Sprogteknologiske ressourcer til informationssøgning

Af Bolette Sandford Pedersen, Patrizia Paggio og Costanza Navarretta

Abstract

Formålet med denne artikel er at præsentere et antal sprogteknologiske ressourcer for dansk i form af ordbaser, ontologier og wordnets for derigennem forhåbentlig at medvirke til at skabe bedre synergi med den beslægtede forskning inden for indholdsbaseret informationssøgning der pågår i biblioteksverdenen. De beskrevne sprogressourcer har alle i større eller mindre grad været anvendt til at eksperimentere med indholdsbaseret informationssøgning, og flere af komponenterne indgår i dag i kommercielle løsninger. I artiklen beskriver vi således også en række eksperimenter hvor anvendelse af sprogteknologiske ressourcer har ført til udvidet funktionalitet eller forbedrede søgeresultater. Dernæst præsenterer vi et dansk wordnet som er under udvikling ved Center for Sprogteknologi og Det Danske Sprog- og Litteraturselskab. Vi diskuterer de mest interessante og udfordrende aspekter ved wordnettet, og vi drøfter perspektiverne i denne store satsning.

Bolette Sandford Pedersen er seniorforsker, bolette@cst.dk;
Patrizia Paggio er seniorforsker og Ph.d., patrizia@cst.dk;
Costanza Navarretta er seniorforsker, costanza@cst.dk - alle ved Center for Sprogteknologi på Københavns Universitet

1. Indledning

Spørgsmålet om hvorvidt sprogteknologiske metoder kan forbedre informationssøgning er blevet undersøgt i en del eksperimenter igennem de sidste fire årtier. Men det er specielt i 90'erne i forbindelse med TREC-konferencerne[1] at interessen for emnet er steget. De første eksperimenter var ikke særlig overbevisende (for en opsummering, se Smeaton, 1997), men bedre resultater kunne påvises efterhånden (Strzalkowski et al., 1997). I 1999 satte man i TREC igen fokus på sprogteknologi i forbindelse med en ny applikation, den såkaldte Question Answering, hvor man opererer med spørgsmål i stedet for enkelte ord i brugerforespørgslerne. Det bedste system til at finde relevante tekstuddrag anvendte rent faktisk navnegenkendelse og en enkel form for syntaktisk analyse (Voorhees & Tice, 2000).

Hvor TREC har været med til at etablere en vis konsensus om anvendeligheden af relativt enkle sproglige manipulationer der bruger morfologisk og noget syntaktisk viden (hvad denne viden består af, vil vi forklare i de næste afsnit), har andre forskere eksperimenteret med semantiske oplysninger bl.a. i form af wordnets – ressourcer hvor ord er sat i relation til hinanden via deres betydning i lighed med tesaurusser. De første undersøgelser (Voorhees, 1994; Smeaton & Quigley, 1996; Gonzalos et al., 1998) syntes ikke særlig lovende, men interessen for at bruge wordnets til informationssøgning er alligevel ikke dalet. Tværtimod har mange forsøgt med andre indfaldsvinkler til problemet også i takt med at res-

sourcerne er blevet større og bedre, samtidig med at wordnets er blevet tilgængelige for forskellige sprog, og de sprogteknologiske værktøjer er blevet bedre til fx at genkende relevante læsninger af tvetydige ord. Mihalcea et al. (2000) opnår således forbedringer i både recall og precision ved at anvende semantisk viden til query expansion såvel som dokumentindeksering. Lignende forbedringer påvises i Magnini & Strapparava (2001), der anvender et wordnet til at finde relevante dokumenter i forhold til en automatisk opbygget brugermodel. Og nye applikationer dukker op: Clough & Stevenson (2004) beskriver således hvordan et wordnet kan bruges til at entydiggøre en forespørgsel inden denne bliver oversat med henblik på tværsproglig søgning.

På Center for Sprogteknologi (CST) ved Københavns Universitet har vi også igennem en årrække forsket i hvordan informationssøgning kan forbedres ved hjælp af sprogteknologiske ressourcer; her har fokus været på dansksprogede tekster. I denne artikel tager vi udgangspunkt i de ressourcer CST har været med til at udvikle for dansk, og som vi har anvendt i flere forskellige forskningsprojekter. Arbejdet er foregået dels i national sammenhæng som et samarbejde mellem flere danske forskningsinstitutioner og virksomheder, dels i international sammenhæng relateret til den nyeste forskning inden for tværsproglig søgning i det semantiske web. Projekterne er allerede hver især blevet beskrevet i videnskabelige artikler hvor vi fokuserer på søgningsfunktionaliteten, og på hvordan brug af sproglig viden og sprogteknologiske metoder kan forbedre den. Formålet med præsentationen her er primært at sprede viden om de danske sprogteknologiske ressourcer og derved bl.a. skabe bedre mulighed for informationsudveksling med den forskning der pågår i biblioteksverdenen inden for informationssøgning. På længere sigt forestiller vi os at flere af de sprogressourcer CST og andre sproginstitutioner, som fx Det Danske Sprog- og Litteraturselskab, kan stille til rådighed, vil komme til at spille en ikke uvæsentlig rolle i fremtidige søgeløsninger.

Selv om vores fokus er på ressourcerne, forklarer vi også hvordan disse anvendes i sprogteknologiske værktøjer, og vi skitserer kort hvordan dette er sket i forbindelse med en række forskningsprojekter. Vi beskriver også med en vis detaljeringsgrad to eksperimenter hvor brugen af ressourcerne til informationssøgning er blevet evalueret, og hvor specielt recall bliver forbedret i forhold til ren keywords-baseret søgning.

Vi indleder (afsnit 2) med en gennemgang af Den danske SprogTeknologiske Ordbase, STO, som med sine morfologiske og syntaktiske oplysninger danner baggrund for en lang række af de sprogteknologiske værktøjer som vi arbejder med og anvender i vores søgeapplikationer, så som lemmatiser, tagger og navnegenkender. Vi ser også nærmere på et specifikt søgeeksperiment med sammensætninger i dansk foretaget i samarbejde med en dansk virksomhed.

Særligt fokus i artiklen har dog de *ontologiske* ressourcer som generelt gennemgår en rivende udvikling i disse år, ikke mindst på grund af teknologien inden for det semantiske web. I afsnit 3 indleder vi med en introduktion til ontologi, og vi bevæger os derfra videre til topic maps, lingvistiske ontologier og wordnets. Vi gennemgår de væsentligste konklusioner fra de nationale og internationale projekter vi har deltaget i på området, og med baggrund heri argumenterer vi for hvorfor vi nu har valgt at satse på et storstilet wordnet-projekt for dansk med genbrug af eksisterende ordbogsdefinitioner fra Den Danske Ordbog.

Endelig beskriver vi i afsnit 4 rammerne for det igangværende wordnet-projekt, DanNet, som foregår i samarbejde med DSL. Med en række eksempler fra udviklingsarbejdet illustrerer vi de problemstillinger der opstår, når ens begrebsmæssige udgangspunkt består af korpusuddragne sproglige data og ikke af et på forhånd velstruktureret ontologisk skelet. Vi skitserer også hvilke semantiske oplysningstyper der synes relevante for informationssøgning inden for ikke bare de konkrete entiteter, men også hændelser, handlinger, egenskaber og abstrakte genstande.

2. Ordbogsressourcer og korpora

De mest grundlæggende ressourcer til sprogteknologiske applikationer som anvendes til at behandle tekster, er ordbogsressourcer og opmærkede korpora. Sprogteknologiske ordbogsressourcer indeholder detaljerede og strukturerede oplysninger om ord som direkte kan bruges af computerprogrammer. Den danske SprogTeknologiske Ordbase, STO, er den største database der indeholder leksikalske oplysninger om danske ord, og som er kodet på en sådan formaliseret måde (Braasch & Olsen, 2004). STO

adskiller sig derfor fra elektroniske leksikalske ressourcer som er tiltænkt mennesker som direkte brugere. STO-ordbasens struktur er modulær og følger internationelle principper for opbygning af leksikalske ordbaser[2]. Ifølge disse principper inddeles oplysninger om ord i tre forskellige beskrivelsesniveauer: ordenes form (morfologi), ordenes konstruktionsmuligheder (syntaks) og deres betydning (semantik), i praksis svarende til tre moduler. Al information om et ord er bundet til selve ordets grundform (lemma). Forskellige applikationer kan anvende forskellige typer oplysninger inden for samme modul eller på tværs af modulerne.

Ordbasens ordforråd stammer hovedsagelig fra almensprog (ca. 68.000 ord) og er fortrinsvis baseret på et aviskorpus. Udvælgelsen af ord er sket ud fra deres frekvens i korpusset, således at kun de mest frekvente ord er blevet medtaget i ordbogen. Det fagsproglige ordforråd (ca. 13.500 ord) dækker seks emneområder: edb, finans, forvaltning, handel/erhverv, miljø og sundhed/helse, og her er valget af ordforråd baseret på domænespecifikke korpora som er blevet indsamlet til formålet. Selv om ordforrådet i STO er udvalgt efter ordenes frekvens i diverse korpora, er frekvens ikke medtaget som en oplysningstype i basen. Ordfrekvens i tekster, både den generelle frekvens i almensprog og frekvensen i bestemte domæner, er en central oplysning for indekserings-, klassifikations- og søgeapplikationer, se bl.a. Salton & McGill (1983) og Manning & Schütze (1998). Derfor arbejdes der nu på at tilføje ord- og bøjningsfrekvens i STO. Frekvensoplysningerne bliver i første omgang uddraget fra Korpus2000 og Korpus1990, da disse er de største balancerede almensproglige korpora for dansk (Andersen et al., 2002). Frekvensangivelsen i STO vil således kunne bruges til bl.a. at tilskrive varierende relevansvægt til ord og ordformer, men også til at spotte om ord i domænespecifikke tekster afviger i frekvens fra deres brug i almensproglige tekster.

2.1 Ordenes form og grammatiske egenskaber
Alle ord i STO (ca. 81.500) er tilknyttet morfologiske oplysninger. Ca. halvdelen af ordene er i øvrigt forsynet med syntaktiske oplysninger (og af disse er ca. 10.000 ord yderligere forsynet med semantiske oplysninger, jf. 3.3). Det morfologiske og det syntaktiske beskrivelsesniveau omfatter følgende oplysningstyper:

- Morfologi: ordklasse, bøjning, stavevarianter ifølge Retskrivningsordbogen, samt, hvor det er relevant, oplysninger vedrørende sammensætning.
- Syntaks: ordets konstruktionsmuligheder inkl. dets valens, dvs. hvor mange sætningsled ordet kan optræde sammen med, samt andre specifikke oplysninger om det enkelte ord, fx bundne partikler.

Ordklasseoplysningen er nyttig, da ordklasserne har forskellige karakteristika. Der skelnes mellem to overordnede typer af ordklasser: funktionsord og indholdsord. Funktionsord dækker ord som har den funktion at forbinde andre sproglige enheder (ord eller større sætningsled). De inkluderer bl.a. præpositioner (forholdsord) og konjunktioner (bindeord). Indholdsord er ord som refererer til bestemte objekter eller entiteter, og inkluderer ordklasser som substantiver (navneord), verber (udsagnsord) og adjektiver (tillægsord). I sætningen *Jeg tænker på sommerferien* har præpositionen *på* kun en funktion i forhold til verbet *tænke*, mens verbet *tænke* og substantivet *sommerferien* i sig selv er betydningsbærende.

Endelig gives der i STO-ordbasen information om hvorvidt et substantiv tager et fugeelement (og i givet fald hvilket) når det indgår som første element i en sammensætning. Fx har substantivet land flere muligheder som det ses i hhv. *landmand*, *land***e***grænse* og *land***s***mand*, mens substantivet *jord* ikke tager fugeelement (*jordlod*). Sådanne oplysninger kan medvirke til automatisk behandling af sammensætninger (afsnit 2.4) og er vigtige idet nye ordsammensætninger opstår hele tiden og derfor ikke kan behandles ved at liste alle mulige forekomster i ordbasen.

2.2 Opmærkede korpora
Ud over sprogteknologiske ordbogsressourcer er opmærkede korpora (elektroniske tekstsamlinger forsynet med især lingvistiske oplysninger) vigtige til at træne værktøjer der automatisk skal genkende og opmærke ukendt tekst (jf. næste afsnit). Denne proces, hvor computeren lærer at behandle nyt materiale ud fra eksisterende eksempler, hedder maskinlæring. Korpora der skal anvendes til maskinlæring skal imidlertid have en vis størrelse og være manuelt valideret hvis opmærkningen er foregået automatisk. Det eneste større danske korpus (ca. 250.000 løbende ord) som er blevet opmærket med morfologiske og

syntaktiske oplysninger og derefter manuelt valideret, er PAROLE-Korpusset (Keson, 1996; Kromann, 2003).

2.3 Anvendelser i informationssøgning

De grundlæggende værktøjer som er nyttige til teksthåndtering, udnytter dels den information der foreligger i de opmærkede korpora, fx PAROLE-Korpusset, dels den der er indkodet i ordbogs-ressourcer som STO (Hansen, 2006). De mest umiddelbart relevante ordbogsoplysninger inden for indeksering og søgning er de morfologiske, som gør det muligt for et system fx at genkende entals- og flertalsformer af det samme ord. Syntaktisk information, først og fremmest ordklasseoplysning, er også vigtig idet det tillader systemet at fokusere på bestemte grupper af ord, fx alle substantiverne i en tekst. Dette kan være relevant i forbindelse med dokumentindeksering. Mere kompleks syntaktisk information kan bruges i systemer der kræver at man genkender sætningsled (grundled, genstandsled mv.), fx informationssystemer der automatisk skal finde alle forekomster af en bestemt type scenarium (fx alle de firmaer der har fusioneret i en bestem periode, samt aktørerne i fusionerne). Endelig er syntaks med til at fjerne tvetydighed når man skal slå ordene op i en ordbase. Fx kan man ud fra den syntaktiske sammenhæng skelne mellem de to forskellige læsninger af verbet *binde* i *binde noget sammen* og *binde noget ind*. Herunder vil vi dog kun fokusere på værktøjer der gør brug af ikke alt for komplekse sproglige oplysninger.

Til indeksering og søgning er det ofte meget relevant at finde og kategorisere navne, som fx firmanavne, personnavne og geografiske steder, som ofte betegner centrale objekter og aktører i teksterne. Fx kan navnegenkendere identificere og opmærke datoer, beløb, komplekse navne (fornavn~mellemnavn~efternavn) som dernæst behandles som simple tokens. Således er det muligt automatisk at udlede at de to substantiver i *Nørrebro Teater* danner navnet på en institution og ikke bør fortolkes som to objekter, dvs. som navnet på en bydel i København og substantivet teater. Programmer der (bl.a. med anvendelse af ordbøger og korpora) identificerer komplekse navne ud fra bestemte mønstre og klassificerer dem ifølge prædefinerede semantiske kategorier, kaldes navnegenkendere. Navnegenkendere for dansk beskrives nærmere i Hansen (2003) og Bick (2004).

En anden opgave hvor leksikalske ressourcer spiller en vigtig rolle, er automatisk opmærkning af ordklasser og evt. af ordenes bøjningsform (fx at et bestemt ord er et substantiv, bøjet i flertal), som foretages ved hjælp af programmer kaldet PoS (Part of Speech)-taggere. Taggere trænes ud fra opmærkede korpora og i nogle tilfælde også ud fra sprogteknologiske ordbaser hvis man vil opnå større precision. Ordklasseoplysningerne i STO bruges til at genkende ordklassen for de ord som er kodet i databasen, men også til at træne taggere der skal kunne kategorisere ukendte ord. Dette er ikke nødvendigvis en nem opgave idet en ordform kan tilhøre forskellige ordklasser afhængig af konteksten. En tagger arbejder typisk ved at undersøge to eller tre ord før og efter det ord der skal opmærkes, og fastlægger derudfra ordklassen. Som vi allerede har påpeget, er en tagger et udmærket redskab til at finde bestemte grupper ord med henblik på dokumentindeksering, fx til automatisk at frasortere funktionsord og udelukkende behandle betydningsbærende ord. Det skal bemærkes at en tagger skal trænes specielt til den type tekster den skal arbejde på. Oplysninger i STO og PAROLE-korpusset er også blevet anvendt til at træne en tagger til opmærkning af ældre danske tekster (Maegaard et al., 2006).

En anden proces der er relevant til informationssøgning, og som gør brug af morfologiske oplysninger, handler om at genkende forskellige former af det samme ord (fx *barnebarn* og *børnebørnenes*) i løbende tekst eller i forespørgsler. I avancerede søgemaskiner er der i dag inkluderet viden om ordenes bøjningsformer, eller programmer som på anden vis kan tage højde for ordformen. På nogle sprog, fx engelsk, fjerner man under søgning typiske endelser fx *-er*, *-ing* eller *-s*. Således kan man automatisk finde ud af at ord som *walk*, *walking* og *walker* er relaterede. Processen kaldes *stemming*. Stemming skal selvfølgelig tilpasses for ord der har uregelmæssige bøjningsformer (*go, went*) men på trods af denne tilpasning leverer den ikke altid ordets korrekte grundform (lemma). Alligevel er stemming nyttig for sprog med simpel og regelmæssig morfologi hvis man ikke har passende sproglige ressourcer til rådighed. For et sprog som dansk virker stemming ikke tilfredsstillende (Pedersen et al., 2005; Navarretta et al., 2006). Dette skyldes ikke kun antallet af uregelmæssige former, men også det faktum at dansk er et mere flekterende sprog end engelsk og har flere bøjningsmønstre. For at forstå dette kan man tænke

på de forskellige måder man på dansk kan danne flertalssubstantiver (*-er, -e* eller intet) sammenlignet med *-s* på engelsk. Derfor er det bedst på dansk at finde ordenes lemma ved hjælp af præcise morfologiske oplysninger. Processen kaldes for lemmatisering. CST-lemmatiseren (Jongejan, 2006; Jongejan & Haltrup, 2001) bruger oplysninger i STO til at gætte grundformen og bøjningsmønstre for alle danske ord. Lemmatiseren bruger også information om ordenes staveformer, fx genkender den at *mayonnaise* og *majonæse* er samme ord således at brugerne undgår eksplicit at skulle søge på ordenes stavevarianter.

De beskrevne værktøjer er blevet anvendt i stort set samtlige CSTs projekter relateret til informationssøgning, og de morfologiske oplysninger i STO er blevet en integreret del af søgegrænsefladen i flere praktiske løsninger, fx hos Sundhed.dk. Værktøjerne kan afprøves online på www.cst.dk/tools.

2.4 Andre anvendelser: informationssøgning og sammensætninger

Et eksperiment med et mere specifikt fokus blev udført inden for VID-projektet (se www.cst.dk/vid) og omhandler søgning på sammensætninger i dansk (Pedersen 2007). Eksperimentet blev lavet i samarbejde med den danske virksomhed Ankiro som netop i deres udviklingsopgaver inden for informationssøgning havde noteret sig store problemer med sammensætninger i dansk. Særligt er det et problem i intranetløsninger hvor datamængden generelt er forholdsvis lille. Her er det særdeles utilfredsstillende at man ofte ikke får nogen hits på sin forespørgsel med sammensætninger udelukkende fordi begrebet er udtrykt anderledes i teksten. Recall er med andre ord alt for lavt ved søgning på sammensætninger (for sprog der sammenskriver disse) i og med at for mange relevante tekstudsnit ikke uddrages, jf. i øvrigt tilsvarende resultater fra svensk: Chen & Gey, 2003 og Dalianos, 2005.

På basis af Ankiros erfaringer samt resultaterne for svensk har vi i vores eksperiment taget det udgangspunkt at ekspansion af søgestrengen til de splittede ordformer er nødvendig for generelt at forbedre et ringe recall. For en sammensætning som *byrådsmedlem* tager vi således for givet at det er nødvendigt også at søge på de enkelte ord (*byråd* og *medlem*) for at sikre at ikke for meget relevant information går tabt. En sådan splitning sikrer dels at fejlskrivninger i to ord – som er meget hyppig på dansk efter påvirkning fra engelsk – fremfindes, dels at de meget hyppige synonyme udtryk til sammensætninger lokaliseres, altså hhv. *byråds medlem* og *medlem af byrådet* som alternativer til *byrådsmedlem*. Et oplagt problem når man splitter sammensætninger under søgning, er imidlertid at precision dermed falder dramatisk; man får altså generelt for meget støj med (jf. Dalianos, 2005) ved at bruge denne metode.

Hvor udgangspunktet for vores eksperiment altså som nævnt er at forbedre et ringe recall, så er det primære formål imidlertid at afdække hvordan dette kan gøres uden at forværre precision for meget. Hertil udnyttes den sproglige kontekst hvori den 'splittede' sammensætning forekommer. Hvis vi fx tager som udgangspunkt at vores søgeforespørgsel til eksemplerne nedenfor er *apoteksovertagelse*, så må udtrykket *overtagelse … af apotek* i eksempel (1) siges at være synonymt hermed. Tekstudtrækket i eksemplet må derfor betragtes som ganske relevant for forespørgslen – også selv om der faktisk er en afstand på 5 ord mellem de to lemmaer *overtagelse* og *apotek*:

1) *Udgifter i forbindelse med [overtagelse, nyanlæg eller flytning af et apotek]*
2) *Han lod ved [overtagelsen] foretage en [optælling af* apotekets *varelager]*

I eksempel (2) er afstanden mellem de to søgeord mindre, alligevel må hittet alt andet lige betragtes som mindre relevant for forespørgslen da der ikke er tale om et synonym til denne. Syntaktisk set adskiller de to eksempler sig fra hinanden ved at søgeordene i det første forekommer inden for samme navnefrase[3] mens de i det sidste optræder i to forskellige fraser. Fraserne er markeret ved hjælp af kantede parenteser. Idet vi antog at der her var tale om et generelt fænomen, fastlagde vi en syntaksbaseret såkaldt 'inden for samme navnefrase'-tærskel for at kunne identificere knap så relevante hits med henblik på en frasortering af disse.

En anden antagelse var at nogle sammensætninger havde større tendens til at forekomme i 'splittet form' end andre slags sammensætninger: sandsynligheden for at *apoteksovertagelse* med det centrale ord *overtagelse* (som stammer fra verbet *overtage* og derfor knytter såkaldte valensled til sig) hyppigt optræder synonymt i splittet form, er stor, hvorimod det ikke ses så hyppigt for en sammensætning som fx *tennisbold*, hvor det centrale ord er *bold*. Derfor an-

tog vi at søgning på splittede sammensætninger ikke ville forbedre søgningen lige så meget ved sådanne (såkaldt avalente) sammensætninger. STOs syntaktiske oplysninger gav os mulighed for automatisk at gruppere vores sammensætninger afhængig af valens og afprøve denne hypotese.

Søgeresultaterne til evalueringen blev udtrukket fra KommuneInformations tekstdatabase, og et søgeresultat blev defineret som et tekstudtræk hvor de to søgeord optrådte inden for en afstand af maks. 10 ord. Søgeresultater for 410 tilfældigt udvalgte splittede sammensætninger (uden kontekst) blev anvendt, og for hver søgning blev der sat en maksimumgrænse på 200, hvilket resulterede i i alt 3.367 søgeresultater. Alle tekstudtrækkene undergik en automatisk syntaksanalyse bl.a. baseret på ordklasseopmærkning hvor navnefraser blev genkendt, og den syntaktisk baserede tærskel blev anvendt til at frasortere formodede irrelevante søgeresultater.

Til brug for evalueringen blev alle hits gennemgået manuelt: hvis evaluatoren skønnede at et hit var relevant for søgeforespørgslen, blev dokumentet registreret som relevant. Evalueringen viste at søgning på splittede sammensætninger gav ganske fine søgeresultater når man anvendte 'inden for samme navnefrase'-tærsklen på valente sammensætninger (som *byrådsmedlem* og *apoteksovertagelse*), som opnåede en precision på 0,94 og et recall på 0,70. For avalente sammensætninger (som fx *tennisbold* og *reaktortank*) var precision på 0,78 ved anvendelse af samme metode, mens recall var på 0,52. Dog var resultaterne for materialet generelt langt bedre end når man ikke sorterede irrelevante hits fra: Uden syntaksbaseret tærskel opnåedes således kun en precision på 0,40 (recall på 1) på det samlede materiale.

Hvis vi afslutningsvis vender tilbage til vores udgangspunkt, kan vi således konkludere tre ting:

- i forbindelse med søgning på sammensætninger er det nødvendigt også at søge på de splittede former hvis ikke alt for meget information skal gå tabt;
- søgning på splittede sammensætninger *kombineret med* en syntaksbaseret tærskel kan markant forbedre den forringede precision som alt andet lige følger med når man inddrager de splittede former;
- nogle typer sammensætninger opnår bedre resultater ved indførelsen af denne tærskel end andre, nemlig sammensætninger hvor det centrale ord er valensbærende som fx i *apoteksovertagelse* og *byrådsmedlem* (i modsætning til *tennisbold*).

Selv om problemstillingen med sammensætninger er særlig for dansk (og andre sprog med sammenskrivning), findes der lignende sproglige udfordringer for dansk såvel som for andre sprog. Til sammenligning kan ses Strzalkowski et al. (1996) og Strzalkowski et al. (1997) som i forbindelse med informationssøgning på engelske tekster anvender syntaktisk analyse til at identificere navnefrasevarianter af typen *weapon proliferation, proliferation of weapons*, og *proliferate weapons.* Generelt for sådanne eksperimentelle tilgange til informationssøgning gælder det at der gøres god brug af både morfologiske og syntaktiske ordbogsdata.

3. Ontologier og wordnets

3.1 Hvad er en ontologi

Hvor ordbøger og elektroniske ordbaser beskriver et sprogs ordforråd, definerer ontologier derimod viden uafhængigt af specifikke sprog. Ontologi betegner i ordets oprindelse en filosofisk retning der beskæftiger sig med tingenes essens. I nutidens terminologi refererer ordet imidlertid også til en abstrakt struktur som repræsenterer viden i form af begreber og relationer imellem begreber, og som har visse formelle egenskaber der gør dem egnede til ikke blot at systematisere domæneviden – noget man også kan opnå ved en tesaurus – men også at gøre det muligt for en computerapplikation at fortolke denne viden. Disse egenskaber følger af at ontologier formelt set bygger på mængdelære og prædikatslogik og derfor tillader at anvende velkendte logiske slutninger. I en formel ontologi er begreber således defineret som klasser (svarende til mængder) hvis ekstension udgøres af et antal individer eller instanser (mængdens entiteter) og for hvilke der gælder visse egenskaber. Egenskaberne kan være atomare eller udtrykke en relation mellem klasser. En helt grundlæggende relation er den der holder mellem en given klasse og dens underklasser, og som hedder inklusion fordi den overordnede klasse *inkluderer* individerne fra dens underklasser. I de næste afsnit vil vi se at inklusionen får en lidt anden betydning når den er brugt i forbindelse med lingvistiske ontologier og wordnets. I øvrigt bruger leksikografer og lingvister tit *taksonomisk relation* i stedet for inklusion, og siger om

det forhold at de egenskaber der gælder for en klasse også gælder dennes underklasser, at de *nedarves*. Ontologier har opnået fornyet popularitet i moderne videnapplikationer, specielt i forbindelse med visionen om det semantiske web (Berners-Lee et al., 2001), en fremtidig version af Internettet hvor computerprogrammer og ikke blot mennesker vil kunne "forstå" *indholdet* af de data – tekster såvel som andre ressourcer – der er tilgængelige på nettet. Der findes efterhånden adskillige ontologier som dækker både almen viden og specifik domæneviden, fx SUMO (Niles & Pease, 2001), DOLCE (Gangemi et al., 2003) og DAML-samlingen (http://www.daml.org/ontologies/). Og der er udarbejdet internationale standarder til at repræsentere ontologier således at XML-baserede applikationer kan fortolke dem. En række formelle sprog som alle bygger på XML – hvor OWL er det seneste skud på stammen – er således blevet udviklet til formålet (Davies et al., 2003), og et vigtigt forskningsemne inden for det semantiske web handler netop om at videreudvikle disse sprog og skabe værktøjer der understøtter deres brug.

Et andet af disse formelle sprog er Topic Maps (Park & Hunting, 2003), som specielt bruges til at understøtte videnbaseret navigering. Et topic map er en struktur hvor instanser af en ontologis semantiske klasser (fx forskellige typer web-ressourcer) er sat i relation til hinanden via såkaldte associationer, dvs. semantiske relationer med to eller flere argumenter der hver spiller sin specifikke rolle. Med andre ord definerer man i et Topic Map alle individerne i en konkret instansiering af et ellers "abstrakt" ontologisk system af klasser og underklasser. Sommetider skelner man således mellem ontologien, der betegner den generelle del af systemet, og videnbasen, som indeholder mere specifik domæneviden. I et domæne hvor et Topic Map beskriver web-ressourcer, ville man have associationer som *skrevet-af* eller *publiceret-af* som fx kunne forbinde bøger med forfattere og forlag.

Et eksempel på et site der anvender Topic Maps er Norges kulturportal Kulturnett.no, hvor det ontologiske skelet bruges til at kategorisere søgeresultaterne. Fx giver forespørgslen "Munch" en lang række hits som ud over "Edvard Munch (Billedkunster)" omfatter "Munchs Hus (Museum)" og "Munch. En biografi (Bog)". Den ontologiske kategorisering er en god hjælp til at sortere irrelevant information fra.

På CST har vi brugt Topic Maps i det europæiske forskningsprojekt MOSES (Atzeni et al., 2004; Paggio et al., in press), hvor vi eksperimenterede med ontologisk modellering og tværsproglig søgning i en 'semantisk web'-platform. Projektet er interessant fordi det forener brugen af ontologier med andre sprogteknologiske ressourcer så som ordbaser og grammatiske regler. Denne kombination af sproglig og ontologisk viden kan fx bruges til at fortolke brugerforespørgslerne, som i MOSES er korte spørgsmål til et netværk af universitetssites. Metoden er realistisk i en domænespecifik sammenhæng hvor den ontologiske viden er rimelig veldefineret, og hvor brugerne stiller ret specifikke spørgsmål.

3.2 Lingvistiske ontologier og wordnets

Hvis vi vender tilbage til de leksikalske ressourcer, så behandles de semantiske aspekter af dette inden for sprogteknologien typisk i såkaldt *lingvistiske* ontologier eller *wordnets* . Lingvistiske ontologier adskiller sig fra de ontologier vi har omtalt ovenfor, ved at være forankret i det sproglige udtryk. Lingvistiske ontologier benyttes altså først og fremmest til at repræsentere leksikalsk-semantisk viden på en struktureret måde, og i sagens natur forholder de sig til konkrete sprog som fx *dansk* eller *spansk*. De modellerer ikke en konkret verden som ontologier i videnbaser typisk gør, og derfor danner de bro mellem begreber og ord snarere end mellem klasser og instanser.

Lingvistiske ontologier er nært beslægtede med wordnets såsom Princeton WordNet, og ofte bruges benævnelserne i flæng. Wordnets består af leksikalske enheder som er forbundet med hinanden ved hjælp af semantiske relationer, som fx hyponymirelationer, del-helhedsrelationer el. lign. Som regel vil man dog forudsætte at et wordnet er forbundet med en sproguafhængig ontologi for at det kan betegnes som en lingvistisk ontologi.

Den leksikalske viden som er udtrykt i lingvistiske ontologier og wordnets er vigtig hvis man vil udvikle søgesystemer der skal kunne håndtere tekst på en nuanceret måde. Også set i et tværsprogligt søgeperspektiv er det vigtigt at have lingvistiske ontologier som refererer til specifikke sprog. På et vist specificeringsniveau har forskellige sprog forskellig ontologisk struktur; på spansk har man fx et fælles overbegreb for fingre og tæer: *dedos*; det har vi ikke

på dansk; der taler vi om *lemmer*, som dog er overbegreb for mere end blot fingre og tæer.

3.3 Eksperimenter med lingvistiske ontologier i informationssøgning på danske tekster

På CST har vi i det seneste tiår forsket en del i anvendelse af lingvistiske ontologier til informationssøgning på dansk tekstmateriale. Her vil vi omtale to projekter: Det første er det tværfaglige projekt OntoQuery (www.ontoquery.dk) som foregik i et samarbejde med DTU, RUC og CBS. Det andet er VID-projektet, som blev introduceret i afsnit 2.4.

Baggrunden for at anvende en lingvistisk ontologi i OntoQuery var at CST havde deltaget i det internationale projekt SIMPLE (Semantic Information for Multifunctional, Plurilingual Lexica, jf. Lenci et al. 2000), hvor ontologiske modeller for formel betydningsbeskrivelse var blevet afprøvet og konsolideret på et bredt ordmateriale; for det danske projekts vedkommende bestående af 10.000 danske begreber i en lingvistisk ontologi, en ressource der nu indgår i STO.

Det undersøgte fagområde inden for OntoQuery indbefattede ernæringsdomænet som det optræder i Den Store Danske Encyklopædi. Projektets formål var at udforske metoder til at opnå en indholdsbaseret tilgang til søgning baseret på ontologisk viden. Et simpelt eksempel er at man via en søgning på *mangelsygdom* automatisk bør kunne identificere tekster med underbegreber hertil, som fx *beriberi*. Dette sker ved at indeksere dokumenterne med relevante begreber fra en underliggende ontologi, og ved at finde de tekster der indeholder termerne i søgestrengen eller termer som er ontologisk relaterede. Det at være ontologisk relateret bliver målt ved at se på hvor langt i forhold til hinanden to termer befinder sig i den ontologiske struktur. I projektet gik man dog videre end til at behandle begreber i isolation; vi ønskede også at udforske specielt navnefraser og de interne relationer der holder mellem de enkelte elementer i disse, fx mellem de to substantiver i en frase som *mangel* på *vitamin*. Hypotesen var at man ved at afdække disse relationer kunne forbedre søgning yderligere bl.a. fordi man mere præcist kunne afdække hvornår to udtryk refererer til det samme begreb, som tilfældet er her med *mangel på vitamin* og *vitaminmangel* (altså en semantisk tilgang til problemet med sammensætninger) eller til begreber der er meget nært beslægtede (fx *tykke børn, børn med overvægt, børn med vægtproblemer*). Til dette formål udviklede man i projektet et algebraisk beskrivelsessprog, OntoLog (Nilsson, 2001), som gav mulighed for på et formelt grundlag at repræsentere denne fælles semantik. For at nå frem til den semantiske analyse anvendte og specialudviklede vi flere af de ovenfor nævnte sprogteknologiske værktøjer i form af bl.a. tagger og navnefrasegenkender. Projektets hypoteser blev afprøvet med positive resultater (Andreasen et al., 2004; Pedersen & Paggio, 2004; Paggio et al., 2003), dog i en begrænset, domænespecifik kontekst.

Hvor vi i OntoQuery primært fokuserede på den teoretiske udvikling af formelle ontologiske informationssøgningsmetoder, arbejdede vi i VID sammen med en konkret bruger: patentvirksomheden Zacco A/S. I samarbejde med Ankiro opbyggede vi en prototype der skulle afprøve hvorvidt sprogteknologiske ressourcer og ontologisk viden kunne forbedre informationssøgning på Zaccos intranet; i dette tilfælde søgning på standarddokumenter inden for patentdomænet. Ankiro implementerede prototypens søgemaskine, som køres gennem internettet med en browser som brugergrænseflade. Prototypen søger efter indholdsord i tekster samt i teksternes XML-metadata ved at berige søgestrengen med lingvistiske og semantiske oplysninger (query expansion). Følgende oplysninger udnyttes i søgning: oplysninger om ordenes bøjningsformer; synonymer og overbegreber; oplysninger om andre ord som er relateret til søgeordene via indirekte, og derfor svagere relationer, fx nærsynonymi; samt relationer der knytter sig til de forskellige dele af en sammensætning. Fx er *patentansøgning* relateret til *ansøgning* via en taksonomisk relation, mens den er svagt bundet til *patent*.

Prototypen returnerer tekster hvor søgeordene, eller ord relateret til søgeordene, er blevet fundet. Søgeresultaterne tilknyttes forskellig vægt afhængigt af de oplysninger der ligger til grund for søgningen. Den højeste vægt tildeles bl.a. resultater opnået ved ekspansion på bøjningsformene, samt resultater fundet ved at følge synonymiske og taksonomiske relationer, mens lavere vægt tildeles ord som har en svagere relation med søgeordene. Prototypen returnerer søgeresultaterne i prioriteret rækkefølge, således at de bedste resultater i forhold til søgemaskines vægtningssystemet står forrest. Prototypen kan også søge i dokumenternes metadata (Dublin Core), hvor bl.a. emnefeltet (subject) er uddraget automatisk fra teksterne via sprogteknologiske metoder og

Resultater	Precision	Recall	F-score
baseline: simpel søgning på strenge	99,7	54,2	76,9
ekspansion med ordformer	98,8	79,2	89
ekspansion med ordformer og semantiske oplysninger	89,2	95,1	92,5

Tabel 1: Evaluering af VID-prototypen

værktøjer. Prototypen er blevet testet i en mindre målestok med 75 virksomhedsrelevante forespørgsler. I tabel 1 vises resultaterne med simpel søgning på strenge (baseline), med ekspansion på bøjningsformerne, og med ekspansion hvor alle lingvistiske oplysninger blev brugt. I den sidste søjle angives den såkaldte F-score, hvor værdierne for precision og recall er samlet i én værdi. I beregningen af F-score er recall og precision vægtet lige, således at F-score = *(Precision+Recall)*/2, selv om recall var vigtigere end precision for brugervirksomheden Zacco A/S.

Resultaterne viser at ekspansion på ordenes bøjningsformer forbedrer recall betydeligt, dog med et lille fald i precision. Semantisk ekspansion forbedrer recall yderligere, mens precision falder lidt i forhold til systemer der kun ekspanderer med ordformer. F-score forbedres betydeligt i begge tilfælde. Mens det kan antages at ekspansion med ordformer forbedrer F-score i alle domæner (antagelsen bekræftes af anvendelsen af morfologisk viden på flere internetsøgemaskiner), kan det samme ikke udledes for ekspansion med semantisk viden, idet ord oftere er tvetydige i almensproglige tekster. For flere detaljer om VID-prototypen henvises til Navarretta et al. (2006).

4. Et wordnet for dansk: DanNet

Som det ses af det ovenstående er de fleste af vores eksperimenter med semantisk viden og informationssøgning foregået på forholdsvis afgrænsede domæner. Ønsket om at eksperimentere med sprogteknologi mere generelt også på almensproglige tekster blev anledningen til at vi i 2005 igangsatte et større ressourceprojekt med en simplere struktur end SIMPLE og – bl.a. af økonomiske hensyn – med en højere grad af genbrug fra traditionelle ordbøgers definitioner. Valget faldt på den internationale standard for wordnets (Vossen et al., 1999) som er langt enklere end SIMPLE (med færre semantiske oplysningstyper) og derfor mere realistisk for et projekt som har høj dækningsgrad som et væsentligt mål.

Det danske wordnet-projekt, DanNet (www.wordnet.dk), udmønter sig som tidligere nævnt som et samarbejdsprojekt mellem CST og DSL. Hvert af disse to miljøer har inden for de senere år afsluttet omfattende leksikalske projekter hvis resultater tilsammen udgør et væsentligt udgangspunkt for udviklingen af det danske wordnet. Den Danske Ordbog (DDO: Hjorth et al., 2005) er en omfattende ressource der på baggrund af korpusundersøgelser beskriver ords betydninger, primært ved hjælp af definitioner og brugseksempler. Erfaringerne fra DDO og SIMPLE i henholdsvis det leksikografiske og det sprogteknologiske miljø giver optimale muligheder for at udarbejde et wordnet for dansk med en høj dækningsgrad (Pedersen & Asmussen, 2006, Asmussen, Pedersen & Trap-Jensen 2007).

Et wordnet består af begreber som, fx 'firehjulet motorkøretøj til brug for persontransport' med tilknyttede leksikaliserede udtryk, fx *bil*, *automobil, dyt, vogn, karet.* En sådan helhed bestående af et begreb og dets leksikaliserede udtryk betegnes et synonymsæt eller kort *synset*. Disse synsets udgør ordnettets byggestene. De enkelte synsets forbindes med hinanden ved at definere de semantiske relationer der hersker mellem dem, fx over- og underbegreber (*motorkøretøj*) – (*bil*, *automobil*..), del-helhed (*bil*, *autmobil*..) – (*motor*) og funktionsrelationer (*motorkøretøj*) – (*køre*). Gennem disse relationer etableres således en eksplicit leksikalsk-semantisk beskrivelse i form af et semantisk net: et wordnet.

Wordnettets skelet udgøres af over- og underbegrebsrelationerne, også kaldet hyponymi. Disse relationer udtrækkes semi-automatisk på baggrund af DDO's definitioner, som overvejende er givet ud fra den leksikografiske *genus-differentia*-model, hvor man først angiver det nærmeste overbegreb og dernæst de adskillende træk. Håndteringen af den ontologiske struktur som det leksikalsk-semantiske wordnet udgør, betragtes som en særlig forskningsmæssig udfordring i DanNet-projektet. Særligt hyponymire-

lationen har en central ontologisk status, og når man udtrækker store grupper af hyponymer fra DDO, bliver det klart at der er tale om en ganske kompleks struktur. Det viser sig tydeligt at hyponymirelationen dækker over forskellige undertyper af relationer hvoraf kun nogle har en egentlig taksonomisk status. Hvor hyponymi kan defineres ud fra sætningen: *X er en Y*, så kan taksonomi forstås som ontologisk set mere præcis relation med definitionen: *X er en slags/type Y* (Cruse, 1991; Cruse, 2002). Taksonomi (svarende til inklusion i formel ontologi) er vellidt i ontologier bl.a. fordi nedarvningsstrukturen er nogenlunde klar; anden slags hyponymi er derimod mere kompliceret at håndtere ud fra et ontologisk synspunkt. Fx ville det være absurd at betragte *egetræ* eller *birketræ* som underbegreber til *vejtræ* selv om både birketræer og egetræer kan fungere som sådan. Det virker også besynderligt at definere *vejtræ* som en slags træ; der er derimod tale om *et hvilket som helst slags træ der står i vejkanten*. Vi må altså konkludere at *vejtræ* er en ikke-taksonomisk sproglig størrelse i modsætning til *birketræ* og *egetræ*, som angiver typer af træer, og den må derfor betragtes som havende en anden ontologisk status (Pedersen & Sørensen, 2006).

Når man arbejder med almensproget, er det slående hvor stor en del af ordforrådet der på denne måde må beskrives som *ikke-taksonomisk*. Ofte er det også langtfra entydigt hvilket perspektiv der skal tages i anvendelse når ordnettet skal opbygges, og ofte kan man overveje at anvende flere perspektiver samtidig. I DanNet har vi valgt en lægmandstilgang til ordnettets overordnede struktur sådan at forstå at vi i udgangspunktet ikke udvikler en dybere hyponymistruktur end hvad der ligger lige for hos en ikke-fagspecialist. De taksonomiske overbegreber for *stol* som er indkodet i DanNet er således *siddemøbel, møbel og genstand*. Dette betyder ikke at andre perspektiver på møbler ikke optræder i wordnettet. I forsikringsmæssig sammenhæng anvendes fx ofte begreberne *bohave* og *indbo* om den samling af ejendele der findes i en bolig; og mere specifikt repræsenterer *løsøre* alle de flytbare genstande som en bolig rummer. Disse begreber forefindes også i ordnettet, men de indgår ikke nødvendigvis i den taksonomiske struktur udover at de tilknyttes et overbegreb; i dette tilfælde *samling* i betydningen en mængde af genstande.

Alle eksemplerne ovenfor omhandler konkrete substantiver – også kaldet *1st order entities* hos Lyons; en klassificering der er overtaget i EuroWordNet-standarden (Vossen et al., 1999). Der er ingen tvivl om at det er til beskrivelsen af disse at ordnettets taksonomi har sin største berettigelse i hvert fald i sin traditionelle form baseret på EuroWordNet-standarden. Denne type oplysninger gør det muligt på en formelt grundlag at måle den semantiske afstand mellem begreberne i et søgeudtryk og begreberne i et dokument og på den baggrund vurdere hvor relevant et hit er - i lighed med den metode der anvendtes i OntoQuery-projektet. Det er straks mere vanskeligt når det kommer til beskrivelsen af hændelser, handlinger, tilstande og egenskaber (*2nd order entities*) som der typisk refereres til med hhv. verber, verbalsubstantiver og adjektiver. Generelt synes den taksonomiske struktur for disse begreber mere uklar, og det er også mere tvivlsomt i hvor høj grad den kan anvendes direkte til at beregne semantisk nærhed. Imidlertid vil en beregning på hændelser, handlinger, tilstande og egenskaber ofte indgå i en dybere semantisk analyse af fraser fx udtrykt i den før omtalte OntoLog-formalisme, og deres semantik må derfor beskrives i en vis udstrækning. Derfor indbygger vi i DanNet de hyppigste verber og adjektiver i en ontologisk struktur som gør det muligt at foretage visse semantiske beregninger. Vi tager udgangspunkt i EuroWordNet og anvender den top-ontologi der er skitseret for disse semantiske kategorier (Vossen et al., 1999:139). Nogle af de egenskaber som indkodes via ontologien er hvorvidt en handling er afsluttet eller uafsluttet (fx *ankomme* vs. *løbe*), intentionel eller ikke-intentionel (fx *cykle* vs. *falde*) og fysisk eller mental (fx *spise* vs. *tænke*).

Abstrakte entiteter (*3rd order entities*) indgår også i DanNet med åndelige begreber som fx *kærlighed* og *idé*, faglige discipliner som fx *teologi* og *datalogi*, tidslige begreber som *time* og *år* mv. Vi ser området som et relevant forskningsemne i projektet idet denne semantiske klasse stort set ikke har været genstand for nærmere undersøgelser i de øvrige wordnet-projekter, og der derfor ikke er tale om et særligt finkornet begrebsapparat til beskrivelse af den. Det vi bl.a. ønsker at undersøge er i hvor høj grad en yderligere klassificering af abstrakte entiteter kan udnyttes i informationssøgning. Bør der under de tidslige begreber skelnes imellem tidsperioder og tidspunkter? Hvilke semantiske relationer er i det hele taget relevante at etablere for dette område? Vil

det fx under de faglige discipliner være givtigt at tilkoble fagterminologier? Osv.

I DanNet er vi særligt her i udviklingsfasen interesserede i en dialog med den del af forskningen og erhvervslivet der arbejder med ontologier, begrebssystemer og informationssystemer; herunder den forskning og udvikling der foregår i biblioteksverdenen. Wordnets for andre sprog er, som vi tidligere har set, blevet anvendt med en vis succes i adskillige søgerelaterede sammenhænge, og det er ønskværdigt at lignende eksperimenter kan gennemføres med DanNet således at projektet kan få den relevante udformning med fokus de steder hvor et semantisk net over almensproget rent faktisk kan gøre en forskel.

5. Konklusion

Ovenfor har vi beskrevet en række sprogteknologiske ressourcer i form af ordbaser der indeholder morfologiske og syntaktiske oplysninger, ontologier og wordnets, og vi har præsenteret en række eksperimenter hvor anvendelse af sprogteknologiske ressourcer har ført til udvidet funktionalitet eller forbedrede søgeresultater i forbindelse med informationssøgning.

Dele af de beskrevne sprogressourcer er ved at have den dækningsgrad der skal til for at de kan udnyttes generelt i praktiske løsninger til informationssøgning på virksomheders intranet såvel som på Internettet. For eksempel bruges de morfologiske oplysninger fra STO i søgegrænsefladen til Sundhed.dk. Andre – navnlig dem der indeholder semantiske oplysninger, har stadig en noget begrænset dækningsgrad, og de systemer hvor de er blevet anvendt, har præg af prototyper og er koncentreret omkring afgrænsede domæner. På trods af at tidligere forskning på området ikke entydigt har kunnet dokumentere en målbar effekt på kvaliteten af informationssøgning ved brug af lingvistiske og semantiske oplysninger, viser nyere forskning som vi så, at dette synes at være i forandring. En forklaring er at de sproglige ressourcer der er tilgængelige i dag (i hvert fald på andre sprog), generelt er på et andet niveau både hvad angår dækningsgrad og konsistens. En anden er at de metoder som anvender ressourcerne er blevet mere raffinerede idet de ofte benytter sig af statistisk såvel som sproglig viden. En tredje forklaring er at nye applikationer, fx Question Answering og tværsproglig søgning, er ved at vinde indpas, og at disse stiller andre krav til søgeteknologien.

Formålet med denne artikel har været at give relaterede forskningsmiljøer mulighed for at stifte bekendtskab med de sprogressourcer som er, eller snart vil blive, tilgængelige for dansk. Vi håber således at flere vil få lyst til at eksperimentere med det præsenterede materiale, og at der vil skabes bredere erkendelse af perspektiverne i dets videre udvikling. STOs ordbase og flere af de beskrevne værktøjer er tilgængelige til forskningsformål og kommerciel anvendelse; de nærmere betingelser for anvendelse forefindes på websiden: cst.dk/sto/distribution/index.html. DanNet forventes afsluttet ved udgangen af 2008 hvorefter ressourcen (som mange af de øvrige wordnets der forhandles gennem European Language Resources Association på www.elra.info) vil kunne anvendes mere eller mindre frit til forskningsformål eller anskaffes til kommerciel anvendelse mod betaling.

Noter

1. TREC står for Text Retrieval Conference.

2. GENELEX (1993;1994a;1994b), EAGLES (1996) mm.

3. En navnefrase, eller nominalsyntagme, er en gruppe af ord som syntaktisk udgør en helhed, og som har et substantiv (navneord) som centralt ord, fx tennisbold, en tennisbold, en god tennisbold, en bold til at spille tennis med.

Litteratur

Andersen, M.S, Asmussen, H, Asmussen, J (2002). The Project of Korpus 2000 Going Public. I: Braasch, A. & Povlsen, C. (eds.): *Proceedings of the Tenth EURALEX International Congress*, Copenhagen 2002.

Andreasen, T, Jensen, PA, Nilsson, J.F, Paggio, P, Pedersen, B.S, Thomsen, H.E (2004). Content-based text querying with ontological descriptors. I: *Database and Knowledge Engineering Journal no. 48*: 199-219, Elsevier Science B.V., Holland.

Asmussen, L, Pedersen, BS & Trap-Jensen, L (2007). DanNet: From Dictionary to WordNet. I: Kunze, C Lemnitzer, L & Osswald, R (eds.) *GLDV-*

2007 Workshop on Lexical-Semantic and Ontological Resources 1-11. Universität Tübingen, Tyskland.

Atzeni, P, Basili, R, Hansen, DH, Missier, P, Paggio, P, Pazienza, M.T & Zanzotto, F.M (2004). Ontology-Based Question Answering in a Federation of University Sites: The MOSES Case Study. *NLDB 2004*, 413–420. ISSN: 0302-9743.

Berners-Lee, T, Hendler, J & Lassila, O (2001). The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American.*

Bick, E (2004). A Named Entity Recognizer for Danish. I: *Proceedings of the 4th International Conference on Language Resources and Evaluation Proceedings 2004,* Lisboa, 305-308.

Braasch, A & Olsen, S (2004). STO: A Danish Lexicon Resource - Ready for Applications. I: *Fourth International Conference on Language Resources and Evaluation, Proceedings, Vol. IV*. 1079-1082. Lisboa.

Chen, A and Gey, F (2003). Combining Query Translation and Document Translation in Cross Language Retrieval. *Proceedings of CLEF Conference* p. 108-121.

Clough, P & Stevenson, M (2004). Evaluating the Contribution of EuroWordNet and Word Sense Disambiguation to Cross-language Information Retrieval. *Proceedings of the Second Global WordNet Conference*. 97–105, Brno, Czech Republic, January 20-23. http://www.fi.muni.cz/gwc2004/proc/73.pdf

Cruse, D.A (1991). *Lexical Semantics* Cambridge: Cambridge University Press.

Cruse, D.A (2002). Hyponymy and Its Varieties. I: Green, R, Bean, C. A & Myaeng, S. H (eds.) *The Semantics of Relationships: An Interdisciplinary Perspective, Information Science and Knowledge Management.* Springer Verlag.

Dalianis, H (2005). Improving search engine retrieval using a compound splitter for Swedish. *Proceedings of Nodalida*, Joensuu, Finland, May 20-21, 2005.

DDO: Hjorth, E, Kristensen, K, Lorentzen, H, Trap-Jensen, L, Asmussen, J, et al. (2005). *Den Danske Ordbog* 1–6. DSL & Gyldendal, København.

EAGLES Consortium (1996). *Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora. A Common Proposal and Applications to European Languages*, ILC, Pisa, May 1996.

Gangemi, A, Guarino, N, Masolo, C & Oltremari, A (1993). Sweetening wordnet with DOLCE. *AI Magazine*, 24(3), 13–24.

GENELEX Consortium (1994a). *Report on the Morphological Layer* V 3.3, November 2, 1994.

GENELEX Consortium (1993). *Report on the Syntactic Layer* V 4.0, December 1, 1993.

GENELEX Consortium (1994b). *Report on the Semantic layer* V2.1, September 30, 1994.

Davies, J, Fensel, D & Van Harmelen, F (2003). *Towards the Semantic Web. Ontology-Driven Knowledge Management.* Wiley& Sons Ltd. Chichester, England.

Gonzales J, Verdejo, F, Peters, C& Calzolari, N (1998). Applying EuroWordNet to Cross-lingual Text Retrieval, in: *Computers and the Humanities Vol. 31*, 185-207, Kluwer Academic Publishers, The Netherlands.

Hansen, H.D. (2003). CST's danske navnegenkender. I: *Årbog for Nordisk Sprogteknologi.* København: Museums Tusculanums Forlag, 159-165.

Hansen, H.D (2006). Sprogteknologiske værktøjer til tekst- og informationshåndtering. I: Braasch, A, Navarretta, C, Nimb, S, Olsen, S, Paggio, P, Pedersen, B. S & Wedekind, J, (eds.): *Sprogteknologi i dansk perspektiv*, 354-369. Reizels Forlag, København.

Jongejan, B (2006). CSTs lemmatiser for dansk. I: Braasch, A, Navarretta, C, Nimb, S, Olsen, S, Paggio, P, Pedersen, B. S & Wedekind, J, (eds.): *Sprogteknologi i dansk perspektiv*, 354-369. Reizels Forlag, København.

Jongejan, B & Haltrup, D. H (2001). The CST Lemmatiser. I: *Technical Report, STO*, Center for Language Technology, Københavns Universitet.

Keson, B (1998). *Vejledning til det danske morfosyntaktisk taggede PAROLE-Korpus*. Det Danske Sprog-og Litteraturselskab, København.

Kromann, M.T (2003). The Danish Dependency Treebank and the DTAG Treebank Tool. *Proceedings from the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, 14-15. Växjö.

Lenci, A, Bel, N, Busa, F, Calzolari, N, Gola, E, Monachini, M, Ogonowski, A, Peters, I, Peters, W, Ruimy, N, Villegas, M & Zampolli, A (2000). SIMPLE – A General Framework for the Development of Multilingual Lexicons. I: Fontenelle, T (ed.) *International Journal of Lexicography.* Vol 13. 249-263. Oxford University Press.

Maegaard, B, Offersgaard, L, Henriksen, L, Jansen, H, Lepetit, X, Navarretta, C & Povlsen, C (2006). The MULINCO corpus and corpus platform. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. 2148-2153. Genova.

Magnini, B. & Strapparava, C (2001). Using WordNet to improve user modelling in a web document recommender system. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh, June. http://www.seas.smu.edu/~rada/mwnw/papers/WNW-NAACL-217.pdf.gz

Manning, C. D & Schütze, H (1998). *Foundations of Statistical Natural Language Processing*. MIT.

Mihalcea, R. & Moldovan, D.I (2000). Semantic Indexing using WordNet Senses. *Proceedings of ACL Workshop on IR & NLP*, Hong Kong, October. http://www.seas.smu.edu/~rada/papers/acl00.nlp_ir.ps.gz

Navarretta, C , Pedersen, S. P, Haltrup D (2006). Language Technology in Knowledge Organisation Systems. I: Tudhope, D & Nielsen, M. L (eds.) *Knowledge Organisation Systems and Services, special issue in: The New Review of Hypermedia and Multidemia,* Vol. 12, no. 1 pp.29-49, Taylor and Francis, UK.

Niles, I & Pease, A (2001). Towards a Standard Upper Ontology. I : *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (FOIS-2001), Chris Welty and Barry Smith, eds, Ogunquit, Maine, October 17–19, 2001.

Nilsson, J F, (2001): A Logico-Algebraic Framework for Ontologies, ONTOLOG. I: P A Jensen, & P. Skadhauge (eds.): *Proceedings of the First International OntoQuery Workshop*, University of Southern Denmark, 11-38.

Paggio, P, Haltrup, D.H & Offersgaard, L (in press). QA with feature structures. In *Proceedings of the 1st International Workshop on Typed Feature Structures (TFSG'06)*. Aalborg, Denmark, June 2006.

Park, J, & Hunting, S (2003). *XML Topic Maps: Creating and Using Topic Maps for the Web*. Addison-Wesley.

Pedersen, B.S (2007). Using shallow linguistic analysis to improve search on Danish compounds. I: Tate, J (ed.) *Natural Language Engineering, Vol. 13, no. 1* 75-90. Cambridge University Press.

Pedersen, B.S & Asmussen, J (2006). DanNet - Fra ordbog til et leksikalsk-semantisk WordNet for dansk. I: *Leda-Nyt*. København, 3-12.

Pedersen, B.S, Navarretta, C & Hansen, D.H (2005). Anchoring Knowledge Organisation Systems to Language. I: Madsen, B.N & Thomsen, H.E (eds.) *Terminology and Content Develpment - Proceedings of 7th International Conference On Terminology and Knowledge Engineering*, København, 419-431.

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

Pedersen, B.S & Sørensen, N (2006). Towards Sounder Taxonomies in Wordnets. I: Oltramari, A, Huang, C, Lenci, A, Buuitelaar, P & Fellbaum, C (eds) *Ontolex 2006 at 5th Interntional Conference on Language Resources and Evaluation*, 9-16. Genova, Italy.

Salton, G & McGill, M.J (1983). *Introduction to Modern Information Retrieval*. New York: MaGraw-Hill.

Smeaton, A.F (1997). *Natural Language Information Retrieval*. In Strzalkowski, T.(ed) *Natural Language Information Retrieval*. Dordrecht: Kluwer Academic Publishers (Text, speech and language technology series, edited by Nancy Ide and Jean Vronis, volume 7).

Smeaton, A & Quigley, I (1996). Experiments on Using Semantic Distances between Words in Image Caption Retrieval. I : *Proceedings of the 19th International Conference on Research and Development in IR*. 174–180. Zurich, Switzerland.

Strzalkowski, T, Gurthrie, L, Karlgren, J, Leistensnider, J, Lin, F, Perez-Carballo, J, Straszheim, T, Wang, J & Wilding, J (1996). Natural Language Information Retrieval: TREC-5 Report p.291-314. *The Fifth Text Retrieval Conference*, NIST Special Publication.

Strzalkowski, T, Lin, F & Perez-Carballo, J (1997). Natural Language Information Retrieval: TREC-6 Report. I : *The Sixth Text Retrieval Conference*, 347-366. NIST Special Publication.

Voorhees, E (1994). Query expansion using lexical-semantic relations. In: Croft, W. Bruce and C. J. van Rijsbergen, eds., *Proceedings of the 17th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 1994, pp. 61 - 69.

Voorhees, E.M & Tice, D.M (2000). The TREC-8 Question Answering track evaluation. *Proceedings of TREC-8.* http://trec.nist.gov/pubs/trec8/t8_proceedings.html

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