Dansk betydningsinventar i et datalingvistisk perspektiv

Af Bolette Sandford Pedersen, Sanni Nimb og Sussi Olsen

In this paper we investigate the Danish sense inventory from a paradigmatic and a syntagmatic perspective, respectively, and we present a collection of related lexical semantic resources that we have developed in collaboration between The Society for Danish Language and Literature and The University of Copenhagen. The resources comprise a Danish wordnet (DanNet), The Danish FrameNet Lexicon, and The Danish Sentiment Lexicon. All three resources are designed to enable semantic processing to be used in digital humanities research as well as more broadly in language-centric technology development. Finally, in order to illustrate the use of the resources when processing running text, we provide some annotation examples of each resource.

1 En datalingvistisk tilgang til leksikalsk semantik

En datalingvistisk beskrivelse af ordenes semantik er ikke bare af sprogvidenskabelig interesse; den er også relevant som basis for sprogcentreret teknologi der skal håndtere ordbetydning og sætningssemantik på en hensigtsmæssig og fleksibel måde (jf. fx Hershcovich & Donatelli 2021). Digitale assistenter på vores mobiltelefoner, sentimentanalyse af brugerholdninger på de sociale medier eller mere avancerede systemer med indbygget kunstig intelligens er eksempler på sådanne digitale anvendelser,[1] men også mere forskningsrettede anvendelser som distant reading og topic modelling der anvendes i digital humaniora (jf. fx Tangherlini & Leonard 2013, Bjerring-Hansen et al. 2019), og som gør det muligt at undersøge temaer og trends i store mængder af litteratur på en kvantitativt systematisk måde, trækker i forskellig udstrækning på semantisk viden. I artiklen udforsker vi det danske betydningsinventar, og vi præsenterer den samling af semantiske sprogressourcer som vi har udviklet i et samarbejde mellem Det Danske Sprog- og Litteraturselskab (DSL) og Center for Sprogteknologi (CST) ved Københavns Universitet, idet vi sætter dem ind i en bredere sprogvidenskabelig sammenhæng.

Den drivende kraft i det forsknings- og udviklingssamarbejde vi beskriver, har været at understøtte at dansk sprogteknologi og digital huma-

1 Se fx sprogteknologirapporten *Dansk Sprogteknologi i verdensklasse – rapport fra sprogteknologiudvalget* (Kirchmeier et al. 2019) for en oversigt over sprogteknologiske anvendelser der kræver semantisk viden.

nistisk forskning på danske data i stedet for udelukkende at være tilpasset fra engelsk baseres på *lokalt forankret viden om sprog og kultur fra det samfund som det interagerer i*. Kongstanken er at eksisterende danske ordbøger og andre danske sprogressourcer er mere end blot systematiske samlinger af ord forsynet med information om morfologi og syntaks; de er kulturelle vidnesbyrd i den forstand at de med deres ordforråd og definitioner beskriver det samfund og den kultur som de er skabt i. Den viden de rummer, er derfor i høj grad relevant for de teknologiske systemer der lige nu er under udvikling til brug overalt i samfundet, ligesom den er essentiel for digitale forskningsanalyser af danske kulturdata. Vores tilgang betyder i al sin enkelthed at vi tager afsæt i allerede anerkendte danske ordbøger og tekstkorpusser som igennem årene er blevet udviklet primært ved DSL, nærmere betegnet de to ordbøger Den Danske Ordbog (ordnet.dk/ddo, herefter DDO) og Den Danske Begrebsordbog (Nimb et al. 2014, herefter Begrebsordbogen) samt de to tekstkorpusser KorpusDK og CLARIN.

I processen med at beskrive betydningsinventaret og den leksikalske semantik på en formaliseret måde går vi dels i bredden ved at indbefatte de væsentligste dele af ordforrådet – baseret på de ovenfor nævnte ressourcer – dels i dybden ved at beskrive og kode både paradigmatiske og syntagmatiske egenskaber for de enkelte ord. Disse egenskaber er for en stor dels vedkommende allerede udtrykt implicit i de leksikografiske værker vi henholder os til (via betydningsdefinitioner og eksempler i DDO og semantiske kategoriseringer i Begrebsordbogen), men de er det på en måde som ikke er umiddelbart anvendelig i sprogteknologi. Det skyldes bl.a. at man under udviklingen af 'almindelige' ordbøger ikke har haft nævneværdig interesse i at overholde internationale datalingvistiske *standarder* der nemt kan integreres i systemudvikling – et aspekt som vi til gengæld har vægtet højt i udviklingen af de teknologiske sprogressourcer som vi beskriver her.

Som et supplement til de udviklede teknologiske sprogressourcer, som i høj grad kan anses for 'håndkodede', dvs. manuelt etableret af datalingvister ud fra læsning af de eksisterende oplysninger i ordbøgerne, forholder vi os også til de distributionelle sprogmodeller, som kan beregnes ud fra meget store tekstmængder i form af statistiske ordprofiler, såkaldte wordembeddings (jf. Mikolov et al. 2013, Devlin et al. 2019 m.fl.). Selv om ordbøger i dag er korpusbaserede i den forstand at leksikograferne via korpuskonkordanser undersøger de ord der beskrives, så har wordembeddings vist sig at være et særdeles stærkt værktøj, som først for nylig er begyndt at finde anvendelse i leksikografisk sammenhæng (som i Sørensen & Nimb 2018, Olsen et al. 2020 og Ahmadi et al. 2020). Ved at sammen-

holde betydningsskel udviklet af mennesker med de meget omfattende automatisk udledte ordprofiler i form af vektorer, kan vi blive klogere på hvilke skel der er umiddelbart verificerbare via den distributionelle kontekst (og derfor umiddelbart håndterbare for computeren), og hvilke der synes mere subtile. De sidstnævnte er måske afledt af den måde hvorpå vi som mennesker – med en veludviklet evne til at anvende metaforik og til dynamisk at skabe overførte betydninger – opfatter og beskriver verden, jf. fx Lakoff & Johnsons ikoniske Metaphors We Live By fra 1980.

Endelig sætter vi i artiklen de beskrevne sprogressourcer ind i en praktisk kontekst ved at eksemplificere hvad hver enkelt sprogressource kan bibringe en automatisk tekstanalyse.

2 Flere dimensioner af betydningsbeskrivelse

2.1 Paradigmatisk beskrivelse

I vores tilgang anskuer vi ordbetydning på en formaliseret måde dels fra en paradigmatisk, dels fra en syntagmatisk vinkel (jf. blandt andre Hjelmslev 1966 og Jakobsen 2008). I det *paradigmatiske* perspektiv betragtes og beskrives et ords betydning primært ud fra hvilke andre ord det pågældende ord kan erstattes med i konteksten, og som det dermed relaterer til på forskellige måder, fx via et synonymi- (*smuk – skøn*), hyponymi- (*dyr – hest*), meronymi- (*finger – hånd*) eller antonymiforhold (*varm – kold*). Herudfra udledes centrale forhold omkring ordets betydning, og ordet (eller rettere begrebet) forstås på denne måde som en del af et komplekst netværk hvor bl.a. den semantiske lighed mellem to begreber formelt kan måles, og hvorfra nedarving af semantiske egenskaber kan beregnes. En central antagelse er således at begreber der ligger tæt på hinanden i det semantiske netværk, også ligner hinanden semantisk – og denne viden er særdeles relevant i mange sprogteknologiske applikationer der indbefatter sprogforståelse i en eller anden form, fx søgemaskiner og digitale assistenter.

Taksonomier og netværk fungerer således som en slags vidensbaser for teknologien, og de kan have forskellig udformning og gå under forskellige betegnelser som *ontologier* (jf. fx Guarino & Musen 2015), *begrebssystemer* (Madsen 2005) eller *wordnets* (Fellbaum and Miller (udg.) 1989). Fælles for vidensbaserne er at de har indkodet over- og underbegreber og andre centrale relationer imellem begreber (eller klasser af begreber), men de divergerer også fra hinanden på flere punkter. Begrebssystemer an-

vendes fx særligt inden for specifikke terminologiske fagområder, mens ontologier kan være både fagspecifikke og almene. Wordnets, som vi interesserer os særligt for i denne artikel, adskiller sig fra ontologier og begrebssystemer ved at være mere sprognære og sprogspecifikke, de indeholder ikke i udpræget grad metasproglige begreber løsrevet fra specifikke sprog.[2] Wordnets beskæftiger sig i øvrigt primært med almensproget og indeholder fx kun ganske få udvalgte proprier. På den måde har wordnets flere ligheder med ordbøger – hvor ontologier og begrebssystemer i højere grad beskriver verdensviden på samme måde som en encyklopædi (Wikipedia har fx en del ontologiske egenskaber) eller et leksikon.

Beslægtet med wordnets er formaliserede leksikografiske netværk som SIMPLE (Lenci et al. 2001, Nimb & Pedersen 2000, Pedersen & Paggio 2004) som er baseret på Pustejovskys teori om *qualiastruktur* (Pustejovsky 1995). Qualiateorien forsøger at indbefatte en præcist afgrænset del af den generelle omverdensviden som anses for relevant i den leksikalsk-semantiske beskrivelse. Dette udgøres af fire roller som til sammen beskriver de mest centrale dimensioner, nemlig 1) *den formelle rolle*, dvs. begrebets tilhørsforhold i en taksonomi (med angivelse af overbegrebet), *den konstitutive rolle* som angiver øvrige paradigmatiske relationer af typen del-helhed, 3) *den teliske rolle* som angiver en genstands formål eller funktion, og endelig 4) *den agentive rolle* som angiver de faktorer der er involveret i frembringelsen af en genstand. Den samlede qualiastruktur beskriver altså mere end blot paradigmatisk viden (den formelle rolle), og i Pustejovskys samlede leksikonstruktur, Det Generative Leksikon, indgår udover qualiastruktur også syntagmatisk viden i form af argumentstruktur og eventstruktur. Hertil kommer en række generative mekanismer, som baseret på qualiastrukturen skal forklare hvordan ords betydning kan variere og transformeres afhængig af de omgivende ord.

Qualiadimensionerne svarer ikke overraskende i store træk til de informationstyper som en klassisk betydningsdefinition i en ordbog indeholder (genus proximum og differentia),[3] og det danske wordnet, DanNet,[4] et

---

2 Dog indeholder fx Princeton WordNet (3.1) udvalgte sammensatte metabegreber af typen *male horse*.

3 Tag fx definitionen på en bog: *trykte eller beskrevne blade af papir indbundet eller på anden måde sammenhæftet i rækkefølge så de danner en helhed, ofte en sammenhængende tekst, beregnet på at blive læst*. Her beskrives overbegreb (i form af dele), tilblivelse samt formål. Jf. også Svensén 2004.

4 Se wordnet.dk.

sprogteknologisk leksikon som er semiautomatisk genereret fra DDO's definitioner (se afsnit 4.1), kan betragtes som en hybrid mellem et wordnet og en SIMPLE-base i og med at det indeholder alle de relationer som er angivet i qualiastrukturen.[5] Et begreb som *kage* beskrives således dels med *den formelle rolle* i form af overbegrebet *bagværk*, som typisk indeholder *mel* og *sukker* (*den konstitutive rolle*), tilbragt ved hjælp af *bagning* (*den agentive rolle*) med det formål at blive *spist* (*den teliske rolle*).

Hvis vi vender tilbage til den førnævnte antagelse om at ord der ligger tæt på hinanden i et semantisk netværk, også ligner hinanden semantisk, ledes vi naturligt videre til den *distributionelle* hypotese om at et ords betydning først og fremmest er en funktion af de kontekster det kan optræde i (jf. Firth 1957, Levin 1993, Lenci 2008 blandt flere). Herudfra må man udlede at ord der ligger tæt på hinanden i det semantiske netværk, også forekommer i lignende kontekster i et korpus – og omvendt. En statistisk tilgang til ordbetydning som komplementerer og supplerer den vidensbaserede i de fleste sprogteknologiske systemer i dag, er de føromtalte wordembeddings, som er beregnet via dyb eller neural læring på meget store korpora.[6] Koblingen til den *distributionelle* hypotese er tydelig i og med at beregningen af wordembeddings er baseret på en antagelse om at ord med lignende vektorer i et vektorrum har stor semantisk lighed.

2.2 Syntagmatisk beskrivelse

Den distributionelle kontekst er også vigtig i det *syntagmatiske* perspektiv hvor ordbetydning anskues ud fra hvordan ordet relaterer sig til de omgivende syntagmer i sætningen. Her er det væsentligt at afgrænse hvilke af disse egenskaber der er *leksikalsk* styrede, altså inhærente for det pågældende ord. Det er typisk valensbærende ord der er genstand for en sådan beskrivelse. Kommunikationsverber knytter fx typisk både en agens der kommunikerer, noget der bliver kommunikeret og evt. også en modtager til sig (*jeg fortalte hende historien*, hvor *jeg* er agens, *historien* det der kommunikeres, og *hende* er modtageren), hvorimod sanseverber typisk knytter en sansende og noget sanset til sig (*jeg indsnusede madduften*, hvor *jeg* er den sansende og *madduften* det der sanses). Udover før-

5 Dette er i modsætning til fx Princeton WordNet som ikke indeholder funktionsrelationer.

6 I sprogmodeller baseret på dyb læring og kunstige neurale netværk modellerer man abstraktioner i sprogdata på et relativt højt (eller dybt) niveau ved at anvende mange proceslag med komplekse strukturer.

nævnte teorier der arbejder med argumentstruktur, er teorien om Frame Semantics velegnet når man vil opmærke sådanne semantiske roller på en systematisk måde. Teorien kombinerer teoretisk semantik med praktisk leksikografi (Fillmore 1968; Fillmore & Atkins 1992), og de konkrete leksikalske databaser udmøntes i såkaldte *framenets* (jf. Berkeley FrameNet: https://framenet.icsi.berkeley.edu/fndrupal og Ruppenhofer et al. 2016; svensk framenet, SweFN, Dannéls et al. 2021), se afsnit 4.2 for beskrivelse af det danske FrameNet-leksikon som vi har samarbejdet om at udvikle, og Bick (2011 og 2017), som også beskriver arbejde med det danske ordforråd og opmærkning af danske tekster baseret på teorien. Grundtanken i teorien er at de valensbærende ord hver især udløser en bestemt, navngiven semantisk 'frame', på dansk ramme, hvori der indgår en række prædefinerede semantiske roller i form af forskellige deltagere og elementer, de såkaldte *frame elements*. På den måde går frame semantics et spadestik dybere end klassisk argumentstruktur: Rammelementerne, der altså udgøres af argumenter eller semantiske roller der er fælles for et antal verber med meget beslægtet betydning, specificeres med særlige egenskaber afhængigt af lige præcis hvilken navngiven ramme de fremkaldes af.

Framenets er, ligesom wordnets, organiseret i et taksonomisk netværk hvor nogle rammer er mere overordnede, mens andre er meget specifikke og arver frame elements fra deres overordnede ramme (fx en overordnet kommunikationsframe vs. en mere specifik *skælde ud*-frame). Som en vigtig del af et framenet hører, udover det leksikon der beskriver rammerne for de enkelte verber (og tilhørende verbalsubstantiver), også et tekstkorpus hvori de leksikalske enheder optræder, og hvor hver enkelt sætning er håndopmærket med frameværdi og frame elements (se mere i afsnit 4.2). Disse korpora bruges bl.a. til at maskinlære rolletilskrivning (*Semantic Role Labeling*) så man efterfølgende automatisk kan beregne hvem der gør hvad, hvor og hvornår i en ytring (se bl.a. Pedersen et al. 2018a).

2.3 Denotation og konnotation

Den paradigmatiske og den syntagmatiske ordbeskrivelse dækker i udgangspunktet ordets eksplicitte betydning, eller dets såkaldte *denotation*. Et udsnit af ordforrådet har imidlertid også en sekundær medbetydning, en *konnotation* som kan være enten positiv eller negativ. Særligt de negativt ladede ord angives i mange ordbøger med en note om at de fx har en nedsættende betydning (jf. Svensén 2004), i DDO har fx 1,5 % af de beskrevne lemmaer en oplysning om at ordet i en af sine betydninger i

en eller anden grad er nedsættende. I en datalingvistisk sammenhæng er det imidlertid nyttigt at angive konnotation for et langt større udsnit af ordforrådet – hvis man fx vil lave sentimentanalyse –, og det er ligeledes nyttigt at kunne angive en systematisk graduering af i hvor høj grad et ord tillægger negativ eller positiv værdi til et udsagn. I afsnit 4.3 beskrives udformningen af den danske sentimentordbog.

3 Betydninger og tekst

3.1 Semantisk opmærkning

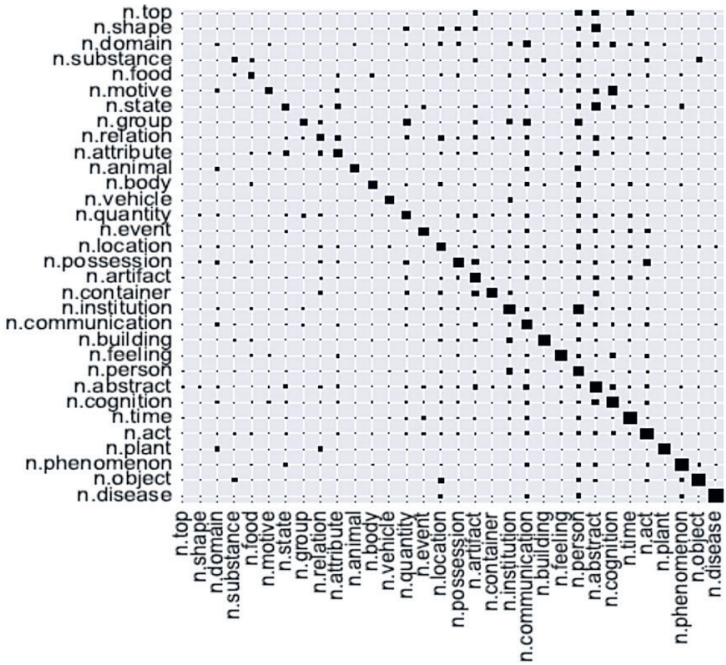
Som vi allerede har antydet ovenfor, udgør tekstmateriale der er håndopmærket på forskellig måde med formaliserede oplysninger om hvad ordene betyder ud fra en given standard, en central semantisk ressource til sprogteknologisk udvikling. Manuelt opmærket tekst bruges til at træne sprogmodeller til at kunne omfatte semantisk viden. De håndopmærkede oplysninger ved de enkelte ord i teksten bruges til at sikre at sprogmodellerne har den nødvendige viden som basis for fx automatisk entydiggørelse af flertydige ord og udtryk og automatisk bestemmelse af de semantiske roller (hhv. såkaldt *sense tagging* og som ovenfor nævnt *Semantic Role Labeling*). Dette vel at mærke kun når der er tilstrækkeligt med håndopmærkede oplysninger til at man maskinelt kan udlede et bestemt mønster i opmærkningerne.

De semantiske opmærkninger som vi anvender, følger de samme principper om paradigmatiske og syntagmatiske oplysningstyper som er beskrevet ovenfor, og i udgangspunktet er der tale om det inventar af betydninger og semantiske rammer der er udviklet for de leksikalske ressourcer i form af DanNet og Det Danske FrameNet-leksikon.[7]

En vis procentdel af materialet bør principielt opmærkes af flere end én annotør for at sikre en vis form for konsensus. For at en gruppe sproglige annotører kan opmærke nogenlunde ensartet, bør der, udover det givne opmærkningsinventar (*annotation scheme*) der skal anvendes, også udarbejdes en kodningsmanual der guider annotørerne når de skal træffe valg i forhold til fx ord- og syntagmeafgrænsning og i andre tvivlspørgsmål. Det er kutyme fx at dobbeltopmærke mindst 2% af korpusmaterialet, men i vores semantiske korpus (SemDaX, Pedersen et al. 2016) har vi dobbelt-

7 https://korpus.dsl.dk/resources/details/framenet.html

opmærket helt op til 60% af materialet. Det har vi gjort for at blive klogere på hvad der ligger til grund for uoverenstemmelser mellem annotørerne. På den måde får vi nemlig både vigtig viden om sprogets forskellige fortolkningsmuligheder i form af fx systematisk polysemi (det at ord med beslægtet betydning udviser samme form for flertydighed[8]) samt vigtig feedback til det annotationsskema der anvendes.



Figur 1: Annotørdiskrepanser for substantiver i SemDax (fra Olsen et al. 2015)

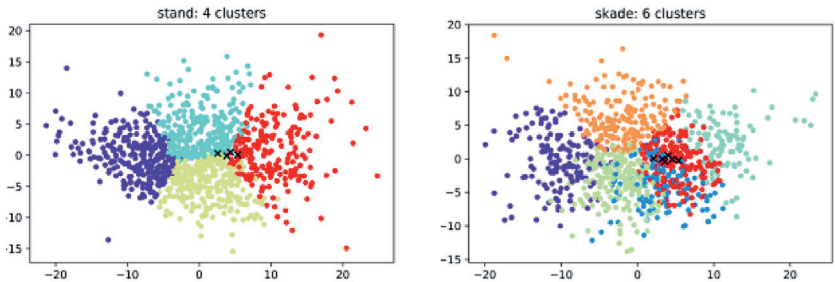
Figur 1 ovenfor viser således uoverensstemmelserne i annotørernes opmærkning af substantivers betydning i SemDaX. Man kan bl.a. se at der er relativ stor enighed mellem dem når det gælder fx opmærkning af sygdomme (én meget stor boks ud for *disease* som derfor er placeret nederst i diagonalen) og mindre enighed når der skal skelnes mellem systematisk polysemi, fx når der skal skelnes imellem om et ord som *skole* i en given kontekst refererer til den gruppe personer der arbejder på skolen, eller mere bredt til skolen som institution (to halvstore prikker ud for hhv. *person* og *institution*).

8 Fx *bog, avis* og *hæfte* refererende til dels indhold, dels fysisk beskaffenhed.

3.2 Wordembeddings og leksikografisk beskrivelse

Man må formode at betydningsnuancer der skaber stor annotøruenighed under håndopmærkningen, er tilsvarende svære at udlede fra de distributionelle wordembeddings; man kan altså antage at deres vektorrum er sværere at skille ud fra hinanden.

I det hele taget er det ikke ligetil at analysere og evaluere resultaterne af wordembeddings da der er tale om meget komplekse, multidimensionale vektorer som i udgangspunktet opererer på karakter- eller ordformsniveau og ikke på betydningsniveau. Nyere embeddingmodeller som BERT (Devlin et al. 2019) opererer dog med såkaldte senseembeddings som i højere grad afspejler de forskellige ordbetydninger. I figur 2 ses to eksempler på de vektorrum som hhv. de polyseme substantiver *stand* og *skade* har i DSL's wordembeddings, beregnet med word2vec og DSL's interne korpus på 1 milliard løbende ord. Ordene har hhv. 4 og 6 betydningsklynger hvor meget tætte betydninger er lagt sammen. De sorte kryds er de DDO-udledte eksempler som klyngerne er genereret ud fra, og dimensionaliteten af vektorrummet er reduceret til 2D af hensyn til visualiseringen.



Figur 2: Betydningsklynger for hhv. stand og skade (Olsen et al. 2020) beregnet fra DK-Korpus.

Af de to figurer kan man udlede at begge ords betydningsbeskrivelse kan verificeres distributionelt, men at *stand*s fire betydninger differentierer sig mere tydeligt i vektorrummet end *skade*s seks betydninger.[9] Dette svarer godt overens med det faktum at enigheden er højere blandt an-

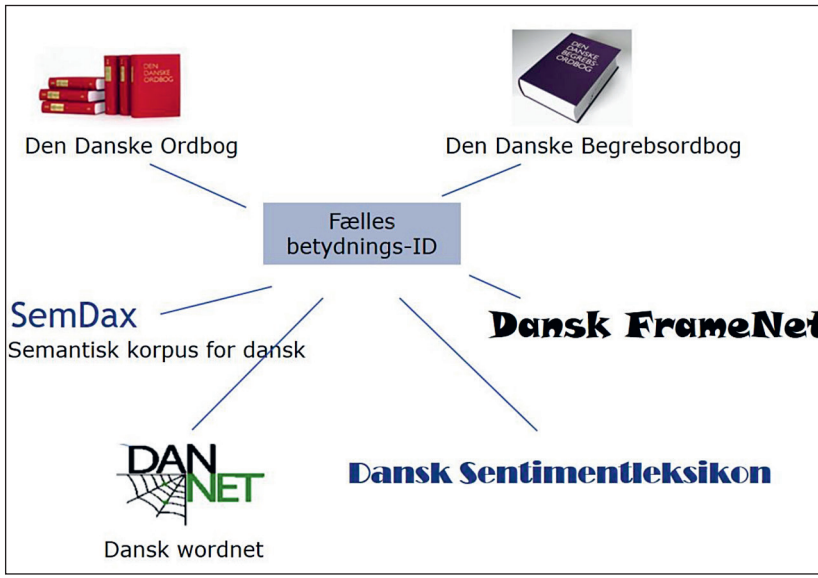
9 Det er særligt differentieringen mellem en fysisk og psykisk skade som volder problemer (røde og blå prikker).

notørerne for *stand* end for *skade* (med en enighed på hhv. 0,86 og 0,65[10]) (Pedersen et al. 2015). Man kan i øvrigt bemærke at eksemplet med *skade* modsiger det generelle billede vi så i figur 1 hvor ord der refererer til sygdomme og lidelser, generelt udviste høj enighed når annotørerne skulle kategorisere dem ud fra løbende tekst. Dette skyldes muligvis at de fleste ord der refererer til lidelser, ikke udviser den dobbelttydighed mht. fysisk eller psykisk affektion som *skade* gør. Man kunne også tolke det som en indikation af at betydningsskellet muligvis ikke er relevant at opretholde idet vi måske netop når vi bruger ordet *skade* ønsker at være uspecifikke mht. om det er fysisk eller psykisk.

4 Flere sprogressourcer med samme betydningsinventar

Den samling af semantiske sprogressourcer til anvendelse inden for sprogteknologi som er udviklet i et samarbejde mellem DSL og CST ved Københavns Universitet, har alle taget udgangspunkt i DDO's betydningsinventar og været nogenlunde tro mod dette, også selvom det ikke er opbygget ud fra formelle og gennemført systematiske kriterier, og selvom dette naturligvis giver visse udfordringer i den formelle sprogbeskrivelse. Figur 3 illustrerer relationen på betydningsniveau mellem sprogressourcerne.

10 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).



Figur 3: Flere sprogressourcer forbundet via de samme betydnings-ID-numre

På ét gennemgående punkt har vi afveget fra DDO i udviklingen af de sprogteknologiske ressourcer, nemlig ved ord der udviser systematisk polysemi; dog henvises også i disse tilfælde entydigt tilbage til betydningsstrukturen i DDO. Vi så med eksempler som *avis*, *bog* og *hæfte* hvordan samme form for flertydighed inden for et bestemt semantisk område går igen på tværs af ordforrådet. De frekvente ord der udviser systematisk polysemi i DDO, er ofte beskrevet med begge betydninger i ordbogen, hvorimod de mindre frekvente ord typisk er beskrevet med kun én betydning med vægt på kun det ene aspekt, idet det andet aspekt er underforstået, bl.a. begrundet i pladshensyn (første udgave af DDO var på tryk). Simpleksordet *bog* er fx udfoldet i to betydninger, én der beskriver genstanden (»trykte eller beskrevne blade af papir indbundet eller på anden måde sammenhæftet i rækkefølge ... «), og en der beskriver indholdet (»tekst der står på disse trykte eller beskrevne, indbundne eller sammenhæftede blade ... «). Substantivet *tegneserie* der udviser samme form for systematisk polysemi som *bog*, har derimod kun én betydning i DDO, nemlig genstandsbetydningen: »række af tegninger der fortæller en historie, typisk anbragt i firkantede rammer, forsynet med talebobler og trykt som selvstændigt hæfte eller som stribe i en avis eller et blad«, også selvom

citatet i stedet afspejler indholdsbetydningen: *Vingummi og tegneserier, hvorfor fylder de sig med den slags? Hvorfor bruger de ikke deres penge til gulerødder og gode bøger...?* Når en formel sprogressource skal opbygges med udgangspunkt i betydningsinventaret i en almindelig ordbog, er den udfoldede løsning som *bog* repræsenterer, klart at foretrække. Det letter arbejdet med at opbygge den sprogteknologiske ressource betydeligt når betydningerne er udskilt i selve ordbogsstrukturen og ikke blot er underforstået i definitionerne (jf. Lorentzen & Nimb 2010). I DanNet har vi bestræbt os på at udfolde systematisk polysemi uanset ordenes frekvens selv om det langt fra er gennemført gennem hele ordforrådet. I det følgende afsnit beskriver vi først arbejdet med at udvikle DanNet, dernæst Det Danske FrameNet-leksikon og endelig det danske sentimentleksikon.

4.1 DanNet

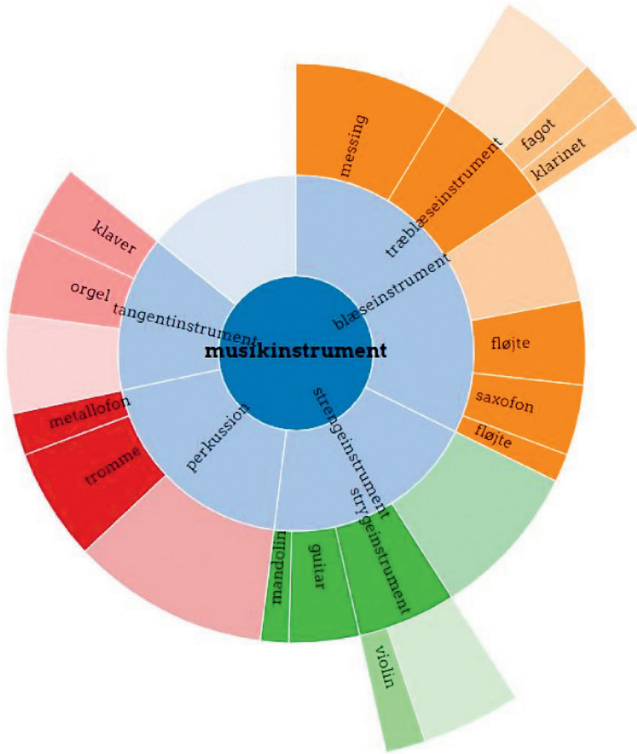
DanNet blev udviklet i løbet af nullerne (Pedersen et al. 2009) under en forskningsrådsbevilling[11] og med udgangspunkt i Princeton WordNet (Fellbaum & Miller (red.) 1998). I modsætning til de fleste andre wordnets er DanNet som nævnt ovenfor udviklet monolingvalt på baggrund af danske data og ikke først oversat og derefter monolingvalt tilpasset. Ressourcen udvikles fortsat under forskellige bevillinger, og i øjeblikket forskes der i at anvende Begrebsordbogen til at udvide antallet af adjektiver semiautomatisk, finansieret af Carlsbergfondet,[12] jf. oplysninger på cst.ku.dk/projekter/dannet hvorfra ressourcen også kan downloades under en opensource licens. Ressourcen omfatter for nuværende knap 70.000 begreber (eller såkaldte *synsets* ≈ synonymsæt) og mere end 300.000 relationer.

Udgangspunktet for arbejdet med at etablere DanNet var oplysninger om nærmeste overbegreb (genus proximum) som var angivet i et særligt felt i xml-manuskriptet til DDO. Disse oplysninger blev udnyttet til at lave et første over-underbegrebsnetværk over ordforråd, jf. fx netværket i figur 4 for *musikinstrumenter* og underliggende begreber (de navngivne felter angiver de underbegreber der selv har underbegreber).

---

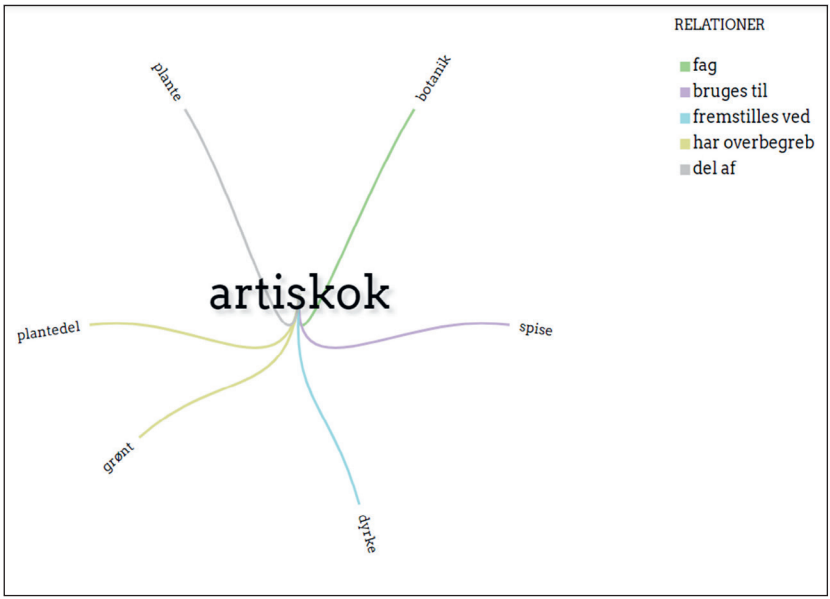
11 DFF Yngre forskere.

12 »Digital udgivelse og sprogteknologisk anvendelse af Den Danske Begrebsordbog« finansieret af Carlsbergfondets infrastrukturmidler 2019-2022



Figur 4: Musikinstrumenter vist med de største grupper af underbegreber. 'Fløjte' optræder to gange i figuren fordi der refereres til to forskellige slags fløjter, nemlig 'et rørformet blæseinstrument af træ, ben, metal, kunststof el.lign., med huller til regulering af tonehøjden' vs. 'et lille instrument af metal eller kunststof der typisk bruges til at give signal med eller som legetøj'.

Mange af de automatisk udtrukne taksonomier er efterfølgende blevet manuelt tilrettet til en mere konsistent struktur. Taksonomierne blev fx harmoniseret mht. artefakter (Nimb 2009) og i forbindelse med kodningen af en række systematisk polyseme madbegreber, jf. forklaringen om etablering af betydninger i afsnittet ovenfor (se også Pedersen et al. 2010). Overbegrebet til *artiskok* er fx *blomsterknop* i DDO, hvor begrebet i DanNet i stedet er kodet med to overbegreber fra hhv. en 'mad'-taksonomi og en biologisk (naturlig) taksonomi: *grøntsag/grønt* på den ene side og *plantedel* på den anden, se figur 5. En sådan beskrivelse sikrer at *artiskok* beskrivelsesmæssigt også indgår i samme semantiske kategori med andre spiselige plantedele som *spinat* og *gulerødder*.



Figur 5: Artiskok og dets semantiske relationer i DanNet; herunder to overbegreber, hhv. plantedel og grøntsag/grønt

DanNet indeholder endvidere semantiske træk i form af fx konnotation og køn på visse substantiver og adjektiver, fx er *vatnisse* kodet med negativ konnotation, mens *formidabel* er kodet med positiv konnotation (Braasch & Pedersen 2010).

Endelig er alle betydninger i DanNet indordnet i en formel ontologi (EuroWordNet Top-Ontology, Vossen 1999) udviklet specifikt til wordnets, og *artiskok* bærer således også det sammensatte ontologiske træk PLANT/PART/COMESTIBLE, der går igen for alle spiselige plantedele, på tværs af de enkelte ords overbegreber. De ontologiske typer udgør således i sig selv en meget grov taksonomi over ordforrådet og gør det muligt på tværs af overbegreber at sammenholde fx alle beklædningsgenstande (ARTIFACT/GARMENT) eller alle møbler (ARTIFACT/FURNITURE).

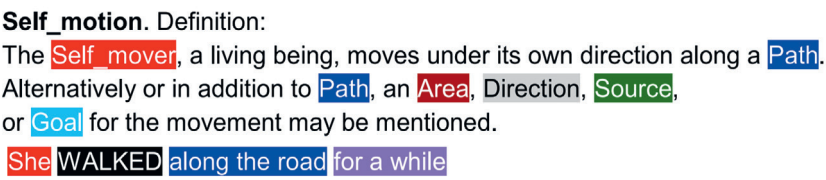
Gennemgående skal det bemærkes at den del af ordforrådet der henviser til konkrete entiteter, ikke overraskende er mere indgående og konsistent beskrevet i wordnettet end abstrakte begreber og egenskaber. Det har med andre ord været vanskeligere både i DDO og i DanNet at angive konsistente overbegreber og andre relationer for mere abstrakte begreber. Men selv ved konkrete genstande er genus proximum-oplysningen i DDO

langtfra altid brugbar i en formel struktur. Den består nemlig af et ord (en streng) taget direkte fra selve betydningsbeskrivelsen i DDO. Lampe (med definitionen »anordning der afgiver lys ved hjælp af elektricitet eller ved afbrænding af fx olie, petroleum eller gas«) har fx genus proximum *anordning* i DDO, men i DanNet har det fået overbegrebet *lyskilde*, ikke *anordning*. Udover at forsyne ordene med overbegreb, blev også andre dele af DDO's betydningsbeskrivelse omformuleret til præcise semantiske relationer i DanNet, fx relationen 'Used for', der for *lampe*s vedkommende peger på verbet *belyse*.

De formaliserede oplysninger gør det muligt for fx en søgemaskine at udtrække ord der er semantisk beslægtet med hinanden, fx alle ord med overbegrebet *lyskilde* eller alle genstande med den ontologiske type ARTIFACT/OBJECT.

4.2 Det Danske FrameNet-leksikon

FrameNet-modellen er udviklet ved Berkeley Universitet ud fra Charles Fillmores teori om Frame Semantics (Fillmore 1968; Baker et al. 1998; Ruppenhofer et al. 2016) og udgør i dag en international standard for formel beskrivelse af rollesemantik. Man har for en lang række sprog opbygget ordlister over de valensbærende ord på det pågældende sprog og annoteret tekstkorpora med udgangspunkt i de frames og det rolleinventar som allerede er opstillet og beskrevet for engelsk i Berkeley FrameNet-projektet, se fx Torrent et al. (2018). Dette sker ud fra en antagelse om at de fleste rammebeskrivelser for engelsk er universelle og gælder på tværs af sprog. I nogle sprog er det nødvendigt at tilføre sprogspecifikke rammer, fx for japansk (Ohara 2012). Berkeley FrameNet beskriver i dag 1075 forskellige leksikalske rammer for engelsk, idet en bestemt betydning af et leksem altid kun kan have én rammeværdi (Rupenhofer et al. 2016). Rammen 'Self_motion' (se figur 6) beskriver fx den helt overordnede betydning for en række engelske verber som *hike, jog, march* og *promenade*, men rammen dækker på samme måde også en lang række danske verber, fx *tøffe, tusse, vimse, gå, vade, vandre* og *valfarte*. De 'frame elements' som er knyttet til rammen i Berkeley FrameNets meget udførlige rammebeskrivelse ('Self_mover', det levende væsen der udfører bevægelsen, og Path / Area / Direction / Source / Goal, der beskriver hvor bevægelsen foregår), dækker også de roller der knytter sig til de nævnte danske verber.



Figur 6: Forklaring på rammen 'Self_motion' med tilhørende 'frame elements' i Berkeley FrameNet

I Berkeley FrameNet er tilskrivning af ramme- og elementværdier til autentisk skriftsprog et meget vigtigt element, og de verber og verbalsubstantiver der kan 'udløse' en bestemt ramme, findes så at sige nedefra og op ved at foretage manuel opmærkning af store mængder tekst med de rammer der bedst dækker sætningens (og verbets) betydning. Vi er i vores arbejde med modellen gået omvendt til værks og har i stedet brugt DDO og Begrebsordbogen som grundlag for at lave en ordliste med oplysninger om hvilke rammer danske verber (og deres tilknyttede verbalsubstantiver) er i stand til at udløse. Ordlisten, et FrameNet-leksikon for dansk,[13] beskriver rammer for 5.300 danske verber (hvilket dækker 80 % af verberne i DDO, dog ikke alle deres betydninger) og derudover 6.490 verbalsubstantiver (se Nimb 2018 og Nimb et al. 2017). Lemmaerne er tildelt 671 forskellige rammeværdier. Med udgangspunkt i leksikonnet er tanken at man i efterfølgende projekter langt nemmere kan opmærke danske sætninger med rammer og roller idet man slår verbet op i ordbogen og ser hvilke rammer der som udgangspunkt er mulige. Når en sådan manuelt opmærket tekst er tilvejebragt, kan den bruges som datagrundlag for *Semantic Role Labeling* som et skridt på vejen til automatisk tekstforståelse af dansksproget materiale.

I etableringen af det danske FrameNet-leksikon udnyttede vi at DSL's to ordbøger for moderne dansk, DDO og Begrebsordbogen er koblet sammen på betydningsniveau, og at de begge er korpusbaserede (Begrebsordbogen indirekte idet den indeholder DDO's lemmaer og betydninger). I grundmanuskriptet til Begrebsordbogen er ordgrupper i de enkelte afsnit opmærket med helt overordnede semantiske kategorier, fx 'handling', 'person' og 'egenskab'. Dette muliggjorde emnebaserede udtræk af verber

13 FrameNet-leksikonnet er udviklet for Carlsbergfondets Infrastrukturmidler i projektet »Fra begrebsordbog til FrameNet« 2016-2017.

inklusive deres verbalsubstantiver ud fra afsnittets navn (fx »Hurtig bevægelse«: *løbe* og *løb*). Tilkobling af verbernes valensmønstre fra DDO gjorde det muligt at udvælge netop den passende ramme ud af de mange der beskriver bevægelse i Berkeley FrameNet, og derpå tildele den til en række af verber og substantiver i afsnittet på en konsistent måde i samme ombæring.

I Tabel 1 ses fx en række verber fra Begrebsordbogens Kapitel 8 »Sted og Bevægelse« kombineret med valensoplysninger fra DDO og med den tildelte ramme 'Self_motion'. Relationen mellem DDO's og Berkeley FrameNets betydningsinventar er ikke én til en, dels fordi DDO's betydningsinventar modsat Berkeley FrameNet ikke er opbygget ud fra formelle og gennemført systematiske kriterier, dels fordi betydningsskel ikke udgør en én til en-relation på tværs af ordbøger, og slet ikke for forskellige sprog. Som det ses, kan flere betydninger af samme verbum i DDO derfor godt have én og samme ramme i FrameNet, fx *løbe,* bet. 1 og 1a i DDO (se ordnet.dk/ddo). Dette gælder også hvis den ene er en overført betydning af den anden – i Berkeley FrameNet har den konkrete og den overførte betydning nemlig typisk samme rammeværdi. Men omvendt kan samme betydning i DDO også tilskrives flere forskellige rammer fra Berkeley FrameNet. Fx har betydning 1 af *løbe* i DDO udover rammen 'Self_motion' også rammen 'Cotheme'; denne betydning fremgår af anden del af valensmønstret: 'ngn løber efter ngn/ngt'

vb.	pendulere	ngn/ngt pendulerer mellem ngt og ngt	Self_motion
vb.	pifte	ngn/ngt pifter advl	Self_motion
vb.	pile	ngn piler retning	Self_motion
vb.	piske	ngn/ngt pisker retning/sted	Self_motion
vb.	løbe	ngn løber (advl); ngn løber efter ngn/ngt	Self_motion
vb.	løbe	ngn løber retning	Self_motion
vb.	pløje	ngn/ngt pløjer (sig) retning	Self_motion

Tabel 1: Verber fra Kapitel 8 'Sted og Bevægelse' i Begrebsordbogen, koblet sammen med deres valensmønstre og tilskrevet en rammeværdi fra Berkeley FrameNet.

Det skal understreges at de rammer der beskrives for et givent verbum i Det Danske FrameNet-leksikon, ikke nødvendigvis er udtømmende. Dels mangler der nogle af de stærkt polyseme verbers betydninger, dels vil der altid være eksempler på atypisk eller nyskabende brug af ord i løbende

tekst der ikke er så etableret i sproget at det er beskrevet i ordbøger og derfor heller ikke i et FrameNet-leksikon der baserer sig på ordbøger og ikke på tekstopmærkning.

4.3 Dansk Sentimentleksikon

Sentimentanalyse (også kaldet 'opinion mining') er en automatisk analyse af tekstlige data hvor mindre eller større tekstmængder analyseres og klassificeres som negative, positive eller neutrale. Udgangspunktet er en antagelse om at ethvert ord har en 'prior sentiment' eller polaritet, dvs. at ordet i sin grundbetydning som en del af sin konnotative betydning er enten neutralt, positivt eller negativt. Sentimentanalyse bruges således til at kategorisere det følelsesmæssige indhold i en given tekst eller mængder af tekster ved at analysere de enkelte ord. Sentimentanalyse anvendes meget bredt til analyse af fx produktanmeldelser, vurdering af kundeservice, overvågning af sociale medier, undersøgelse af politiske holdninger (for dansk, se Enevoldsen og Hansen 2017), men kan eksempelvis også bruges til at vise stemnings- eller holdningsmæssig udvikling i et eller flere litterære værker, se Liu (2015).

Der er flere tilgange til sentimentanalyse. Den rent vidensbaserede tilgang klassificerer tekster ud fra polaritetskategorier baseret på tilstedeværelsen af entydige polaritetsbærende ord taget fra en ordliste, et såkaldt sentimentleksikon, men kan også tildele sandsynlig polaritet til vilkårlige ord. Rent statistiske metoder kan være *bag of words*, wordembeddings og andre værktøjer der gør brug af dyb eller neural læring helt uden støtte af sproglige ressourcer. Endelig anvendes også hybride metoder der kombinerer sentimentleksika med statistiske metoder (Jacobs 2019).

Sentimentanalyse har været udbredt i mange år især inden for kunstig intelligens, og man kan finde beskrivelser af sentimentleksika og deres udarbejdelse for mange sprog. Vi har særligt interesseret os for det svenske sentimentleksikon, som ligesom vores sentimentleksikon er baseret på en tesaurus, nemlig det svenske SenSALDO-leksikon (Rouces et al. 2018a og b). Ud fra den information man kan udlede fra den topologiske placering af ordene og deres indplacering i semantiske grupper i SALDO, blev de ord som man vurderede indeholdt polaritet, udvalgt. Man medtog i første version kun særligt hyppige substantiver, verber, adjektiver og interjektioner baseret på frekvensmålinger i et stort tekstkorpus for mo-

derne svensk. Resultatet var en ordliste på knap 2000 ord med polaritet. Den seneste version af SenSALDO, er udvidet med langt flere ordbetydninger, heraf også mange neutrale.[14]

For dansk har det mest udbredte sentimentleksikon hidtil været AFINN (Nielsen 2018). Ordlisten er oversat fra engelsk, bearbejdet til dansk og omfatter godt 3000 unikke lemmaer opmærket med polaritetsværdi (negativ/positiv) samt en angivelse af polaritetsgrad fra -5 til +5, men i modsætning til SenSALDO ingen neutrale ord. Et nyere sentimentleksikon for dansk, SENTIDA (Lauridsen et al. 2019) blev udviklet i 2019 baseret på DSL's liste over de 10.000 mest frekvente danske ord.[15] Listens substantiver, adjektiver, verber, adverbier og interjektioner blev opmærket med AFINNs polaritetsgrad fra -5 til +5 (inkl. 0 for neutral) af tre annotører, en fælles værdi blev fundet, og også i dette tilfælde blev rent neutrale ord fjernet. Ordlisten blev derefter sammenholdt og suppleret med ord fra AFINN i tilfælde af manglende leksikalsk dækning. Dernæst blev ordlisten udsat for automatisk 'stemming', hvor hvert lemma reduceres til »den del af lemmaet som ikke ændres, men som stadig bevarer ordets mening« (Lauridsen et al. 2019), fx reduceres 'spille', 'spillende', 'spiller' til 'spil'. Resultatet er en liste på ca. 5.200 indgange, der if. forfatterne svarer til ca. 35.000 ordformer.[16] Forfatterne har udført forskellige eksperimenter som viser at SENTIDA giver bedre resultater end AFINN.

Vores tilgang til et nyt sentimentleksikon er som den svenske at tage udgangspunkt i en tesaurus, nemlig Begrebsordbogen, hvis manuskript løbende udvides med nye ord. Etableringen af leksikonnet er en del af et projekt finansieret af Carlsbergfondet der har til formål at undersøge hvordan Begrebsordbogens data kan udnyttes i sprogteknologi, jf. 4.1, udvidelsen af DanNet med adjektiver.[17] Begrebsordbogens manuskriptet dækker på nuværende tidspunkt ca. 95 % af DDO's 150.000 lemmaer og betydninger. Dermed opnås et sentimentleksikon med høj leksikalsk dækningsgrad. Udgangspunktet er en manuel opmærkning af Begrebsordbogens 888 sektioner. En fjerdedel af de 888 sektioner er, baseret på

---

14 Den nye version (v.0.2) indeholder 12.287 ordbetydninger, heraf 4.273 negative og 2.013 positive, resten neutrale. SenSaldo er også udgivet som fuldformsliste med i alt knap 85.000 ordformer, se https://spraakbanken.gu.se/resurser/sensaldo.

15 korpus.dsl.dk/resources/details/freq-lemmas.html.

16 De tekster som analyseres med den 'stemmede' liste, bliver ligeledes underlagt 'stemming'.

17 »Digital udgivelse og sprogteknologisk anvendelse af Den Danske Begrebsordbog« finansieret af Carlsbergfondets infrastrukturmidler 2018-2022.

sektionsnavnet, blevet vurderet til at indeholde polaritetsord. 122 er blevet opmærket som negative (fx 'Uvigtig' og 'Tristhed'), 80 som positive (fx 'Vigtig', 'Beundre' og 'Venskab'), og 12 er blevet vurderet som mere uklare tilfælde, hvorfra ordforrådet kunne være relevante at medtage i et sentimentleksikon (fx 'Omdømme', og 'Protest, opstand'). Værdierne for hver sektion er blevet overført til hvert enkeltord i sektionen (faste udtryk er udeladt), og derpå er de enkelte ord (ordbetydninger) blevet vurderet og opmærket manuelt med enten positiv eller negativ værdi. Ord der på trods af afsnittets polaritetsværdi ikke formidler polaritet af nogen art, er blevet tildelt værdien 0 og herefter, som i AFINN og SENTIDA, men modsat SenSALDO, frasorteret.[18] Datasættet er desuden suppleret med dels betydninger af lemmaer fra AFINN der ikke var dækket (jf. SENTIDAs metode), dels med lemmaer med valør 'nedsættende' i DDO der ikke var indeholdt i de annoterede begrebsordbogsafsnit.

Resultatet er et betydningsleksikon med 17.883 betydninger med polaritetsværdi. Betydningerne stammer fra 14.444 lemmaer; forinden blev 2.250 lemmaer der havde 0-værdier for alle opmærkede betydninger, frasorteret. Også i etableringen af sentimentleksikonnet udnyttede vi at DSL's to ordbøger for moderne dansk, DDO og Begrebsordbogen er koblet sammen på betydningsniveau. Betydningerne blev gennemgået manuelt idet vi udnyttede links til DDO og tilkoblede oplysninger om frekvens, fag og kronologi. Formålet var at fjerne sjældne betydninger. Men udover dette ønskede vi også at udlede et supplerende leksikon der kun angiver én polaritetsværdi for hvert lemma. Et sådant er mere praktisk anvendeligt i tekstanalyse fordi det ikke kræver entydiggørelse i de tilfælde hvor et ord har flere betydninger. I tilvejebringelsen af dette supplerende lemmaleksikon blev de lemmaer der udover en betydning med polaritetsværdi, også havde en udbredt neutral betydning, helt udeladt (fx *gås* der på trods af den negative betydning 'dum kvinde' ikke blev medtaget), jf. SENTIDA's fremgangsmåde. Også den promille af lemmaerne i betydningsleksikonnet der havde decideret modstridende polaritetsværdier for de enkelte betydninger, blev gennemgået. Halvdelen af disse blev forkastet pga. tvetydigheden (fx *frelst*, *sej*, *skarp*, *overlegen* og *glat*); de

18 De leksikalske data i hvert afsnit blev kun annoteret af én person fordi vi indledningsvis dobbeltannoterede 400 tilfældigt udvalgte ord og konstaterede at der var meget høj overensstemmelse (92,5 %), og at præcise retningslinjer og mulighed for at anføre tvivlstilfælde til fælles diskussion ville føre til et pålideligt resultat – desuden var der en del gengangere på tværs af afsnittene så mange ord reelt endte med at være dobbeltannoteret.

øvrige er bevaret med den værdi som den absolut hyppigste betydning af lemmaet har.[19] Der er fx blevet set bort fra en negativ betydning af *vigtig*: »som har overdrevent høje tanker om sig selv« fordi den er markeret som sjælden i DDO, og på den måde kunne *vigtig* med positiv polaritet bevares i leksikonnet.

Resultatet blev et supplerende lemmaleksikon med 13.859 lemmaer med kun én polaritetsværdi. Af de 13.859 lemmaer har 8575 (62 %) negativ polaritet og 5284 (38 %) positiv polaritet, hvilket er nogenlunde samme fordeling som i SenSALDO, jf. fodnote 14, altså en klar overvægt af negative lemmaer. I litteraturen diskuteres dette fænomen, som ser ud til at gælde generelt for sentimentleksika, og der sammenlignes med fordelingen i løbende tekst hvor forholdet menes at være omvendt, se fx Devitt & Ahmad (2013). Det nyudviklede danske sentimentleksikon vil bl.a. kunne anvendes til at lave sådanne undersøgelser for dansk.

En udfordrende del af opgaven med at etablere et sentimentleksikon er at finde en objektiv måde at inkludere skalerbare værdier i polaritetsopmærkningen som en supplerende oplysning til de rene negative og positive værdier. Det er svært for en annotør at være konsistent med netop denne type opmærkning på tværs af ordforrådet, for hvad er mest negativt: *harmdirrende*, *hjerteskærende* eller *krumspring*? Og hvad er mest positivt: *gennemslagskraft* eller *idealisme*? I udarbejdelsen af den første version af det svenske sentimentleksikon SenSaldo anvendte man derfor en metode der kaldes Best-Worst Scaling. Ud af fire tilfældige ord med polaritet skulle ikke blot én, men fire forskellige annotører udvælge de to ord der var mest værdiladede, dvs. mest i hver sin ende af en positiv-negativ-skala, før man fastlagde den endelige værdi i leksikonnet.[20] Vi har i stedet taget udgangspunkt i de allerede påsatte grader i AFINN-leksikonnet, hvor der opereres med 5 positive og 5 negative værdier, men vores leksikon anvender kun 3 grader af positiv, henholdsvis negativ polaritet.[21] Vi har sammenlignet lemmaer i vores lister fra Begrebsordbogens afsnit med polaritetsgraden af samme ord i AFINN. Da AFINN har en væsentlig lavere leksikalsk dækningsgrad (3000 ord), havde kun få ord i hvert afsnit en forstærket værdi. Men i sammenligningen har vi bevaret

19 De øvrige betydninger var sjældne, gammeldags eller faglige ifølge DDO.

20 I den nye, udvidede version af SenSaldo der omfatter langt flere ord, er gradstildelingen udeladt

21 Graden 2 i AFINN svarer også til graden 2(px/nx) i vores, mens de stærkeste grader 3, 4 og 5 i AFINN er slået sammen til én særligt høj grad, værdien 3 (pxx/nxx), i vores.

den semantiske rækkefølge inden for afsnittet i Begrebsordbogen, og på den måde kan den forstærkede værdi i AFINN, udover at blive overført til samme lemma i vores liste, samtidig også ekspanderes manuelt til de nærmeste synonymer og nærsynonymer i afsnittet der hvor det er relevant. Der er også tilfælde hvor værdien i AFINN ikke er blevet overført, men ændret, fx blev *ophidset* der var positiv i AFINN, ændret til negativ i vores leksikon. Af de 13.859 lemmaer (i lemmaleksikonnet) er 4 % vurderet til at være særdeles positive (fx *glædesudbrud* og *gennemsolid*), 12 % til at være meget positive (fx *suveræn*, *strålende* og *supergod*) og 22 % til at være let positive (fx *appetitlig* og *ambition*). 7 % er vurderet til at være særdeles negative (fx *lortevejr*, *ækelhed* og *løgnhals*), 27 % til at være meget negative (fx *ørkenvandring* og *ukultiveret*) og 28 % til at være let negative (fx *apparatfejl* og *beslaglægge*). Se tabel 2 for flere eksempler.

lemma	**grundpolarity**	**gradspolarity**	**lemma-id-nummer, DDO**
dannet	p	p	11008370
danselyst	p	p	53006103
dåre	n	nx	11010433
dårlig	n	nxx	11010436
darling	p	pxx	11008417
dårskab	n	n	11010438
dase	p	p	11008421
dåselatter	n	n	11010440
dasen	p	p	51003840
daske	n	n	11008423
databedrageri	n	nx	30001009

Tabel 2: Udsnit af data i sentimentleksikonnet på lemmaniveau. Til venstre lemmaet, dernæst ses den grundlæggende polarityværdi p (positiv) eller n (negativ), til højre ses gradsværdien (skalaen går fra minus 3 (nxx) til plus 3 (pxx)). Ordklasse og fuldformer vil blive tilføjet vha. lemma-id-nummer i DDO.

Vi planlægger at overføre polaritetsinformation fra betydningsleksikonnet til DanNet, der som tidligere nævnt allerede indeholder denne information for en lille del af ordforrådet, og baserer os her igen på de fælles id-numre på betydningsniveau på tværs af vores ressourcer. På den måde

kan konnotativ viden om de enkelte ord der er indeholdt i et synset i DanNet, udnyttes i automatiske analyser med DanNet-leksikonnet. Hvis informationen skal anføres på det samlede synset i DanNet, skal der udvikles en semiautomatisk metode; et synset kan nemlig godt indeholde ord med forskellige grader af polaritetsværdier, endda med modstridende værdier, da synset-begrebet bygger på et synonymibegreb der udelukkende er baseret på den denotative betydning af ord og som udgangspunkt ikke tager højde for den konnotative betydning, fx polaritetsværdien.

5 Semantisk opmærkning og analyse med brug af de tre ressourcer

I dette afsnit illustrerer vi hvordan opmærkning med de tre ressourcer kan afdække forskellige semantiske lag i dansk tekst. Vi tager udgangspunkt i et litterært eksempel, nemlig Peter Seebergs novelle ''Forsøg med vrede'' fra bogen *Om fjorten dage*, der udkom i 1982, dvs. inden for den periode i dansk sprog som beskrives i DDO og dermed også i Begrebsordbogen, DanNet og sentimentleksikonnet, hvis ordforråd direkte bygger på DDO. Vi eksemplificerer annoteringen på nedenstående tekstuddrag bestående af tre sætninger: (1) *Ingen skulle nogensinde få hende fra, at trafikkontrollør Jensen var et nederdrægtigt svin, en gemen svinepels, et lortesvin, en lort, en skid, et røvhul.* (2) *Nu gik han igen henad perronen i morgenmørket, med sin krumme, kvalmende pibe, sin lange uniformsfrakke, og bukserne hevet op med selerne til langt op på skinnebenene og hænderne på ryggen som en præst, skønt han nærmest var det modsatte, og med det uopsættelige ærinde i hovedet: at forrette sin nødtørft på kontorpersonalets lokum før alle andre, når gårsdagens lugte havde forenet sig med to ugers samlede stank.* (3) *Mens hun pumpede vand op i spandene ved gavlen, forsvandt han ind i den privilegerede helligdom efter at have ført nøglen ind i hullet og drejet den rundt og havde lukket slåen efter sig.*

Som vi skal se, bidrager de tre leksikalske ressourcer hver især med meget forskellige typer af formaliseret semantisk viden om teksten og supplerer dermed hinanden i forsøget på at identificere viden om betydningsindholdet i teksten. Sentimentleksikonnet indkredser allerede i sætning 1 og 2 at tekststykket er negativt ladet, se figur 7. Figuren illustrerer hvilke polaritetsværdier der kan påsættes ord i titlen og de to første sætninger i novellen alene ud fra direkte oplysninger om opslagsord i sentimentlek-

sikonnet beskrevet ovenfor. Hele 9 ord er negative i stærkeste grad, yderligere 2 er negative og kun 2 er positive (ingen af dem i stærkeste grad).

Forsøg med Vrede
Ingen skulle nogensinde få hende fra, at trafikkontrollør Jensen var et nederdrægtigt svin, en gemen svinepels, et lortesvin, en lort, en skid, et røvhul.

Nu gik han igen henad perronen i morgenmørket, med sin krumme, kvalmende pibe, sin lange uniformsfrakke, og bukserne hevet op med selerne til langt op på skinnebenene og hænderne på ryggen som en præst, skønt han nærmest var det modsatte, og med det uopsættelige ærinde i hovedet: at forrette sin nødtørft på kontorpersonalets lokum før alle andre, når gårsdagens lugte havde forenet sig med to ugers samlede stank.

-3: stærkest negativt	+1: let positivt
-2: stærkere negativt	+2: stærkere positivt
-1: let negativt	+3: stærkest positivt

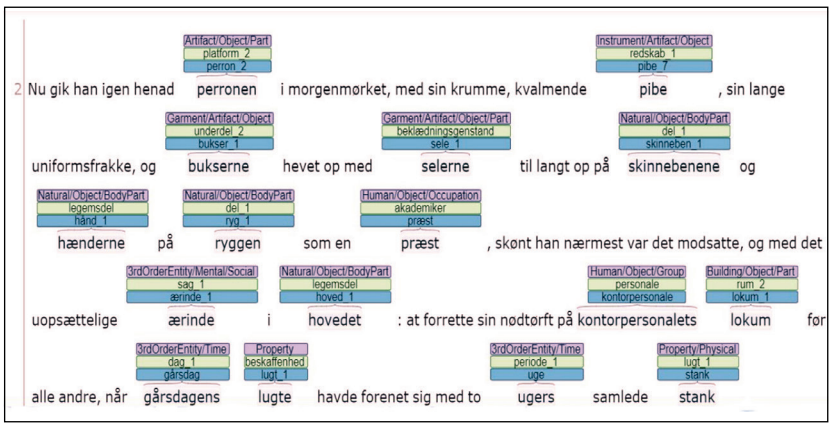
Figur 7: *Polaritetsværdier for ord i indledningen af Seebergs »Forsøg med Vrede«: nederdrægtig* -3*, svin* -3*, gemen* -3*, svinepels* -3*, lort* -3*, skid* -2*, røvhul* -3*, kvalmende* -3*, hive* -1*, modsat* -1*, uopsættelig* +1*, forene* +2*, stank* -3.

Eksemplet viser også at sentimentleksikonnets dækningsgrad er meget høj, idet kun de negativt ladede lemmaer *lortesvin* og *lokum* ikke er med i det; *lortesvin* dog af den enkle grund at ordet ikke er med i DDO og må betragtes som en ad hoc-sammensætning.[22] I den praktiske anvendelse af leksikonnet er det oplagt at tilføje et modul til automatisk opsplitning af ukendte sammensætninger der ikke er at finde blandt opslagsordene, fx i figur 7 *lortesvin*. Ud fra en opsplitning af sammensætningen kan den negative polaritet udledes automatisk af første- og sidsteleddenes polaritet i leksikonnet.

DanNet- og FrameNet-opmærkningerne indkredser i højere grad hvad de tre sætninger handler om. DanNet-værdierne på substantiverne i sætning 2 viser at der er en del ord om tøj, kroppen og lugte. Det er i høj grad substantivernes ontologiske type i DanNet der leder til denne semantiske sammenhæng mellem ord i teksten: Tre har den ontologiske type BODYPART fælles (*skinneben, hånd, ryg*), og to har typen GARMENT fælles: *bukser, seler*. To andre har den ontologiske type TIME fælles: *gårsdag* og *uge*, jf. figur 8. Men DanNet bidrager også med relevante oplysninger om andre relationer mellem ordene i teksten, *stank* har

22 *Morgenmørke*, der for nylig er medtaget i DDO, kandiderer også til at komme med i sentimentleksikonnet evt. som let negativt lemma.

således *lugt* som overbegreb. Used_for-relationen angiver de involverede artefakters funktion, som i *ryge* for *pibe* og *gå på toilettet* for *lokum*. Vi kan konstatere at DanNets dækningsgrad er høj hvad angår betydningen af substantiverne i teksten. Kun to lemmaer er ikke i DanNet, nemlig *morgenmørke* og *uniformsfrakke*; det skyldes også i dette tilfælde at lemmaerne ikke var medtaget i DDO da DanNet blev udviklet.



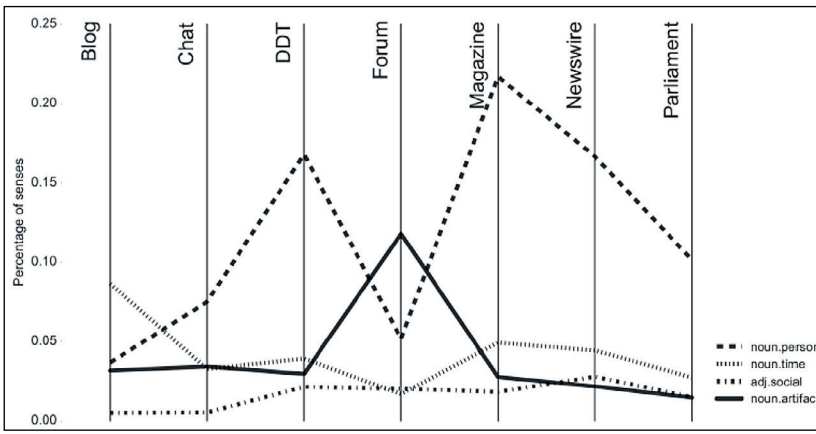
Figur 8: Sætning 2 annoteret med DanNet-oplysninger for substantiverne. Nederste annotationslag angiver betydnings-id, mellemste angiver nærmeste danske overbegreb og øverste ontologisk type.

FrameNet er i højere grad velegnet til at indkredse den semantiske kontekst, hvem der agerer hvor og hvornår, og hvad der i det hele taget foregår, jf. figur 9.

Figur 9: Sætning 2 og 3 opmærket med frames og semantiske roller (frame elements).

Ud fra opmærkningerne er det tydeligt at tekstuddraget har meget fokus på bevægelse. Det omhandler dels en aktørs egen bevægelse, illustreret ved rammerne Self_motion (*gå henad* og *komme*), Departing (*forsvinde (ind i)*) og Halt (*holde*), dels flytning eller bevægelse af ting udført af en agent, udledt af rammerne Cause_motion (*hive*), Cause_fluidic_motion (*pumpe op*), Cause_to_move_in_place (*dreje rundt*), Placing (*føre ind i*) og Closure (*lukke*). Dækningsgraden er også i dette tilfælde høj. Udover bevægelsesrammer optræder rammen Amalgamation (*forene sig*) og Excreting (*forrette sin nødtørft*).

Det er vores formodning at opmærkninger af typen vist her udover at bidrage til næranalyse også kan bibringe interessante semantisk betingede statistiske mønstre på tværs af større tekstmængder. Man kan følge et forfatterskab og se om det ændrer sig over tid, man kan lave bred genreanalyse og se på hvordan forskellige genrer udtrykker sig forskelligt, eller man kan se på hvordan nye emner dukker op og bliver behandlet over tid i en given litterær epoke. Fra vores eget korpus SemDaX viser figur 10 fx en helt enkel, men dog interessant statistik over hvordan ontologiske typer har temmelig forskellig forekomst afhængig af teksttyper og domæne; i ugeblade forekommer mennesker (Person) fx langt hyppigere som type end i de øvrige teksttyper.



Figur 10: Forskellig forekomst af ontologiske typer i de forskellige teksttyper i SemDaX (Martínez Alonso et al. 2015)

Automatisk semantisk analyse af store tekstmængder til brug for fx forskning i litterære trends er kun i sin absolutte vorden, særligt inden for det danske felt. Det skyldes bl.a. at de relevante sprogressourcer og digitale

datasæt først nu er ved at blive udviklet og tilgængeliggjort. Flere store forskningsprojekter er imidlertid igangsat for nylig med det formål at besvare nye forskningsspørgsmål med basis i store tekstmængder, og hvor tanken er at anvende semantisk processering på digitale tekstkorpora i større eller mindre grad.[23] Det er vores håb at vores tre ressourcer kan bidrage til denne forskning.

6 Konkluderende bemærkninger – den tværsproglige dimension, opskalering og kobling til andre oplysningstyper

Alle sprog har deres unikke (omend dynamiske) betydningsinventar, som typisk repræsenteres gennem monolingvale ordbøger. Men skæringsfeltet mellem hvordan morfologisk, syntaktisk og semantisk information udtrykkes, varierer, ligesom forholdet mellem udtryk og indhold varierer. Graden af homonymi, polysemi og synonymi i et givent betydningsinventar synes altså ikke at bero på en fuldstændig sprogneutral naturlov men på idiosynkratiske forhold i de enkelte sprog. Dog genkender man fx systematisk polysemi i de fleste sprog (jf. fx Cruse 1989, Copestake & Briscoe 1996) ligesom mange andre leksikalsk-semantiske egenskaber går igen særligt inden for grupper af beslægtede sprog. De germanske sprog er fx karakteriseret ved at have en stor mængde partikelverber og verber med valensbundne præpositioner hvoraf nogle er betydningsmæssigt transparente (fx den konkrete betydning af *gå ud*), mens andre er mere eller mindre uigennemskuelige og leksikaliserede (*slå op*, *se op til*). Hvor bevægelse med retning i de romanske sprog typisk er leksikaliseret (fx på spansk: *salir* ('gå ud'), *entrar* ('gå ind')), anvender de germanske sprog i høj grad disse partikelkonstruktioner til at udtrykke retning (fx på tysk *ausgehen* ('gå ud'), *gå ind*) jf. Talmy (1985).

---

23 *FabulaNet*-projektet finansieret af Velux Fonden ved Aarhus Universitet https://interactingminds.au.dk/news/enkelt/artikel/fabula-net-a-deep-neural-network-for-automated-multidimensional-assessment-of-literacy-fiction-and/ er et af de nyere større projekter med digital litteraturanalyse, sammen med *Measuring Modernity*-projektet ved KU, finansieret af Carlsberg Fondet https://www.carlsbergfondet.dk/da/Forskningsaktiviteter/Bevillingsstatistik/Bevillingsoversigt/CF19_0661_Jens-Bjerring_Hansen.

I det flersproglige EU-projekt ELEXIS (elex.is, Krek et al. 2018, Pedersen et al. 2018) forsøger vi vha. fælles standarder og leksikografiske værktøjer der kan anvendes tværsprogligt, at sammenkæde betydningsinventarer på tværs af sprog samt semantisk at opmærke parallelle, flersproglige korpora med disse leksikalske ressourcer. Formålet er dels at kvalificere de ressourcer som opbygges, dels at tilvejebringe flere flersproglige korpusressourcer der kan anvendes til træning af sprogmodeller. Projektet har således både leksikografisk og sprogteknologisk relevans idet det overordnede formål er at videreudvikle begge felter og sikre at velstrukturerede leksikalske ressourcer af høj kvalitet i højere grad indgår i de sprogteknologiske løsninger der udvikles.

I denne artikel har vi behandlet de centrale *semantiske* oplysningstyper i leksikalske ressourcer til datalingvistiske formål. Det er klart at morfologiske og syntaktiske oplysningstyper også er relevante for sprogprocessering til sprogteknologi og digital humanistisk forskning, og at de ideelt set skal være umiddelbart tilgængelige, gerne direkte via betydningsinventaret. Det har ikke været tilfældet for de tre ressourcer vi har beskrevet her (wordnet, framenetleksikon og sentimentordbog for dansk). Som bruger af ressourcerne har man altså måttet indhente sine morfologiske og syntaktiske oplysninger andre steder fra.

I et nyligt igangsat projekt er det målet i et samarbejde mellem Digitaliseringsstyrelsen, Dansk Sprognævn, DSL og CST på Københavns Universitet at integrere de ovenfor beskrevne ressourcer i én samlet databasestruktur hvori indgår morfologi – bl.a. via Retskrivningsordbogen – og på længere sigt også syntaks, bl.a. via STO-ordbasen (Braasch & Olsen 2004).[24] Målet er et Centralt OrdRegister, COR, som skal kunne downloades frit til forsknings- og udviklingsformål, og som løbende og dynamisk skal kunne udvides til fx også at rumme terminologiske ressourcer i form af fagspecifikke termbaser. Sammen med en generel opgradering og validering er det vores håb at de semantiske ressourcer hermed kan integreres og komme endnu mere bredt i spil både blandt forskere og i danske virksomheder.

---

24 COR-projektet: https://sprogteknologi.dk/blog/udarbejdelsen-af-et-centralt-ordregister-skydes-i-gang

Tak

Vi vil gerne takke revieweren for meget relevante kommentarer, uddybende bemærkninger og supplerende litteraturhenvisninger, som vi efter bedste evne har indarbejdet i den endelige artikel. Også tak til Jeppe Barnwell, DSL/KU, for at udvælge relevant tekst til annoteringen og til Thomas Troelsgård, DSL, for dataudtræk til brug for det leksikografiske arbejde med at etablere både FrameNet- og Sentiment-leksikonnet. Endelig en tak til de øvrige kolleger som igennem årene har bidraget til de beskrevne ressourcer og projekter. De vil være at finde i referencehenvisningerne, som vi har bestræbt os på at angive i rigt mål undervejs i artiklen.

Ordbøger og sprogteknologiske ressourcer:

Asmussen, Jørg og Jakob Halskov (2011)*, DK-CLARIN Reference Corpus of General Danish,* CLARIN-DK-UCPH Centre Repository*,* http://hdl.handle.net/20.500.12115/36.

DanNet, det danske wordnet: Kan downloades fra http://wordnet.dk. Kan browses fra http://wordties.cst.dk/wordties-dannet/. Forskningsreference: Pedersen et al. 2009.

Den Danske Begrebsordbog: Nimb, Sanni, Henrik Lorentzen, Thomas Troelsgård, Liisa Theilgaard, Lars Trap-Jensen (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab, København, Danmark.

Det Danske FrameNet-leksikon. https://korpus.dsl.dk/resources/details/framenet.html. Forsknings-reference: Nimb 2018, Pedersen et al. 2018a, samt Nimb et al. 2017

Den Danske Ordbog (DDO). Tilgået på: http://ordnet.dk/ddo, Det Danske Sprog- og Litteraturselskab, København, Danmark.

KorpusDK: https://ordnet.dk/korpusdk; https://korpus.dsl.dk/resources/details/korpusdk.html

Retskrivningsordbogen, 4. udgave, 2012 Dansk Sprognævn og Forlaget Lindhardt og Ringhof.

SemDax, det semantisk opmærkede korpus kan downloades fra github: https://github.com/coastalcph/semdax. Forskningsreference: Pedersen et al. 2016.

STO – SprogTeknologisk Ordbase. CLARIN-DK-UCPH Centre Repository: http://hdl.handle.net/20.500.12115/21 (samt /22, /23, /26). Forskningsreference: Braasch et al. 2004.

Litteratur:

Ahmadi, Sina, Sanni Nimb, John P. McCrae, Nicolai H. Sørensen (2020). Towards automatic linking of lexicographic data: the case of a historical and a modern Danish dictionary. *Congress of the European Association for Lexicography: EURALEX XIX: Lexicography for Inclusion*.

Baker, Collin F., Charles J. Fillmore, John B. Lowe (1998). The Berkeley FrameNet project. I: *Proceedings of the COLING-ACL*. Montreal, Canada.

Bick, Eckhard (2011). A FrameNet for Danish. I: *Proceedings of NODALIDA 2011*, May 11-13, Riga, Latvia. NEALT Proceedings Series, Vol. 11, pp. 34-41. Tartu: Tartu University Library

Bick, Eckhard (2017). Propbank Annotation of Danish Noun Frames, I: *ACL Anthology*. W17, 69, p. 6.

Bjerring-Hansen, Jens, Frank Fischer, Torben Jelsbak & Nicolai Hartvig Sørensen (2019). Nodes and Edges in Literary History. Modelling 19th Century Literary Landscapes. I: *DH2019 Proceedings*, Utrecht.

Braasch, Anna & Bolette Sandford Pedersen (2010). Encoding Attitude and Connotation in Wordnets. I: Proceedings of the XIV Euralex International Congress.

Braasch Anna, Sussi A. Olsen (2004). STO: A Danish Lexicon Resource - Ready for Applications. I: *Proceedings of the 4th International Conference on Language Resources and Evaluation,* pp. 1079-1083. Lisboa, Portugal 2004

Calzolari, Nicoletta, Antonio Zampolli, Alessandro Lenci (2002). Towards a Standard for a Multilingual Lexical Entry: The EAGLES/ISLE Initiative. I: *CICLing 2002: Computational Linguistics and Intelligent Text Processing* pp. 264-279.

Copestake, Ann & Ted Briscoe (1996). Semi-productive polysemy and sense extension. I: *Lexical Semantics: The Problem of Polysemy*, J. Pustejovsky & B. Boguraev (red.). New York: Clarendon Press. pp. 15-67.

Cruse, David A. (1989). *Lexical Semantics*. Cambridge University Press.

Dannéls, Dana, Lars Borin & Karin Friberg Heppin (eds) (2021). *Swedish FrameNet++. Harmonization, integration, method development and practical language technology applications*. Amsterdam: John Benjamins.

Devitt, Ann & Khurshid Ahmad (2013). Is there a language of sentiment? An analysis of lexical resources for sentiment analysis. I: *Language Resources & Evaluation 47*: 475-511. doi 10.1007/s10579-013-9223-6.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. I: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics

Enevoldsen, Kenneth C. & Lasse Hansen (2017). Analysing Political Biases in Danish Newspapers Using Sentiment Analysis. *Journal of Language Works – Sprogvidenskabeligt Studentertidsskrift, 2(2)*, pp.87-98.

Fellbaum, Christiane, and George Miller (1989) *WordNet. An electronic lexical database*. Cambridge, MA: MIT Press; 1998.

Fillmore, Charles. (1968). The case for case. I: Bach & Harms (red.): *Universals in Linguistic Theory*. New York: Holt, Rinehart, and Winston. pp. 1-88.

Fillmore, Charles, Sue Atkins (1992). Toward a frame-based lexicon: The semantics of RISK and its neighbours. I: *Frames, Fields, and Contrast: New Essays in Semantics and Lexical Organization*, A. Lehrer and E. Kittay, red., pp. 75-102, Lawrence Erlbaum Associates.

Firth, John R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis.* Oxford, Blackwell.

Guarino, Nicola & Mark Musen (2015). Applied ontology: The next decade begins. I: *Applied ontology 10*, IOS Press.

Hershcovich, Daniel & Lucia Donatelli (2021). It's the Meaning that Counts: The State of the Art in NLP and Semantics. *Künstliche Intelligenz Special Issue on NLP and Semantics*.

Hjelmslev, Louis (1966). *Omkring sprogteoriens grundlæggelse*. København: Akademisk Forlag.

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,* pp. 389-397. The Association for Computational Linguistics.

Jacobs, Arthur. (2019). Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics. I: *Frontiers in Robotics and AI. 6*. 53. 10.3389/frobt.2019.00053.

Jakobsen, Lisbeth Falster (2008). Paradigmatisk og associationsdannelse som grundlæggende funktioner i sproget. *Ny Forskning i Grammatik 15*, pp. 49-68

Kirchmeier, Sabine, Peter Juel Henrichsen, Philip Diderichsen og Nanna Bøgebjerg Hansen (2019). *Dansk Sprogteknologi i verdensklasse – rapport fra sprogteknologiudvalget*, https://sprogtek2018.dk/?p=409

Krek, Simon, Iztok Kosem, John McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, & Tanja Wissik (2018). *European Lexicographic Infrastructure (ELEXIS).* 881-892. Paper presented at the XVIII EURALEX International Congress, Ljubljana, Slovenia.

Krippendorff, Klaus (2011). Agreement and Information in the Reliasbility of Coding. I: *Communication Methods and Measures 5 (2)* pp: 93-112.

Lakoff, George & Mark Johnson (1980). *Metaphors We Live By.* University of Chicago Press.

Lauridsen, G., Dalsgaard, J., & Svendsen, L. (2019). SENTIDA: A New Tool for Sentiment Analysis in Danish. *Journal of Language Works – Sprogvidenskabeligt Studentertidsskrift*, *4*(1), 38-53. https://tidsskrift.dk/lwo/article

Lenci, Alessandro. (2008). Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1-31.

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 gene-

ral framework for the development of multilingual Lexicons. I: *International Journal of Lexicography*, 13(4), pp. 249-263.

Levin, Beth (1993). *English verb classes and their alternations. A preliminary investigation*. University of Chicago Press.

Liu, Bing (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139084789

Lorentzen, Henrik & Sanni Nimb (2010). Fra ordbog til wordnet - Hvordan udmøntes en traditionel ordbogsdefinition i en formaliseret wordnetbeskrivelse? Nordiska Studier i Lexikografi 10, 2010, pp. 329-344. *Rapport från Konferens om lexikografi i Norden*, Tammerfors 3.-5. juni 2009

Madsen, Bodil Nistrup (2005). *Håndbog i begrebsarbejde*. Sundhedsdatastyrelsen.

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,* ACL Anthology, Linköping University Electronic Press, Sweden.

McCarthy, Diana, Marianna Apidianaki, Katrin Erk (2016). Word Sense Clustering and Clusterability. I: *Computational Linguistics*, Vol. 42, no. 2.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean (2013). Distributed representations of words and phrases and their compositionality. I: *Advances in neural information processing systems*, pp. 3111-3119.

Nielsen, F. Å. (2018). *Danish resources*. https://bit.ly/2NDHcbW

Nimb, Sanni (2018). The Danish FrameNet Lexicon: Method and lexical coverage. I: *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructions*, pp. 48-52.

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, pp. 25-59.

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

Nimb, Sanni (2009). The Semantic Relations of Artifacts in DanNet. I: B.S. Pedersen, A. Braasch, S. Nimb, R. Vatvedt Fjeld (red.): *Proceedings of the NODALIDA 2009 workshop. WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. NEALT Proceedings Series, Vol. 7.

Nimb, Sanni & Bolette S. Pedersen (2016). Fra begrebsordbog til sprogteknologisk ressource: verber, semantiske roller og rammer – et pilotstudie. I: *Skrifter / Nordisk forening for leksikografi, Vol. 14,* 2016, pp. 405-415.

Nimb, Sanni & Bolette S. Pedersen (2000). Treating Metaphoric Senses in a Danish Computational Lexicon – Different cases of regular polysemy. I: *Proceedings of the 9th EURALEX International Congress*.

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

Nimb, Sanni, Nicolai Hartvig Sørensen (2018). Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. *Proceedings from Euralex 2018*, Ljubliana, Slovenia, 2018

Ohara, Kyoko (2012). Semantic annotations in Japanese Framenet: comparing frames in Japanese and English. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*) Istanbul, Turkey.

Olsen, Ida R., Asad Sayeed, & Bolette S. Pedersen (2020). Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources. I: *Globalex Workshop on Linked Lexicography: LREC 2020 Workshop Language Resources and Evaluation Conference* (pp. 45-52). Marseille, France: European Language Resources Association

Olsen, Sussi, Bolette S. Pedersen, Héctor Martínez Alonso & Anders T. Johannsen (2015). Coarse-grained sense annotation of Danish across textual domains. I: *Proceedings of the Workshop on Semantic resources and Semantic Annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015* (pp. 36-43). Sweden: Linköping University Electronic Press. Linköping Electronic Conference Proceedings.

Pedersen, Bolette S., John McCrae, Carole Tiberius, & Simon Krek (2018). ELEXIS - a European infrastructure fostering cooperation and information exchange among lexicographical research communities. I: *Proceedings of Global WordNet Conference 2018* Singapore.

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

Pedersen, Bolette S., Manex Agirrezabal, Sanni Nimb, Sussi Olsen, Ida Rørmann Olsen (2018b). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. I: *Proceedings of Global WordNet Conference 2018*, Singapore.

Pedersen, Bolette S., Anna Braasch, Anders Johanssen, Héctor Martínez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard, Nicolai Hartvig Sørensen (2016). The SemDaX Corpus – sense annotations with scalable sense inventories. I: *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, Portorož, Slovenia.

Pedersen, Bolette, Sanni Nimb, Anna Braasch, Sussi Olsen (2016a). Betydningsinventarer – i ordbøger og i løbende tekst. In: *Skrifter/Nordisk Forening for Leksikografi, Vol. 14,* pp. 417-429.

Pedersen, Bolette S., Sanni Nimb, Anna 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 the Seventh International Conference on Language Resources and Evaluation* 2010 p. 3181-3186. Malta.

Pedersen, Bolette S., Krister Lindén, Kadri Vider, Markus Forsberg, Neeme Kahusk, Jyrki Niemi, Lars Nygaard, Mitchell Seaton, Heili Orav, Lars Borin, Kaarlo Voionmaa, Niklas Nisbeth and Eirikur Rögnvaldsson (2013). Nordic and Baltic wordnets aligned and compared through »WordTies«. I: *Proceedings from the 19th Nordic Conference on Computational Linguistics (NODALIDA).* Linköping Electronic Conference Proceedings; Volume 85 (ISSN 1650-3740)

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. *Language Resources and Evaluation, Computational Linguistics Series*, pp.269-299.

Pedersen Bolette S., Sussi Olsen, Sanni Nimb, Anna Braasch (2015). Betydningsinventar - i ordbøger og i løbende tekst. I: *13. Konference om Leksikografi i Norden*, København, Denmark.

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

Pedersen, Bolette Sandford, Anna Braasch, Anders Trærup Johannsen, Héctor Martinez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard, Nicolai Sørensen (2016). The SemDaX Corpus - sense annotations with scalable sense inventories. In *Proceedings of the 10th conference of the Language Resources and Evaluation Conference* (pp. 842-847). Portorož, Slovenia.

Pustejovsky, James (1995). *The Generative Lexicon*, Cambridge, MA: MIT Press.

Rouces Jacobo, Lars Borin, Nina Tahmasebi, Stian R. Eide (2018a). Defining a gold standard for a Swedish sentiment lexicon: Towards higher-yield text mining in the digital humanities. I: *CEUR Workshop Proceedings vol. 2084. Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference Helsinki*, Finland, March 7-9, 2018. Edited by Eetu Mäkelä Mikko Tolonen Jouni Tuominen.

Rouces Jacobo, Lars Borin, Nina Tahmasebi, Stian R. Eide (2018b). SenSaldo: Creating a sentiment lexicon for Swedish. I: *Proceedings of LREC 2018, Eleventh International Conference on Language Resources and* Evaluation, si. 4192-4198, Miyazaki, ELRA.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher. R. Johnson, Collin F. Baker, Jan Scheffczyk (2016). I: *FrameNet II: Extended Theory and Practice* https://framenet.icsi.berkeley.edu/fndrupal/the_book.

Scarlini, Bianca, Tommaso Pasini and Roberto Navigli (2020). SensEmBERT: Context-Enhanced Sense Embeddings for Multilingual Word Sense Disambiguation. I: *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020).*

Svensén, Bo (2004). *Handbok i lexikografi. Ordböker och ordboksarbete i teori och praktik*. Norstedts Akademiska Förlag.

Sørensen, Nicolai H. & Sanni Nimb (2018). Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. I: *Proceedings from Euralex 2018*, Ljubliana, Slovenien.

Talmy, Leonard (1985). Lexicalization patterns: Semantic structure in lexical forms. In T. Shopen (Ed.), *Language typology and syntactic description* (pp. 36-149). Cambridge: Cambridge University Press.

Tangherlini Timothy & Peter Leonard (2013). Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research. *Poetics* 41(6): 725-749.

Torrent, Tiago, Lars Borin, Collin F. Baker (2018). Multilingual Framenets and Constructions. *Proceedings of the LREC 2018 Workshop International FrameNet Workshop 2018*. Miyazaki, Japan

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