Semantisk processering og leksikografi

Bolette Sandford Pedersen

The paper argues that lexicographical knowledge is crucial for the development of intelligent language technology, i.e. technology that deals with language in a nuanced and all-compassing way. In this context, an overview is given of some of the developments and acknowledgements that we have achieved through a number of cooperative projects at The University of Copenhagen (Center for Language Technology) and The Danish Society for Language and Literature. Main achievements encompass a number of interlinked Danish language technology resources that rely on the same sense inventory as *Den Danske Ordbog*, and which comply with international standards. Reference is also made to our work on developing principled methods for sense clustering as well as to our experiments with partial language transfer from English to Danish when developing the Danish FrameNet.

1. Intelligent teknologi kræver sprogviden – ikke mindst leksikalsk viden

Intelligent teknologi står foran det helt store gennembrud, og semantisk processering udgør en vigtig komponent i denne udvikling. Semantisk processering kan helt kort defineres som behandling af indholdsdimensionen i talt eller skrevet sprog. Det kan indbefatte automatisk entydiggørelse af flertydige ord i kontekst, emneklassifikation og beregning af hvem i teksten der gør hvad, hvor og hvornår. Øvelser der kan lyde banale og ligetil, men som kun de allerfærreste sprogtjenester og robotter i dag mestrer på bare et minimalt niveau og kun for de store sprog. Virtuelle assistenter, informationssøgerobotter og automatisk resumering er alle eksempler på nogle af de systemer der kræver sprogviden, og som allerede nu er ved at blive en del af dagligdagen.

I det inviterede foredrag ved NFL 2017 i Reykjavík argumenterede jeg for vigtigheden af at den viden vi repræsenterer og genererer i de leksikografiske miljøer, gøres egnet til og tilgængelig for brug i teknologien; herunder særligt i systemer der arbejder med semantisk viden, ofte omtalt som systemer med kunstig intelligens. Ordbøger er ikke bare systematiske samlinger af ord med information om ordforrådets bøjning og syntaks; de er i sig selv kulturelle vidnesbyrd om det samfund de er udviklet i. Dén viden bør på længere sigt indgå med alle nuancer i den teknologi vi udvikler til brug i både privat- og arbejdsliv.

I denne artikel vil jeg i det store og hele følge foredragets tematik idet jeg vil argumentere for hvorfor det er vigtigt at vi i de relativt små sprogsamfund som de

nordiske er særligt bevidste om dette fokus på udnyttelse og tilgængeliggørelse – og samtidig forsøge at beskrive nogle af de muligheder og udfordringer der ligger i denne tilgang.

For mange af sprogmiljøerne i Norden udgjorde de såkaldte METANET-rapporter fra 2012[1] noget af en øjenåbner – eller i hvert fald et referencepunkt for hvordan vi på dét tidspunkt mente at vi var stillet rent sprogligt i forhold til den udbredte digitalisering og den nye teknologi i vores samfund. Vi vurderede at flere af de europæiske sprog var temmelig ringe stillet i forhold til de digitale udfordringer, og at der var et stort behov for opgradering. Dette gjaldt i høj grad de nordiske sprog der gennemgående lå i feltet med "ringe eller fragmentarisk støtte". Siden rapporterne er der sket en hel del udvikling også for de nordiske sprog, men gabet til engelsk er imidlertid stadig voksende.

Mit udgangspunkt for artiklen er baseret på mit eget arbejde i de senere år med udvikling af teknologiske sprogressourcer for dansk. Arbejdet er udført i et mangeårigt samarbejde mellem Center for Sprogteknologi på Københavns Universitet (KU) og Det Danske Sprog- og Litteraturselskab (DSL). Afsnit 2 præsenterer i kort form de ressourcer der er tale om, og forklarer overordnet hvordan de forholder sig til *Den Danske Ordbog* (DDO) og *Den Danske Begrebsordbog* (Begrebsordbogen) som er udviklet ved DSL. I afsnit 3 kommer jeg ind på begreberne *standardisering* og *sprogtransfer* – igen med udgangspunkt i de erfaringer vi har gjort os med vores danske data, særligt DanNet og Dansk FrameNet. Afsnit 4 omhandler specifikt udfordringerne omkring betydningsinventaret og beskriver vores arbejde med at udvikle en principiel metode til generering af betydnings*klynger*; en tilgang som letter den automatiske entydiggørelse betragteligt. Endelig kommer jeg i afsnit 5 ind på udviklingen i Europa og beskriver perspektiverne i et nyt infrastrukturprojekt, ELEXIS, som skal se på netop nogle af de problemstillinger som jeg har rejst i artiklen.

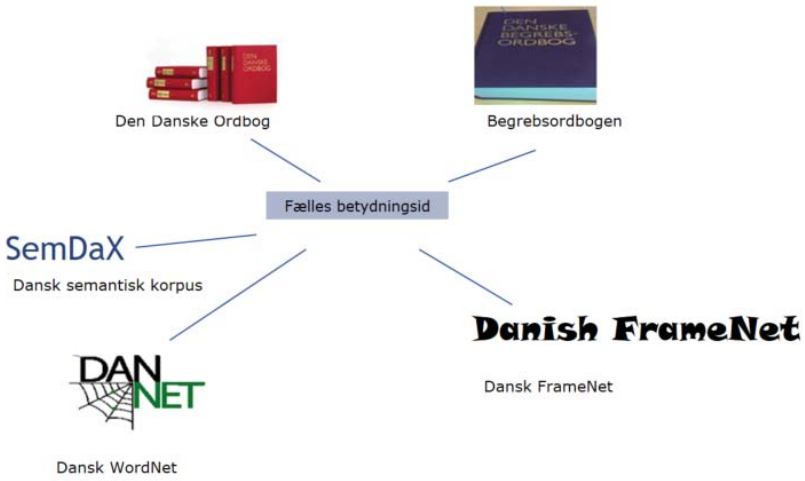
2. Flere sprogressourcer – samme betydningsinventar

Med udgangspunkt i flere forskningsrådsprojekter[2] er KU og DSL i gang med at udarbejde en portefølje af sprogteknologiske ressourcer der i større eller mindre grad hænger sammen med og/eller trækker på DDO og Begrebsordbogen. Fælles for disse ressourcer er at de er udarbejdet med reference til de samme betydnings-id'er forstået på den måde at alle ressourcerne tager udgangspunkt i det betydningsinventar og den betydningsstruktur som er anlagt i DDO (se Figur 1).

[1] Metanetrapporterne kan ses på <http://www.meta-net.eu/whitepapers/overview>.

[2] Følgende fonde har støttet følgende involverede projekter: "DanNet – et leksikalsk-semantisk ordnet for dansk" 2005–2008: Det Danske Forskningsråd, "DK-CLARIN" 2008–2011: Forsknings- og Innovationsstyrelsens nationale pulje for forskningsinfrastruktur, "META-NORD" – EU-projekt 2011–2013 (ICT PSP Program), "Semantic Processing across Domains" 2013–2017: Det Danske Forskningsråd og "Fra Begrebsordbog til Framenet" 2016–2017: Carlsberg Fondet.

Det betyder ikke at der ikke i flere af ressourcerne arbejdes med grovere betydningsinventarer, i wordnettet DanNet[3] kan der fx godt være flere underbetydninger i samme *synset*,[4] og fx er det semantiske korpus, SemDaX,[5] opmærket med et gradueret betydningsinventar spændende fra meget grove betydningsdistinktioner til meget finkornede. Men referencen til DDO's struktur er intakt, dvs. de grovere inventarer udgøres af *klynger* af DDO-betydninger sådan at linket til den samme basis ikke går tabt. Det fælles referencepunkt gør det langt nemmere at udvikle ressourcerne i samme takt, og det gør nemmere at bygge nye fordi man i højere grad kan genbruge den information som skal være tilgængelig i mere end én ressource.



Figur 1: Sammenhængende danske sprogressourcer med et fælles betydningsinventar/id.

Jeg vil ikke her komme nærmere ind på de enkelte sprogressourcers opbygning og struktur men henviser til følgende artikler der hver især beskriver dem i nærmere detaljer: *DanNet*: Pedersen et al. (2009), *SemDaX*: Pedersen et al. (2016), og *Dansk FrameNet*: Nimb & Pedersen (2016) og Nimb et al. (2017).[6]

---

[3] *DanNet*, det danske wordnet, kan downloades fra <http://wordnet.dk>. Det kan browses fra <http://wordties.cst.dk/wordties-dannet/>.

[4] Synsets er sæt af synonymer, som udgør den bærende enhed i wordnets.

[5] *SemDaX*, det semantisk opmærkede korpus, kan downloades fra github: <https://github.com/coastalcph/semdax>.

[6] Man kan desuden læse om Begrebsordbogens udnyttelse af og sammenhæng med DanNet og *Den Danske Ordbog* i Nimb (2013, 2016).

3. Sprogtransfer og standardisering af sproglige data

I forbindelse med at resultaterne fra METANET-rapporterne igennem de forløbne år er blevet analyseret og fortolket i forskellige sproglige og politiske sammenhænge, er begreberne *sprogtransfer* og *standardisering af sproglige data* blevet endnu mere centrale end tidligere. Nok er vores ordbøger unikke, og nok indeholder de meget viden om det samfund de er udarbejdet i, som kan være relevant også i teknologien, men alligevel er der sproglige genveje man både kan og bør slå når et givent sprogsamfund skal rustes til den digitale tidsalder.

På KU og DSL har vi arbejdet bevidst med disse to begreber idet vi allerede i begyndelsen af vores samarbejde var meget opsatte på at overholde eksisterende standarder for ressourcer til sprogteknologi. Således er de danske WordNet og FrameNet direkte kompatibelt med de tilsvarende ressourcer der bygges for andre sprog.[7] Det giver en række fordele idet der er en hel del principielle detaljer man ikke behøver tage stilling til, og fordi man kan overtage og lære en del af de ressourcer der allerede er bygget. Endelig indgår man ved at anvende en international standard i et forsknings- og udviklingsfællesskab på internationalt niveau som hele tiden bevæger sig videre og afprøver nye tilgange på ens data, hvilket er en stor fordel fordi de hermed efterprøves og valideres.

Lad mig også give to eksempler på vores arbejde med *transfer* af sprogviden fra engelsk til dansk, men vel at mærke i en kontekst hvor grunddataene stadig er danske. I Nimb & Pedersen (2016) og Nimb et al. (2017) viser vi at semantiske rolleinventarer (som Communicator, Recipient mv.) identificeret på baggrund af engelske korpusanalyser fungerer fint på danske data.

Her har vi taget udgangspunkt i Begrebsordbogens xml-opmærkede grundmanuskript og udnyttet de synonymklynger man finder for verberne til relativt hurtigt og effektivt at generere et dansk FrameNet hvor grupper af synonyme danske verber tilskrives associerede semantiske rammer (frames) adapteret fra Berkeley Framenet (Ruppenhofer et al. 2016). I Pedersen et al. (2018b) går vi skridtet videre med sprogtransfer-tilgangen idet vi også eksperimenterer med at kombinere de danske FrameNet-data med rolleopmærkede engelske korpora.

Vi viser her at det er muligt til en vis grad at maskinlære fra engelske rolleopmærkede data til dansk, men at læringen af dansk forbedres betydeligt når den kombineres med træning på danske data (i dette tilfælde med opmærkede data fra SemDaX).

Dette kan lyde ukontroversielt for professionelle sprogfolk, men i dag hvor meget sprogteknologi i praksis *kun* anvender sprogtransfer (og fx benytter sprogmodeller trænet på engelsk til danske systemer), er det vigtigt at kunne påvise at ja, vi kan maskinlære på tværs af sprog, men kvaliteten højnes betragteligt når vi kombinerer læringen med sprogspecifikke data på målsproget – også selv om der som i vores tilfælde kun er temmelig begrænsede data til rådighed.

---

[7] Der forestår dog stadig standardiseringsarbejde, fx konvertering af DanNet til LMF (jf. <http://ww.lexicalmarkupframework.org/>).

4. Betydningsinventaret – graduering af detaljeringsgrad

Jeg har nævnt at vi forholder os til den samme betydnings-id i alle vores ressourcer (nemlig DDO's), og at vi betragter dette som en styrke ved vores samlede portefølje. Det betyder dog ikke at vi har løst problemet med hvordan vi definerer det rette niveau af betydningsnuancering til sprogteknologiske formål. En væsentlig hindring for en direkte og helt umiddelbar udnyttelse af almindelige ordbøger til sprogteknologi er lige præcis deres ofte meget finkornede og ikke altid lige systematiske betydningsinddelinger. Betydningerne er meget svære at skelne imellem automatisk. Faktisk udgør betydningsinventarets kompleksitet i konventionelle ordbøger efter min vurdering den største hindring overhovedet for automatisk udnyttelse af den semantiske viden der ligger gemt her.

Dette virker i nogen grad paradoksalt idet den menneskelige bruger som regel føler sig rigtig godt informeret når han/hun konsulterer en ordbog; hvorfor er det da så vanskeligt at udnytte i teknologiske sprogtjenester? Sagen er selvfølgelig at betydningsinventaret udgør et kontinuum af svært afgrænselige betydninger som selv i konkrete teksteksempler er svære at blive enige om og at definere entydigt. Dette faktum lever vi som regel helt fint med, og vi opfatter ikke umiddelbart ordbogens klart afgrænsede hoved- og underbetydninger som værende i uoverensstemmelse med sprogets praksis. Snarere tværtimod: ordbogen hjælper os som regel fint til at få overblik over spændvidden af et ords betydning. Men når vi arbejder med ordbøgernes betydningsinddelinger i direkte sammenhæng med løbende tekst eller tale, bliver det klart at der på ingen måde er tale om et én-til-én-forhold.

I de sprogteknologiske miljøer har man derfor i de sidste årtier eksperimenteret meget med hvordan betydningsinventaret kan gøres anvendeligt til semantisk processering (se fx Izquirdo et al. (2009), McCarthy et al. (2016) og Pedersen et al. (2018a)). Forskningen går ad flere veje: fra udtræk og sammenlægning af leksikografiske betydningsbeskrivelser fra forskellige kilder (også kaldet vidensbaseret "sense clustering") til en ren korpusbaseret tilgang hvor man forsøger at drage slutninger fra de enkelte betydningsforekomster og herudfra inducere ordets betydningsprofil (også kaldet "word sense induction", som i McCarthy et al. (2016)). Ved KU og DSL har vi arbejdet med den først nævnte tilgang, og vi har arbejdet dels med at håndopmærke korpuseksempler, dels med automatisk entydiggørelse ved hjælp af maskinlæring (Martinez et al. 2015). Øvelsen går ud på at finde det rette nuanceringsniveau, ikke mere finkornet end at vores annotører og algoritmer kan håndtere det, og ikke mere grovkornet end at det stadig giver betydningsmæssig mening. Det vil sige så der stadig skelnes mellem de betydninger som er væsentlige for at sprogtjenesterne kan fungere tilfredsstillende.

I vores arbejde har vi taget spørgsmålsvar-systemer (a la Siris personlige assistent på iPhonen eller a la Watson, IBM's verdenskendte robot) som vores *prototypiske* case for hvilket detaljeringsniveau det er relevant at kunne håndtere. Spørgsmålsvar-systemer udmærker sig ved at kræve en mere nuanceret afdækning af hvad der foregår i en ytring, end fx almindelige søgetjenester. Omvendt

er det ikke nødvendigt for et spørgsmålsvar-system at fange alle de sproglige nuancer som der fx kræves i oversættelse, hvor den fulde betydning ideelt set skal gengives på målsproget.[8]

De seneste resultater (Pedersen et al. 2018a) tyder på at det er muligt at finde et passende leje mellem DDO's fine betydningsdistinktioner og de ontologiske "primitiver" som er angivet i DanNet, og dermed skabe et noget grovere inventar som annotører kan blive enige om når de tekstopmærker, og som også algoritmerne kan håndtere.

	Eksperiment 1 **Antal DDO- betydninger og enighed**		**Eksperiment 2** **Antal betydningsklynger og enighed**		**Eksperiment 3** **Antal betydningsklynger og enighed**	
Selskab	10	0,57	6	0,69	5	**0,91**
Plads	13	0,73	9	0,79	6	**0,74**
Slag	17	0,80	11	0,81	9	**0,79**
Skud	12	0,61	12	0,61	11	**0,69**
Skade	6	0,73	5	0,68	4	**0,80**
Kort	10	0,71	4	0,77	3	**0.78**
Vold	9	0,33	7	0,67	5	**0,77**
Hul	14	0,72	11	0,61	7	**0,78**
Blik	7	0,59	6	0,61	4	**0.61**
Model	8	0,74	7	0,84	6	**0,89**

Tabel 1: Tre opmærkningseksperimenter med polyseme substantiver med forskellig detaljeringsgrad og forskellig annotørenighed (IA).

I vores eksperimenter satte vi to annotører til at opmærke ti meget polyseme og samtidig højfrekvente substantiver (*selskab, plads, slag, skud, skade, kort, vold, hul, blik* og *model*) i tekster fra forskellige teksttyper og domæner. Vi udtrak 100 + 15n eksempler for hvert af de polyseme ord, hvor *n* repræsenterer antallet af betydninger i DDO. Det vil sige at jo flere betydninger der er beskrevet for et ord, desto flere eksempler opmærkes der.

I det første eksperiment indgår alle DDO's betydninger i tagsættet. I det andet eksperiment reducerer vi derimod antallet af betydninger ved at klynge underbetydninger inden for én hovedbetydning sammen hvis de er af samme ontologiske type, fx Human.[9] Det betyder at DDO-betydningerne selskab_2 og selskab_2a der hhv. dækker betydningerne 'gruppe af personer som foretager sig noget i fællesskab' og 'gruppe af personer som er samlet til en fælles social aktivitet, fx en fest eller en middag', lægges sammen. I det sidste eksperiment slår vi også betydninger sammen som går på tværs af hovedbetydninger hvis de er af samme ontologi-

[8] Til maskinoversættelse bruger man desuden særlige teknikker til entydiggørelse, idet man her har en særlig "videnbase" at trække fra så at sige, nemlig tidligere oversatte tekster.

[9] DanNet anvender EuroWordNets topontologi (jf. Vossen et al. 1991).

ske type, fx PhysicalLocation ved *plads* i betydningen 'plads til noget' (plads_1) og 'siddeplads' (plads_2).

For alle tre eksperimenter beregnede vi annotørenigheden ved hjælp af Krippendorffs alpha (Krippendorff 2011). Et af de særlige træk ved denne beregningsmetode i forhold til fx almindelig procentberegning er at enighedskoefficienterne her korrigeres for sandsynlighed. Dvs. man modregner det faktum at det alt andet lige er nemmere at blive enige om få betydninger end mange. Det betyder at de tre annotørenighedskolonner i tabel 1 skulle være direkte sammenlignelige.

Værdierne varierer fra 0 til 1 hvor 0 er fuldstændig uenighed og 1 er fuldstændig enighed. Det ses tydeligt at enigheden øges for de grovere betydningsklynger (markeret med fed). Faktisk er det først i tredje eksperiment at annotørenigheden nærmer sig noget praktisk anvendeligt (0.78) idet man inden for semantisk processering antager at resultater med $\alpha \geq .67$ er nogenlunde acceptable.

Vores grundpræmis i denne metodeudvikling er at maskinen ikke kan forventes at skelne bedre mellem betydninger end de menneskelige annotører kan. Og vi har med den menneskelige annotation påvist at der i sandhed er tale om en temmelig vanskelig opgave siden vi faktisk kun for ét af ordene (*selskab*) kommer over 0.90 i enighed selv med de grove betydningsklynger. Maskinlæringseksperimenterne understøtter vores overordnede antagelse; også her opnås de bedste resultater (også sammenlignet med en tilfældig sammenligning (*random baseline*)) med de reducerede betydningsklynger (Pedersen et al. 2018b:5).

Vores undersøgelse ser kun på et udvalg af meget polyseme substantiver, og selv om vores klyngemetode er principiel, dvs. kan udføres automatisk på hele ordforrådet, så ved vi endnu ikke om den er praktisk anvendelig for den største del af substantiverne (som gennemgående ikke er nær så polyseme som dem vi har undersøgt her). Vi ved heller ikke hvordan metoden virker på adjektiver og verber; måske skal man her identificere nogle andre træk at lave klynger ud fra end blot de ontologiske.

5. Europæiske tiltag

Hvordan ser det ud på europæisk plan mht. at genanvende viden i eksisterende ordbøger til teknologi? På trods af flere store europæiske projekter op gennem nullerne initieret af netop de datalingvistiske miljøer, så som EAGLES, PAROLE og SIMPLE (Lenci et el. (2000), Calzalori et al. (2002) m.fl.) går det ikke imponerende godt. Selv om disse projekter bragte initiativer som the Text Encoding Initiative[10] og Lexical Markup Framework[11] med sig, har forhindringer i form af intellektuelle rettigheder og meget individuelle tilgange i de forskellige leksikografiske miljøer endnu vanskeliggjort det helt store gennembrud.

[10] TEI: <http://www.tei-c.org/>.

[11] LMF: <http://www.lexicalmarkupframework.org/>.

I 2013 oprettedes imidlertid det europæiske netværk e-Lexicography (EneL)[12] som en såkaldt EU COST Action. Dette initiativ blev igangsat for at forbedre adgangen til anbefalede ordbøger og gøre dem mere almindeligt kendte for et større publikum. Netværket lagde ud med 34 medlemmer fra 20 lande, men voksede i 2017 til 285 medlemmer fra 31 lande. I forbindelse med netværket opstod der et klart behov for en bredere og mere systematisk udveksling af ekspertise til etablering af fælles standarder og løsninger til udvikling og integration af leksikografiske ressourcer og for at udvide anvendelsesområdet til at indbefatte sprogteknologi og digital humaniora.

Som en udløber af ENeL-bestræbelserne blev der derfor etableret et konsortium, og i august 2017 blev forslaget til en infrastruktur under navnet ELEXIS udvalgt til finansiering som et projekt under Horizon 2020. De vigtigste mål for denne nye infrastruktur er i) at fremme samarbejde og vidensudveksling mellem forskellige leksikografiske forskningsmiljøer bl.a. for at mindske kløften mellem de mindre sprogmiljøer og dem med avanceret e-leksikografisk erfaring, ii) at arbejde med strategier, værktøjer og standarder for uddragning, strukturering og sammenkobling af leksikografiske ressourcer, iii) at lette adgangen til standarder, metoder, leksikografiske data og værktøjer, og sidst men ikke mindst iv) at tilskynde til en open access-kultur i den leksikografiske verden i overensstemmelse med de henstillinger der nu fremføres både fra EU-Kommissionens side og på nationalt plan mange steder, bl.a. i det danske forskningsministerium.

Det sprogteknologiske aspekt har et særligt fokus i projektet, og da både KU og DSL deltager som danske partnere i projektet, forventer vi at resultaterne fra det hidtidige danske samarbejde i høj grad vil kunne udnyttes og videreudvikles de næste 4 år. Vi håber desuden at kunne bidrage til udvikling af nye metoder og værktøjer til harmonisering af ordbogsformater og til automatisk segmentering og identifikation af ordbogsstruktur der muliggør udtræk og sammenkædning af ordbogsindhold – også på tværs af sprog. Endelig indgår der i projektet en arbejdspakke der skal se på netop metoder for at udvikle graduerede betydningsinventarer som er håndterbare til automatisk entydiggørelse.

Jeg vil gerne afslutningsvis nævne to andre europæiske initiativer under EU-programmet "Connecting Europe Facility" som også Dansk Sprognævn deltager i sammen med KU. Det drejer sig om European Language Resource Coordination[13] og eTranslation TermBank; begge initiativer der skal fremme genbrug og open access til sprogressourcer særligt til brug for EU's maskinoversættelsessystem.

Til sammen vidner alle disse initiativer om en stigende interesse og forståelse for hvorfor det er vigtigt at gøre både leksikografiske, terminologiske og andre sprogdata direkte tilgængelig for teknologi.

---

12 Jf. <http://www.elexicography.eu/>.

13 ELRC-SHARE: <https://elrc-share.eu/>.

Litteratur

Ordbøger og sprogteknologiske ressourcer

DDO = *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab. <http://ordnet.dk/ddo> (1.11.2017).

Begrebsordbogen = Nimb, S., H. Lorentzen, T. Troelsgård, L. Theilgaard & L. Trap-Jensen (2014): *Den Danske Begrebsordbog*, København: Det Danske Sprog- og Litteraturselskab.

Anden litteratur

Calzolari, Nicoletta, Antonio Zampolli & Allessandro Lenci (2002): Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. I: *CICLing 2002: Computational Linguistics and Intelligent Text Processing*. Berlin/Heidelberg: Springer, 264–279.

Izquirdo, Rubén, Armando Suárez & German Rigau (2009): An empirical study on class-based word sense disambiguation. I: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. The Association for Computational Linguistics, 389–397.

Krippendorff , K. (2011): Agreement and Information in the Reliability of Coding. I: *Communication Methods and Measures* 5(2): 93–112.

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* 13(4): 249–263.

Martínez Alonso, Héctor, Anders Johannsen, Sussi Olsen, Sanni Nimb, Nicolai Hartvig Sørensen, Anna Braasch, Anders Søgaard & Bolette Sandford Pedersen (2015): Supersense tagging for Danish. I: *Proceedings of the 20th Nordic Conference of Computational Linguistics. NODALIDA 2015*. Linköping Electronic Conference Proceedings #109*,* Linköping: ACL Anthology/Linköping University Electronic Press, 21−29.

McCarthy, D., M. Apidianaki & K. Erk (2016): Word Sense Clustering and Clusterability. I: *Computational Linguistics* 42(2):245−275.

Nimb, Sanni (2013): Leksikalsk-semantisk information i en ny dansk begrebsordbog. I: Dorthe Duncker, Anne Mette Hansen & Karen Skovgaard-Petersen (red.), *Betydning og Forståelse*. *Festskrift til Hanne Ruus*. København: Selskab for Nordisk Filologi, 251–266.

Nimb, Sanni (2016): Der er ikke langt fra tanke til handling. Om semantiske typer og systematisk polysemi i Den Danske begrebsordbog. I: *Danske Studier* 2016. Universitets-Jubilæets danske Samfund, 25–59.

Nimb, Sanni & Bolette S. Pedersen (2016): Fra begrebsordbog til sprogteknologisk ressource: verber, semantiske roller og rammer – et pilotstudie. I: Asgerd Gudiksen & Henrik Hovmark (red.), *Nordiske Studier i Leksikografi* 13. Skrifter udgivet af Nordisk forening for Leksikografi, Skrift nr. 14, 405–415.

Nimb, Sanni, Anna Braasch, Sussi Olsen, Bolette Sandford Pedersen & Anders Søgaard (2017): From Thesaurus to FrameNet. I: *Electronic Lexicography in the 21st century*. *Proceedings of eLex 2017 conference*, Leiden, 1–22.

Pedersen, Bolette.S, Sanni Nimb, Jørg Asmussen, Nicolai Sørensen, Lars Trap-Jensen & Henrik Lorentzen (2009): DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. I: *Language Resources and Evaluation* 43(3):269–299.

Pedersen, Bolette S., Lars Borin, Markus Forsberg, Neeme Kahusk, Krister Lindén, Jyrki Niemi, Niklas Nisbeth, Lars Nygaard, Heili Orav, Eirikur Rögnvaldsson, Mitchell Seaton, Kadri Vider & Kaarlo Voionmaa (2013): Nordic and Baltic Wordnets Aligned and Compared through "WordTies". I: *Proceedings from the 19th Nordic Conference on Computational Linguistics (NODALIDA 2013)*. Linköping Electronic Conference Proceedings 85. Linköping: Linköping University Electronic Press.

Pedersen, Bolette S., Anna Braasch, Anders Johanssen, Héctor Martínes Alonso, Sanni Nimb, Susi Olsen, Anders Søgaard & Nicolai Hartvig Sørensen (2016): The SemDaX Corpus – sense annotations with scalable sense inventories. I: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož: European Language Resources Association (ELRA), 842−847.

Pedersen, Bolette S., Manex Agirrezabal, Sanni Nimb, Sussi Olsen, Ida Rørmann (2018a): Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In: *Proceedings of Global WordNet Conference 2018*, Singapore. <http://compling.hss.ntu.edu.sg/events/2018-gwc/pdfs/GWC2018_paper_27.pdf> (1.4.2018).

Pedersen, Bolette S., Sanni Nimb, Anders Søgaard, Mareike Hartmann & Sussi Olsen (2018b): A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier. Accepteret ved *LREC 2018*, Japan.

Ruppenhofer, J., M. Ellsworth, M. R. L. Petruck,, C. R. Johnson, C. F. Baker & J. Scheffczyk (2016). *FrameNet II: Extended Theory and Practice*. <https://framenet.icsi.berkeley.edu/fndrupal/the_book> (1.11.2017).

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

Bolette Sandford Pedersen
Professor, ph.d.
Center for Sprogteknologi, Institut for Nordiske Studier og Sprogvidenskab
Københavns Universitet
Njalsgade 136, DK-2300 København S
bspedersen@hum.ku.dk