Evaluering af sprogforståelsen i danske sprogmodeller – med udgangspunkt i semantiske ordbøger

BOLETTE SANDFORD PEDERSEN, NATHALIE C. HAU SØRENSEN, SUSSI OLSEN & SANNI NIMB

ABSTRACT

Artiklen beskriver hvordan vi har udviklet en række datasæt – et såkaldt benchmark – til at evaluere forskellige aspekter af sprogforståelse i danske sprogmodeller. Vores antagelse er at den viden der allerede er beskrevet i en række eksisterende danske ordbøger, kan opfattes som 'ground truth' for semantikken i det danske ordforråd. Vores metode går derfor ud på at 'vende' de semantiske ordbøger om og bruge dem til at generere et benchmark der afprøver modellernes evne til at forstå dansk. Mere specifikt undersøger vi hvor godt modellerne i) forstår synonymi, nærsynonymi, og hvornår noget er semantisk associeret, ii) skaber inferens i relation til begrebsmæssig viden og nedarvning af egenskaber fra overbegreb til underbegreb, iii) laver korrekte følgeslutninger i forbindelse med specifikke handlinger og hændelser, iv) skelner mellem centrale betydninger af et ord i kontekst og v) håndterer positiv og negativ konnotation eller 'sentiment' i løbende tekst. Vi afprøver vores datasæt på ChatGPT 3.5 turbo og på ChatGPT 4.0 og kan se at datasættene har en passende sværhedsgrad i forhold til hvad modellerne er i stand til at håndtere, om end ChatGPT 4.0 opnår særdeles gode resultater for flere af datasættene.

EMNEORD: danske sprogmodeller; evaluering; sprogforståelse; semantiske ordbøger; benchmark; ChatGPT

1 STORE SPROGMODELLER – HVOR GODT FORSTÅR DE DANSK?

Store sprogmodeller indgår som den helt centrale komponent i de nye intelligente sprogtjenester der er ved at revolutionere vores videnssamfund. Chatbotter som ChatGPT er taget i brug i store dele af samfundet – også på dansk – på trods af at vi endnu ikke helt forstår hvordan de virker, og hvor høj deres abstraktionsevne egentlig er.

 DOI 10.7146/nys.v1i65.143072

Selv om de danske tekster der genereres af chatbotterne, ved første øjekast ser imponerende og temmelig overbevisende ud, afslører et nærmere eftersyn hurtigt en del huller i forhold til deres evne til at håndtere det danske ordforråd og dets semantik på en nuanceret og sprogligt kohærent måde. En del af disse mangler skyldes de sproglige og kulturelle bias der er opstået på grund af de engelsksprogede tekster som modellerne i høj grad er trænet på. Man kan derfor antage at disse problemer vil mindskes og måske med tiden forsvinde når og hvis sprogmodellerne i højere grad trænes på dansksproget materiale, og i mindre grad baseres på oversættelse eller 'sprogtransfer' fra engelsk og andre store sprog.

Det samme gælder for så vidt den problematik der opstår fordi modellerne primært er trænet på løbende tekst. I tekst formulerer vi typisk forgrundsviden, dvs. det der træder frem og er ekstraordinært og kontroversielt, hvorimod den baggrundsviden vi mennesker har og bruger når vi læser og forstår tekst, ikke træder så tydeligt frem da den forudsættes kendt. En del faktuel baggrundsviden kan findes i ordbøger og encyklopædier, men den er typisk sværere tilgængelig hvad angår udnyttelse i store sprogmodeller bl.a. på grund af betalingsmure og ophavsret. Et ofte anvendt eksempel til at illustrere sådanne statistiske bias hvor baggrundsviden er fraværende, er *det sorte får*. Sprogmodeller har længe fejlagtigt svaret *sort* hvis man spurgte hvilken farve et får har, selv om der i ordbøger og encyklopædier står at et får typisk er hvidligt. Det skyldes at det sorte får er det der slår igennem statistisk i teksterne. Udtrykket, som i sig selv er en metafor for noget der stikker ud (og ikke lever op til fællesskabets forventninger), bliver altså en ny metafor for statistisk bias eller skævvridning i sprogmodellerne, hvor væsentlig baggrundsviden ikke får nok vægt. På samme måde kunne man fejlagtigt tro at komikere typisk er kvinder fordi vi (stadig) synes det er relevant at nævne det eksplicit i teksten når de lige netop er kvindelige[1].

Andre problemer bunder i højere grad i sprogmodellernes tekniske design hvor vi endnu mangler helt at forstå hvordan og i hvor høj grad

1 Disse skævvridninger udjævnes hen ad vejen når datamaterialet vokser. Problematikken består imidlertid stadig i mindre, dansktrænede modeller, som vi potentielt også adresserer med vores datasæt så de netop kan bruges til at identificere modellernes begrænsninger.

de abstraherer og 'forstår'[2] de implicitte referencer vi mennesker anvender når vi bruger og fortolker sprog. Fx når det gælder overført og metaforisk sprog, når vi anvender ironi og sarkasme, eller når vi bruger ordforråd med negativ og positiv konnotation. Når den danske metafor *flødebolle* forklares af ChatGPT som *en der måske er blød, blid, eller ikke alt for modig eller robust* kan man som danskkyndig hurtigt konkludere at det er en temmelig skæv fortolkning; muligvis en fejloverførsel fra det engelske *creampie* idet metaforen på dansk snarere refererer til en (oftest mands)person med meget høje tanker om sig selv uden at have så meget at have det i, med reference tilbage til det oppustede æggehvideskum som udgør flødebollens indre. Et andet eksempel på fejloverførsel fra engelsk er æbleskive der af modellen tolkes til at være mad fra en plante, (dvs. bogstaveligt en skive af æble) hvor fænomenet ikke kendes som den kuglerunde kage vi især spiser i Danmark ved juletid.

Hvad enten de sproglige mangler skyldes for lidt dansk input i træningsmateriale, manglende baggrundsviden eller modellernes manglende abstraktionsevne mere generelt, er der brug for store og varierede evalueringsdatasæt, dvs. datasæt der systematisk kan belyse hvor godt sprogmodellerne virker, og hvornår og ved hvilke typer af udvikling de faktisk forbedres. Sådanne datasæt benævnes ofte *benchmarks*.

Det er vores antagelse at det er vigtigt at disse datasæt udvikles med et monolingvalt udgangspunkt fra det sprogsamfund de skal interagere i, og ikke, som det ofte ses, genereres via tilsvarende engelske datasæt. Ved oversættelse af datasæt sker der nemlig ofte det at bias igen introduceres via kildesproget, og at nuancer og mindre hyppigt forekommende eller særegent ordforråd i målsproget overses og dermed ikke afprøves. I denne sammenhæng er det vigtigt at understrege at relevante benchmarks også i høj grad handler om at ramme præcist i forhold til såkaldt 'aboutness' (jf. fx Hershcovich et al. 2022); dvs. de skal teste netop de aspekter der er særligt relevante for det givne sprog og det givne sprogsamfund, og ikke i hovedsagen evaluere aspekter som er mere relevante i fx et engelsk sprogsamfund.

2 Når vi i denne tekst anvender termerne 'forstå' og 'ræsonnere', refererer vi til sprogmodellens evne til at for-holde sig nogenlunde adækvat til specifikke sprogforståelsesopgaver. Vi går således ikke nærmere ind i den sprogfilosofiske fortolkning af hvad sprogforståelse i øvrigt indbefatter af fx hensigt, kontekst osv.

Vores benchmark tager udgangspunkt i det faktum at de allerede etablerede semantiske ordbøger og ressourcer der findes på dansk, udgør valideret semantisk viden; vi betragter dem således som en slags 'ground truth' eller 'guldstandard'. I vores undersøgelse bliver de så at sige 'vendt om' og brugt til at generere et benchmark der afprøver udvalgte sprogmodellers evne til at forstå dansk. Og som gør det på en omfangsrig og struktureret måde der også inddrager de mindre prototypiske dele af sproget og fx de mindre hyppige ord.

Vi regner med at vores benchmark som i øjeblikket indbefatter i alt knap ca. 4200 testeksempler fordelt over 26 datasæt[3], kan anvendes til at afdække i hvor høj grad sprogmodellerne i) forstår synonymi, nærsynonymi og hvornår noget er semantisk associeret, ii) skaber inferens i relation til begrebsmæssig viden og nedarvning af egenskaber fra overbegreb til underbegreb og iii) laver korrekte følgeslutninger i forbindelse med handlinger og hændelser, iv) skelner mellem centrale betydninger af et ord og v) kan håndtere positiv og negativ konnotation på en adækvat måde.

Via vores mangeårige arbejde med udvikling af danske semantiske ordbøger ved vi at netop disse informationstyper er indeholdt i de leksikalske beskrivelser på en nogenlunde systematisk måde som giver mulighed for semiautomatisk at generere omfangsrige evalueringsdata. Som det vil fremgå, fokuserer vi på danske data og evaluering af modeller der fungerer på dansk, men vi antager at vores metodeudvikling og erkendelser i den forbindelse også vil være relevante for andre sprog. Særligt for de mindre og mellemstore sprog er der et udtalt behov for at få genereret sådanne datasæt semiautomatisk (og dermed med lave omkostninger) og i stor skala, så udviklingen af sprogmodeller for disse kan følges og vurderes nøje.

Vi indleder i afsnit 2 med en kort introduktion til sprogmodellernes virkemåde og til hvordan man typisk evaluerer dem i en sprogteknologisk kontekst; dels på dansk, dels på andre sprog. Herefter introducerer vi i afsnit 3 de semantiske ordbøger som vi betragter som 'ground truth' for vores benchmark. I afsnit 4 beskriver vi hvordan vi udvikler de enkelte datasæt på baggrund af information fra ordbøger-

3 Data er tilgængelige på https://github.com/kuhumcst/danish-semantic-reasoning-benchmark og vil løbende blive opdateret.

ne, og for hver sprogforståelsesopgave diskuterer vi desuden hvor godt de løses af udvalgte 'state of the art'-modeller. I afsnit 5 sammenligner vi sprogmodellernes løsning af de syv sprogforståelsesopgaver, og i afsnit 6 konkluderer vi og beskriver de videre perspektiver for at udvikle benchmarket.

2 HVORDAN VIRKER SPROGMODELLERNE, OG HVORDAN EVALUERER VI DEM?

En central bestanddel i store sprogmodeller er de såkaldte statistiske ordprofiler, også kaldet *word embeddings*, som udgøres af numeriske vektorer, en slags lister med numeriske værdier der repræsenterer hvilke ord et givent ord *typisk optræder sammen med* i et tekstkorpus. Hver ordprofil repræsenterer med sin vektor et punkt i et flerdimensionelt vektorrum, og ord med afvigende distribution vil ligge langt fra hinanden mens ord med lignende distribution og dermed lignende betydning[4] vil ligge tæt på hinanden i vektorrummet. Synonyme ord som fx *sprinte* og *spurte* har således ordprofiler der ligger tæt på hinanden fordi de typisk optræder sammen med lignende ord i et korpus[5].

Efterhånden som man i de senere år har øget antallet af parametre til lagring af oplysninger om ordenes distribution, har man også mangedoblet den sproglige information som sprogmodellen kan regne sig frem til og gør brug af. Det betyder i korte træk at modellerne repræsenterer mere tekstlig kontekst på et højere generaliseringsniveau, og at der i højere grad kan tages højde for langdistanceafhængigheder i det tekstlige input. Vi ser således at det hyppige danske fænomen med partikelverber hvor partiklen ofte strander langt fra sit verbum, nu kan fortolkes langt bedre af modellerne, som i eksemplet: *du kan slå alle de ord du ikke kender, op i ordbogen online.* Her er pointen at *op* skal fortolkes som en del af partikelverbet *slå op*, og at betydningen af verbet således ikke er gennemsigtig i forhold til betydningen af ordets to enkeltdele, hhv. *slå* og *op*.

4 Vi henholder os her til den distributionelle hypotese om at ord der har lignende distributionel kontekst, også har lignende betydning, jf. fx Firth (1957), Levin (1991) og relateret til NLP fx Lenci & Sahlgren (2023).

5 For en mere uddybende forklaring for lægmand kan vi henvise til fx Lee & Trott (2023) eller andre umiddel-bart tilgængelige beskrivelser i blogposts og på wikipedia.

Modellens abstraktionsevne understøttes af det *lagdelte* design (som også kaldes neurale lag efter inspiration fra den menneskelige hjerne), hvor hvert lag tager en sekvens af ordprofiler som input, beregner yderligere information ud fra distributionen og videregiver disse udvidede vektorer til næste lag. En simpel googlesøgning afslører at ChatGPT 3.5 fx indeholder 175.000.000.000 sådanne parametre på tværs af 93 neurale lag, mens ChatGPT 4.0 indeholder ca. 1.8.000.000.000.000 parametre på tværs af 120 lag. GPT-Modellerne benævnes også *transformermodeller,* hvilket indikerer at der for hvert lag transformeres en inputsekvens til en outputsekvens med et højere generaliseringsniveau, og at dette sker ved hjælp af såkaldte 'self attention heads', som forstærker signalet mellem to eller flere relaterede ord (reelt tokens) i en sætning og dermed muliggør at mere generel information genereres og deles på tværs af inputtet. Hvert lag giver plads til endnu et abstraktionsniveau, og dermed øges den generelle viden om bl.a. ordenes morfologiske, syntaktiske og semantiske egenskaber, om forholdet mellem ord og ordsekvenser, og ikke mindst om hvilket ord der typisk følger efter som det næste. Det er væsentligt at bemærke at de modeller der i øjeblikket fungerer bedst for dansk, er flersproglige, og at de rent faktisk lærer en del om dansk sprog via andre sprog, særligt engelsk. Dette er en meget stor fordel da mange sproglige fænomener går igen fra sprog til sprog; ulempen er så at der opstår sproglige bias fordi det er svært at styre modellerne så de kun overfører *relevant* viden fra det ene sprog til det andet. Generelt gælder det, ikke overraskende, at jo mere tekst der trænes eksplicit med for det pågældende sprog, desto bedre fungerer modellen for netop det sprog.

Udover at analysere tekst bruges sprogmodeller i dag typisk produktivt til netop at forudsige hvilke ordsekvenser der hyppigt kommer efter et givent ord. Så taler vi om en *generativ* eller *selvproducerende* anvendelse af sprogmodellen, som er den vi ser i netop ChatGPT, hvor modellen selv skaber nyt indhold på en relativt selvstændig og kreativ måde, dog i udgangspunktet naturligvis baseret på den tekst som modellen er trænet med. Generative sprogmodeller og også generativ kunstig intelligens mere generelt kan altså forstås som systemer der selv skaber indhold, i modsætning til systemer der kun *analyserer* data.

Evaluering af sprogmodeller og relevante eksisterende benchmarks

Når man udvikler og vedligeholder sprogmodeller, er det almindeligt at anvende standardiserede og omfangsrige evalueringsdatasæt så man løbende kan afprøve i hvor høj grad og på hvilke områder modellerne udvikler sig. Datasættene giver mulighed for automatisk og systematisk at måle hvor godt de hver især løser en lang række opgaver, bl.a. med henblik på at sammenligne dem med hinanden. Opgaverne kan være designet så de tester modellernes evne til at udføre forskellige praktiske sprogtjenester ('extrinsic evaluation'), som fx at oversætte mellem to sprog eller at lave et resumé af en tekst. Men de kan også designes så de mere generelt tester modellernes sprogforståelse fx i forhold til at kunne 'ræsonnere' om forskellige forhold, til at forstå nuanceforskelle mellem ord, og til at forstå konsekvensen af en beskrevet handling (samlet set også kaldet 'intrinsic evaluation'). Vores datasæt fokuserer på sidstnævnte af to grunde: dels mangler vi i høj grad sådanne datasæt for dansk, hvilket vanskeliggør en nuanceret evaluering af hvor godt danske sprogmodeller udvikler sig; dels er det her at vi antager at vores semantiske ordbøger udgør en helt særlig ressource.

Der findes altså allerede forskellige evalueringsdatasæt for dansk der løser praktiske sprogopgaver som fx tekstklassifikation, tekstresumering og maskinoversættelse. Flere af disse er tilgængelige via Digitaliseringsstyrelsens sprogteknologiportal 'sprogteknologi.dk', og de samles desuden i den skandinaviske evalueringsplatform *ScandEval* (Nielsen 2023) i de tilfælde hvor der findes tilsvarende datasæt for svensk og norsk.

Der eksisterer også datasæt rundt omkring i virksomheder (fx medievirksomheder) og ved de forskellige forskningsinstitutioner som kan bruges til at evaluere delelementer i sprogmodellernes performans. Disse indbefatter datasæt for leksikalsk-semantisk nærhed (fx Nielsen & Hansen 2017, Schneidermann et al. 2020) samt betydningsopmærkede korpora der kan anvendes til at vurdere hvor godt sprogmodellerne skelner mellem ordbetydninger i kontekst (Pedersen et al. 2016, 2023).

I det internationale miljø findes der pakker af datasæt og evalueringsplatforme der samlet set indbefatter begge typer evaluering. Det gælder fx GLUE- og SUPERGLUE-testsuiterne (Wang et al. 2018, 2020) samt den nyligt udviklede svenske SUPERLIM-platform (Berdicevskis et al. 2023) lavet med SUPERGLUE som forbillede. Udfordringen er

at modellerne hele tiden forbedres; derfor må testapparatet følge med og gøres sværere. Således indeholder den seneste version af SUPERGLUE otte relativt komplekse sprogforståelsesopgaver, herunder beregning af koreference, entydiggørelse af flertydige ord, udredning af følgeslutninger mv. For norsk findes der for nuværende den såkaldte *NorBench* (Samuel et al. 2023), der blev præsenteret ved NODALIDA-konferencen i Thorshavn i 2023. Den indeholder dog først og fremmest et antal sprogtjenesteopgaver af typen maskinoversættelse og spørgsmål–svar og ikke specifikke sprogforståelsesdatasæt som dem vi præsenterer her.

3 HVILKEN RELEVANT VIDEN RUMMER DE SEMANTISKE ORDBØGER?

De semantiske ordbøger som vi afprøver som grundlag for vores benchmark, er *Den Danske Begrebsordbog* (Nimb et al. 2014, Nimb 2016), det danske *wordnet*, *DanNet* (Pedersen et al. 2009), *Det Danske FrameNet-leksikon* (Nimb et al. 2017)[6], og den semantiske komponent i *Det Centrale OrdRegister for dansk*, også kaldes *COR.SEM*[7] (Nimb et al. 2022a). Alle de her nævnte ordbøger indeholder formaliserede semantiske oplysninger der mere eller mindre eksplicit bygger på beskrivelser af lemmaer i *Den Danske Ordbog* (Hjort & Kristensen 2003-2005; https://ordnet.dk/ddo).

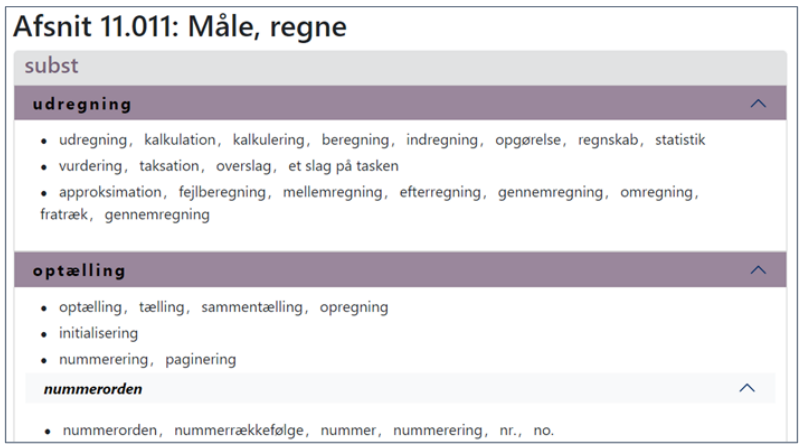
Fra **Den Danske Begrebsordbog** kan vi udlede hvilke ord mennesker anser for at være synonyme og nært semantisk og tematisk beslægtede. Ordbogen indeholder ca. 100.000 lemmaer og 130.000 betydninger fra Den Danske Ordbog fordelt over 22 kapitler og 888 afsnit. Hvert afsnit er navngivet tematisk og indeholder en række ordklasseopdelte og dernæst semantisk inddelte grupper der hver især består af betydningsmæssigt relaterede ord og udtryk, hvor den indbyrdes afstand afspejler graden af semantisk nærhed. Nogle af grupperne indledes af såkaldte 'nøgleord', der fungerer som en slags overskrift. Et nøgleord er i langt de fleste tilfælde et overbegreb til de følgende ord, men det kan i nogle tilfælde også blot være det mest udbredte ord indenfor en mindre semantisk gruppe. Der opereres med to niveauer af nøgleord: overord-

6 En samlet beskrivelse på dansk af disse ordbøger og principperne bag dem kan findes i Pedersen et al. (2021).

7 https://ordregister.dk

nede og underordnede, hvor de overordnede har virkefelt over alle efterfølgende grupper, også dem der indledes med et underordnet nøgleord.

FIGUR 1: ET EKSEMPEL FRA DEN DANSKE BEGREBSORDBOG. OVERORDNEDE NØGLEORD HAR LILLA BAGGRUND (*UDREGNING, OPTÆLLING*) OG UNDERORDNEDE NØGLEORD HAR GRÅ BAGGRUND (*NUMMERORDEN*). HVER PRIK ADSKILLER BETYDNINGSGRUPPER.



I figur 1 har det overordnede nøgleord *udregning* virkefelt over tre betydningsgrupper frem til det næste overordnede nøgleord *optælling,* som har virkefelt over fire betydningsgrupper, heraf én gruppe som også hører under det underordnede nøgleord *nummerorden*. Tættest semantiske lighed finder man indenfor samme betydningsgruppe (uanset om den indledes med eget nøgleord eller ej), dernæst blandt de øvrige betydningsgrupper med samme nøgleord i hele det navngivne afsnit eller i det samlede kapitel som afsnittet tilhører. Strukturen er udnyttet i præsentationen af afgrænsede grupper af semantisk beslægtede ord fra *Begrebsordbogen* i online-udgaven af Den Danske Ordbog (https://ordnet.dk/ddo) i form af funktionen 'Ord i nærheden' (Nimb et al. 2018).

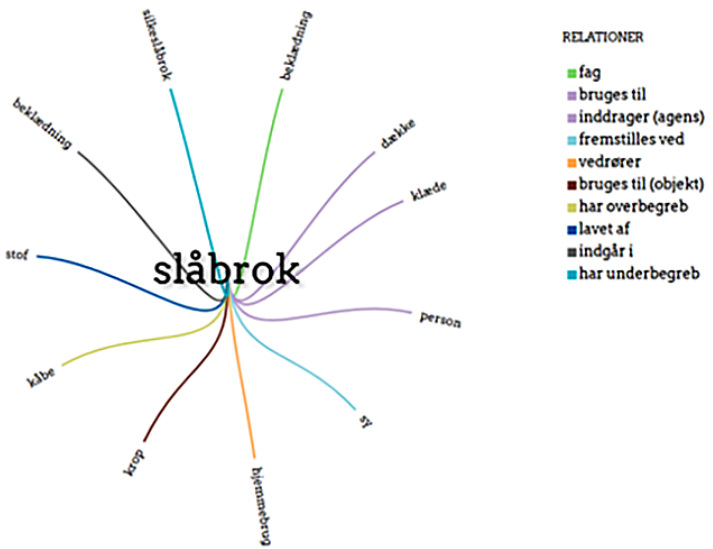
Ud fra DanNets, det danske wordnets, ontologi[8] og semantiske relationer kan vi udlede en række prototypiske egenskaber (i alt er 70.000 lemmaer beskrevet), dels dem der er specifikke for hvert begreb, dels dem der nedarves fra mere overordnede begreber (fx at fødevarer typisk

8 DanNets ontologi bygger på *EuroWordNet*-ontologien (Vossen et al. 1999), men har derudover en række mere specificerede typer, jf. Pedersen et al. (2009).

har en nærende funktion). Med disse benchmarkdata vil vi gerne afprøve i hvor høj grad sprogmodellerne har tilegnet sig en tilsvarende viden om semantiske nedarvningsrelationer.

DanNet giver også information om begrebers kerneegenskaber angivet på baggrund af Pustejovskys (1998) teori om begrebers såkaldte *qualiastruktur*. Det er fx anført hvordan en entitet er blevet til (et brød fx ved bagning), og hvad nogets typiske funktion eller rolle er (fx at *lyse* for en lampe, eller at *undervise* for en lærer). Figur 2 illustrerer som eksempel en række kerneegenskaber for beklædningsgenstanden *slåbrok* som i udgangspunktet er udledt semiautomatisk fra begrebets definition i Den Danske Ordbog og kombineret med de nedarvede egenskaber fra dets overbegreber (fx at beklædningsgenstande typisk er blevet syet, og at de typisk bruges af mennesker for at tildække kroppen). Denne viden tester vi om sprogmodellerne besidder.

FIGUR 2: BEGREBET SLÅBROK OG DETS KERNEEGENSKABER SOM BESKREVET I DANNET BL.A. PÅ BAGGRUND AF DEN DANSKE ORDBOGS DEFINITIONER. ('FAG' SKAL HER FORSTÅS SOM EMNEOMRÅDE.)



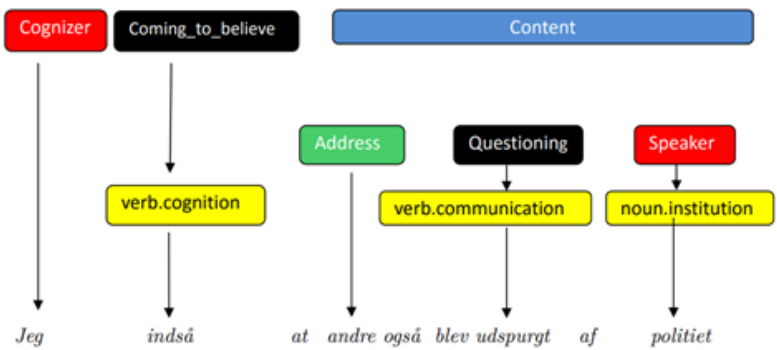
Fra **Det Danske FrameNet-leksikon** kan vi udlede hvilke slags handlinger, transaktioner og situationer danske verber og verbalsubstantiver refererer til, og kombineret med *Berkeley FrameNets* udfoldning af de

semantiske frames (http://framenet.icsi.berkeley.edu/) har vi viden om de semantiske roller der typisk forekommer i en ytring med en specifik frame, dvs. hvem der gør hvad med hvem. I en kommunikationsframe er der fx typisk nogen der kommunikerer, nogen der lytter, og noget information der bliver kommunikeret; for bevægelsesframes nogen eller noget der bevæger sig på en bestemt måde og evt. fra et sted til et andet, dvs. en kilde, en sti og et mål.

Tilsvarende kan andre handlinger eller hændelser have særlige følgevirkninger, dvs. en ny situation udløst af at handlingen har fundet sted. Gode eksempler er frames som CUTTING med verberne *klippe* eller *skære og* framen COMMERCE_BUY med verber for at købe. Vi afprøver her i hvor høj grad sprogmodellerne kan tage højde for disse følgevirkninger, og om de inddrager de semantiske roller som de forskellige verber aktiverer i en ytring.

Figur 3 illustrerer hvordan verberne *indse* og *udspørge* udløser hver deres semantiske frame for hhv. kognition og kommunikation (indeholdt i FrameNet-leksikonet), og hvordan de semantiske roller for hver frame realiseres i den givne ytring. Når denne kognition har fundet sted, har 'jeg'et' erkendt noget, som det ikke vidste før, nemlig at også andre var blevet udspurgt af politiet.

FIGUR 3: OPMÆRKNING MED FRAMES (I SORT) OG SEMANTISKE ROLLER FOR HHV. EN KOGNITIONS- OG EN KOMMUNIKATIONSFRAME, UDLØST AF HHV. INDSE OG UDSPØRGE (FRA NIMB ET AL. 2017: 12)

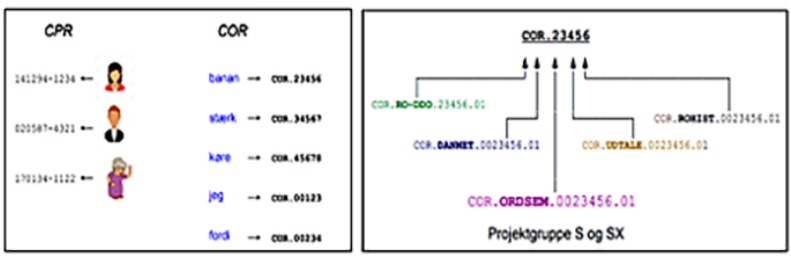


COR.SEM er en helt ny leksikalsk ressource, som i stor udstrækning samler, strømliner og forenkler semantisk information fra de før om-

talte ordbøger. Ordbogen rummer for nuværende 34.000 lemmaer og deres almene betydninger, hvor fokus har været på at beskrive de lemmaer der er enten begrebsmæssigt eller emnemæssigt centrale i dansk. Betydningsinventaret er en forenklet udgave af betydningsinventaret i Den Danske Ordbog (DDO) idet mange betydninger fra ordbogen enten er sammenlagt eller helt frasorteret ud fra en række systematiske principper og en nøje analyse af hvert lemma (jf. Pedersen et al. 2022 for en redegørelse af denne forenkling).

Ressourcen er således særlig værdifuld i forhold til vores benchmark idet den er velegnet til at evaluere modellernes evne til at *skelne mellem de centrale ordbetydninger i kontekst*, og den fungerer umiddelbart som en bedre guldstandard for dette end fx DDO's mere finkornede betydningsinventar. COR.SEM udgør en del af Det Centrale Ordregister[9] som er et standardiseret ordregister som sikrer at alle lemmaer i dansk tildeles et unikt nummer på tværs af ordbøger og fagterminologier således at det er nemmere at samkøre forskellige leksikalske ressourcer til sprogteknologisk anvendelse, jf. figur 4.

FIGUR 4: DET CENTRALE ORDREGISTER: LIGESOM BORGERE I DANMARK ALLE HAR ET CPR-NR, FÅR ALLE LEMMAER I DANSK TILDELT ET UNIKT NUMMER PÅ TVÆRS AF ORDBØGER (FRA HENRICHSEN 2021).



COR.SEM indeholder også en mere detaljeret version af *Det Danske Sentimentleksikon* (Nimb et al. 2022b). Hvor *Det Danske Sentimentleksikon* mere forenklet angiver positiv eller negativ konnotation på lemmaniveau fra -5 (meget negativ) til +5 (meget positiv) og udelader lemmaer der er tvetydige (fx adj. *fed*), er konnotation i COR.SEM angivet for hver *betydning* (*fed* har fx 2 negative og 3 positive betydninger), dog

9 *Det Centrale OrdRegister* administreres af Dansk Sprognævn og er åbent tilgængeligt: https://ordregister.dk.

kun inden for en konnotationsskala fra -3 til +3, hvor -3 altså angiver meget negativ konnotation og +3 meget positiv konnotation. Da vi også har brugseksempler i form af citater fra DDO, kan vi med denne oplysning evaluere et ords sentiment i dets kontekst som en del af vores benchmark.

4 FRA INFORMATION I DE SEMANTISKE ORDBØGER TIL SPROGFORSTÅELSESOPGAVER

Vores metode går altså som nævnt ud på at anskue de semantiske ordbøger som vores guldstandard eller 'ground truth' som vi kan evaluere sprogmodellernes evne til sprogforståelse[10] op imod. Vi genererer datasættene semiautomatisk ud fra ordbøgerne og afprøver dem på i første omgang GPT3.5 Turbo og GPT4 via såkaldt prompting[11]. Det gør vi for at få en første fornemmelse af om vores datasæt er tilstrækkelig udfordrende for modellerne. Vi stiller en række stilistisk stramme spørgsmål (jf. eksemplet i figur 5), hvor svarene kan evalueres automatisk idet sprogmodellen skal vælge svarmuligheder fra et lukket vokabular. Det kan være et eller flere ord, en sandhedsværdi eller en numerisk værdi. Til hver opgave skriver vi en prompt som er delt op i en overordnet instruktion og en skabelon hvori dataeksempler kan indsættes automatisk. Den overordnede instruktion indeholder en kort forklaring af input til modellen, formålet med opgaven, og hvordan outputtet skal se ud. Nedenfor ses et eksempel fra en af opgaverne (synonymi, figur 5).

FIGUR 5: OPGAVEN TIL CHATGPT SOM DEN ER FORMULERET FOR SYNONYMI

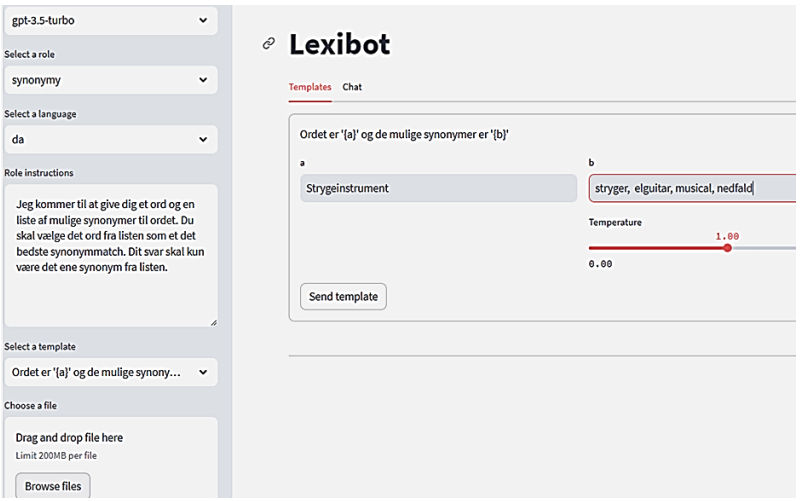
Jeg kommer til at give dig et ord og en liste af mulige synonymer til ordet. Du skal vælge det ord fra lisen som er det bedste synonymmatch. Dit svar skal kun være det ene synonym fra listen. Output skal være et enkelt ord og intet andet.

10 Igen er der tale om sprogforståelse på et niveau der angår leksikalsk semantisk og sætningssemantik. Man kan naturligvis forestille sig en lang række andre sprogforståelsesopgaver som inddrager funktionelle, tekstse-mantiske eller andre forhold.

11 Flere af disse forsøg er også blevet præsenteret ved LREC-konferencen i Torino i 2024, jf. Pedersen et al. (2024).

Instruktionen er en enkelt sætning med en række pladsholdere som dataeksemplerne kan udskiftes med. Til synonymiopgaven fra figur 5 giver vi sætningen: 'Ordet er {a}, og de mulige synonymer er {b}, hvor {a} er et enkelt ord og {b} en liste af fire mulige synonymer', hvorfra sprogmodellen skal vælge det rigtige. Formålet med skabelonen er automatisk at kunne behandle et stort antal dataeksempler gemt i en fil. Vi har derfor udviklet en grænseflade som i baggrunden har API-adgang til *OpenAI*'s generative sprogmodeller (ChatGPT 3.5 Turbo og ChatGPT 4.0). Igennem grænsefladen kan man vælge en fil med en række dataeksempler som bliver behandlet, og resultatet bliver gemt. Grænsefladen tillader også at man nemt kan skifte mellem model og opgavetype.

FIGUR 6: SCREENDUMP FRA LEXIBOT, VORES API-GRÆNSEFLADE TIL CHATGPT.[12]



12 Sprogmodellers 'temperatur' er en parameter hvormed man kan angive graden af tilfældighed og kreativitet i modellens output. En lav temperatur giver konsistente og mere ensartede svar; når temperaturen øges, stiger variationen og kreativiteten i outputtet. Ved højere temperatur øges også risikoen for irrelevante output. Her har vi kun lavet forsøg med standardindstillingen 1.00. Vi vil i senere forsøg undersøge om en lavere temperatur ændrer på resultaterne. Da vi instruerer sprogmodellerne i kun at måtte svare med et enkelt ord og dermed har sat en grænse for hvor kreativt et output de kan levere, er vores hypotese at en lavere temperatur ikke vil medføre større ændringer.

Det skal nævnes at denne opsætning blot er et eksempel på hvordan benchmarket kan bruges på én type sprogmodel; i praksis vil benchmarket kunne bruges på andre sprogmodeller med andre typer opsætninger[13].

I det følgende gennemgås hver sprogforståelsesopgave for sig, og for hver opgave undersøger vi hvor godt den løses af de to udvalgte sprogmodeller.

Synonymi

Synonymiopgaven går ud på at finde det mest passende synonym til et målord fra en liste med fire mulige ord. Ordenes grad af synonymi varierer i forhold til målordet på en skala fra højeste grad af semantisk beslægtethed, 'snæver' synonymi, som er det korrekte svar, til mindste grad af semantisk beslægtethed (se nedenfor mht. eksempler). Skalaen er automatisk beregnet ud fra strukturen med kapitler, afsnit, betydningsgrupper og nøgleord i Den Danske Begrebsordbog, se ovenfor.

På baggrund af strukturen udtrækkes ordpar i form af to direkte naboord i Begrebsordbogen inden for samme betydningsgruppe. De ordpar der samtidig er synonymer i DDO, repræsenterer snæver synonymi og den højeste grad af semantisk beslægtethed. Når vi har identificeret et synonympar automatisk i DDO-manuskriptet, skal vi finde de resterende tre ord som modellen skal vælge imellem. For at opgaven skal have en vis sværhedsgrad, udvælger vi kandidater som i varierende grad er semantisk beslægtede med synonymparret. Den nærmeste kandidat udvælges blandt ord i en anden betydningsgruppe, men stadig under samme nøgleord i samme afsnit i Begrebsordbogen. Fx kan vi i det navngivne afsnit 'musikinstrumenter' under nøgleordet 'strengeinstrument' finde synonymparret *strygeinstrument* og *stryger* og i en anden betydningsgruppe under samme nøgleord finde *elguitar s*om en mulig tæt semantisk beslægtet kandidat. Den tredje kandidat findes i et af de andre navngivne afsnit inden for samme kapitel, fx i dette tilfælde kandidaten *musical* i afsnittet 'Teater, optræden', der ligesom afsnittet 'musikinstrumenter' ovenfor også hører under kapitlet 'Kunst og kultur'. Den fjerde og sidste kandidat findes i et af de øvrige 21 kapitler i Begrebsordbogen, som fx *nedfald* fra kapitlet 'Sted og bevægelse', idet

13 Ikke alle sprogmodeller instrueres med prompting. Nogle modeller kræver at man først transformerer data til et format modellen kender (fx vektorrepræsentation), og andre modeller klarer sig bedre hvis de først er blevet trænet eller fintunet til den specifikke opgave.

vi forinden har frasorteret de afsnit der er tematisk beslægtede. Den endelige liste kan ses nedenfor i figur 7.

FIGUR 7: EKSEMPEL PÅ DATA FRA SYNONYMIDATASÆTTET.

Målord	Kandidat 1 (korrekt svar)	Kandidat 2	Kandidat 3	Kandidat 4
Strygeinstrument	Stryger	Elguitar	Musical	Nedfald
Kunst og Kultur; Musikinstrument; Strengeinstrument	**Kunst og Kultur;** Musikinstrument; Strengeinstrument	**Kunst og Kultur;** Musikinstrument; Strengeinstrument	**Kunst og Kultur;** Teater, optræden; Teaterform	**Sted og bevægelse;** Ned; Fald

Alle lister der anvendes i opgaven, består af ordkandidater udvalgt tilfældigt fra de overstående kriterier idet rækkefølgen af de fire kandidater også angives tilfældigt. I alt ender vi med 317 lister (hver med fire kandidater) fra 43 afsnit i Den Danske Begrebsordbog.

Begge modeller klarer opgaven godt med 93,6 % korrekt identificerede synonymer hos ChatGPT 3.5 Turbo og 96,2 % korrekte synonymer for ChatGPT 4.0. En del af ChatGPT 3.5 Turbos fejl skyldes at modellen ikke følger instruktionen og giver selve målordet som output i stedet for at vælge et ord fra listen. Med henblik på at analysere fejlene spørger vi modellerne med en ny prompt der samtidig beder modellen om at forklare sit valg. Her ser vi to mønstre af fejl: 1) den har ikke kendskab til betydningen af et af ordene i synonymparret (fx *mø* i parret '*mø*, *jomfru*', *minut* i parret '*minut, bueminut*', *citronmelisse* i parret '*citronmelisse*, *hjertensfryd*' og det historiske militære udtryk *retræte* i parret '*retræte*, *tappenstreg*'), eller 2) den nægter at svare pga. stødende eller nedsættende sprog (fx i tilfældene '*aboriginer*, *australneger*' og '*tyskerpige*, *tyskertøs*'). Det sidste sker også når ChatGPT 4.0 anvendes, dog kun med synonymerne '*sort*, *neger*'.

I de resterende 11 tilfælde af fejl som ChapGPT 4.0 laver, vælger den ni gange den næst-tætteste ikke-synonyme kandidat (samme afsnit og nøgleord, men anden betydningsgruppe under nøgleordet). I ét tilfælde, for ordet *felt*, vælger den fejlagtigt det ord som er mindst semantisk beslægtet idet det er fra et helt andet kapitel, nemlig ordet *følelsesliv* blandt mulighederne '*område*, *flig*, *overlegenhed*, *følelsesliv*'. Det rigtige

synonym er *område*. Dette kan skyldes en klassisk transferfejl hvor behandlingen af det danske eksempel på en uhensigtsmæssig måde går via engelsk. Modellen sammenblander det engelske verbum *to feel* ('at føle'), som bøjes *felt* i datid, og det danske substantiv *felt* i betydningen 'afgrænset flade eller område'.

Semantisk nærhed i et bredere perspektiv

Vi evaluerer også sprogmodellernes evne til at vurdere grader af semantisk lighed i videre forstand end den rent snævre synonymirelation der afspejles i DDO's synonymangivelser. Det gør vi ved at stille modellerne overfor to såkaldte *word intrusion*-opgaver (Nielsen & Hansen 2017), hvor et afvigende ord skal identificeres i en ordgruppe der i øvrigt består af nært beslægtede ord. Hvert testeksempel består med andre ord af en kernegruppe af nære ord og en afviger idet alle ordene udtrækkes automatisk fra Begrebsordbogen. Efterfølgende udvælger vi manuelt de bedste testeksempler til brug for opgaverne.

Formålet med den første *word intrusion*-opgave er at evaluere semantisk nærhed. I denne opgave er hovedgruppen med de nært beslægtede ord derfor synonymer eller nærsynonymer. De stammer fra den samme betydningsgruppe i Begrebsordbogen, dvs. fra en gruppe af ord der ifølge dens struktur er så tæt beslægtet semantisk som ord kan være. Herefter suppleres den med et afvigende ord fra en anden betydningsgruppe der dog stadig hører under samme overordnede nøgleord, dvs. er forholdsvis tæt semantisk beslægtet med ordene i kernegruppen, men helt sikkert ikke er et synonym.

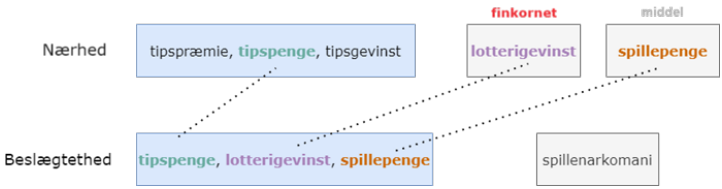
Vi opdeler opgaven i en 'finkornet' og 'grovkornet' underopgave, idet hypotesen er at den finkornede underopgave er sværest at løse. Baggrunden for denne niveauopdeling er at vi gerne vil udfordre sprogmodellerne mest muligt, og vi kender i forvejen ikke til modellernes nuværende niveau. De to underopgaver adskiller sig kun fra hinanden ved at have to forskellige afvigende ord, kernegruppen er den samme som det kan ses i figur 8.

FIGUR 8: EKSEMPEL PÅ EN KERNEGRUPPE OG ET AFVIGENDE ORD FRA HHV. ET GROV- OG FINKORNET DATASÆT OM SEMANTISK NÆRHED

Nært beslægtede ord	**Afvigende ord**	
tipspræmie, tipspenge, tipsgevinst	Finkornet:	*lotterigevinst*
	Grovkornet:	spillepenge

I etableringen af det finkornede datasæt tages det afvigende ord fra en anden betydningsgruppe der hører direkte under det samme overordnede nøgleord og derfor må formodes at være forholdsvis lidt afvigende fra de øvrige ord. I det grovkornede datasæt tages det afvigende ord på samme måde fra en af de betydningsgrupper der hører under det samme overordnede nøgleord, men i dette tilfælde hører det desuden under et underordnet nøgleord. Ud fra strukturen i Begrebsordbogen må man derfor antage at det drejer sig om et betydningsmæssigt lidt mere afvigende ord end i det første datasæt. I begge datasæt vil det afvigende ord dog være forholdsvis tæt semantisk beslægtet med de tre øvrige ord da alle fire ord har samme overordnede nøgleord.

FIGUR 9: SAMMENLIGNING AF KERNEGRUPPEN (BLÅ) I DATASÆTTENE OM NÆRHED OG BESLÆGTETHED. BEMÆRK AT AFVIGENDE ORD FRA DATASÆTTET 'NÆRHED' (GRÅ KASSER) GODT KAN FREMTRÆDE I DEN SAMME KERNEGRUPPE I DATASÆTTET 'BESLÆGTETHED'.

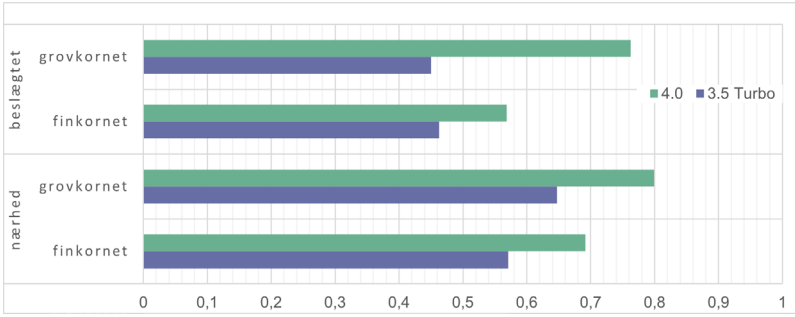


I den anden *word intrusion*-opgave, som vi betegner 'beslægtethed' (eller association), indgår der også to forskellige underopgaver. Endnu engang opstiller vi både en finkornet og en grovkornet underopgave. I det første tilfælde tages det afvigende ord fra en ordgruppe der, modsat ovenfor, hører under et andet overordnet nøgleord, men dog står i samme afsnit. Inden for dette datasæt vil der derfor være tale om en større semantisk afvigelse mellem kernegruppen og det afvigende ord end i den første word intrusion-opgave. Det kan fx være den forskel man finder mellem ord der beskriver handlinger, og ord der beskriver personer (se figur 9).

Til brug for den anden underopgave indenfor 'beslægtethed', den grovkornede underopgave, etablerer vi endnu et datasæt, denne gang bestående af en gruppe ord der blot har det til fælles at de står i samme afsnit i Begrebsordbogen, og hvor det fjerde afvigende ord tages tilfældigt fra et helt andet afsnit, dog et afsnit der stadig hører under samme kapitel. Vi opnår dermed et datasæt med tre ord der er tematisk tæt beslægtet[14], hvorimod det fjerde ord er fjernere beslægtet tematisk set, men dog stadig associativt beslægtet.

Det er vigtigt at understrege at vi manuelt har valideret alle de datautræk der indgår i opgaverne for at sikre os at et menneske kan adskille det afvigende ord fra kernegruppen i alle de dataeksempler der anvendes.

FIGUR 10: SVARKORREKTHED I PROCENT FOR CHATGPT 3.5 TURBO (BLÅ) OG CHATGPT 4.0 (GRØN) PÅ TVÆRS AF DATASÆTTENE SORTERET EFTER GRADEN AF SEMANTISK SAMMENFALD. ØVERST ER DER MINDST SAMMENFALD (BESLÆGTETHED, GROVKORNET) OG NEDERST ER DER MEST SAMMENFALD (NÆRHED, FINKORNET).



I figur 10 ses resultaterne af en kørsel af datasættene på hhv. ChatGPT 3.5 Turbo (blå) og ChatGPT 4.0 (grøn). Datasættene er sorteret således at der er mindst semantisk sammenfald øverst (beslægtethed, grovkornet) og mest sammenfald nederst (nærhed, finkornet). Som forventet ser vi generelt at de opgaver vi betegner som 'grovkornede' under begge underopgaver er mindre udfordrende for modellerne. Når vi kigger på selve underopgaverne 'nærhed' versus 'beslægtethed', ser vi at model-

14 Kernegruppen på dette niveau har nu et emneområde (fx hører *togskinne*, *lokomotiv* og *sporarbejde* alle under emneområdet *jernbane*) eller en kategori (*lilje*, *myrte* og *citrontimian* er alle en type *planteart*) til fælles.

lerne klarer sig bedst ved 'nærhed'. Det kan skyldes at betydningsgrupperne er klart afgrænsede. I datasættene af typen 'beslægtethed' er der langt mere variation mellem ordene da de er blevet udledt automatisk af strukturen i Begrebsordbogen hvorimod de øvrige datasæt er baseret på ordbogens betydningsgrupper og dermed manuelt grupperet ud fra semantisk lighed da den blev skrevet. Flertydighed skaber desuden ikke overraskende problemer i genereringen af kernegrupperne i begge underopgaver af typen 'beslægtethed' hvilket man skal være opmærksom på når man analyserer resultaterne.

Entydiggørelse af ord i kontekst
For at evaluere sprogmodellernes evne til at forstå og adskille betydninger af polyseme ord i en kontekst, genererer vi et såkaldt *Ord-i-Kontekst-datasæt* (af engelsk *Word-in-Context*, Pilehvar & Camacho-Collados 2019). I denne opgave får modellen et målord og to sætninger som begge er korpuseksempler der indeholder målordet. Opgaven går ud på at svare på om målordet er brugt med den samme betydning eller to forskellige betydninger i de to sætninger. Nedenfor ses et eksempel med to forskellige betydninger for målordet *kost*.

*a) Fødevarer, som indeholder fedt og sukker, er ikke sund **kost**.*
*b) Han kom fra en familie, hvor skænderier og drillerier var daglig **kost**.*

I sætning a) refererer *kost* til en konkret betydning af *mad/madvare*, mens sætning b) er en overført betydning i DDO og COR.SEM (*noget som man oplever, skal sætte sig ind i eller på anden måde forholde sig til*). For at kunne svare korrekt på om sætning a) og b) indeholder den samme eller to forskellige betydninger af *kost*, skal man altså være i stand til entydiggøre målordets betydning i hver sætning og sammenligne resultatet.

Vi baserer betydningerne i denne opgave på det i mange tilfælde forenklede betydningsinventar i COR.SEM sammenlignet med DDO, som beskrevet ovenfor. De brugseksempler der angives for størstedelen af COR.SEM's betydninger, anvendes som sætninger i vores Ord-i-Kontekst-datasæt. Vi opretter to versioner. Først laver vi

et rent monosemt datasæt, hvilket måske kan undre da der jo så kun kan være brugseksempler med samme betydning. Men de monoseme data gør det muligt at teste om en model er konsekvent ved entydiggørelse af et målord. Dertil burde de monoseme ord være nemmere at behandle for modellen end de polyseme da de ikke har den samme asymmetri mellem form og betydning. Denne 'lettere' opgave kræver med andre ord at man først entydiggør målordet, og et dårligt resultat vil for dette datasæt være et tegn på at modellen ikke løser opgaven efter hensigten. Dette skete netop for ChatGPT 3.5 Turbo, som konsekvent svarede at sætningerne i det monoseme datasæt indeholdt forskellige betydninger.

For de polyseme ord indsamler vi alle brugseksempler på tværs af betydninger for hvert målord og genererer samtlige mulige kombinationer af sætningspar. Da dette antal varierer både i forhold til antallet af målordets betydninger og i forhold til hvor mange brugseksempler hver betydning har, afgrænser vi antallet af eksempler til maksimalt tre for hver kategori (samme betydning og forskellig betydning).

Ved anvendelsen af ChatGPT 4.0 opnår vi henholdsvis 88 % svarkorrekthed for de monoseme data og 66 % for de polyseme.

Sentimentværdi af ord i kontekst

Vi evaluerer også sprogmodellernes fortolkning af ords såkaldte sentimentværdi, dvs. deres positive eller negative konnotation. I denne opgave giver vi et målord som input suppleret med et brugseksempel med målordet. Opgaven går ud på at svare på om målordet i den givne kontekst er brugt negativt eller positivt på en skala fra -3 (stærkt negativt) til +3 (stærkt positivt). Vi genererer datasættet automatisk baseret på COR.SEM ved at udtrække alle betydninger der både har sentimentværdi og mindst ét brugseksempel. Vi har i samme omgang også manuelt udvalgt en række neutrale betydninger som ikke har sentimentværdi i COR.SEM. Vi validerer manuelt alle eksempler (altså et målord med tilhørende brugseksempel og sentimentværdi) for at være sikre på at sentimentværdien kan identificeres af et menneske i brugseksemplet.

Nedenfor ses eksempler på målord (fed) i kontekst med målordets sentiment.

a) *Landboforeningerne øjner en historisk **chance** for at få afskaffet jordskatten.* (+2)
b) *Undervisningen i engelsk som tilvalgsfag **påbegyndes** i 2. semester.* (0)
c) *Danmark er blevet meget rost for at være én af de første nationer, der afskaffede **slaveriet**.* (-3)

En væsentlig distinktion i denne opgave udgøres af sentimentværdi på hhv. ordniveau og sætningsniveau. Et målord kan forekomme i en kontekst som i sin helhed har en anden sentimentværdi end målordet, fx som i sætning c) hvor målordet *slaveriet* har en sentimentværdi på -3 mens selve sætningen har en positiv konnotation. Netop de eksempler hvor målordets konnotation ikke svarer overens med konnotationen af den samlede kontekst, har givet problemer for modellerne i vores eksperimenter.

Svarene i denne opgave afviger fra svarene i de andre typer opgaver da vi her arbejder med en *skala*. Vi beregner derfor en såkaldt Pearsons korrelationskoefficient, dvs. den lineære sammenhæng mellem to variable. De to variable er i vores tilfælde det korrekte svar (sentimentværdien i COR.SEM) og modellernes output. ChatGPT 3.5 Turbo opnår med denne tilgang en koefficient på 69,6 %, mens ChatGPT 4.0 opnår en koefficient på 81,9 %, hvor det ideelle resultat altså er en koefficient på 100 %.

Inferens om begrebsmæssigt tilhørsforhold

Sprogforståelsesspørgsmål i forhold til begrebsmæssigt tilhørsforhold og semantiske relationer genererer vi automatisk ud fra DanNet-ontologien og de relationer som de ontologiske begreber har indkodet. Vi anvender i første omgang kun skabeloner af typer 'X er en Y', 'X er en slags Y', 'Man bruger X til Y', 'Man laver X ved at Y', 'X kan have en Y' og 'X er en del af Y'. Variablene udfyldes for forskellige ontologiske typer, både konkrete og abstrakte[15].

Figur 11 nedenfor viser eksempler på forskellige slags autogenererede forståelsesspørgsmål på basis af disse skabeloner der altså dels refererer

15 Begreber der tilhører konkrete ontologiske typer, er fx *hus* og *træ*, mens begreber der tilhører abstrakte ontologiske typer, kunne være *kærlighed* og *filosofi*.

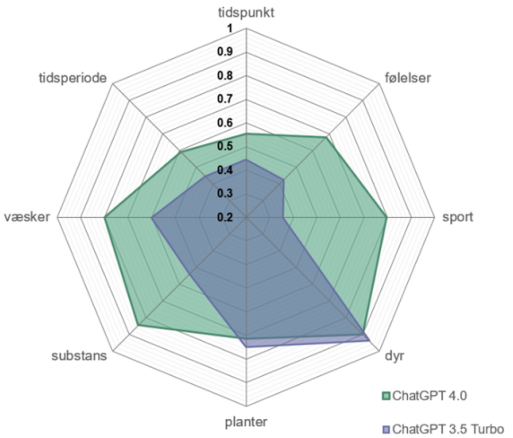
til hvilken overordnet ontologisk kategori et begreb tilhører, dels hvad noget typisk bruges til, hvordan det er blevet til, hvilke dele det har, og/eller hvad det selv er en del af. Udsagnene i del 1 bruges til at forberede chatbotten på spørgsmålstypen, hvor del 2 angives som et 'query' som chatbotten skal besvare som enten sandt eller falsk. Bemærk at der stilles både positive og negative spørgsmål (på en randomiseret måde), noget som sprogmodellerne traditionelt har haft lidt svært ved at skelne mellem.

FIGUR 11: AUTOGENEREREDE UDSAGN OM ONTOLOGISKE TYPER OG SEMANTISKE RELATIONER.

typer	Udsagn del 1	Udsagn del 2 (query)
Feeling; Creature	*Sympati er en følelse; Tryghed er en følelse*	*Q: Spøgelse er en følelse* (falsk)
Liquid; Quantity	*En bouillon er en væske; En drik er en væske*	*Q: En slurk er ikke en væske* (sandt)
Semiotic; Artifact	*Man laver en roman ved at skrive den; Man laver et essay ved at skrive det*	*Q: Man laver ikke en hat ved at skrive den* (sandt)
Food; Liquid	*Man laver et tog ved at fremstille det; Man laver en ret ved at tilberede den*	*Q: Man laver te ved at pochere den* (falsk)
Garment; Artifact	*Man tager en frakke på for at holde sig varm; Man tager en hue på for at holde sig varm*	*Q: Man tager en ring på for at holde sig varm* (falsk)
Instrument	*Man bruger en kniv til at skære med; Man bruger en hammer til at hamre med*	*Q: Man bruger ikke et rivejern til at rive med* (falsk)
BodyPart; Part	*En hånd kan ikke have et øje; Et ansigt kan have en mund*	*Q: Et fly kan have en propel* (sandt)

De autogenererede udsagn kan af og til virke pudsige, særligt når flertydige ord er i spil. Fx vil de færreste ved første øjesyn mene at en kylling er en væske, men ordet har faktisk den betydning om (indholdet i) en lille flaske spiritus. Vi overvejer i hvor høj grad vi i kommende eksperimenter manuelt bør luge ud i sådanne pudsigheder og mere sjældne betydninger, eller om de bør være med netop for at dække ordforrådet i alle dets afskygninger.

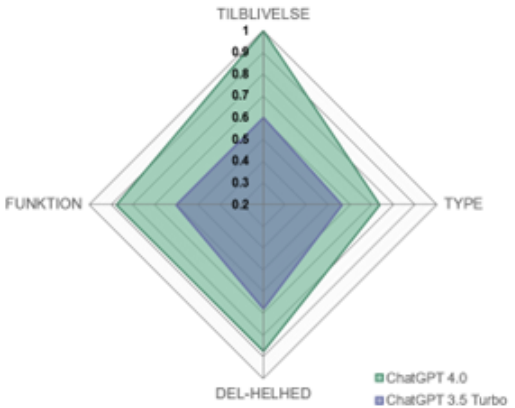
FIGUR 12: SPROGMODELLERNES (CHATGPT3.5 TURBO (BLÅ) OG CHATGPT4S (GRØN)) FORSTÅELSE AF HVILKE BEGREBER DER TILHØRER HVILKE ONTOLOGISKE TYPER.



En kørsel med datasættet på hhv. ChatGPT3.5 Turbo (blå) og ChatGPT4 (grøn) viser i figur 12 ikke så overraskende at forståelsen overordnet set er bedst for mere konkrete begreber som dyr og planter, hvorimod tidslige enheder og abstrakte typer som følelser er vanskeligere for modellerne at håndtere. Selv om ChatGPT3.5 Turbo fungerer en anelse bedre på dyr og planter, så ses der generelt en meget stor forbedring fra ChatGPT3.5 Turbo til ChatGPT4.

Figur 13 angiver hvordan de forskellige semantiske relationer (inkl. ontologisk type) forstås af de samme to sprogmodeller (ChatGPT.5 Turbo (blå) og ChatGPT4 (grøn)). Her ses generelt en god forståelse af hvordan noget er blevet til, hvorimod de andre semantiske relationer er lidt vanskeligere – bl.a. pga. af de mere abstrakte ontologiske typer som trækker resultatet ned. Igen opnår den nyeste af de to modeller de bedste resultater.

FIGUR 13: CHATGPT3.5 TURBOS (BLÅ) OG CHATGPT4S (GRØN) FORSTÅELSE AF SEMANTISKE RELATIONER.



Følgeslutninger vedr. handlinger og hændelser

For at afprøve modellernes evne til at foretage følgeslutninger efter at forskellige hændelser og handlinger har fundet sted, udarbejder vi for hver frametype i FrameNet-leksikonet to standardsætninger. Den ene eksemplificerer en handling/hændelse som et givent verbum eller verbalsubstantiv udløser, og den anden angiver resultatet og fremsættes som et 'query' som er enten sandt eller falsk. Standardsætningerne anvendes til at autogenerere udsagn med de verber og verbalsubstantiver der falder inden for den samme ramme. Igen blandes positive og negative udsagn på en randomiseret måde for at teste sprogmodellernes evne til at håndtere negation.

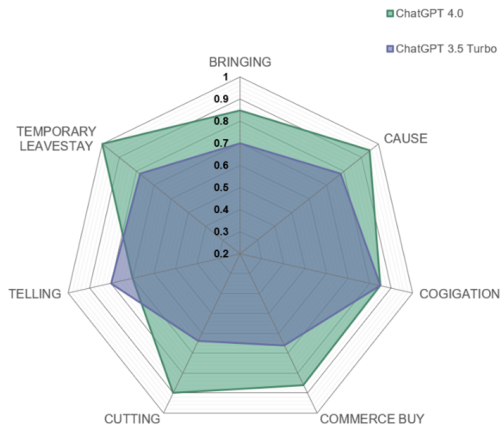
FIGUR 14: UDSAGN FOR FORSKELLIGE FRAMES. ANDET UDSAGN INDIKERER RESULTATET AF HANDLINGEN/HÆNDELSEN OG FREMSÆTTES SOM ET 'QUERY' (SANDT ELLER FALSK).

FRAME	UDSAGN
BRINGING (bringe)	*Peter bringer mad ud til Pia* *Q: Pia har nu mad.* (sandt)
CAUSE (forårsage)	*Eksplosionen medførte svære skader på bygningen.* *Q: Efter eksplosionen var bygningen ubeskadiget* (falsk)
COMMERCE_BUY (handel_køb)	*Peter købte bogen af Anne* *Q: Nu ejer Anne bogen* (falsk)
CUTTING (skære)	*Pia klipper rebet over* *Q: Pia har nu to kortere reb* (sandt)
TELLING (meddele)	*Peter fortalte Pia om forlovelsen* *Q: Nu kender Pia ikke til forlovelsen* (falsk)
TEMPORARY_LEAVESTAY (midlertidigt fravær/ophold)	*Per tog på ferie i uge 7* *Q: Per var ikke på arbejde i uge 7* (sandt)
COGITATION (overvejelse)	*Peter tænkte på Hanne hele tiden* *Q: Hanne var aldrig i Peters tanker* (falsk)

Hvis vi afprøver vores datasæt for følgeslutninger på de samme to sprogmodeller som før, ser vi en generelt god forståelse særligt ved anvendelsen af ChatGPT4 (grøn) som er overraskende god til at tackle 'queries' vedr. konkrete handlinger og hændelser. Kommunikationshandlinger (som fx i tilfældet Telling) er den type der volder størst problemer, hvilket ikke undrer. Disse er mindre entydige hvad resultat angår; fx er et budskab ikke nødvendigvis forstået selv om det er afsendt (se figur 15)[16].

16 Vi har i opgaven slået to frames sammen: Temporary_Leave og Temporary_Stay til TemporaryLeaveStay. Framen Temporary_Leave beskriver verbale udtryk som fx *tage orlov* og *skulke* mens Temporary_Stay be-skriver ord som *ophold*, *gøre holdt* og *gå i hi*. Framen Cogitation dækker over udtryk for tankevirksomhed såsom *tænke* og *analysere*, men også *skumle* og *dagdrømme*.

FIGUR 15: CHATGPT3.5 TURBOS (BLÅ) OG CHATGPT4'S (GRØN) HÅNDTERING AF FØLGESLUTNINGER VED HANDLINGER OG HÆNDELSER.



5 SAMMENLIGNING AF DE SYV SPROGFORSTÅELSES-OPGAVER

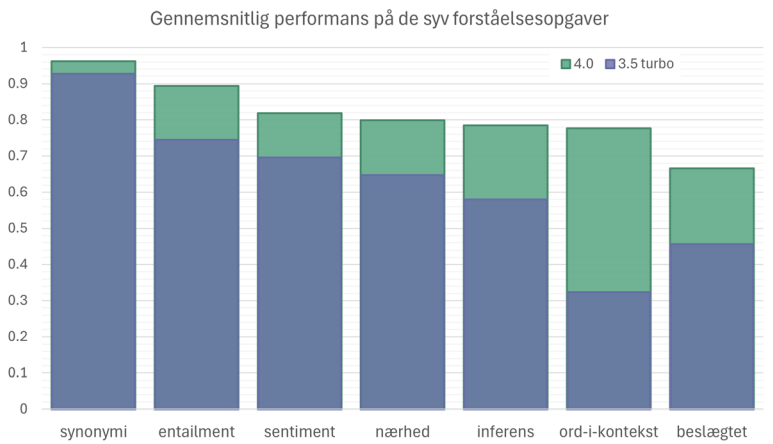
Figur 16 viser hvordan ChatGPT har præsteret gennemsnitligt for hver af de syv opgaver. Synonymi er den opgave som modellerne klarer bedst. Dette er ikke så overraskende da netop semantisk nærhed er noget af det som modellerne er særligt skarpe til at håndtere via de såkaldte word embeddings. Modellerne klarer dog opgaven så godt at vi skal videreudvikle vores testsæt i forhold til sværhedsgraden for reelt at kunne evaluere præstationen. Fx kan man tilføje flere kandidater som semantisk er tættere på synonymiparret i listen af muligheder eller tilføje valør til opgaven således at det korrekte synonym også skal matche valøren i DDO (fx uformelt eller nedsættende).

Det kan umiddelbart overraske at beslægtethed er så relativt svær en opgave for modellerne at løse, men det skyldes at det er svært at skelne imellem om noget ligner hinanden semantisk eller blot er associativt beslægtet. Hvor *kop* og *kaffe* er associativt beslægtet, dvs. nært relateret, er *kaffe* og *te* semantisk tættere på hinanden idet begge refererer til væsker vi indtager, og dermed har samme ontologiske type og samme overbegreb. De sproglige, dvs. de ordmæssige og ordfølgemæssige kontekster for *kop* og *kaffe* er imidlertid næsten de samme som for *kaffe* og *te* (i fx et tekstkorpus). Dette vanskeliggør opgaven for modellerne.

Vi ser også at forståelse af ord i kontekst er en af de svære opgaver for modellerne at løse, også selv om det betydningsinventar der evalueres op imod, er relativt grovkornet (fra COR.SEM). Her er det relevant at nævne at også vi mennesker generelt er uenige i mange tilfælde når vi skal opmærke løbende tekst med ordbetydning (jf. fx Pedersen et al. 2016 om semantisk opmærkning på dansk).

ChatGPT 4 klarer til gengæld inferensopgaven og særligt opgaven om følgeslutninger (kaldet entailment i figur 16) overraskende godt. Også her må vi tænke i mere udfordrende opgaver i fremtidige datasæt.

FIGUR 16: DE SYV FORSTÅELSESOPGAVER AFPRØVET PÅ TO VERSIONER AF CHATGPT.



6 KONKLUSION OG VIDERE PERSPEKTIVER

Denne artikel beskriver de første forsøg med at udvikle et benchmark i stor skala som kan bruges til at evaluere danske sprogmodellers evne til at håndtere leksikalsk-semantiske forhold i dansk sprog. Udgangspunktet har været at udnytte de data som vi allerede har tilgængelige via de semantiske ordbøger vi har udviklet gennem flere år, og at betragte dem som vores 'ground truth' eller guldstandard. Vores eksperimenter har vist at metoden er realiserbar, og at de benchmarkdatasæt der udvikles, rent faktisk udfordrer de nyeste sprogmodeller på et passede niveau. Vi planlægger at skalere op så en større del af ordforrådet dækkes, og

så vi opnår en større variation og dermed også en højere sværhedsgrad i de stillede opgaver. På sigt vil vi også gerne kunne evaluere hvor godt sprogmodellerne fortolker mere avancerede og subtile dele af det danske ordforråd, fx metaforer og overført sprog. Vi antager at det især er inden for disse områder at sprogtransfer fra engelsk slår negativt igennem på dansk og genererer upræcist eller forvrøvlet sprog, jf. eksemplerne fra afsnit 1 om *flødebolle* og *æbleskive*. Hvor godt forstår modellerne fx sammenhængen mellem et ords grundbetydning og dets metaforiske søsterbetydning sådan som vi mennesker intuitivt gør når vi fx siger at noget er en *en følelsesmæssig rutsjebane*, eller at noget er *et vindue til verden*? Til dette formål ønsker vi at udnytte data fra Den Danske Ordbog, hvor mindst 6000 lemmaer er beskrevet med dels deres grundbetydning, dels deres afledte metaforiske betydning.

Vores forsøg har imidlertid også givet os anledning til at stille en række metodiske spørgsmål som det vil være interessant at efterforske i det videre arbejde. Ét spørgsmål vedrører i hvor høj grad vores benchmarks i deres nuværende form svarer til hvad almindelige mennesker ville svare på når de bliver stillet over for samme sprogforståelsesopgaver. Konkurrerer sprogmodellerne her i virkeligheden mod 'supermennesker' som præsterer langt bedre end den almindelig sprogbruger? Dette spørgsmål drøftes bl.a. i Tedeschi et al. (2023), og vi overvejer at undersøge det ved at lave nogle stikprøver med en række informanter der ikke er lingvistisk eller leksikografisk skolede.

Et andet spørgsmål som kunne være interessant at undersøge nærmere, vedrører de monolingvalt genererede benchmarks' påståede suverænitet overfor oversatte benchmarks. Her må der tænkes lidt over hvordan vi rent faktisk kan afprøve om vores benchmarks er de oversatte benchmarks overlegne. Vi forestiller os at en grundig analyse af evalueringsresultater baseret på fx oversatte sentimentdata sammenlignet med resultater baseret på vores egne monolingvalt genererede sentimentdata kan være med til at belyse dette.

Endelig kan man stille sig selv det spørgsmål om evnen til at løse generelle sprogforståelsesopgaver modsvares af evnen til at løse praktiske sprogtjenesteopgaver a la resumeringsopgaver, spørgsmål-svar-opgaver eller maskinoversættelse. I og med at vi stadig mangler at forstå i dybden hvordan de meget komplekse modeller arbejder, og hvilket abstrak-

tionsniveau de egentlig opnår, er det ikke givet at der er en fuldstændig overensstemmelse mellem praktisk anvendelighed og evne til at klare sprogforståelsesopgaver. En model løser måske ikke en sprogforståelsesopgave særlig godt, men kan sagtens lave et meningsfuldt resumé af en tekst, fx, eller vurdere til perfektion om en anmeldelse er positiv eller negativ. Hele dette spørgsmål vil vi gerne undersøge nærmere på længere sigt, og en af vejene til det er at koble vores datasæt med andre mere taskdrevne datasæt udviklet fx i danske virksomheder.

TAK
Tak til Simon Gray, Center for Sprogteknologi, Institut for Nordiske Studier og Sprogvidenskab, KU, for at udvikle automatiseringsrutiner til DanNet-udtræk og udformning af udsagn til inferensdatasættet.

Bolette Sandford Pedersen, professor
Center for Sprogteknologi, Institut for Nordiske Studer og Sprogvidenskab
Københavns Universitet
bspedersen@hum.ku.dk

Nathalie C. Hau Sørensen, assisterende redaktør
Det Danske Sprog- og Litteraturselskab
nats@dsl.dk

Sussi Olsen, videnskabelig medarbejder
Center for Sprogteknologi, Institut for Nordiske Studer og Sprogvidenskab
Københavns Universitet
saolsen@hum.ku.dk

Sanni Nimb, seniorredaktør
Det Danske Sprog- og Litteraturselskab
sn@dsl.dk

LITTERATUR

Berdicevskis, A. et al. 2023. Superlim: A Swedish language understanding evaluation benchmark. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8137–8153. Singapore: Association for Computational Linguistics. DOI:10.18653/v1/2023.emnlp-main.506.

Firth, J.R. 1957. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*. Oxford: Blackwell.

Henrichsen, P.J. 2021. Det Centrale OrdRegister, Del 1. Oplæg ved Sprogteknologisk Konference 2021. Københavns Universitet. https://cst.ku.dk/kalender/sprogteknologisk-konference-2021/

Hershcovich, D. et al. 2022. Challenges and strategies in cross-cultural NLP. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long papers), 6997–7013. Dublin: Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.482.

Hjorth E. & K. Kristensen (red.). 2003–2005. *Den Danske Ordbog*. København: Det Danske Sprog- og Litteraturselskab & Gyldendal. https://ordnet.dk/ddo.

Lee T. & S. Trott. 2023. A jargon-free explanation of how AI large language models work. *ArsTechnica*. https://arstechnica.com/science/2023/07/a-jargon-free-explanation-of-how-ai-large-language-models-work/

Lenci, A. & M. Sahlgren. 2023. *Distributional semantics*. Cambridge: Cambridge University Press. DOI: 10.1017/9780511783692.

Levin, B. 1991. *English verb classes and alternations – a preliminary investigation*. Denver: University of Colorado Press.

Nielsen, D.S. 2023. ScandEval: A benchmark for Scandinavian natural language processing. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 185–201, Tórshavn: University of Tartu Library. https://aclanthology.org/2023.nodalida-1.20

Nielsen, F.Å. & L.K. Hansen. 2017. Open semantic analysis: The case of word level semantics in Danish. *Human language technologies. Challenges for computer science and linguistics*, 415–419. Mannheim: Springer.

Nimb, S. et al. 2014. *Den Danske Begrebsordbog*. Odense: Det Danske Sprog- og Litteraturselskab & Syddansk Universitetsforlag.

Nimb, S. 2016. Der er ikke langt fra tanke til handling. *Danske Studier 2016*, 25–59.

Nimb, S. et al. 2017. From thesaurus to framenet. Electronic lexicography in the 21st century. *Proceedings of eLex 2017*, 1–22. Brno: Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper01.pdf.

Nimb, S., N.H. Sørensen & T. Troelsgård. 2018. From standalone thesaurus to integrated related words in the Danish dictionary. *Proceedings of the XVIII EURALEX International Congress: lexicography in global contexts*, 916–923. Ljubljana: Ljubljana University Press, Faculty of Arts. https://euralex.org/publications/from-standalone-thesaurus-to-integrated-related-words-in-the-danish-dictionary/.

Nimb, S. et al. 2022a. COR-S – den semantiske del af Det Centrale OrdRegister (COR). *LexicoNordica* 29. https://tidsskrift.dk/lexn/article/view/134776.

Nimb, S. et al. 2022b. A thesaurus-based sentiment lexicon for Danish: The Danish Sentiment Lexicon. *Proceedings of the 13th Language Resources and Evaluation Conference (LREC2022),* 2826–2832. Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.302.

Pedersen, B.S. et al. 2009. DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language resources and evaluation* 43. 269–299. DOI: 10.1007/s10579-009-9092-1.

Pedersen, B.S. et al. 2016. The Semdax corpus. Sense annotations with scalable sense inventories. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 842–847. Paris: European Language Resources Association. https://aclanthology.org/L16-1136.

Pedersen, B.S., S. Nimb & S. Olsen. 2021. Dansk betydningsinventar i et datalingvistisk perspektiv. *Danske Studier 2021*. 72–106. https://danskestudier.dk/wp-content/uploads/2021/07/danske-studier-2021.pdf.

Pedersen, B.S. et al. 2022. Compiling a suitable level of sense granularity in a lexicon for AI purposes: the open source COR-Lexicon. *Proceedings of the 13th Language Resources and Evaluation Conference (LREC2022)*, 51–60. Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.6.

Pedersen, B.S. et al. 2023. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European languages. *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, 11–18. Tórshavn: Association for Computational Linguistics. https://aclanthology.org/2023.resourceful-1.2.

Pedersen, B.S. et al. 2024. Towards a Danish semantic reasoning benchmark - compiled from lexical-semantic resources for assessing selected language understanding capabilities of large language models. *Proceedings of the Fourteenth Language Resources and Evaluation Conference (LREC24)*. Torino: European Language Resources Association.

Pilehvar, M.T. & J. Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, vol. 1*, 1267–1273. Minneapolis: Association for Computational Linguistics. DOI: 10.18653/v1/N19-1128.

Pustejovsky, J. 1998. *The generative lexicon*. Cambridge: The MIT Press. DOI: 10.7551/mitpress/3225.001.0001

Samuel, D. et al. 2023. NorBench – a benchmark for Norwegian language models. *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, 618–633. Tórshavn: University of Tartu Library. https://aclanthology.org/2023.nodalida-1.61.

Schneidermann, N.S., R. Hvingelby & B.S. Pedersen. 2020. Towards a gold standard for evaluating Danish word embeddings. *Proceedings of the 12th Language Resources and Evaluation Conference (LREC2020)*, 4754–4763. Marseille: European Language Resources Association. https://aclanthology.org/2020.lrec-1.585/.

Tedeschi, S. et al. 2023. What's the meaning of superhuman performance in today's NLU? *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (vol. 1: long papers), 12471–12491, Toronto: Association for Computational Linguistics.

Vossen, P. (red.). 1999. *EuroWordNet, A multilingual database with lexical semantic networks*. Amsterdam: Kluwer Academic Publishers. DOI: 10.1017/S1351324903223299

Wang, A. et al. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*, 353–355. Brussels: Association for Computational Linguistics. DOI: 10.18653/v1/W18-5446.

Wang, A. et al. 2020. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems 32* (Proceedings, *NeurIPS 2019*). Vancouver. https://proceedings.neurips.cc/paper_files/paper/2019/hash/4496bf24afe7fab6f046bf4923da8de6-Abstract.html.