Betydningsinventarer – i ordbøger og i løbende tekst

Bolette Sandford Pedersen, Sanni Nimb, Anna Braasch & Sussi Olsen

We examine a set of highly polysemous nouns in Danish in order to understand how well word senses in Den Danske Ordbog (DDO) and the Danish wordnet, DanNet, correlate with the meanings found in examples of running text. The overall goal of the task is to provide adequate training data for automatic word sense disambiguation. To this end, we annotate a corpus with a full sense inventory from DDO and compare with the annotations provided on the basis of a *clustered* inventory derived automatically from DDO and DanNet's ontological classes. The results show that our hypothesis of clustered senses being more manageable and providing higher annotation agreement (apart from what can be expected from a higher chance agreement when there are less annotation options), holds for a vast majority of the studied cases. Further, the investigations provide a valuable assessment of the organization of senses in the studied dictionaries.

1. Et betydningsopmærket korpus – hvorfor?

I forbindelse med et samarbejdsprojekt mellem Københavns Universitet og Det Danske Sprog- og Litteraturselskab[1] betydningsopmærker vi en offentligt tilgængelig korpusressource med betydningsdistinktioner i dansk. Det praktiske, sprogteknologiske formål med denne opmærkning er at skabe sprogressourcer der muliggør træning af computermodeller til at kunne entydiggøre ord på dansk og derigennem forbedre teknologiske sprogtjenester som søgemaskiner, maskinoversættelse og spørgsmål-svarsystemer.

Imidlertid har vi under opmærkningen særligt fokus på at udforske forholdet mellem de betydningsskel der etableres i den konventionelle ordbog, og de betydningsskel som man intuitivt kan genkende og opnå konsensus om når man opmærker betydninger i løbende tekst. For at belyse dette område

[1] Projektet "Semantic Processing across Domains" finansieres af Forskningsrådet for Kultur og Kommunikation i perioden 2013-2016 (jf. Pedersen et al. 2014 og <cst.ku.dk/projekter/semantikprojekt/>).

bedst muligt er alle de korpusopmærkninger vi diskuterer i artiklen, opmærket af flere annotører, og enigheden er beregnet.

Vi opmærker korpusset med betydningsinventarer af varierende grovhed. Disse betydningsinventarer fremkommer ved systematisk at kombinere den hierarkiske struktur fra en konventionel ordbog (Den Danske Ordbog; DDO, jf. ordnet.dk) bestående af hoved- og underbetydninger med den ontologiske viden der er indarbejdet i betydningerne i det leksikalsk-semantiske wordnet, DanNet (Pedersen et al. 2009, jf. www.andreord.dk) i form af typer som HUMAN, EVENT og PROPERTY. Ved at sammenligne annoteringerne med et grovkornet betydningsinventar med annoteringer med et mere finkornet, prøver vi at komme nærmere en forståelse af hvor det er svært at skelne for mennesker – og hvor det derfor sandsynligvis også er svært for en computermodel. Helt overordnet er målet at nærme os et "passende" niveau af finkornethed; et som er håndterbart for de menneskelige annotører og for de maskinlæringssystemer der efterfølgende skal processere dem, men som samtidig er tilstrækkelig nuanceret til at kunne forbedre eksisterende teknologiske sprogtjenester.

Undervejs i denne proces opnås i tillæg en større erkendelse af hvilke grundlæggende problemstillinger man står overfor i det leksikografiske arbejde når man skal lægge sig fast på hvilke principper man skal anvende i beskrivelsen af betydningsinventaret. Vi indleder i afsnit 2 med en beskrivelse af hhv. DDO's og DanNets betydningsstruktur og redegør på baggrund heraf for hvordan vi automatisk kan udforme klynger af betydninger. Endelig redegør vi for den udvalgte empiri i form af 20 meget polyseme substantiver. I afsnit 3 går vi i dybden med annoteringsarbejdet og beskriver nogle af de divergenser der opstår mellem annotørerne. I afsnit 4 evaluerer vi det samlede annoteringsarbejde.

2. Betydningsstruktur og etablering af betydningsklynger

Inden vi omtaler undersøgelsen, vil vi kort fremlægge vores hypoteser ud fra teoretiske beskrivelser af systematiske sammenhænge mellem bestemte typer af betydning af samme ord, og den beskrivelse af ordforrådet man finder i DDO og DanNet. Cruse (2000:111) introducerer flere typer af relationer der beskriver sammenhæng mellem betydninger af samme ord (modsat semantiske relationer der beskriver sammenhæng mellem forskellige ord, fx de rela-

tionstyper der anvendes i wordnets). Disse ”interne” relationer er oplagte at tage udgangspunkt i når man ønsker at sammenlægge betydninger i en ordbog med henblik på at opnå et mindre detaljeret betydningsinventar. Autohyponymi kalder han fx den systematiske sammenhæng der er indenfor ord der både har en bred og en mere indsnævret betydning med samme overbegreb (fx *drikke* (væske) over for *drikke* (for meget alkohol)). Det modsatte fænomen, sammenhængen fra en snæver til en udvidet betydning af samme ord, og hvor de to betydninger har samme overbegreb, betegner han autosuperordination (fx *mand* (person af hankøn) over for *mand* (person uden tanke på køn)). Hvis sammenhængen mellem et ords to betydninger bygger på en helhed set i forhold til en del af samme helhed, bruger han to termer: automeronymi (del for helhed, fx *dør* i betydningen ’åbning’ hhv. ’plade’) og autoholonymi (helhed for del, fx *krop* i betydningen ’hele legemet’ over for *krop* i betydningen ’kun den centrale del af kroppen’, ’torsoen’).

I DDO's struktur afspejles de systematiske sammenhænge inden for samme ord som Cruse observerer og navngiver, i princippet via ordbogens hierarkiske organisering i hoved- og underbetydninger. *Krop* = ’legeme’ er fx hovedbetydning (overbetydning) til underbetydningen *krop* = ’torso’ i DDO. Dette fremgår også af DDO’s brugervejledning der fastslår at en underbetydning kan være enten ”en faglig betydning, en delbetydning, en overført betydning eller en udvidet betydning i forhold til overbetydningen” (DDO, bind 1, s. 27). Da en faglig betydning vil være en indsnævret betydning med samme overbegreb som overbetydningen, er relationen mellem dem af typen autohyponymi. Relationen mellem en delbetydning og en overbetydning svarer til automeronymi. Det som i DDO betegner udvidet betydning, kan manifestere sig både i form af autosuperordination (samme overbegreb som overbetydningen, men bredere betydning) og i form af autoholonymi (overbetydningen betegner en del af noget, men underbetydningen er bredere og betegner helheden).

Man finder dog også eksempler på at de forskellige sammenhænge er usynliggjort i DDO’s struktur. Substantivet *dør* er fx beskrevet med to hovedbetydninger selv om der er en automeronymirelation mellem dem, det samme er de to betydninger af verbet *drikke*, selv om der er en autohyponymirelation mellem dem. Og substantivet *mand* (person af hankøn), *mand* (i den

indsnævrede betydning 'ægtemand = autohyponymi') og *mand* i den udvidede betydning 'person/menneske uden tanke på køn' (autosuperordination) er beskrevet som tre hovedbetydninger og ikke som en overbetydning med to underbetydninger. Dette skyldes at DDO-redaktionen af formidlingsgrunde foretrak en flad og let tilgængelig struktur når det føltes logisk, fx når den betydning der principielt skulle have været en underbetydning, var mest frekvent i sproget. En yderligere, ikke-ubetydelig faktor er at der i leksikografisk arbejde altid er en generel usikkerhedsfaktor mht. hvor langt et udvidet eller indsnævret betydningsaspekt må fjerne sig fra sin kernebetydning før det bør udløse en ny hovedbetydning (jf. Svensén 2009:212).

DanNet er et wordnet for dansk der er baseret på DDO's betydningsinventar og de semantiske relationer mellem de forskellige ords betydninger, og som samtidig grupperer synonyme og nærsynonyme betydninger i såkaldte synsets. I DanNet kan man derimod intet udlede om relationen mellem et enkeltords flere betydninger, ej heller hvilken af de måske mange betydninger der er mest prominent, idet alle betydninger ligestilles.Til gengæld er hvert synset i DanNet forsynet med en ontologisk type baseret på EuroWordNets topontologi (Vossen et al. 1999), som er en hierarkisk ontologi der består af ca. 40 kategorier af typen HUMAN, EVENT, PROPERTY, DISEASE, ARTIFACT, BUILDING mv. Denne oplysningstype danner sammen med DDO's hoved- og underbetydninger udgangspunktet for en automatisk klyngedannelse på et velunderbygget grundlag.

På trods af at DDO's struktur langtfra i alle tilfælde afspejler de systematiske sammenhænge der er mellem betydninger, har vi alligevel en vis forventning om at især en stor del af de indsnævrede betydninger er beskrevet som underbetydninger der via viden om et fælles overbegreb (udtrykt via DanNets ontologi) med god mening kan slås sammen med deres hovedbetydning til én betydning. Man finder fx substantivet *område* i DDO beskrevet med en hovedbetydning 'landområde' (LOCATION) og en indsnævret underbetydning 'administrativt landområde'; disse to betydninger kan lægges sammen automatisk via DanNets ontologiske oplysninger, hvilket intuitivt giver god mening. Vi forventer også en del automatiske sammenlægninger i de tilfælde hvor del/helhedsrelationer er afspejlet i den hierarkiske struktur i DDO og samtidig kan identificeres via DanNets ontologiske oplysninger om betydningerne. Fx beskrives substantivet *sten* i DDO med hovedbetydningen 'fast, hårdt materiale' og derefter med to underbetydninger 'stykke af et

sådant materiale' og 'stykke der anvendes som byggemateriale'; her vil alle tre betydninger blive slået sammen til kun én via den autoholonymi-relation der kan udledes af DanNets ontologiske typer, hvilket umiddelbart også giver god mening.

Ud fra et ønske om at arbejde med et så stort antal sammenlagte betydninger som overhovedet muligt udvalgte vi til vores undersøgelse et antal substantiver med mange betydninger der samtidig i høj grad var hierarkisk organiseret, dvs. at antallet af underbetydninger skulle være relativt højt i forhold til antallet af hovedbetydninger. 20 substantiver blev udvalgt på dette grundlag idet det samtidig var et krav at de havde en høj forekomst i tekster, målt i frekvens i det korpus som dannede grundlaget for beskrivelsen af lemmaerne i DDO, nemlig Korpus90. På den måde sikrede vi os at korpusset vi anvender, sandsynligvis ville indeholde det nødvendige antal teksteksempler der skulle bruges i opmærkningsarbejdet beregnet ud fra en generel tommelfingerregel om at man bør opmærke 100 eksempler + 15 x antal betydninger for et givent lemma (jf. fremgangsmåden anvendt i Senseval-projektet: www.senseval.org).

Efter sammenlægningen af betydninger ud fra de ontologiske typer der var tildelt i DanNet, opnåede vi at 18 af de 20 substantiver fik reduceret deres antal af betydninger, se Tabel 1; gennemsnitligt set reduceres antallet af betydninger med 23,5 procent.

	Antal betydninger i DDO	**Antal klynger**
selskab	10	6
plads	13	9
slag	17	11
plade	11	10
skud	12	12
ansigt	7	6
skade	6	5
stykke	18	16
kontakt	9	5
stand	9	7
top	8	6
kort	10	4
vold	9	7
hul	14	11
hold	12	10
lys	13	9

blik	7	6
model	8	7
kurs	3	3
tang	Udeladt pga. for få forekomster i korpus	

Tabel 1: Udvalgt i empiri: 20 polyseme substantiver og deres betydningsreduktioner – i snit 23,5 % reduktion.

Som illustration af hvordan klyngedannelsen kommer til udtryk i de enkelte leksikalske indgange, ses den leksikalske indgang for *selskab* i Figur 1 hvor betydningerne 1, 1a og 1b klynges sammen idet de har den samme ontologiske type i DanNet: RELATION. Underbetydning 1c udgør derimod sin egen klynge da den har den ontologiske type PERSON. Hovedbetydning 2 er også af typen PERSON, men udgør igen sin egen klynge ud fra princippet om at hovedbetydninger aldrig klynges sammen. Hovedbetydning 3 er af typen EVENT (altså en begivenhed). Hovedbetydning 4 og underbetydning 4a giver igen anledning til en klynge da typen her er INSTITUTION, hvorimod hovedbetydning 5 udgør sin egen klynge. De i alt 10 betydninger reduceres altså hermed til 6 klynger. Herudover følger en række faste udtryk som alle bibeholdes og udgør hver deres egen klynge.

HOVEDBETYDNING 1
det at være eller foretage sig noget sammen med en eller flere andre personer
1a. samvær eller fællesskab med en eller flere andre, ofte af underholdende eller adspredende karakter
1b. det at noget forekommer sammen med noget andet
1c. personkreds eller miljø som man knytter sig til eller bliver forbundet med af personlige eller professionelle årsager
HOVEDBETYDNING 2
gruppe af personer som foretager sig noget bestemt i fællesskab
2a. gruppe af personer som er samlet til en fælles social aktivitet, fx en fest el. middag
HOVEDBETYDNING 3
festlig sammenkomst med mange mennesker, mad og drikke og evt. underholdning
HOVEDBETYDNING 4
handels-, industri- eller erhvervsvirksomhed
4a. sammenslutning mellem to eller flere juridiske personer som har til formål at drive en handels-, industri- eller anden virksomhed fx et aktieselskab eller et anpartsselskab
HOVEDBETYDNING 5
sammenslutning af personer med en fælles, oftest faglig eller akademisk interesse

Figur 1: *Selskab*s betydninger i DDO.

3. Om annoteringsarbejdet

De tekster vi opmærker, er udtrukket fra Det Danske CLARIN Reference Corpus (Asmussen 2012) som er et almensprogligt korpus på 45 mio. løbende ord, med hovedvægten på avistekster (48 %), men korpusset omfatter også mange andre forskellige teksttyper. Vi har lagt vægt på at have et bredt udsnit af teksttyper repræsenteret i vores korpus da en optimering af sprogprocessering *på tværs* af teksttyper er et vigtigt mål for vores projekt. I Figur 2 kan man se en liste over de teksttyper vi har opmærket.

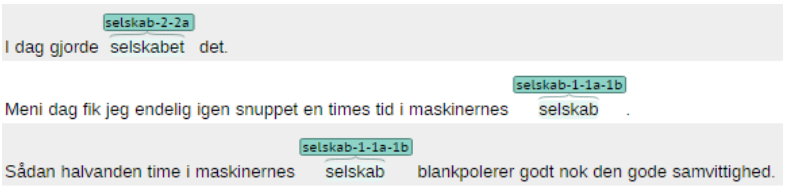
Titel	**Beskrivelse**
Bentes blog	En blog skrevet af en kvinde i fyrrerne
Selvhenter	Et chatforum for især unge mennesker
Mangamania	Et chatforum for mangafans
Se og Hør	Ugeblad
Folketingstaler	Nedskrevne folketingsdebatter
Politiken	Avis

Figur 2: Teksttyper der opmærkes.

For hvert ord udvælges et antal sætninger som beskrevet i afsnit 2.4 Eksempelvis annoteres for *selskab* 100 + 15 x 7 = 205 teksteksempler. Vi har ved udvælgelsen af sætninger valgt at lægge særlig vægt på at få repræsenteret de sociale medier, og alle sætninger indeholdende de aktuelle ord medtages fra disse teksttyper. Det faktum at der er vis uoverensstemmelse mellem vores valgte korpusmateriale og det korpus som DDO-betydningsbeskrivelserne er baseret på, giver nogle problemer. DDO's interne korpus på 40 mio. tokens (svarende nogenlunde til Korpus90, jf. Asmussen (2006)) er for det første ældre, for det andet var det vægtet bredere hvad angår teksttyper end Clarin-korpusset (først og fremmest i form af langt flere litterære tekster), for det tredje indeholdt det af gode grunde ikke tekster fra sociale medier. Vi mener dog at fællesnævneren i form af relativt meget avisstof i begge korpora til en vis grad udjævner de forskelle der er. Fx optræder de polyseme substantiver, vi opmærker i denne sammenhæng, og som er velrepræsenterede i Korpus90, ikke særligt hyppigt i tekster fra de sociale medier, så til trods for denne vægtning er især avistekster og folketingsdebatter statistisk set i overtal i vores korpusser.

Annoteringen af de udvalgte substantiver foretages både som *finkornet opmærkning*, dvs. med det fulde, detaljerede DDO-betydningsinventar og som *klyngeopmærkning*, dvs. med bredere betydningsklynger der er dannet ved sammenlægning af betydninger (jf. ovenfor, afsnit 2). Formålet er at undersøge hvorvidt den finkornede eller den klyngebaserede opmærkningsmetode giver den højeste annotørenighed, idet graden af annotørenighed er en rimelig god målestok for metodens anvendelighed.

En gruppe på tre studerende har stået for opmærkningen af teksterne. Derudover har flere af forskerne i projektet opmærket et antal filer. Alle tekster er blevet opmærket af mindst to personer, så det er muligt at måle enigheden mellem annotørerne og undersøge de tilfælde hvor de divergerer. Teksterne opmærkes i opmærkningsværktøjet WebAnno (se Figur 3), et webbaseret værktøj udviklet af Technische Universität Darmstadt for CLARIN (Yimam et al. 2013). WebAnno kan tilpasses mange opmærkningstyper og gør det let at bevare overblikket over fremdriften, lave en korrigeret version af de annoterede filer samt ikke mindst beregne kvaliteten af opmærkningerne målt ud fra enigheden annotørerne imellem.



Figur 3: Annotering i WebAnno, *selskab* med klyngeinventar.

Der synes generelt at være to grundlæggende forhold der volder problemer for annotørerne, og som kommer til udtryk som uenigheder i annoteringerne:

- graden af *kompleksitet* i et givet ords betydningsstruktur (også afspejlet i dets reducerede betydningsklynger), og
- antallet og arten af de *faste udtryk.*

Hvis vi indledningsvis ser på kompleksitet i betydningsstruktur, kan vi se at annoteringen generelt forløber uproblematisk når der er tale om en simpel betydningsstruktur med dels klare skel mellem (få) hovedbetydninger og

dels relativt få underbetydninger. Et godt eksempel herpå er *model*, idet ordet i DDO har seks klart adskillelige hovedbetydninger (med i alt to underbetydninger) og et enkelt fast udtryk (*stå model til)*; disse udgør det finkornede opmærkningsinventar. De i alt otte finkornede hoved- og underbetydninger er slået sammen til syv betydningsklynger, hvilket selvsagt er en ganske lille reduktion; det faste udtryk følger uændret med. Hver af de to opmærkningsmetoder resulterer i relativ høj grad af annotørenighed (hhv. 0,66 og 0,76 i henhold til Krippendorffs α[2]; se også afsnit 4). Hypotesen om at et mindre detaljeret inventar giver en mere konsistent opmærkning, bekræftes altså her.

Opmærkningen af *kontakt* viser et mere sammensat mønster: Ved konkrete betydninger med klare betydningsskel er der stor enighed, som i eksempel 1:

1) *Afblegning af hår må kun finde sted i striber eller så afblegningsmidlet ikke er i **kontakt** med huden*

Sætninger af denne type er af alle opmærket med betydningsklyngen **kontakt_2_2a** → "det at to genstande, dele, flader el.lign. er i fysisk forbindelse eller berøring med hinanden". Ligeledes i eksempel 2:

2) *For selv om det er bedst at slukke på **kontakten** for ting, du ikke bruger, så efterlever de færreste rådet fuldt ud*

Sætninger af denne type er af alle opmærket med **kontakt_3_3a** → "teknisk indretning som […]giver mulighed for at tilslutte og afbryde en elektrisk strøm i en ledning og dermed tænde eller slukke for noget".

Men enigheden ophører så snart vi bevæger os over i abstrakte eller overførte betydninger, hvor skellet er utydeligt, især hvis ordets kontekst er kort og/eller kan fortolkes på forskellig vis, som vist i eksempel 3:

3) *Gennem hårdt arbejde og omfattende kontakter fik Udenrigsministeriet lokaliseret danskeren og fik konsulær adgang*

[2] Vi anvender Krippendorffs α til beregning af annotørenighed som bl.a. modregner det faktum at det alt andet lige er nemmere at opnå enighed om få betydninger end om mange (Krippendorff 2011).

Her opstår der uenighed mellem klyngerne **kontakt_1_1a** → ”forhold mellem to parter som indebærer indbyrdes kommunikation”, **kontakt_1b** → ”person eller part som man har forbindelse med eller kommer i berøring med fagligt”, og **kontakt_1c_1_d** → ”det at sætte sig i forbindelse med en anden part, ofte for at indlede et samarbejde eller for at opnå noget eftertragtet; OVERFØRT det at der er forbindelse og dermed grundlag for kommunikation, indsigt”.

Meget ofte er der desuden delvis enighed om at et ord er brugt i overført betydning, men hvilken af dem den enkelte annotør vælger, synes lidt tilfældigt, som belyst i eksempel 4.

4) *I dette lys er det ikke betryggende at opleve repræsentanter […] tage afstand fra menneskerettighederne og ligefrem tale om[….]*

Eksemplet er opmærket med hhv. **lys_1f → ”**OVERFØRT opmærksomhed**,** herunder: søgelys, rampelys” (DDO), med **lys_1e → ”**OVERFØRT viden; (religiøs) indsigt**,** herunder: forklarelsens lys, erkendelsens lys” (DDO), og endelig med **lys_F_i-et-nyt-(andet, -…)-lys** → som (variant af) fast udtryk (listet i DDO).

Opmærkning af *faste vendinger* burde i princippet være helt uproblematisk da man blot skal forholde sig slavisk til den liste af vendinger som er opstillet i DDO. Dette forhold ses også meget tydeligt i opmærkningen af *stand:* dels er enigheden meget høj med både fuldt og klyngebaseret inventar, dels nås næsten den samme α-værdi ved de to metoder. Dette skyldes at langt størstedelen af eksemplerne (79 af 100) udgøres af faste udtryk af typen *i stand til* og *ude af stand til*, og kun 16 forekomster er opmærket med egentlig betydning (fordelt på fire betydninger). Der er dog nogen usikkerhed/uenighed hos annotørerne med hensyn til faste udtryk når 1) DDO’s liste mangler et etableret og hyppigt forekommende fast udtryk, 2) det faste udtryk i korpus er en syntaktisk eller leksikalsk variant af et udtryk på DDO’s liste (med semantisk nærliggende betydning), 3) korpusforekomsten er en kontamination af to faste udtryk, og endelig 4) der er syntaktisk og leksikalsk sammenfald mellem ordets forekomst i et kompositionelt udtryk og i et fast udtryk.

4. Evaluering og konklusion

Den overordnede konklusion på annoteringseksperimentet kan sammenfattes til at en ontologisk baseret klyngetilgang til betydningsinventaret, hvor man opmærker med klynger af betydninger, opnår en bedre enighed mellem annotørerne beregnet ved Krippendorffs α end opmærkning med det fulde DDO-inventar. De enkelte polyseme ords resultater kan aflæses i Tabel 2. Således fremgår det at hypotesen bekræftes i 67 % af tilfældene og afkræftes i 22 % af tilfældene, mens 11 % af de polyseme ord fremviser samme enighed med uden og klyngeannotering.

	Enighed v. finkornet opmærkning	Enighed v. klynge-opmærkning	Bekræftelse af hypotese
selskab	0.48	0.81	Ja
plads	0,51	0,63	Ja
slag	0,50	0,51	Ja
plade	0,048	0,051	Ja
skud	0,29	Ingen reduktion	Neutral
ansigt	0,33	0,35	Ja
skade	0,59	0,65	Ja
stykke	0,55	0,62	Ja
kontakt	0,50	0,60	Ja
stand	0,80	0,86	Ja
top	0,57	0,46	Nej
kort	0,48	0,51	Ja
vold	0,38	0,57	Ja
hul	0,52	0,42	Nej
hold	0,64	0,60	Nej
lys	0,64	0,56	Nej
blik	0,45	0,53	Ja
model	0,66	0,76	Ja
kurs	0,84	Ingen reduktion	Neutral
tang	For få belæg		

Tabel 2: Samlet evaluering af de udvalgte polyseme substantiver beregnet med Krippendorffs α.

Men som det også fremgår af de udvalgte eksempler i afsnit 3, forekommer der meget store udsving fra det ene polyseme ord til det andet, og antallet af betydninger synes ikke i sig selv at være afgørende for hvor vanskeligt det er at opnå annotørenighed. Ikke overraskende udgør ord med mange abstrakte og metaforiske betydninger et problem både med det finkornede betyd-

ningsinventar og med det klyngebaserede. I flere tilfælde kunne det faktisk se ud til at det udvalgte ordforråd har fået nye betydningsnuancer siden det blev registreret i DDO ud fra et snart 25 år gammelt korpus. Det gælder især den udbredte brug af metaforisk sprog, samt forholdet til hvilke af de faste udtryk der er gængse og hyppige.

I vores videre arbejde kunne vi tænke os at kaste et mere kritisk blik på hovedbetydningernes status i forhold til underbetydningerne, jf. diskussionen i afsnit 2. Det kunne fx være interessant at belyse om en sammenlægning af hovedbetydninger med samme ontologiske type generelt vil føre til for stort informationstab, eller om det faktisk vil styrke ressourcen i forhold til opmærkning og senere automatisk entydiggørelse. Et sådant undersøgelsesarbejde ville i tillæg kaste yderligere lys over konsistensniveauet i de anvendte leksikografiske data.

Tak

Tak til studenterannotørerne Sarah Lee Naldal, Selma Rosenfeldt-Olsen, Ida Hauerberg Wolthers, samt til Nicolai Hartvig Sørensen og Héctor Martínez Alonso for teknisk støtte til korpus- og annotationsarbejdet.

Litteratur

Asmussen, J. (2006): Towards a methodology for corpus-based studies of linguistic change. Contrastive observations and their possible diachronic interpretations in the Korpus 2000 and Korpus 90 Corpora of Danish. I: Dawn Archer/Paul Rayson/Wilson (eds.): *Corpus Linguistics Around the World*. Amsterdam: Rodopi, 33-48.

Asmussen, J. (2012): *CLARIN-Referencekorpus*. Foredrag ved Sprogteknologisk Workshop, Københavns Universitet, 31. oktober 2012. <cst.ku.dk/Workshop311012/sprogtekno2012.pdf>.

Cruse, D. A. (2000): *Meaning in language*. Oxford: Oxford University Press.

DDO = *Den Danske Ordbog* (E. Hjorth et al.) 2003-2005. Det Danske Sprog- og Litteraturselskab & Gyldendal: København.

Fersøe, H. B. Maegaard & B. S. Pedersen (2013): Humanities eInfrastructure initiatives in Denmark. I: *Proceedings of the workshop on Nordic language research infrastructure at NODALIDA 2013*. Linköping University Electronic Press, Vol. 089.

Pedersen, B.S., S. Nimb, S. Olsen, A. Søgaard & N. Sørensen (2014): Semantic Annotation of the Danish CLARIN Reference Corpus. I: *Proceedings from isa-10, 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, Reykjavík, 25-29.

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. I: *Language Resources and Evaluation, Computational Linguistics Series* 43(3), 269-299.

Krippendorff, K. (2011): Agreement and information in the reliability of coding. I: *Communication Methods and Measures* 5(2), 93-112.

Svensén, B. (2009): *A Handbook of Lexicography*. Cambridge: CUP.

Vossen, P. (ed.) (1999): *EuroWordNet, A Multilingual Database with Lexical Semantic Networks*. Netherlands: Kluwer Academic Publishers.

Yimam, S.M., I Gurevych, R. Eckart de Castilho & C. Biemann (2013): WebAnno: A Flexible, Web-based and Visually Supported System for Distributed Annotations. I: *Proceedings of ACL-2013*, demo session, Sofia, Bulgaria.

Bolette S. Pedersen
professor, institutleder
Nordisk Forskningsinstitut
Københavns Universitet
Njalsgade 136, bygn. 27. 2
DK-2300 København S
bspedersen@hum.ku.dk

Sanni Nimb
seniorredaktør
Det Danske Sprog- og Litteraturselskab
Chr. Brygge 1
DK-1219 København K
sn@dsl.dk

Anna Braasch
seniorforsker emeritus
Center for Sprogteknologi, NFI
Københavns Universitet
Njalsgade 140, bygn. 25.4.
DK-2300 København S
braasch@hum.ku.dk

Sussi A. Olsen
videnskabelig medarbejder
Center for Sprogteknologi, NFI
Københavns Universitet
Njalsgade 140, bygn. 25.4
DK-2300 København S
saolsen@hum.ku.dk