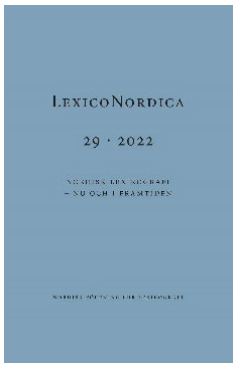
LexicoNordica

Titel: COR-S – den semantiske del af Det Centrale OrdRegister (COR)

Forfatter: Sanni Nimb, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen & Thomas Troelsgård

Kilde: LexicoNordica 29, 2022, s. 73-95

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



COR-S – den semantiske del af Det Centrale OrdRegister (COR)

Sanni Nimb, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen & Thomas Troelsgård

We present the formal lexicon *COR-S*, which constitutes the semantic part of a Danish computational lexicon project called *COR*. *COR-S* is based on linked data, but apart from transferring, adjusting, and validating the information from existing Danish lexicons and dictionaries, the goal is also to compile an AI suitable sense granularity level. Based on the fine-grained sense inventory of *Den Danske Ordbog* (DDO), senses are clustered, partly by hand according to a set of principles, and partly by means of automatic NLP methods.

1. Baggrund og introduktion til COR

COR står for Det Centrale OrdRegister for dansk og er et nystartet sprogteknologisk ordbogsprojekt der har til formål at etablere en leksikalsk ressource som er egnet til anvendelse i kunstig intelligens og andre teknologiske applikationer der arbejder med dansk sprog. Projektet blev igangsat i 2021 i et samarbejde imellem på den ene side to af de ledende udviklere af danske ordbøger, Dansk Sprognævn og Det Danske Sprog- og Litteraturselskab (DSL), og på den anden side en sprogteknologisk partner, Center for Sprogteknologi (CST) ved Københavns Universitet. Det løber frem til udgangen af 2023 og udgør en del af den sprogteknologisatsning der er blevet igangsat i forbindelse med den danske regerings AI-strategi fra 2019, og hvor det påpeges at der er et behov for at styrke indsatsen omkring frit tilgængelige danske sprogressourcer.

En af grundtankerne i COR er at skabe analogi mellem Det Centrale PersonRegister (CPR) og Det Centrale OrdRegister. Alle

oplysninger om danske ord kobles således til samme stabile og fremtidssikrede register, både de semantiske oplysninger der allerede nu udvikles i COR-S og fremtidige leksikalske data. Kongstanken er at man ved at stille et indekseret system med unikke og stabile id-numre på alle ordformer til rådighed kan sikre en mere effektiv deling af danske leksikalske ressourcer. Frit tilgængelige oplysninger om stavning, bøjning og formelle betydningsbeskrivelser koblet til en central del af ordene vil gøre det langt nemmere for danske offentlige institutioner og virksomheder at arbejde med dansk sprogforståelse (jf. AI-strategiens målsætning). COR giver fx mulighed for at forbedre sprogcentrerede AI-systemer der oprindeligt er udviklet til engelsk, idet de leksikalske beskrivelser tager afsæt i anerkendt, lokalt forankret viden om dansk sprog og kultur som matcher det samfund som systemerne skal fungere i.

I COR-projektet indarbejdes og forenkles ordbogsmateriale og sprogressourcer der tidligere er udviklet ved de tre samarbejdende institutioner, som alle oplever en stigende efterspørgsel på ordbogsdata med basale oplysninger om udtale, bøjning og betydning beskrevet på en standardiseret og kompatibel måde. Offentlige institutioner og virksomheder der i dag arbejder med danske, digitale sprogdata, finder det udfordrende og tidskrævende at afklare hvad der findes, hvor det findes, og hvor tilgængeligt det i praksis er, blandt andet pga. mangel på sproglig ekspertise i virksomheden og pga. ressourcernes ofte fragmenterede natur. Ofte rejses spørgsmålet om hvordan de spiller sammen med virksomhedens egen fagterminologi, og i praksis fører det til at man i mange tilfælde udelukkende arbejder tekstbaseret og undlader at inddrage vigtig viden fra leksikalske ressourcer i sine sprogteknologiske komponenter. Men da de tekstmængder der er til rådighed, ofte er for sparsomme til at opnå gode resultater, giver det rigtig god mening også at inddrage leksikalsk viden, særligt hvis den er tilgængelig i en overskuelig og formaliseret form.

Artiklens fokus er den semantiske del af COR-projektet, COR-S, der lanceres sidst i 2023. I næste afsnit beskriver vi eksisterende bagvedliggende sprogressourcer der udnyttes i arbejdet med at udvikle COR-S, og vi giver eksempler på indholdet af en ordbogsindgang baseret på overførsel af data fra disse ressourcer. I afsnit 3 beskriver vi den leksikografiske fremgangsmåde i projektet, og i afsnit 4 de metoder der udvikles med henblik på at automatisere en del af arbejdet med at opnå et betydningsinventar for polyseme lemmaer der er håndterbart i sprogteknologi.

2. Bagvedliggende ordbogsressourcer

I COR-S samles en lang række formaliserede semantiske oplysninger om danske lemmaer der i dag kun er offentligt tilgængelige i adskilte ressourcer. Oplysningerne er gennem mange år opbygget i et tæt samarbejde mellem DSL og CST om at udvikle semantiske ressourcer til anvendelse i forskning i automatisk sprogforståelse. Ressourcerne er baseret på internationale standarder, og samarbejdet blev indledt med udviklingen af det danske WordNet *DanNet* i 2004 (Pedersen 2009 et al.), kort efter at den trykte udgave af *Den Danske Ordbog* (Hjorth & Kristensen 2003-2005, herefter DDO) var færdiggjort. Hver ny ressource har undervejs i forløbet muliggjort den næste og bidraget med afgørende ny information (Pedersen et al. 2018a, Pedersen, Nimb & Olsen 2021). Også *Den Danske Begrebsordbog* (herefter Begrebsordbogen), udgivet af DSL i 2015 (Nimb et al. 2015), er udviklet efter dette princip. Begrebsordbogen bygger videre på informationer i både DDO og DanNet, og efterfølgende blev dens indhold og emnestruktur kombineret med oplysninger i DDO og udnyttet til at udvikle to andre semantiske ressourcer. Den første var et FrameNet-leksikon der angiver en eller flere semantiske rammer for størstedelen af DDO's verber (se Nimb et al. 2017, Nimb 2018), og nogle år efter fulgte et senti-

mentleksikon der angiver om et lemma har positiv eller negativ konnotation (Nimb et al. 2022, Pedersen, Nimb & Olsen 2021).

Fælles for ressourcerne er at de alle er koblet til DDO's unikke og stabile betydningsnumre og dermed også til selve DDO-lemmaet. Det er det der nu udnyttes til fulde i COR-projektet, idet DDO's lemmaer kobles til *Retskrivningsordbogens* lemmaer der udgør basis for COR-registret. Den trykte DDO er siden 2005 videreudviklet som onlineordbog, og de data der er tilføjet siden første udgave, indgår også i ressourcerne. Med udgangspunkt i DDO-betydningsnumrene kan vi i dag kombinere alle typer af data på kryds og tværs: DDO-artiklens mange oplysninger vedrørende betydningen, DanNets oplysninger om ontologisk type og semantiske relationer, FrameNet-leksikonets semantiske rammer, oplysninger vedrørende positiv eller negativ konnotation fra sentimentleksikonet og endelig oplysninger om beslægtede ord, nøgleord og emne fra Begrebsordbogen.

Fælles for ressourcerne er også at de, modsat DDO, alle er opbygget ud fra en ontologisk tilgang som traditionelt anvendes i arbejdet med at opbygge formelle semantiske leksika (se Geeraerts 2002, Pustejovsky 1995). DanNet strukturerer således betydninger af DDO-lemmaer i en række semantiske relationer, primært i over- og underbegrebsrelationer. Ud fra betydningernes placering i træstrukturen kan man fx udlede automatisk om de er tæt på hinanden eller ej. WordNets er udviklet for en lang række sprog i verden ud fra samme format og standard som oprindelig blev etableret ved Princeton University (Fellbaum 1998), se hjemmesiden for Global WordNet Association. WordNets udgør grundlæggende ordbogsressourcer i international forskning inden for udviklingen af metoder til automatisk sprogforståelse. Det samme gør sig gældende for FrameNets, der findes for engelsk, svensk og en lang række andre sprog. De baserer sig på en standard der er udarbejdet (og stadig videreudvikles) ved University of California, Berkeley, USA (Ruppenhofer et al. 2016), se også FrameNets hjem-

meside. I projektet fastlægges og navngives en lang række semantiske rammer og deres tilhørende semantiske roller for engelsk. Korpustekster opmærkes med disse oplysninger så man dermed efterhånden får etableret et leksikon med de lemmaer der "udløser" en ramme. For dansk blev leksikonet i stedet udarbejdet ud fra en opmærkning af semantiske grupper af verber og verbalsubstantiver i Begrebsordbogen idet valensmønstre fra DDO samtidig blev koblet til de enkelte forekomster.

Den ontologiske tilgang i udviklingen af både DanNet og det danske FrameNet-leksikon har som konsekvens at ikke alle betydninger af et givet DDO-lemma nødvendigvis er repræsenteret i dem. Fokusset har ikke ligget på polysemi og betydningsinventar, men i stedet på taksonomier, begreber og relationer mellem betydninger på tværs af lemmaer. Da COR-projektets grundidé er at koble sproglige ressourcer sammen via indekserede lemmalister, må vi nødvendigvis vende tilbage til det semasiologiske udgangspunkt fra DDO og forholde os til alle DDO-betydninger. En formel ressource med fuld betydningsdækning for lemmaerne vil samtidig give nye muligheder for forskning i automatisk entydiggørelse af ordbetydninger, en af de helt store udfordringer i automatisk analyse af sprog. Vi ved fra tidligere forskningsprojekter at DDO's betydningsinventar er meget finkornet og ikke umiddelbart velegnet til formålet (Pedersen et al. 2016, Pedersen 2018, Pedersen et al. 2018b). Automatiske metoder er afhængige af at et lemmas forskellige betydninger udviser distributionelle forskelle i korpora. Betydningerne må i brug omgive sig med temmelig forskellige naboord for at kunne skelnes automatisk fra hinanden. En af de væsentlige opgaver i projektet er at opbygge et betydningsinventar med dette for øje, uden dog at gå på kompromis med de betydningsskel der tydeligt opfattes af mennesker. COR-S bliver med andre ord en forenklet udgave af DDO, hvor kun de væsentligste betydninger af lemmaer er repræsenteret, vel at mærke udtrykt ved hjælp af værdier inden for et begrænset formelt inventar (on-

tologisk type, semantisk ramme) samt et overbegreb i form af en præcis ordbetydning fra COR-S, ikke blot en tekststreng. Samtidig bliver der givet forklarende oplysninger til den menneskelige læser og it-bruger af COR-S. I tabel 1 ses et eksempel på en COR-S-ordbogsindgang med formaliserede betydningsoplysninger der kan trækkes direkte ud fra eksisterende data i DanNet, FrameNet-leksikonet og Begrebsordbogen.

Verb *bemærke*	**Betydning 1**	**Betydning 2**
Ontologisk type	Act+Mental	Act+Communication
Overbegreb	opfatte_COR_1	ytre_COR_1
Semantisk ramme	Becoming_aware	Mention
Definition	blive opmærksom på; lægge mærke til	gøre opmærksom på; nævne
Stikord	få øje på, observere	nævne, omtale
Eksempel (fra DDO)	*Flere naboer bemærkede en kraftig banken på et stuevindue …*	*Ja, fru Nielsen er flink, bemærkede Linda adspredt*

Tabel 1: Et eksempel på et verbum i COR-S hvor de formelle oplysninger om betydning stammer fra eksisterende sprogteknologiske ordbøger der er koblet til DDO.

Den ontologi der anvendes i COR-S, er i udgangspunktet magen til DanNets (og EuroWordNets), se Pedersen et al. (2009) og Vossen (1999). Den er dog forenklet, ikke kun hvad angår selve betegnelserne (’3rdOrderEntity’ hedder fx i stedet ’Abstract’), men også hvad angår omfanget af typer. Mange typer i DanNet er sammensat af adskillige betydningselementer (fx både ’Purpose’ og ’Social’ i typen ’Dynamic+Agentive+Purpose+Social’). I COR-S er de forenklet, baseret på hvor ofte de sammensatte typer reelt er anvendt i DanNet, og leksikografen må i stedet vurdere hvilket betydningselement der er vigtigst. Både *teselskab* og *hverv* har fx ovennævnte type i DanNet, men i COR-S har *teselskab* fået tildelt

typen 'Act+Social', mens *hverv* har fået tildelt 'Act+Purpose'. Et skel i DanNet-ontologien mellem uafsluttet og afsluttet handling/hændelse er helt fjernet i COR-S; skellet var oprindeligt møntet på romanske sprog i EuroWord-ontologien, men det er sjældent leksikaliseret i dansk. I alt er antallet af ontologiske typer reduceret med 36 % fra 204 i DanNet til 130 i COR-S. Se eksempler i tabel 2.

DanNet	COR-S	Eks.
UnboundedEvent	Event	*ske*, *hænde*, *foregå*
BoundedEvent		
UnboundedEvent+Agentive	Act	*gøre*, *handle*, *handling*
BoundedEvent+Agentive		
Dynamic+Agentive		
3rdOrderEntity+Mental+Purpose	Abstract +Purpose	*formål*, *mål*
3rdOrderEntity+Mental+ Purpose+Manner		
BoundedEvent+Agentive+ Purpose+Possession	Act+ Possession	*overdrage*, *give*
BoundedEvent+Agentive+ Purpose+Possession+Social		

Tabel 2: Eksempler på ontologiske typer i DanNet og COR-S.

Overbegrebet i tabel 1 overføres fra DanNet og tilrettes semiautomatisk til det modsvarende COR-S-betydningsnummer når det ligger fast ved projektets afslutning. Semantisk ramme overføres direkte fra FrameNet-leksikonet der beskriver rammer for en eller flere betydninger af 5.300 verber og 6.490 verbalsubstantiver i DDO. I alt er der anvendt 671 forskellige værdier fra Berkeley FrameNet, og de kan slås op i en ordbog på projektets hjemmeside, hvor deres betydning og tilhørende semantiske roller beskrives. Definitionerne overføres fra DanNet (hvor de består af "klippede" definitioner fra DDO). De suppleres med stikord i form af ordet umiddelbart til venstre for DDO-betydningen i Begrebsordbogen

samt det nærmeste nøgleord til venstre. Stikordene udtrækkes med samme algoritme som anvendes i funktionen *Ord i nærheden* i DDO (Nimb, Sørensen & Troelsgård 2018), se figur 1.

Figur 1: Stikordene *få øje på* og *observere* for betydning 1 af verbet *bemærke*, taget fra *Ord i nærheden*, ordnet.dk/ddo.

I tilfælde af uanvendelige stikord fjernes de i redigeringsprocessen. Endelig overføres eksemplet i tabel 1 fra citatmaterialet i DDO.

2.1. Hvilket ordforråd?

Det er et krav at de "væsentligste" betydninger i dansk skal være repræsenteret i første version af COR-S. Faste udtryk medtages ikke, og vi fokuserer på de åbne ordklasser, primært substantiver, verber og adjektiver. Vores lemmaselektion bygger på viden om det danske ordforråd som vi har opnået i arbejdet med andre ordbøger og ressourcer. For det første medtager vi de danske lemmaer der via DanNet i mindst én betydning er udpeget som ækvivalent til et af de 5.000 centrale begreber i Princeton WordNet (Pedersen et al. 2019). Der er tale om 4.600 DDO-lemmaer som vi betegner CBC-lemmaer (forkortelse for 'Core/Base Concepts'; for udvælgelse af disse se Global WordNet Associations hjemmeside). Nogle eksempler er *abe*, *acceptere*, *adgang*, *ekspert*, *elegant*, *knække*, *spise*. Tre fjerdedele af CBC-lemmaerne er polyseme i DDO, og selvom de kun udgør ca. 3,5 % af ordbogens lemmaer, dækker de ca. 11 % af betydningerne i den, endda uden at medregne betydninger fra de mange faste udtryk som en del af lemmaerne indgår i. For det andet medtager vi alle DDO-lemmaer der i mindst én betydning

optræder som nøgleord i Begrebsordbogen, dvs. er fremhævet som indledende overskrift for en semantisk gruppe af nærsynonymer og/eller synonymer i et af de 888 navngivne tematiske afsnit. På den måde sikrer vi en bred dækning af emner i COR-S-ordforrådet, og vi sikrer at fremtidige brugere af COR-registret med stor sandsynlighed kan relatere deres egne ord til et COR-lemma, fx inden for et fagligt område. Omkring 11.500 lemmaer optræder som nøgleord i Begrebsordbogen, dog er ca. 3.100 af dem allerede udvalgt som CBC-lemma, men i alt opnår vi på denne måde ca. 13.000 "væsentlige" og ofte polyseme lemmaer. Første version af COR-S indeholder også mange af de øvrige polyseme lemmaer i DDO hvoraf mindst én betydning er med i DanNet. Vi har manuelt opmærket og sammenlagt 2.600 af disse på linje med CBC-ordene, mens vi påregner at behandle 5.000 automatisk, se afsnit 4. Desuden omfatter første version af COR-S alle monoseme DDO-ord der er i både DanNet og *Retskrivningsordbogen*, i alt 18.000 lemmaer. Disse kan forholdsvis nemt overføres fra DanNet og kan fungere som sikre forankringer for automatiske semantiske analyser. I alt vil første version af COR-S omfatte mindst 35.000 lemmaer.

3. Leksikografisk fremgangsmåde

De færreste polyseme lemmaer i DDO opfører sig som verbet *bemærke* ovenfor, hvor vi anser begge betydninger for at være væsentlige og berettiget til at blive repræsenteret i COR-S-ressourcen. Mange har i stedet en eller flere ikke-centrale betydninger eller betydninger der kan anses som en indsnævret (eller udvidet) variant af en hovedbetydning. I nogle tilfælde er en sådan variant beskrevet som en ny hovedbetydning for at undgå alt for dybe betydningshierarkier i DDO. En stor del af det manuelle leksikografiske arbejde i COR består derfor i at analysere de enkelte lemmaers

betydninger med henblik på en enklere repræsentation. Arbejdet foretages i flere trin og optimeres ved inddragelse af automatiske metoder som beskrives i afsnit 4.

Første trin er selve udvælgelsen af de leksikografiske informationer der er relevante ved analysen af betydningsinventaret for hvert lemma, se tabel 3. Vi har her inddraget oplysninger fra DDO (fx definition, brugs- og fagmarkeringer og valensmønster), Begrebsordbogen (fx naboord og ordets evt. status og frekvens som nøgleord), DanNet (overbegreb, ontologisk type) og FrameNet-leksikonet (semantisk ramme). Derudover er der beregnet et pointtal for hver betydnings "tyngde" ud fra hvor mange citater, kollokationer og andre oplysningstyper i det hele taget der er knyttet til betydningen i DDO. Der er seks leksikografer involveret i arbejdet, og der arbejdes i online regneark; nogle er lavet til de enkelte leksikografer, andre er fælles, men de har alle ens opsætning.

Verbet *blomstre*	**semant. ramme**	**nøgle-ord**	**point**	**hovedbet. = 1, underbet. = 2**
bet.1. 'have blomster der er sprunget ud'		0	56	1
bet. 1.a 'trives og udfolde sig, være el. komme i god udvikling'	Thriving	2	72	2
bet.1.b 'være sund og smuk'		0	17	2

Tabel 3: Eksempel på oplysninger for verbet *blomstre*. De to underbetydninger 1.a og 1.b. lægges sammen til én i COR-S.

Arbejdet med at udvælge og sammenlægge DDO-betydninger, herunder fravælge betydninger man mener er for perifere eller sjældne, foregår ud fra en række principper og udføres af både studentermedhjælpere og erfarne leksikografer. Fra tidligere projekter med manuel opmærkning af leksikalske data har vi gode

erfaringer med at udarbejde detaljerede regler for at sikre at alle annotører, både studentermedhjælpere og erfarne leksikografer, arbejder ud fra samme retningslinjer. I arbejdet med at forenkle betydningsstrukturen udledte vi på baggrund af opmærkning af 25 % af CBC-ordene en række principper der baserer sig på de tilgængelige leksikografiske oplysninger. Betydninger der er markeret som sjældne eller historiske i DDO, bør fx fravælges, det samme gælder betydninger med meget lavt pointtal i det datasæt der er opstillet. Indsnævrede og udvidede underbetydninger lægges sammen med deres hovedbetydning, hvorimod overførte betydninger bevares. Konkrete betydninger bevares så vidt muligt, også selv om de er indskrænkede eller udvidede underbetydninger. Vi er også meget opmærksomme på at sikre en række faste principper vedr. systematisk polysemi (se fx Pustejovsky 1995) så den samme type polysemi behandles ens for alle ord. Fx bibeholdes begge betydninger ved mønstret dyr vs. madvare (fx *kylling*, *kalv*), mens de lægges sammen ved bygning vs. institution (fx *skole*). Ved mønstret proces vs. resultat bevarer vi begge betydninger når resultatet er konkret (som i *byggeri*), men når resultatet er abstrakt, lægger vi derimod de to betydninger sammen (som i *udtalelse*). I alt har vi registreret og opstillet regler for 35 forskellige mønstre, og information om mønstret vil komme med i ressourcen.

Der er udarbejdet skemaer med beskrivelser af de enkelte ontologiske typer og eksempler, herunder også en grafisk illustration til at overskueliggøre ontologien. Ikke kun ontologiske typer og semantiske rammer, men også overbegreber betragtes i øvrigt som lukkede inventarer. En liste over alle overbegreber der er anvendt i DanNet sorteret efter frekvens sikrer at tildelingen af overbegreber til nye betydninger strømlines og holdes inden for lemmaer der i forvejen er udvalgt til COR-S.

Hvad angår oplysninger om ontologisk type og overbegreb, kompliceres genbrug af data fra DanNet når betydninger lægges sammen, og det leksikografiske arbejde består derfor også i at ju-

stere de overførte værdier så de passer på den sammenlagte betydning. Samme justering skal i øvrigt foretages for synonymer fra DDO og for de positive/negative konnotationsværdier fra sentimentleksikonet. Synonymer tilføjes evt. i stedet som et særskilt modul til COR-registret.

Derudover skal de polyseme lemmaer forsynes med oplysninger ved relevante betydninger der ikke i forvejen er med i DanNet. Endelig skal eksisterende DanNet-oplysninger om ontologisk type og overbegreb ved de monoseme lemmaer valideres og tilpasses den nye, forenklede COR-S-ontologi.

Datasættet med polyseme CBC-ord spiller en central rolle i COR-S. De manuelle sammenlægninger der annoteres i datasættet, danner nemlig også udgangspunkt for de automatiske metoder til sammenlægning af betydninger ved resten af det polyseme ordforråd (se afsnit 5). Vi validerer derfor hinandens arbejde, dels ved at notere svære tilfælde der efterfølgende tjekkes af endnu en leksikograf, dels ved systematisk at validere 2 % af alle lemmaer med 6 eller færre DDO-betydninger i datasættet. Annotørenigheden er på 88 %, hvilket regnes for højt når der er tale om semantiske opmærkninger. Desuden er alle lemmaer med mere end 6 betydninger i DDO gennemgået af mindst to leksikografer, se fx verbet *støtte* i tabel 4, hvor man konfererede om sammenlægningerne. Resultatet af arbejdet er en meget klar reduktion af den finkornede betydningsinddeling i DDO. Antallet af betydninger er reduceret med 43 %, fra et gennemsnit på 4,3 betydninger pr. CBC-lemma i DDO til 2,4 betydninger pr. lemma i COR-S.

DDO-bet.	***støtte*, vb.**	COR-bet.
1	'yde moralsk, økonomisk eller anden hjælp og bistand'	1
1.a	'give sin tilslutning til; bakke op; gå ind for'	1

1.b	'underbygge yderligere; gøre mere troværdig eller overbevisende'	2
2	'bære eller holde noget oppe så det ikke falder ned eller vælter'	3
2.a	'hjælpe nogen med at holde sig oprejst, rejse sig, bevæge sig af sted el.lign. ved at lade vedkommende holde fast i eller hvile med sin vægt mod en'	1 (= 'yde hjælp') eller 3 (= 'holde/støtte fysisk')?
3	'hvile med sin vægt mod eller på noget; læne sig op ad'	4
3.a	'lade en legemsdel hvile mod eller på et fast underlag el.lign. så den holdes oppe'	4

Tabel 4: DDO-underbetydninger lægges ofte sammen med DDO-hovedbetydninger. Tvivlstilfælde diskuteres med de øvrige leksikografer, her 2.a af verbet *støtte*.

4. Automatiserede metoder

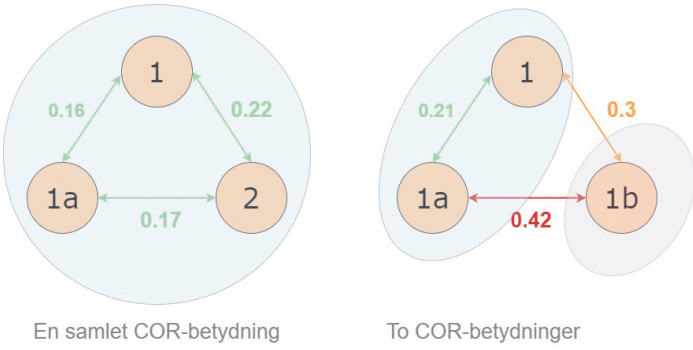
Selvom de leksikografiske principper skaber et godt grundlag for sammenlægning af betydninger, er den manuelle sammenlægning stadig en omstændelig og tidskrævende proces. Vi eksperimenterer derfor med at automatisere en del af processen, særligt for de knap så polyseme lemmaer (jf. Pedersen et al. 2022 for en fuld teknisk beskrivelse af disse eksperimenter). Til denne automatisering kan vi genbruge vores håndannotationer som både trænings- og valideringskorpus, med andre ord fungerer CBC-datasættet samt de øvrige håndopmærkede polyseme lemmaer som vores "guldstandard".

Vi undersøger tre tilgange til automatisk sammenlægning: en regelbaseret tilgang der er baseret på de leksikografiske principper og to statistiske tilgange med tekstdata fra DDO. Fælles for dem er at vi bruger den samme metode. Først beregner vi en afstandsscore

mellem alle kombinationer af et lemmas betydninger. Det er her tilgangene afviger mest fra hinanden, da vi bruger forskellige modeller til at bestemme semantisk nærhed. I de statistiske tilgange bruger vi trænede word embeddings der udregner en vektorrepræsentation ud fra distributionen af et lemmas forekomster i et træningskorpus[1]. Disse vektorrepræsentationer sammenlignes med henblik på at beregne en afstand imellem dem. I den regelbaserede tilgang får resultater der afspejler de udarbejdede principper, i stedet den bedste score. I andet trin benytter vi en algoritme til at bestemme hvilke betydninger der lægges sammen på baggrund af afstandsscoren. Når mere end to betydninger lægges sammen, sikrer vi dermed at de kun lægges sammen hvis afstanden mellem alle betydningerne er tilstrækkelig lille. Et eksempel på automatisk sammenlægning ses på figur 2 som viser to lemmaer der begge har tre betydninger som udgangspunkt.

Hvis alle afstandsscorerne er lave (til venstre i figur 2), kan alle betydningerne lægges sammen. Det er tilfældet for et lemma som *tøj* hvor definitioner for alle tre betydninger indeholder lemmaet *stof* og andre lemmaer relateret dertil (fx *beklædningsgenstande*, *klæde*). Lemmaet *sport* opfører sig derimod som eksemplet til højre på figuren hvor en betydning har mindst én høj afstandsscore til de andre. Betydning 1.b er nemlig i dette tilfælde overført, hvilket også kommer til udtryk i citatet (*for mange drenge er det en sport at skrive bilnumre op*). Citaterne for betydning 1 og 1.a omhandler dyrkning af sportsgrene, hvorimod citatet fra 1.b beskriver ordet *sport* om en sjov aktivitet eller leg. Betydning 1 og 1.a kan derfor lægges sammen, mens 1.b får sin egen betydning.

1 Word embeddings tager udgangspunkt i den distributionelle hypotese (Firth 1957, Harris 1954). Ifølge denne kan man udlede et ords semantik fra den omkringliggende kontekst. Dette kan afbildes statistisk som en vektor i et vektorrum via en word embedding-model, fx word2vec (Mikolov, Yih & Zweig 2013). Ord der ligger tæt på hinanden i vektorrummet, deler distributionel information og kan derfor antages at have betydninger der ligner hinanden.



Figur 2: Sammenlægning baseret på afstandsberegning.

4.1. Eksperiment 1: Regelbaseret tilgang

Den regelbaserede tilgang udregner semantisk nærhed efter de leksikografiske principper, jf. afsnit 3. To betydninger anses for at være semantisk tæt på hinanden hvis tre kriterier er opfyldt: (1) de hører under samme hovedbetydning, (2) ingen af betydningerne er overførte ifølge DDO, og (3) de har samme ontologiske type ifølge DanNet. Det sidste kriterium er betinget af at begge betydninger findes i DanNet.

Fordelen ved denne tilgang er at den er teoretisk og praktisk ligetil, dog med den ulempe at vi ikke har eksplicitte principper for hvornår betydninger på tværs af hovedbetydninger kan sammenlægges.

4.2. Eksperiment 2: word embeddings – statistiske ordprofiler

Traditionelle word embedding-modeller kan i udgangspunktet ikke adskille betydninger. Flertydige ord vil derfor få én samlet vektor der indeholder distributionel information om alle betyd-

ningerne. For at kunne udnytte de leksikalske informationer i COR-S har vi brug for at kunne splitte word2vec-vektorerne i betydninger baseret på DDO (jf. fx Olsen, Pedersen & Sayeed 2020). Konkret bruger vi en word2vec-model trænet af DSL på basis af DSL's korpus (Sørensen & Nimb 2018). For hver betydning vi har udtrukket fra DDO, beregner vi en kombineret word2vec-vektor ud fra definitioner og citater i DDO. I eksemplet ved figur 2 er det netop overlappet mellem definitioner (fx *tøj*) og citater (fx *sport*) der dannede baggrund for at beregne afstanden mellem betydninger. Afstanden måles ved hjælp af cosinus og læren om trekanters vinkler.

4.3. Eksperiment 3: kontekstualiserede embeddings

Kontekstualiserede embeddings er en nyere og mere kompleks metode som omgår problemet med flertydige repræsentationer ved at skabe en vektor for hver token (streng) i en sætning. Da kontekstualiserede embeddings repræsenterer et givent lemma i en specifik kontekst, kan vi antage at vektorerne her i højere grad afspejler betydninger.

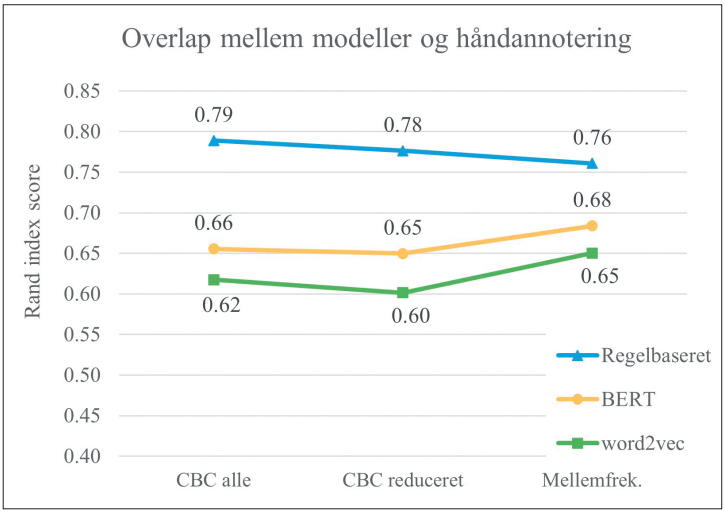
Vi bruger en såkaldt BERT-model (Devlin et al. 2019)[2], der er fortrænet af firmaet Certainly på et dansk tekstkorpus på cirka 1,6 milliarder ord fra Common Crawl, Danish OpenSubtitles, Danish Wikipedia og anden tekst fra internettet. Man kan yderligere finindstille modellen til et bestemt formål ved hjælp af en ekstra træningsopgave, og i vores tilfælde tilpasser vi modellen på vores håndannoterede datasæt. Inputtet er ligesom ved word2vec-eksperimentet definitioner og citater fra DDO. En ekstra fordel ved denne tilgang er at vi direkte kan anvende outputtet fra modellen

2 BERT står for Bidirectional Encoder Representations from Transformers og er en tilgang der er mere fleksibel end fx den tidligere såkaldte word2vec-tilgang, idet den processerer teksten forfra og bagfra på samme tid og på den måde indfanger konteksten i beregningerne på en mere nuanceret måde.

som en afstandsscore mellem betydninger og på den måde afgøre om betydningerne ligger semantisk tæt på hinanden.

4.4. Foreløbige resultater

I figur 3 ses en score for hver tilgang anvendt på tre forskellige dele af de håndannoterede data, nemlig et sæt med alle de centrale CBC-lemmaer, et med kun de polyseme CBC-lemmaer der har 5 eller færre DDO-betydninger, og et med ikke-centrale polyseme lemmaer i DDO (dvs. lemmaer der ikke er i CBC-udvalget) der tilsvarende har 5 eller færre betydninger. Den regelbaserede tilgang udviser foreløbig de bedste resultater med over 0,7 på alle datasæt.



Figur 3: Score for de tre tilgange sammenlignet med håndannoteringen.

Af de statistiske modeller fungerer BERT bedst; modellen opnår konsekvent bedre scorer end word2vec-modellen. Dog er forskellen mindst hvad angår de almindelige (ikke-centrale) polyseme lemmaer med mellem 2 og 5 betydninger. Her opnår begge modeller deres bedste score. Det viser også at statistiske modeller har

svært ved at håndtere meget finkornede betydningsadskillelser, og det skyldes bl.a. et klassisk problem med dataknaphed. Kun ét citat og én definition fra DDO er simpelthen ikke fuldt tilstrækkeligt datamateriale til statistisk beregning. Et udsnit af de automatiske sammenlægninger er gennemgået manuelt, og sandsynligvis vil kun de lemmaer der har 4 eller færre betydninger i DDO, kunne indsættes direkte i COR-S, de øvrige skal behandles manuelt.

5. Konklusioner

Det er en stor udfordring at samle leksikalske data fra de mange forskellige eksisterende ordbogsressourcer og ikke mindst at forenkle det betydningsinventar som har dannet grundlag for deres tilblivelse. Det kræver en del leksikografisk arbejde, særligt for stærkt polyseme ord, men de automatiske metoder vi har udviklet i projektet, giver gode resultater for lemmaer med 4 eller færre betydninger i DDO.

Det leksikografiske arbejde har givet os nye indsigter i det danske ordforråd, fx hvad angår mønstre af systematisk polysemi, og hvad angår fordelingen af ontologiske typer blandt monoseme ord. De grundige analyser af betydningsinventaret og de manuelle opmærkninger vil efterfølgende kunne anvendes i mange sammenhænge i redigeringen af DDO. Vi har fx registreret en del betydninger der i dag er blevet sjældne og gammeldags, 25 år efter at den første udgave af DDO blev redigeret. Det forenklede COR-S-betydningsinventar (der stadig er koblet til DDO-betydningerne) vil kunne danne grundlag for DSL's arbejde med at udgive en forenklet udgave af DDO, fx til skolebrug.

Planen er at COR-S-ressourcen efter projektets afslutning skal opdateres årligt med nye ord baseret på *Retskrivningsordbogens* løbende udvidelse. Vi håber at der bliver mulighed for at arbejde videre med et modul der beskriver de mange faste udtryk i DDO når

første version af COR-S er færdig; disse er naturligvis vigtige i en sprogteknologisk ressource. Også syntaktiske oplysninger fra *STO*, en ordbog med formaliserede syntaktiske oplysninger om en stor del af *Retskrivningsordbogens* og DDO's lemmaer, skal på længere sigt kobles på. Eksterne brugere af COR-registret vil forhåbentlig finde stor nytte i at tilkoble domænespecifikke ordbøger og i den sammenhæng få gavn af at almenordforrådet i forvejen er dækket af COR-S.

6. Litteratur

Ordbøger og digitale ressourcer

Begrebsordbogen = *Den Danske Begrebsordbog*. Sanni Nimb (hovedred.), Henrik Lorentzen, Liisa Theilgaard & Thomas Troelsgård (2015). Det Danske Sprog- og Litteraturselskab og Syddansk Universitetsforlag.

DanNet, se: <cst.ku.dk/projekter/dannet/>, <andreord.nors.ku.dk> (april 2022).

DDO = *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab, se: <ordnet.dk/ddo> (april 2022).

FrameNet, se: <framenet.icsi.berkeley.edu/fndrupal/> (april 2022).

Global WordNet Association, se: <globalwordnet.org> (april 2022).

Princeton Wordnet, se: WordNet Base Concepts.

Retskrivningsordbogen. Dansk Sprognævn, se <dsn.dk/ordboeger/retskrivningsordbogen/> (april 2022).

STO = Sprogteknologisk Ordbase. Anna Braasch et al. (red.) København: Center for Sprogteknologi. <cst.dk/cgibin/defisto> (april 2022).

WordNet Base Concepts, se: <globalwordnet.org/resources/gwa-base-concepts/> (april 2022).

Anden litteratur

Delvin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. I: *Proceedings of the 2019 Conference of NAACL: Human Language Technologies*, Volume 1. Minneapolis, Minnesota: Association for Computational Linguistics, 4171-4186.

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

Firth, John Rupert (1957): *Studies in Linguistic Analysis.* Special Volume of the Philological Society. Oxford: Blackwell.

Geeraerts, Dirk (2002): The theoretical and descriptive development of lexical semantics. I: Leila Behrens & Dietmar Zaefferer (eds.): *The Lexicon in Focus. Competition and Convergence in Current Lexicology*. Frankfurt: Peter Lang Verlag, 23-42.

Harris, Zellig (1954): Distributional structure. I: *Word* 10 (23), 146-162.

Hjorth, Ebba & Kjeld Kristensen (red.) (2003-2005): *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab og Gyldendal.

Mikolov, Tomas, Wen-tau Yih & Geofffrey Zweig (2013): Linguistic regularities in continuous space word representations. I: *Proceedings of the 2013 conference of NAACL: Human language technologies*, Atlanta, Georgia, 746-751.

Nimb, Sanni (2018): The Danish FrameNet Lexicon: method and lexical coverage. I: *Proceedings of the International FrameNet Workshop at LREC 2018*, Miyazaki, Japan, 51-55.

Nimb, Sanni, Sussi Olsen, Bolette S. Pedersen & Thomas Troelsgaard (2022): A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. I: *Proceedings of the 13th LREC conference*, Marseille, Frankrig, 2826-2832.

Nimb, Sanni, Nicolai H. Sørensen & Thomas Troelsgård (2018): From standalone thesaurus to integrated related words in the Danish Dictionary. I: *Proceedings from Euralex 2018*, Ljubliana, Slovenien, 916-923.

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

Olsen, Ida R., Bolette S. Pedersen & Asad Sayeed (2020): Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources. I: *Proceedings of the 2020 Globalex Workshop on Linked Lexicography*, Marseille, Frankrig, 45-52.

Pedersen, Bolette Sandford (2018): Semantisk processering og leksikografi. I: Ásta Svavarsdóttir, Halldóra Jónsdóttir, Helga Hilmisdóttir & Þórdís Úlfarsdóttir (red.): *Nordiske Studier i leksikografi* 14. Reykjavík: Nordisk forening for leksikografi, 18-28.

Pedersen, Bolette S., Nathalie C. H. Sørensen, Sanni Nimb, Ida Flörke, Sussi Olsen, Thomas Troelsgård (2022): Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open-Source COR Lexicon. I: *Proceedings of the 13th LREC Conference*, Marseille, Frankrig, 51-60.

Pedersen, Bolette S., Sanni Nimb og Sussi Olsen (2021): Dansk betydningsinventar i et datalingvistisk perspektiv. I: *Danske Studier* 2021. Odense: Syddansk Universitetsforlag & Universitets-Jubilæets danske Samfund, 72-106.

Pedersen, Bolette S., Sanni Nimb, Ida Rørmann Olsen, Sussi Olsen (2019): Merging DanNet with Princeton WordNet. I: *Proceedings of the 10th Global WordNet Conference2019 Proceedings*, Wroclaw, Poland: Oficyna Wydawnicza Politechniki Wroclawskiej, 125-134.

Pedersen, Bolette S., Sanni Nimb, Sussi Olsen & Nicolai H. Sørensen (2018a): Combining Dictionaries, Wordnets and other

Lexical Resources – Advantages and Challenges. I: *Globalex Proceedings 2018*, Miyasaki, Japan, 102-105.

Pedersen, Bolette S., Manex Aguirrezabal Zabaleta, 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: Global WordNet Association, 182-189.

Pedersen, Bolette S., Anna Braasch, Anders Johannsen, 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 LREC conference.* Portorož, Slovenien: European Language Resources Association (ELRA), 842-847.

Pedersen, Bolette S., Sanni Nimb, Jørg Asmussen, Nicolai Hartvig 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, 269-299.

Pustejovsky, James (1995): *The Generative Lexicon: A Theory of Computational Lexical semantics*. Cambridge: MIT Press.

Ruppenhofer, Josef, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker & Jan Scheffczyk (2016): *FrameNet II: Extended Theory and Practice*. <framenet.icsi.berkeley.edu/fndrupal/the_book> (april 2022).

Sørensen, Nicolai Hartvig & Sanni Nimb (2018): Word2Dict – Lemma Selection and Dictionary Editing Assisted by Word Embeddings. I: *Proceedings from Euralex 2018*. Ljubliana, Slovenien: Ljubljana University Press, 819-824.

Vossen, Piek (1999): *EuroWordNet General Document*. <archive.illc.uva.nl//EuroWordNet/docs.html> (april 2022).

Sanni Nimb, ledende redaktør, ph.d.
Ida Flörke, assisterende redaktør, cand.mag.
Thomas Troelsgård, seniorredaktør, cand.mag.
Det Danske Sprog- og Litteraturselskab
Christians Brygge 1
DK-1219 København
{sn, if, tt}@dsl.dk

Bolette Sandford Pedersen, professor, ph.d.
Nathalie Carmen Hau Sørensen, videnskabelig assistent, cand.mag.
Sussi Olsen, videnskabelig medarbejder, cand.mag.
Center for Sprogteknologi
Københavns Universitet
Emil Holms Kanal 2
2300 København S
{bspedersen, nmp828, saolsen}@hum.ku.dk