizia Paggio 2004: The Danish SIMPLE Lexicon and its Application in Content-based Querying. I: *Nordic Journal of Linguistics 27(1)*, 97–127.

Pedersen, Bolette S. & Nicolai H. Sørensen 2006: Towards Sounder Taxonomies in Wordnets. I: Oltramari, Alessandro, Chu-Ren Huang, Alessandro Lenci, Paul Buuitelaar og Christiane Fellbaum (eds.): *Ontolex 2006* at 5th International Conference on Language Resources and Evaluation. Genova, 9–16

Trap-Jensen, Lars 2005a: Virtuelle perspektiver for ordbogsredigering: muligheder, strategier og virkelighedens begrænsning. I: *LexicoNordica 12,* 109–122.

Trap-Jensen, Lars 2005b: Kommentar til Henning Bergenholtz og Vibeke Vrang: Den Danske Ordbog bind 2 (E-H) og 3 (I-L) – en ordbog for folket eller for akademikere? I: *LexicoNordica 12*, 189–198.

Vossen, Piek (red.) 1999: *EuroWordNet. A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers.

Vossen, Piek (red.) 2005: *EuroWordNet General Document.* University of Amsterdam.

Sanni Nimb
seniorredaktør, f. 1961
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK–1219 København K
sn@dsl.dk

Bolette Sandford Pedersen
seniorforsker, f. 1962
Center for Sprogteknologi, Københavns Universitet
Njalsgade 80
DK–2300 København S
bspedersen@hum.ku.dk

Lars Trap-Jensen
ledende redaktør, f. 1960
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK–1219 København K
ltj@dsl.dk