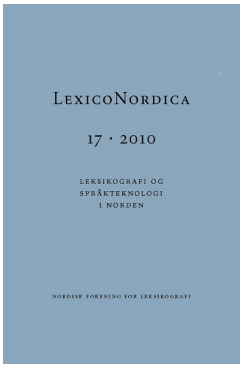
LexicoNordica

Titel: Semantiske sprogressourcer - mellem sprogteknologi og leksikografi

Forfatter: Bolette Sandford Pedersen

Kilde: LexicoNordica 17, 2010, s.163-180

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



Søgbarhed
Artiklerne i de ældre LexicoNordica (1-16) 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.

Semantiske sprogressourcer – mellem sprogteknologi og leksikografi

Bolette Sandford Pedersen

This paper discusses the synergy between lexicography and semantic language resources meant for computational use; before, now and in the future. On the basis of a brief historical overview of the backgrounds for language technology and lexicography, respectively, I analyze why the two fields have not always cooperated as closely as one would think useful. I give a recent example of a project that has exploited the similarities between the two fields by reusing a monolingual dictionary for the compilation of a Danish wordnet for technological use: DanNet. I describe some areas where modifications have been necessary in the reuse process; this regards in particular the adjustment of hyponymy hierarchies and the spelling-out of underspecified information. I conclude that the two fields will most presumably be much more connected in the future due to recent corpus and editing tools which help exploit more radically the intersection between the two areas.

1. Indledning

I udgangspunktet må det være helt naturligt at sprogteknologien straks vender sig mod leksikografien når der opstår behov for leksikografiske data i større mængder. Selvom almindelige ordbøger er beregnet til mennesker mens maskinanvendelige ordbaser har computeren som mellemled i form af et program der udnytter dem, må mængden af fællesviden være stor. I denne artikel fokuserer jeg på semantisk viden og på i hvor høj grad der er synergi mellem den semantiske viden vi udtrykker ved hjælp af definitioner og brugseksempler i menneskeordbøger, og den semantik vi udtrykker i maskinanvendelige ordbaser i form af unifikationsba-

serede netværk. Mit udgangspunkt for denne redegørelse er DanNet, det leksikalsk-semantiske wordnet som vi på Københavns Universitet og Det Danske Sprog- og Litteraturselskab har udviklet i årene 2005-2010 og som kan downloades som open source fra wordnet.dk (se også Pedersen et al. 2009)[1].

I udviklingen af denne leksikalske ressource har vi i høj grad forsøgt at udnytte den formodede fællesmængde af viden i de to typer ressourcer idet vi har tilstræbt genanvendelse af definitioner og oplysninger om nærmeste danske overbegreb fra Den Danske Ordbog (DDO), som er en moderne, korpusbaseret ordbog for dansk.

Artiklen indledes med et historisk tilbageblik på sprogteknologiens udspring i det formelle sprogsyn med det formål at afdække hvorfor synergien mellem leksikografi og sprogteknologi ikke altid har været åbenbar. Jeg refererer bl.a. til en artikel af Ide & Véronis (1995) som har fået en vis betydning for synet på genanvendelse af leksikografiske data til sprogteknologiske formål, og jeg ser bl.a på Pustejovskys generative leksikon som et moderne bud på hvordan leksikografiske data behandles i den formelle sprogtradition.

Dernæst beskrives DanNets tilblivelse med udgangspunkt i DDO. Der fokuseres på to typer af justeringer som har været nødvendige for at sikre at den semantiske viden fra DDO får en passende form i et formelt baseret wordnet: Tilpasning af inkonsistente eller underspecificerede hyponymier samt tilføjelse eller eksplicitering af manglende information. Endelig beskrives i sidste afsnit nogle anvendelsesperspektiver for semantiske sprogressourcer.

1 DanNet-projektet er støttet af Statens Humanistiske Forskningsråd, og ressourcen videreudvikles nu under DK-CLARIN-projektet støttet af Forskningsrådet for Kultur og Kommunikation.

2. Sprogteknologiens udspring i det formelle sprogsyn

I modsætning til leksikografien, som udspringer af et beskrivende sprogsyn forankret i en humanistisk tradition, så udspringer datalingvistik/sprogteknologi og den deraf afledte datamatiske leksikografi i højere grad af det såkaldt formelle sprogsyn. Et karakteristikum for det formelle sprogsyn er at man her anser sproget som et produktionssystem som fx udtrykt i Chomskys generative grammatikker. Disse systemer har nær tilknytning til matematikkens og logikkens udvikling af formelle, logiske systemer, og med den hastige udvikling af computeren i sidste halvdel af forrige århundrede blev denne et håndgribeligt udtryk for hvordan man kan formalisere sprog og tænkning. Man kan altså sige at udgangspunktet her i højere grad er naturvidenskabeligt eller i hvert fald placerer sig i et grænseområde mellem datalogi og naturvidenskab på den ene side og humaniora på den anden (jf. bl.a. Prebensen 2006 og Spang-Hanssen 2006). Traditionelt har (datamatisk) leksikografi ikke spillet nogen stor rolle i de formelle teorier; først med de leksikalistiske teoriers fremkomst i firserne (Kaplan & Bresnan 1982, Pollard & Sag 1989) kom ordforrådet og dets informationer i fokus. Trenden gik nu fra at betragte ordbogen som et mere eller mindre uinteressant appendiks til den formelle grammatik til at indeholde langt mere strukturel information om fx valens, argumentstruktur og senere selektionsrestriktioner.

2.1. Semantikken i det formelle sprogsyn

Hvis vi ser mere specifikt på semantikken i den formelle sprogtradition, bliver det også klart at interesseområderne inden for dette felt relaterer sig stærkt til logikken og de genstande som typisk behandles her. Cann (1993) udtrykker fx at det at kende betydningen

af en ytring er at forstå dens sandhedsforudsætninger. Ytringers sandhedsværdi er et centralt parameter i formel semantik sammen med aspekter som *referens* (fx pronominel referens), *virkefelt* (fx negationers virkefelt) og *kausalitet* (fx mellem hoved- og bisætning når de forbindes med forskellige konjunktioner). Interesseobjekter er med andre ord typisk de elementer i sproget der har en parallel til logikkens elementer; det vil primært sige funktionsordene, så som determinatorer og ubestemte pronominer der minder om kvantorer i logikken, samt konjunktioner (*og, eller, hvis, så*) og visse adverbier. Indholdsordene, derimod, som er leksikografiens hovedfelt, altså især substantiver, verber og adjektiver, har derimod været stedmoderligt behandlet i den formelle semantik; oftest blot repræsenteret ved hjælp af det underspecificerede semantiske mærke (') placeret efter et indholdsord som reference til dets indhold[2].

Vel hjulpet af Partee som i høj grad dannede bro imellem formel og leksikalsk semantik, indledte Boguraev & Briscoe (1989), Copestake & Briscoe (1995), Calzolari (1988) m.fl. i firserne og begyndelsen af halvfemserne en ny æra i og med at de begyndte at opstille formelle, unifikationsbaserede teorier for en leksikalsk semantik der havde indholdsordene i centrum og kunne interagere med andre teorier inden for formel syntaks. AQUILEX-projektet (jf. fx Copestake 1990) var et EU-forskningsprojekt der fik stor betydning i denne sammenhæng idet det var et af de første projekter som samtidig med teoriudvikling reelt påbegyndte udviklingen af større leksikalske ressourcer med semantisk information til maskinel anvendelse.

Pustejovskys formelle beskrivelsesapparat til håndtering af leksikalsk semantik (1995) fik også stor gennemslagskraft. Det

2 Fx som hos Cann (1993:38): *poison' (ethel', the_cat')* hvor der primært fokuseres på relationen mellem prædikat og argumenter mens verbets og substantivernes øvrige betydningskomponenter underspecificeres.

blev udarbejdet med henblik på dels at kunne håndtere forskellige betydningsdimensioner og opløse flertydighed, dels at kunne beregne semantisk nærhed og afhængighedsrelationer imellem begreber. Pustejovsky omtaler ordenes kernesemantik som "qualiastruktur": en struktur der indbefatter dimensioner som *form*, *oprindelse* og *formål*. Denne kernesemantik menes at interagere med en række generative mekanismer som således forklarer hvordan begreber tager farve af hinanden i kontekst. Når vi fx omtaler en *bil* som *hurtig*; refererer vi til dens køreegenskaber, mens når vi omtaler en *hurtig læser* refererer til læsehastighed. Dette fænomen omtales som selektiv binding; egenskaben *hurtig* "bindes til" kernesemantikken i det substantiv det lægger sig til som for substantiverne *bil* og *læser* indbefatter hhv. *køre* og *læse*.

Pustejovskys teorier dannede udgangspunkt for det lidt senere EU-projekt SIMPLE, som implementerede centrale dele af qualiastrukturen (Lenci et al. 2000). SIMPLE havde en dansk (se Pedersen & Paggio 2004) og en svensk aflægger (Kokkinakis et al. 2000) og senere en norsk, og den meget omfattende semantiske struktur i dette projekt dannede i høj grad forbillede for DanNet, som indbefatter aspekter fra SIMPLE der ikke indgår i traditionelle wordnets (bl.a. elementer fra qualiastrukturen).

2.2. Den leksikalske semantik og det mentale leksikon

Nogenlunde samtidig med denne udvikling inden for det formelle sprogparadigme påbegyndte Miller og Fellbaum (Fellbaum 1998) deres psykolingvistiske eksperimenter omkring det mentale leksikon i form af wordnets. Med udgangspunkt i to sprogpsykologiske hypoteser, nemlig adskillelseshypotesen og mønsterhypotesen, påbegyndte man udviklingen af det såkaldte Princeton WordNet, som er et netværk af begreber lagret i en databasestruktur. Adskillelseshypotesen går ud fra en antagelse om at det mentale leksikon har selvstændig status i forhold til fx vores evne til at producere

grammatisk korrekte sætninger, og at det derfor giver mening at studere ordforrådet, altså det mentale leksikon, uafhængigt af vores øvrige sprogevne, fx vores grammatiske. Mønsterhypotesen bygger på antagelsen om at menneskers viden om ords betydninger er lagret i form af netværk med indbyrdes semantiske relationer imellem begreber. Wordnets med sådanne semantiske relationer – eller leksikalsk-semantiske net som de også kaldes – er efterfølgende blevet en stor succes inden for især sprogteknologien og udvikles nu for en lang række sprog inklusive flere af de nordiske.

2.3. Hvor kommer leksikografien ind?

Som allerede nævnt kan man undre sig over at almindelige menneskeordbøger og den semantik de indeholder, ikke fra starten er blevet anvendt langt mere når der skulle bygges datamatiske ordbaser og leksikalske net til datamatisk anvendelse. Ordbaserne har i lang tid udgjort en flaskehals i sprogteknologiske applikationer bl.a. fordi deres dækningsgrad langtfra har været god nok, og hvad var mere naturligt end at genbruge fra andre, store ordressourcer?

Selvom Boguraev & Briscoe (1989) allerede i slutningen af firserne beskriver muligheder for genbrug af almindelige ordbøger, og selvom genbrugsaspektet indgår i den første fase af AQUILEX-projektet, så evalueres genbrugsaspektet i 1995 negativt af Ide & Véronis. Deres vurdering er relativt nedslående, og denne evaluering har muligvis påvirket udviklingen negativt da der er tale om et af de eneste bredere studier af området fra den periode. Forfatterne konkluderer bl.a. at den information man kan udlede fra ordbøger, er for inkonsistent og usystematisk til at det kan betale sig at forsøge at udtrække den[3]. Udtrækning af hierarkier og se-

3 Det er værd at bemærke at selvom ordbogsproduktion generelt har været inde i en rivende udvikling bl.a. på grund af moderne editerings- og korpusværktøjer, og ordbøgerne således er blevet langt mere konsistente

mantiske relationer giver altså et for ujævnt resultat som kræver for meget efterredigering efter deres skøn, og de foreslår derfor at man i stedet anvender viden fra korpora, som da også var baggrundsressourcen for den anden fase af Aquilex-projektet. Meget semantisk information er også underforstået i ordbøger fordi man regner med sprogbrugerens forhåndsviden, og den eksplicitering som er nødvendig for computeranvendelse er blevet anset for at være for omfattende til at genbrug kunne betale sig.

En anden, relateret diskussion går på i hvor høj grad de betydningsdistinktioner der ses i almindelige ordbøger, kan genanvendes til maskinelle formål, som beskrevet i Kilgarriff (1997) og senere i Ide & Wilks (2007). Er de for finkornede? Er de korpusbaserede i tilstrækkelig grad? Bør betydningsdistinktioner beskrives på en helt anden måde end ved at opsplitte i betydninger og underbetydninger? Diskussionerne er mange, og også i wordnetkredse er der en livlig debat om hvorvidt det kan betale sig at tage udgangspunkt i almindelige, monolingvale ordbøger når man skal udvikle wordnets for nye sprog, eller om man hellere skal oversætte Princeton WordNet til de pågældende sprog for derefter at efterjustere monolingvalt. På NFL's årlige symposium på Schæffergården i 2010 fremgik det at der i øjeblikket udvikles wordnets for de nordiske sprog med begge metoder. I det følgende afsnit beskriver jeg fordele og ulemper ved anvendelse af den monolingvale metode, som er den vi har valgt for udviklingen af DanNet.

3. Genbrug af en monolingval ordbog i sprogteknologisk sammenhæng

Selvom det har været en stor hjælp og altså efter vores bedste overbevisning besværet værd at genbruge en eksisterende ordbog, skal

end tidligere, så vil der være tale om en vis forsinkelse inden dette slår igennem inden for sprogteknologien.

der ikke lægges skjul på at der har været behov for en del tilpasninger undervejs i stil med dem som andre allerede har erfaret, og som er kort skitseret ovenfor. I det følgende vil jeg opridse nogle af de vigtigste justeringer og kompletteringer.

3.1. Tilpasning af inkonsistente eller underspecificerede hierarkier

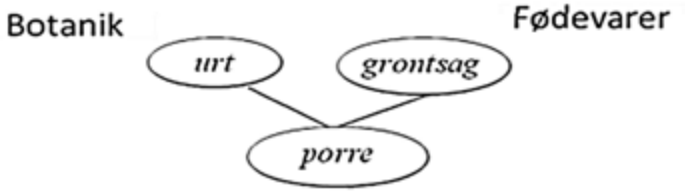
Det vigtigste genbrugsaspekt i udviklingen af DanNet har været brug af genus proximum-oplysninger i DDO til skabelsen af den overordnede struktur i det semantiske net. Hvis vi fx ser på et semantisk område som frugt og grønt vil man se at udgangspunktet i DDO er noget forskelligartet idet der for denne gruppe er fire forskellige beskrivelsesstrategier (jf. Pedersen, Nimb & Braasch 2010):

1. Lemma med kun én betydning i DDO; overbegreb er grøntsag eller rodfrugt: Fx *avokado, majroe*
2. Lemma med kun én betydning i DDO; overbegreb er plante: Fx *artiskok, spinat*
3. Lemma med to betydninger i DDO, med hhv. plante og grøntsag som overbegreb. Fx: *tomat*, *græskar*
4. Lemma med over- og underbetydning i DDO, med hhv. plante og plantedel som overbegreb: Fx: *gulerod, jordskok*

Det man ser af denne forskelligartede måde at beskrive grøntsager i DDO på er at man – ud over at vakle mellem overhovedet at beskrive en madbetydning som en selvstændig betydning – vakler imellem hvorvidt man skal beskrive fødevarebetydningen af en plante eller plantedel ud fra en *fødevarekategorisering* som er tilfældet i 1 og 3 (*grøntsag, rodfrugt* som overbegreb), eller man skal anvende den *botaniske taksonomi*, som det er tilfældet i 4 (*plantedel*). For at få en bedre struktur på dette i netværket forsøger vi i Dan-

Net at etablere to parallelle taksonomier: en botanisk taksonomi og en fødevaretaksonomi. Den første kalder vi i henhold til Cruse (2002) for en naturlig (biologisk) taksonomi, mens fødevaretaksonomien er en funktionstaksonomi hvor genstandens anvendelse (i dette tilfælde som fødevare) er det strukturerende princip[4].

Overbegreber der kun indgår i fødevaretaksonomien, er begreber som *grøntsag, suppeurt,* og *krydderurt*, mens begreber som fx *rod, stenfrugt* og *skærmplante* kun figurerer i den botaniske taksonomi. I et forsøg på at opnå en mere systematisk struktur i DanNet end den man finder i DDO, bestræber vi os på at repræsentere de planter eller plantedele som vi typisk spiser i vores kultur i *begge* taksonomier. Således *porre* som i DanNet dels har overbegrebet



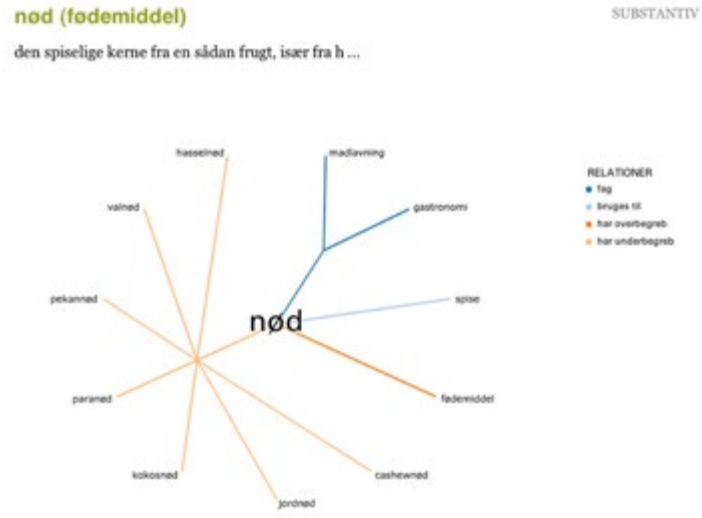
grøntsag/grønt fra fødevaretaksonomien, dels overbegrebet *urt* fra den botaniske taksonomi; se figur 1:

Figur 1: *porre* nedarver dels fra den botaniske taksonomi, dels fra fødevaretaksonomien

For at gøre forvirringen komplet findes der også en række udtryk som refererer til forskellige begreber i de to taksonomier, og som viser at lægmand og fagmand (fx botanikeren) har forskellige strukturelle opfattelser af domænet. *Bær, nødder* og *frugter* udpeger fx forskellige genstande hos hhv. en botaniker og en lægmand. For en botaniker er en *tomat* således en *frugt* og et *jordbær* en *nød* (som igen er en *frugt*), hvilket går imod den almindelige opfattelse

4 Wierzbicka (1996) har en lignende skelnen mellem *naturlige* og *kulturelle* typer.

hvor vi anser en *tomat* for at være en *grøntsag* og et *jordbær* for et *bær*. "Falske venner" som *bær*, *nød* og *frugt* holdes adskilt ved at etablere én betydning i hver taksonomi; dog skal det bemærkes at DanNet langt fra afbilder en komplet biologisk taksonomi på disse områder da resursen generelt behandler almensproget og derfor fx



beskriver fødevarer mere uddybende. Figur 2 angiver *nød* i fødevaretaksonomien.

Figur 2: Fødevarebetydningen af *nød* i DanNet

En mere uddybet redegørelse for hvordan fødevarer, botanik og zoologi er behandlet i DanNet gives i Pedersen, Nimb & Braasch (2010).

3.2. Reorganisering af synonymibegrebet – en pragmatisk tilgang

DanNet har også et andet og mere pragmatisk synonymibegreb end DDO. Det er typisk for wordnets at synonymer tolkes noget

bredere end i traditionelle ordbøger. I DanNet har vi i høj grad set på paralleliteten i søsterstrukturen som den forekommer i DDO. Når begreberne *fag*, *videnskab*, og *lære* i DDO har parallelle "søstre", nemlig hhv. *informatik*, *bromatologi* under *lære*, *samfundsfag* under *fag* og *datalogi* som undertype til *videnskab*, så er det en indikation om at overbegreberne bruges mere eller mindre i flæng. Fra et praktisk, sprogteknologisk synspunkt er det mest nærliggende at foretage en sammenlægning i ét såkaldt "synset" (= mængde af synonymer der betegner samme begreb): {*lære, fag, videnskab*} og have de forskellige fag som søstre under samme overbegreb. *Anordning* og *indretning* er andre eksempler på overbegreber i DDO som parallelle søsterbegreber refererer til som nærmeste overbegreb, fx hhv. *støttefod* og *lampeskærm*.

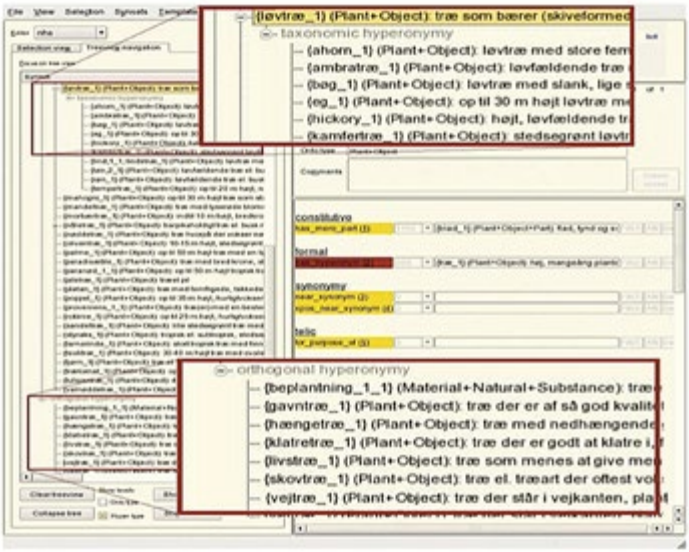
I praksis er disse typer af synonymi imidlertid ikke altid lige lette at identificere. Derfor vil man stadig finde begreber i DanNet som set fra en sprogteknologisk synsvinkel med fordel kunne sammenlægges. Dette er en af de opgaver som vi håber at kunne udføre i DK-CLARIN-projektets sidste fase.

3.3. Tilføjelse af manglende information

Hvor tilpasning af hyponymier og sammenlægning af synonyme begreber i teorien ligger nogenlunde lige for (selvom det i praksis er en meget stor opgave at gennemføre det konsistent for hele ordforrådet), så er spørgsmålet om tilføjelse af manglende information i wordnets noget vanskeligere at håndtere. Som det også redegøres for i Pedersen et al. 2009, så er der meget semantisk information i ordbøger for mennesker som tages for givet og derfor ikke ekspliciteres.

For det første ekspliciteres det ikke i definitionerne hvilken type hyponymirelation der er tale om, når der angives et overbegreb. Når det i en DDO-definition angives at *klatretræer* og *egetræer* er *træer*, så er det vigtigt i DanNet at skelne mellem *trætyper* hvor-

til *eg* hører, og træer der snarere opfylder forskellige funktioner (*vejtræer, klatretræer, suttetræer*). Figur 3 viser hvorledes denne skelnen angives i DanNet i og med at kun typer af træer henføres som taksonomiske, hvorimod de andre anføres som ortogonale



(se også Pedersen & Sørensen 2006 for en nærmere redegørelse for denne skelnen).

Figur 3: Skelnen i DanNet mellem taksonomiske begreber og øvrige begreber

En anden oplysning som kun sjældent angives i DDO, er hvem der er den typiske bruger eller frembringer af en genstand. At en maler har malet et billede, at det typisk er en apoteker der driver et apotek, og en læge der bruger en skalpel, og at det typisk er kvinder der bruger stiletter, er ofte oplysninger som regnes for kendte af ordbogsbrugeren eller som må udledes af brugseksempler. Ligeledes angives det heller ikke altid hvad en genstand bruges til; af og til beskrives kun genstandens udseende eller dele, og funk-

tionen skal igen udledes af brugseksemplerne. I DanNet er målet at disse oplysninger angives konsekvent for alle artefakter selvom dette ikke er fuldt gennemført endnu. Nedarvningsmekanismen er en af de funktioner der hjælper DanNet-redaktøren med at identificere sådanne manglende relationer. Hvis der for et professionsbegreb nedarves at en person arbejder, vil man som redaktør blive promptet til at beskrive alle professioner mere specifikt: At en lærer arbejder ved at *undervise*, en læge ved at *behandle* og så fremdeles. En vanskelig opgave i denne henseende er grænsedragningen mellem lingvistisk viden og mere generel omverdensviden som jo i princippet er uendelig og ikke kan rummes i et wordnet. Hvor sætter man grænsen for hvad der skal med? Pustejovskys qualiastruktur med de semantiske relationer den bringer med sig (som fx hvordan noget er blevet til, og hvad det bruges til), taler for at holde sig til den lingvistisk relevante information, og det er i første omgang denne ramme vi forsøger at holde os inden for, svarende til det skel som leksikografer forsøger at drage i forhold til encyklopædisk viden.

4. Anvendelser af semantiske sprogressourcer

Det har tilsyneladende ikke ligget helt lige for at anvende DanNet i kommercielle produkter; man kan sige at markedet har ladet vente på sig. På nuværende tidspunkt har vi kendskab til at DanNet anvendes i OpenOffices danske skrivehjælp som en udvidet synonymifunktion hvor det udnyttes at man kan fremfinde ikke kun "rigtige" synonymer men også nært beslægtede begreber. Samt som en ekstra semantisk visualiseringsfunktion i DDO's egen opslagsfacilitet på nettet (ordnet.dk) hvor man gøres bekendt med beslægtede ord til et givent opslagsord.

DanNet har i skrivende stund været open source i et års tid (siden marts 2009), og vi håber at det blot er et spørgsmål om tid før

flere virksomheder får øjnene op for anvendelsespotentialet. Ved IT- og Telestyrelsens konkurrence i januar 2010 om offentlige data i spil søgte to danske virksomheder om støtte til at eksperimentere med DanNet til forskellige anvendelser, bl.a. til fleksibel søgning i offentlige databaser. Desværre vandt ingen af dem konkurrencen. Vores formodning er at det er en tidskrævende proces for virksomhederne at undersøge potentialet af ressourcen.

På Københavns Universitet har vi de sidste tre år afholdt kurser i sprogteknologi og informationssøgning hvor vi anvender DanNet til semantisk udvidelse af søgeforespørgsler. Ressourcen er afprøvet på tre forskellige domæner (ernæring, uddannelse, biler), og på alle tre tekstsamlinger viser de studerendes eksperimenter en forbedring af "recall" (systemets evne til at udtrække relevante dokumenter) – uden en tilsvarende kritisk forværring af "precision" (systemets evne til at afvise irrelevante dokumenter).

Til forskningsformål anvendes ressourcen på flere niveauer. Der er et lingvistisk anvendelsespotentiale for ressourcen idet databaseformatet gør det muligt at udtrække statistiske oplysninger om leksikalsk semantik på dansk. Det er således muligt at drage mere eller mindre dristige konklusioner på basis heraf: fx at kvinder primært vurderes på deres udseende og seksuelle tilbøjeligheder, mens mænd primært vurderes på deres opførsel. Denne kontroversielle slutning drages i Braasch & Pedersen (2010) ved at undersøge antallet af værdiladede benævnelser for hhv. mænd og kvinder (fx *tøjte* og *vatpik*) og udlede statistik for hvilke egenskaber disse begreber hver især specificerer (seksualitet, opførsel, udseende, formåen mv.). Eller man kan udlede noget om dansk madkultur gennem tiden ved at udtrække to konkurrerende taksonomier for *oste*: Én baseret på en traditionel opfattelse af *ost* som pålæg i form af enten *skæreost* eller *smøreost*, og en anden (eller flere!) baseret på importerede italienske, franske og spanske oste som anvendes i madlavningen på en anden måde. Eller man kan udtrække en stor mængde udtryk for talehandlinger på dansk.

Mulighederne er mange.

Hvordan ressourcen på længere sigt vil indgå i sprogteknologiske forskningsprojekter vil fremtiden vise. Jeg har nævnt informationssøgning, og under dette felt bør også nævnes semantisk annotering af tekst til brug for søgning, resumering og data mining. Semantisk annotering er et af de felter som vi på Center for Sprogteknologi og Det Danske Sprog- og Litteraturselskab gerne vil udforske i fremtiden hvis vi kan opnå fornøden bevilling til opgaven. I den forbindelse skal det afprøves i hvor høj grad DanNets detaljerings- og dækningsgrad er passende til opmærkning af løbende ord i dansk tekst, og i hvor høj grad den rigtige betydning af et ord i en given kontekst kan udledes automatisk ud fra håndopmærkede træningskorpora.

5. Konkluderende bemærkninger

Ressourcer som DanNet er med til at påvise at der er en tæt synergi imellem sprogteknologi og leksikografi idet fællesmængden af relevant viden er stor. Det faktum at moderne leksikografi via nyere korpus- og redigeringsværktøjer har udviklet sig drastisk de seneste tiår og nu producerer mere konsistente og korpusnære resultater, har nok generelt styrket tendensen til at de to felter er rykket nærmere hinanden. Jeg tror yderligere at sprogteknologiske værktøjer som DeepDict (Bick 2009 og i dette bind) og Sketch Engine (Kilgarriff et al. 2004) der ud fra store analyserede korpora genererer typiske leksikalske mønstre, vil styrke denne tendens i de kommende år således at det i fremtiden vil virke besynderligt at udarbejde sprogteknologiske ressourcer uafhængigt af allerede eksisterende leksikografisk materiale og vice versa. Og der er ingen tvivl om at begge felter vil have gavn af denne udvikling.

Litteratur

Bick, E. 2009: *DeepDict – A Graphical Corpus-based Dictionary of Word Relations*. I: *Proceedings of NODALIDA 2009. NEALT Proceedings Series Vol. 4*. Tartu: Tartu University Library, 268-271.

Boguraev, B. & T. Briscoe (eds) 1989: *Computational Lexicography for Natural Language Processing*. London & New York: Longman.

Braasch, A. & B. S. Pedersen 2010: Encoding Attitude and Connotation in Wordnets. *Proceedings of the 14th EURALEX Conference*, Leeuwarden, Holland.

Calzolari, N. 1988: The dictionary and the thesaurus can be combined. I: M. Evens: *Relational models of the lexicon. Studies in natural language processing*, Cambridge: Cambridge University Press, 75-96.

Cann, R. 1993: Formal semantics: an introduction. Cambridge University Press.

Copestake, A. 1990: An approach to building the hierarchical element of a lexical knowledge base from a machine readable dictionary. I: *Proceedings of the First International Workshop on Inheritance in Natural Language Processing*, Tilburg, 19-29.

Copestake, A. & T. Briscoe 1995: Semi-productive Polysemy and Sense Extension. I: *Journal of Semantics, Vol 12, 1,*15-67.

Cruse, D.A. 2002: Hyponymy and Its Varieties. I: R. Green, Bean, C.A. & Myaeng, S. H. (eds.), *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management*. Springer Verlag, 2-21.

DDO = Hjorth, E., Kristensen, K. et al. (eds.). (2003-2005). *Den Danske Ordbog 1-6*. Gyldendal & Det Danske Sprog- og Litteraturselskab. www.ordnet.dk/ddo.

Fellbaum, C. (ed.) 1998: *WordNet – An Electronic Lexical Database*. Cambridge, Massachusetts, London, England: The MIT Press.

Ide, N. & J. Véronis 1995: Knowledge Extraction from Machine-Readable Dictionaries: An Evaluation. I: Steffens, P. (ed.), *Machine Translation and the Lexicon, Third International EAMT Workshop*, *Heidelberg 1993, Proceedings*. Lecture Notes in Computer Science 898. Springer.

Ide, N. & Y. Wilks 2007: Making Sense About Sense. I: E. Agirre & P. Edmonds eds.) *Word Sense Disambiguation – Algorithms and Applications*. Springer, 47-75.

Kaplan, R. M. & J. Bresnan 1982: Lexical Functional Grammar: A Formal System for Grammatical Representation. I: J. Bresnan: *Mental Representation of Grammatical Relations*, MIT Press.

Kilgarriff, A. 1997: I don't believe in word senses. I: *Computers and the Humanities 31,* 91-113.

Kilgarriff, A., P. Rychly, P. Smrz 2004: The Sketch Engine. *Proceedings of EURALEX 2004*, Lorient, France, 105-116.

Kokkinakis, D., M. Toporowska Gronostaj M. & K. Warmenius 2000: Annotating, Disambiguating & Automatically Extending the Coverage of the Swedish SIMPLE Lexicon, Språkdata, Göteborg. I: *Proceedings of the 2nd LREC (Language Resources and Evaluation Conference)*, Athens, Hellas, 1397-1405.

Lenci, A. , N. Bel, F. Busa, N. Calzolari, E. Gola, M.Monachini, A.Ogonowski, I. Peters, W. Peters, N. Ruimy, M.Villegas, A. Zampolli 2000: SIMPLE: A general framework for the development of multilingual lexicons. I: *International Journal of Lexicography 2000 13(4)*, 249-263.

Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen 2009: DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation vol. 43*, 269-299.

Pedersen, B.S, S. Nimb, A. Braasch 2010: Merging specialist taxonomies and folk taxonomies in wordnets – a case study of plants, animals and foods in the Danish wordnet. I: *Proceedings from Language Resources and Evaluation Conference*. Malta.

Pedersen, B., Paggio, P. 2004: The Danish SIMPLE Lexicon and its Application in Content-based Querying. I: *Nordic Journal of Linguistics Vol 27:1*, 97-127.

Pedersen, B.S., N. Sørensen 2006: Towards Sounder Taxonomies in Wordnets. I: A. Oltramari, Chu-Ren Huang, A. Lenci, P. Buitelaar, C. Fellbaum (eds): *Ontolex 2006 at 5th International Conference on Language Resources and Evaluation*. Genova, Italy, 9-16.

Pollard, C. & I. Sag 1994: *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Prebensen, H. 2006: Formens fascination. I: A. Braasch, C. Navarretta, S. Nimb, S. Olsen, P. Paggio, B. Pedersen (red.): *Sprogteknologi i dansk perspektiv – En samling artikler om sprogforskning og automatisk sprogbehandling*. København: Reitzels Forlag, 51-69.

Pustejovsky, J. 1995: *The Generative Lexicon*. MIT Press.

Spang-Hanssen, E. 2006: Sprogteknologi og humaniora. I: A. Braasch, C. Navarretta, S. Nimb, S. Olsen, P. Paggio, B. Pedersen (red.): *Sprogteknologi i dansk perspektiv – En samling artikler om sprogforskning og automatisk sprogbehandling*. København: Reitzels Forlag, 39-51.

Wierzbicka, A. 1996: *Semantics: Primes and Universals*. Oxford: Oxford University Press.

Bolette Sandford Pedersen
professor, ph.d.
Københavns Universitet
Njalsgade 140
DK-2300 København S
bspedersen@hum.ku.dk