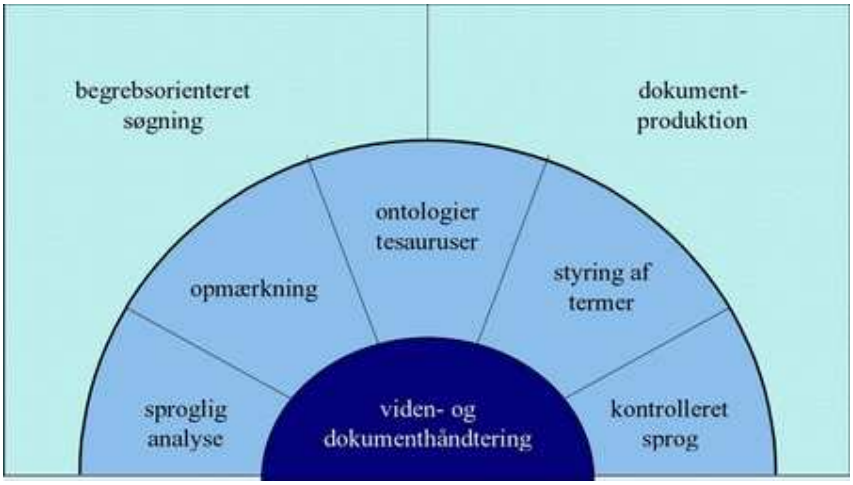


Ontologibaseret teksthåndtering

– med sprogteknologi

Bolette S. Pedersen, Costanza Navarretta & Dorte Haltrup Hansen

VID-rapport nr. 6

Center for Sprogteknologi

Marts 2005

Rapporten kan fås ved henvendelse til CST, cst@cst.dk, eller hentes fra CST's hjemmeside www.cst.dk.

VID-projektet er støttet af Center for Informationsteknologi (nu overgået til Forskningsstyrelsen).

Om VID: Viden- og Dokumenthåndtering med sprogteknologi

Der er et udtalt behov hos danske virksomheder for at kunne supplere deres eksisterende sproglige kompetence og viden med sprogteknologiske værktøjer og metoder som dels kan støtte medarbejderne, dels forankre viden og processer i virksomhedens IT-systemer, dels danne grundlag for den udvikling der kræves hvis virksomhederne skal overleve og vokse i den stadigt mere globaliserede økonomi.

VID-projektet er et forsknings- og udviklingsprojekt der har til formål at udforske de forskellige muligheder som sprogteknologi frembyder inden for informationssøgning og dokumentproduktion, og at understøtte de deltagende virksomheder i at udvikle værktøjer til bedre udnyttelse af egen viden, samt til bedre og mere effektiv produktion af dokumentation, herunder flersproget dokumentation. Foruden CST omfatter projektet på den ene side virksomhederne Bang & Olufsen A/S, Zacco A/S og Nordea A/S, som i dette projekt udgør teknologiens brugere, på den anden Navigo Systems A/S og Ankiro, som er teknologiproducenter. Projektet omfatter følgende forskningsopgaver:

- analyse af de tekstuelle data virksomhederne skal kunne håndtere for at kunne fastlægge tesauruser/ontologier for de relevante semantiske domæner, undersøgelse af den bedst egnede formalisme/teknologi til at udtrykke disse;

- afdækning og videreudvikling af sprogteknologiske komponenter til brug for automatisk tekstklassifikation og begrebsorienteret informationssøgning, indbefattende tilpasning af sprogteknologiske 'basismoduler' til opmærkning af tekst;

- udforskning af flertydighed i tekstuelle data som kan vanskeliggøre informationssøgning; ligeledes den omvendte problematik: at samme indhold kan udformes forskelligt rent sprogligt og derfor kan være svær at fremfinde i store datamængder;

- forskning inden for kontrolleret sprog - også set i et flersproget perspektiv - til brug for dokumentproduktion; herunder analyse af den sprogstil og tone som virksomhederne ønsker at anvende, samt opstilling af modeller for dette sprog;

- undersøgelse af hvilke sprogteknologiske metoder der kan anvendes til denne kvalitetssikring af dokumentproduktionen i form af f.eks. termstyring og grammatikkontrol.

Projektet er støttet af Center for IT-forskning og løber i perioden 2003-2004.

Indhold

Indhold V
1 Indledning 1
2 Præproces 1: Indeksering med sprogteknologi - identifikation af nøgleord 2
2.1 Indeksering 2
2.1.1 Den traditionelle indekseringsproces 2
2.1.2 Identifikation af nøgleord med sprogteknologi 3
2.1.3 Forsøg I: Indeksering af AMU-kurser 4
2.2 Fra konkrete forekomster af ord til "lignende" ord og udtryk 7
2.2.1 Forsøg II: Opløsning af sammensætninger 8
2.2.2 Forsøg III: Fra "førsteled" til dokumentemne i forskellige teksttyper 10
2.3 Konklusion 11
3 Præproces 2: Sprogbaseret ontologiopbygning 12
3.1 Ekspertviden versus viden i tekster 12
3.2 OWL i Protégé-2000 13
3.3 Ontologiens top 13
3.4 Ontologiens nederste lag 15
3.5 Ontologiens mellemlag 16
3.6 Ontologiens tværgående relationer 17
4 Syntaksbaseret søgefunktionalitet: synonyme udtryk i søgning med fokus på sammensætninger 18
4.1 Sprogbaseret søgning 18
4.2 Sammensætninger og synonyme udtryk 18
4.3 Tekstmaterialet og søgningerne 20
4.4 Manuel analyse: forskellige kategorier af sammensætninger 20
4.5 Automatisk analyse af tekstmaterialet med sprogteknologisk opmærkning 22
4.6 Evaluering 25
4.6.1 Metode 25
4.6.2 Sammensatte ord hvor der ingen hits er 27
4.6.3 Evaluering af beregningerne på synonymer til deverbale sammensætninger 27
4.6.4 Evaluering af beregningerne på synonymer til andre relationelle sammensætninger 29
4.7 Konkluderende bemærkninger 29
5 Prototype til søgning og dokumenthåndtering med ontologi og metadata 31
5.1 En første evaluering af prototypen og nogle konkluderende bemærkninger 37
6 Konklusion 39
Referencer A-1
Bilag A Cass-grammatik 3

1 Indledning

Denne VID-rapport (nr. 6) omhandler de resultater som er opnået i VID-projektet i relation til ontologibaseret teksthåndtering. Rapporterne 2 og 3 udgør forløberne for denne rapport, idet vi i Rapport 2 beskriver de teoretiske og praktiske overvejelser bag anvendelsen af ontologier og metadata, mens vi i Rapport 3 beskriver de rent praktiske processer i forbindelse med håndtering af de involverede virksomheders tekstmateriale. Endelig beskriver vi i Rapport 7 de eksperimenter vi har foretaget i forbindelse med statistisk ontologiopbygning.

I denne rapport undersøger vi ved hjælp af en række praktiske eksperimenter i hvilke sammenhænge sprogteknologi kan spille en tidsbesparende og/eller kvalitetsforbedrende rolle når det gælder 1) opbygning af videnbaser i form af ontologier og metadata, og 2) udtrækning af viden fra sådanne videnbaser. Alle eksperimenter er i en eller udstrækning relateret til eller udsprunget af en bestemt 'case', nemlig Zaccos ønske om at opbygge en videnbase i forbindelse med et dokumenthåndteringssystem. En af Zaccos målsætninger var at denne videnbase skulle være tilgængelig på en fleksibel måde, dvs. ved hjælp af forespørgsler i naturligt sprog.

Kapitel 2 og 3 omhandler nogle af de præprocesser der ligger forud for opbygningen af en videnbase: de relevante tekster skal håndteres og opmærkes med metadata, heriblandt nøgleord der beskriver tekstens emne (kapitel 2). Eksperimentet med nøgleord er udført på opfordring fra teknologiudvikleren Navigo, som i forbindelse med deres indekseringsarbejde ønskede at få klarhed over i hvilken udstrækning identifikation af metadata – herunder nøgleord – kunne automatiseres. Derudover skal termer udtrækkes fra teksterne, og der skal på basis af disse udarbejdes en ontologi (kapitel 3).

De to næste kapitler beskriver forskellige sprogteknologiske tiltag til at gøre søgning på en videnbase mere fleksibel: dels syntaktiske tiltag, dels semantiske. I kapitel 4 beskriver vi et søgeeksperiment med ekspansion på synonyme udtryk med fokus på sammensætninger. Målet for dette eksperiment, som er udført efter opfordring fra Ankiro og i samarbejde med dem, er at afgøre hvornår det kan betale sig at ekspandere på sammensætningen i form af søgning på de enkelte ord, og hvordan man med simpel syntaksanalyse automatisk kan frasortere dårlige hits. I kapitel 5 beskriver vi den prototype på et søgesystem i Zaccos tekstmateriale som vi har udarbejdet sammen med Ankiro. Her demonstreres brug af metadata og ontologibaseret søgning i praksis. Ved at ekspandere på søgestrengen ud fra bl.a. semantiske kriterier sikres at dokumenter fremfindes også selv om de kun indeholder et underbegreb (hyponym) eller et synonym til de begreber der forekommer i søgestrengen.

2 Præproces 1: Indeksering med sprogteknologi - identifikation af nøgleord

I mere end et halvt århundrede har man arbejdet med og forsket i indeksering til informationssøgning (se fx Luhn 1957 allerede tilbage i 50'erne). Forskningen er traditionelt foregået inden for biblioteksvidenskab og datalogi, mens de sprogteknologiske miljøer først er kommet med inden for de senere år. Det betyder at hovedvægten af den forskning der er foregået, har været lagt på udvikling af algoritmer til frekvensberegning af ord, se fx van Rijsbergen (1999). Da informationssøgning og indeksering imidlertid har med tekst og naturligt sprog at gøre, er det nærliggende at eksperimentere med at anvende sprogteknologiske metoder til indeksering og emnebestemmelse af tekst for at se om de automatiske resultater kan forbedres.

2.1 Indeksering

Ved indeksering forstår vi det at udvælge et antal ord som tilsammen beskriver emnet i et dokument. Processen vi vil beskrive her, drejer sig om på en simpel måde at repræsentere et dokuments indhold ved hjælp af nøgleord skabt med sprogteknologiske metoder. Det 'simple' indebærer at vi i den automatiske proces i første omgang ikke inddrager viden om det domæne vi behandler, eller resurser om ords semantik.

2.1.1 Den traditionelle indekseringsproces

Inden for informationssøgning arbejder man med forskellige former for normalisering af tekst. Den simpleste måde er at fjerne alle højfrekvente ord, stopord, der optræder på en stopordsliste. Listen kunne fx indeholde præpositioner, pronominer, adverbier og de mest frekvente verber. Det næste niveau i normaliseringsprocessen er at lave *stemming* på tekstens ord, dvs. at fjerne endelserne fra ordene (Lovins 1968, Porter 1980). I den simpleste form for stemming, kaldet *weak stemming*, fjernes interpunktionstegn og meget frekvente bøjningsendelser, fx –ed og –ing på engelsk. Ved *strong stemming* fjernes desuden afledningsendelser som fx –able og –ability. Det har vist sig at weak stemming er bedst i forhold til informationssøgning på engelsk, formodentlig fordi weak stemming ikke går ind i selve ordet og "ødelægger" stammen; men muligvis også fordi fjernelse af afledningsendelser kan føre til en uheldig sammenblanding af forskellige ordbetydninger. Det ville fx være i orden at fjerne afledningsendelsen *-able* fra *readable*; men hvis man gør det samme ved *capable* får man *cap* hvilket ikke er ønskværdigt fordi det betyder noget helt andet. Tanken bag strong stemming er at hvis man finder de fælles stammer for forskellige ord, finder man også den fælles betydning. Som man kan se i eksemplet ovenfor, er den antagelse for simpel og skaber derfor mange fejl.

Sidste niveau i normaliseringsprocessen er at tildele de fundne nøgleord (eller indekstermer) en vægt i forhold til deres forekomst.

2.1.2 Identifikation af nøgleord med sprogteknologi

Vores måde at identificere nøgleord ved hjælp af sprogteknologi indeholder samme elementer som den traditionelle tilgang, nemlig: 1) normalisering af ordformer til grundformen, 2) udvælgelse af betydningsbærende ord og 3) vægtning og dermed udvælgelse af nøgleord i forhold til deres forekomst. Vi forfølger altså den traditionelle idé og ser på hvor langt vi kan komme i beskrivelsen af dokumentindhold uden at have adgang til resurser om domæneviden eller ordbetydning.

Vi foretager normalisering af ordformerne med en lemmatiser udviklet på CST. Lemmatiseren er en sofistikeret form for stemmer der fører en ordform fx *børnenes* tilbage til lemmaet *barn* ved dels at fjerne bøjningsendelsen´, dels at ændre stammen hvis det er nødvendigt. CST's lemmatiser er altså en form for weak stemmer fordi den kun fjerner bøjningsendelser; men da dansk har en rigere og mere kompleks morfologi end fx engelsk, er det ikke nok blot at skære endelsen fra. I dansk kan ords endelser og stamme være flettet så tæt sammen at det er nødvendigt at beregne det bagvedliggende lemma. De ordformer som samles under samme lemma ved hjælp af lemmatiseren har med stor sikkerhed den samme betydning, hvilket er en del af målet i indekseringsprocessen.

Lemmatiseren er bygget op omkring den SprogTeknologiske Ordbog (STO) (Braasch & Pedersen 2002) der indeholder ½ mio. ordformer svarende til 81.000 lemmaer. Ud fra forholdet mellem ordformer og lemmaer er der automatisk genereret 45.000 regler der for enhver ordform i en tekst, kendt såvel som ukendt, beregner lemmaet. Den bedste performans fås hvis teksten er POS-tagget, dvs. at hvert ord har fået tilskrevet sin ordklasse. Ved at behandle hver ordklasse for sig fjerner man den flertydighed der kan være imellem især verber og substantiver, fx mellem verbet *hoppe* og substantivet *hoppe*. For en korrekt POS-tagget tekst beregner lemmatiseren det rette lemma for 97,8 % af alle ord; mens 94,5 % får korrekt lemma hvis teksten ikke er POS-tagget. Den gode performans uden POS-tags er en stor fordel hvis man ikke har en god POS-tagger, fordi en fejlprocent på blot 5% for POS-taggeren samlet set vil give et dårligere resultat. (se Jongejan & Haltrup 2001)

I den traditionelle indekseringsproces udelukkes en række ord fra at være nøgleordskandidater ved hjælp af en stopordsliste. Vi anvender den modsatte tilgang ved at udvælge de formodede betydningsbærende ord, substantiver, som nøgleordskandidater. Ved ikke kun at frasortere de lukkede ordklasser, men også de åbne og dermed potentielt uendelige ordklasser adjektiver og verber, indskrænkes mængden af mulige nøgleord.

Vægtningen af de fundne nøgleordskandidater er sidste trin mod de egentlige nøgleord. Luhn beskrev i 1957 sine ideer i forhold til ords signifikans i en tekst og selvom hans formål var automatisk resummering, er hans ideer blevet meget brugt inden for informationssøgning (Luhn, 1957). Luhns hypotese er, forenklet sagt, at tekstens relevante ord hverken findes blandt de få meget højfrekvente ord eller blandt de mange lavfrekvente ord; men derimod blandt de forholdsvis frekvente ord i midtergruppen. De

højfrekvente ord svarer groft set til de ord man fjerner ved hjælp af stopordslister, mens grænsen for de lavfrekvente ord er svære at definere.

Ved informationsøgning hvor nøgleordene skal bruges til at adskille ét dokument fra andre, er lemmaernes absolutte frekvens i dokumentet ikke nok, frekvenstallene må relativeres i forhold til de andre dokumenter i samlingen. Der findes en mængde algoritmer til beregning af lighed og forskellighed mellem dokumenter, se bl.a. van Rijsbergen (1999); men da det falder uden for fokus i dette kapitel, vil vi ikke komme nærmere ind på det her.

I næste afsnit vil vi illustrere vores tilgang til indeksering med et forsøg hvor nøgleord er fundet i dokumenter om AMU-kurser.

2.1.3 Forsøg I: Indeksering af AMU-kurser

Vi har lavet et forsøg med 185 dokumenter om AMU-kurser (Arbejdsmarkedsuddannelser) fra Undervisningsministeriet, hvor hvert dokument beskriver ét fagområde. Der er tale om meget specialiserede tekster der på den ene side stammer fra samme domæne idet de alle handler om AMU-kurser; mens de på den anden side er helt forskellige fordi handler om forskellige fagområder. Det at dokumenterne alle er om AMU-kurser, viser sig ved at termer som *jobområde*, *medarbejder*, *kompetence, kvalifikationskrav* og *arbejdsplads* er højfrekvente i alle dokumenter. Disse termer vil altså ikke kunne bruges som nøgleord til at differentiere mellem dokumenterne; men tværtimod til at samle dem i en fælles gruppe.

Første trin i processen er at tokenisere teksten, dvs. at adskille hvert token med *space*. Et token er et ord, tal, tegn e.lign. Interpunktionstegn skilles på den måde fra ord og flerordsforbindelser som fx *i forhold til*, samles til *i_forhold_til*.

I næste trin kan man vælge at identificere egennavne og andre faktuelle fakta som fx, datoer, telefonnumre, adresser mm., hvis det skønnes vigtigt i ens domæne. I eksemplet er navn og adresse på en person identificeret og klassificeret:

Kontakt: /*Ole=Skov=Jensen PERSON*/
/*Vestergade STREET*/ /*1 NUM*/
/*5000 NUM*/ /*Odense C CITY*/

Tredje trin er POS-tagging, hvor hvert ord får tilskrevet en ordklasse og enkelte morfologiske træk. I eksemplet ses en tekststump før og efter POS-tagging:

Opgaverne inden for blikkenslager og håndværkskunst udføres af en lille del af arbejdsstyrken i jobområdet.

Opgaverne/N_INDEF_PLU inden_for/PRÆP
blikkenslager/N_INDEF_SING og/SKONJ
håndværkskunst/N_INDEF_SING udføres/V_PRES_PAS af/PRÆP
en/PRON_UBST lille/ADJ del/N_INDEF_SING af/PRÆP
arbejdsstyrken/N_DEF_SING i/PRÆP jobområdet/N_DEF_SING.

I fjerde trin lemmatiseres teksten og lemmaernes frekvens beregnes. I eksemplet ses en række lemmaer efterfulgt af parentes med ordenes konkrete forekomster i teksten:

3 blikkenslager (1 Blikkenslageren, 1 Blikkenslagerens, 1 blikkenslagere)
2 blikkenslagerarbejde (1 Blikkenslagerarbejde, 1 blikkenslagerarbejdet)
1 blikkenslagerfirma (1 blikkenslagerfirmaer)
1 blikkenslageropgave (1 blikkenslageropgaver)
2 vvs-område (2 Vvs-området)
4 vvs-branche (3 Vvs-branchen, 1 Vvs-branchens)

I femte trin udtrækkes alle substantiver som mulige nøgleordskandidater.

Sidst beregnes lemmaernes relativfrekvens hvor der bl.a. ses på hvor mange andre dokumenter pågældende lemma optræder i. Tanken er at et lemma kan være et godt og karakteriserende nøgleord selvom det har lav frekvens *hvis* det blot ikke optræder i mange andre dokumenter. Samtidig siger denne beregning at et højfrekvent lemma ikke altid er karakteriserende for dokumentet, nemlig hvis det er højfrekvent i hele samlingen.

Skemaet i figur 2.1 viser de automatisk fundne nøgleord fra dokumentet om AMU-kurset for *blikkenslagerfaget*. Skemaet viser både frekvens for lemmaerne i pågældende dokument og i hele dokumentsamlingen.

Lemma	Frekvens i teksten	Frekvens i hele domænet	Forekommer i antal dokumenter
bygningsdel	14	19	5
inddækning	7	13	3
tårn	6	6	1
spir	5	5	1
ovenlys	5	5	1
kuppel	5	5	1
skorsten	5	12	4
ventilationskanal	4	4	1
tagudluftning	4	4	1
karnap	4	4	1
facadegennembrud	4	4	1
dækning	4	4	1
overgang	4	8	4
decoration	4	15	8
vvs-branche	4	8	6
blikkenslager	4	7	4
kvist	4	5	2
renoveringsopgave	3	6	4
reparationsarbejder	3	5	3
udsmykningsdetalje	3	3	1
tyndpladebearbejdning	3	3	1
håndværkskunst	3	3	1
afvandingssystem	3	3	1

afdækningsprofil	3	3	1
vvs-område	2	11	7
projektstyring	2	7	6
samlingsmetode	2	7	5
montageopgave	2	9	5
vækstområde	2	5	4
bygningselement	2	11	4
oplægning	2	3	2
fremstillingsopgave	2	5	3
nybyggeri	2	3	2
lampe	2	3	2
brugsgenstand	2	5	2
tyndplade	2	2	1
skifer	2	2	1
formstykke	2	2	1
blikkenslagerarbejde	2	2	1
afvanding	2	2	1

Figur 2.1: Automatisk fundne nøgleord fra dokumentet om AMU-kurset for *blikkenslagerfaget.*

Som man kan se er de fleste af de 40 nøgleord ganske beskrivende for blikkenslagerfaget, dog er ord som *projektstyring, lampe, vækstområde* og *brugsgenstand* ret vage. Dokumentets mest frekvente substantiver *jobområde, medarbejder,* og *kompetence* er ikke medtaget som "dokumentkarakteriserende" nøgleord, da de er højfrekvente i alle AMU-dokumenterne.

Vi har sammenlignet vores resultater med de resultater som Navigo har opnået for samme dokument. I deres proces findes først en råliste af ord fra dokumentet som derefter bliver behandlet manuelt. Rålisten er en blanding af forskellige bøjningsformer og forskellige ordklasser. I den manuelle bearbejdning af listen slettes irrelevante ord mens andre relevante ord tilføjes, fx: alle bøjningsformer af de fundne ord, enkelte synonymer og beslægtede ord samt en række flerordstermer, som fx *supplerende kurser*. Den manuelle bearbejdning af rålisten tager selvsagt en del tid. Sammenligningen viser at alle de nøgleord vi har fundet, også er på Navigos liste; mens Navigo også har andre ord som fx *AMU-kursus* og *efteruddannelse* med. Ved hjælp af sprogteknologi finder vi altså i første omgang langt færre og meget mere præcise nøgleord end med Navigos metoder. I semiautomatisk indeksering hvor et menneske, fx forfatteren til et dokument, skal vælge relevate nøgleord fra en liste, er en lille og præcis liste langt mere overskuelig. Ved fuldautomatisk indeksering er det åbenlyst at mere præcise nøgleord giver bedre performans.

Forsøget viser at brug af simple sprogteknologiske metoder giver færre og bedre nøgleord. Uden at have adgang til resurser om domæneviden eller ordbetydning, kan man altså komme et stykke vej mod beskrivelsen af et dokuments indhold. Men som eksemplerne ovenfor viser, kunne det være ønskværdigt hvis vi kunne samle lemmaer som fx *vvs-branche* og *vvs-område* samt *blikkenslager* og *blikkenslagerarbejde* fordi de åbenlyst har fælles indholdselementer, nemlig hhv. *vvs-* og *blikkenslager*. I næste afsnit vil vi beskrive et forsøg vi har lavet med at splitte sammensætninger.

2.2 Fra konkrete forekomster af ord til "lignende" ord og udtryk

Den næste udfordring, og den store udfordring i forhold til den traditionelle adgang til indeksering, ligger i at bevæge sig fra de konkrete ord der forekommer i dokumentet, til ord eller udtryk der har samme eller lignende betydning. Med andre ord: at finde den bagvedliggende betydning. Igen prøver vi i første omgang at tage den mest simple tilgang og ser på hvor langt vi kan nå uden brug af resurser om ordbetydning og domæneviden. I kapitel 4 vil vi se på hvordan synonyme udtryk kan bruges i informationssøgning og i kapitel 5 vil vi undersøge hvordan opbygning af en informationsmodel, en ontologi, over et givent domæne kan hjælpe i indeksering og søgning.

I sidste afsnit antydede vi at forskellige sammensætninger åbenlyst kan have fælles indholdselementer i form af det første eller det sidste led i sammensætningerne. Groft sagt angiver sidste led overbegrebet eller arten, i fx last*vogn*, person*vogn*, eller gods*vogn* er der i alle tilfælde tale om vogne. Første led specialiserer derimod indholdet ved i det konkrete eksempel at beskrive hvilken slags vogn der er tale om. I vores jagt på repræsentation af indholdet i dokumenter er det i første omgang førsteleddene i sammensætninger der interesserer os. Hypotesen er at førsteleddene tilsammen indsnævrer hvilke emner dokumentet handler om.

For alle ord der kan være første led i en sammensætning, er det i STO-ordbogen angivet hvilket bogstav, fugeelement, der er krævet mellem første og andet led. Denne oplysning udnytter vi til automatisk at opløse sammensætninger. Vi har på den måde udtrukket godt 55.000 ord – både simpleks og almindelige sammensætninger – som vi bruger til at styre hvad der kan være førsteled i sammensætningen. De sammensætninger der findes på listen, giver ikke i sig selv problemer. De afsluttes alle med et fugeelement og kan derfor ikke optræde som selvstændige ord i en tekst. Andetleddene findes på en liste af de knapt ½ million ordformer i STO.

Der er forskellige strategier i forhold til at opløse en sammensætning: man kan enten splitte forfra eller bagfra, matche på første kendte ord eller på længste kendte ord. Desuden kan man vælge om et eller begge af led skal være eksisterende ord. Vi har skønnet at det bedste resultat kommer ved at man opløser ordet forfra (fordi vi der har styr på fugeelementet) matcher på længste kendte ord samt kontrollerer at begge leddene er eksisterende ord. Det kan illustreres med følgende eksempel:

pension-s-afkast-beskatning-s-lov

Længste eksisterende ord forfra er:	*pensionsafkastbeskatning*
Det fremkomne sidste led er et eksisterende ord:	*lov*
Altså kan ordet opløses i:	*pensionsafkastbeskatning+s +lov*

Man kunne i stedet argumentere for at splitte ordet efter længste sidsteled, hvorved man vil få *pensionsafkast* + *beskatningslov*; men aldrig *pension* + *afkastbeskatningslov*.[1]

Det at kræve at begge led skal være eksisterende ord, minimere fejl; men til gengæld finder man ikke nye ord - hverken i første eller sidste led. Støder man fx på termen *EPO-kontrol* eller *smølfekontrol* opløses sammensætningerne kun hvis *EPO-* eller *smølfe* findes på listen over førsteled. (Pt. findes *smølfe* men ikke *EPO-* på listen). Det man generelt kan sige om vores tilgang er, at kvaliteten af listerne over første- og andetled er altafgørende for processen. Antallet og arten af ordene har stor betydning i forhold til om man vil splitte forfra eller bagfra. Desuden skal man sikrer sig at der ikke tilfældigvis er præ- eller suffikser på listen så som *under-*,*–lig*, *-hed*, eller *–skab*.

I litteraturen om sammensætninger beskæftiger man sig hovedsagligt med forholdet mellem ordets to led (se fx for dansk Ørsnes, 1995). Det vigtigste for os her, er dog i første omgang at sammensætninger kan splittes i to dele, og at førsteled der optræder med en vis frekvens, relaterer sig til dokumentets emne. Det vil være interessant at udforske om førsteleddets natur har indflydelse på denne relation; men det falder desværre udenfor rammerne i dette kapitel. I næste afsnit vil vi i stedet koncentrere os om at vise den sammenhæng der er mellem førsteled og emnerne i en række dokumenter.

2.2.1 Forsøg II: Opløsning af sammensætninger

Til undersøgelsen har vi igen brugt dokumenterne om AMU-kurser fra undervisningsministeriet. Vi har udvalgt en række dokumenter der har beslægtede emner, nemlig to dokumenter om teknik: flyteknik og elevatorteknik, to dokumenter om "detail": detailhandel og detailforarbejdning samt fire dokumenter om transport: godstransport, godschauffør, buschauffør og taxichauffør. Vi vil nu undersøge om vi ved at opløse sammensætninger og se på de første led der har frekvens på mere end en, kan få et klart og entydigt indtryk af dokumenternes emner.

Skemaerne i figur 2.2 og 2.3 viser resultatet. For hvert dokument ses alle fundne første led samt deres frekvens. De ord der går igen i alle dokumenter er markeret med kursiv, de der giver en god karakteristik af dokumentets indhold, dokumentets emne, er markeret med fed, og de ord der er fælles for grupperne to og to er understregede.

Flyteknik		**Elevatorteknik**		**Detail**		**Detailhandel**	
Frq	**1. led**	**Frq**	**1. led**	**Frq**	**1. led**	**Frq**	**1. led**
46	**fly**	24	*job*	30	*job*	25	**vare**
16	*job*	18	**elevator**	19	**slagter**	15	*job*
14	*arbejde*	12	*arbejde*	18	**vare**	13	salg
12	**reparation**	6	*kvalifikation*	18	**detail**	12	*arbejde*
9	sikkerhed	5	kvalitet	14	*arbejde*	12	**detail**
8	kvalitet	5	el	12	**kød**	11	**kunde**
6	drift	4	*certifikat*	11	kæde	6	**marked**
4	**motor**	3	**fejlfinding**	9	**fødevare**	4	nøgle

[1] Se også beskrivelsen af sammensætninger i Kapitel 4

4	**værksted**	2	programmering	7	**lager**	3	**økonomi**
4	miljø	2	Rulle	7	**pris**	3	kæde
4	*kvalifikation*	2	Drift	5	**kunde**	3	koncept
4	erhverv			6	**salg**	3	kasse
4	*certifikat*			6	**butik**	3	information
3	**luftfart**			5	*kvalifikation*	3	branche
3	værdi			5	*certifikat*	2	uddannelse
3	reserve			4	service	2	**markedsføring**
3	information			3	vej	2	**indkøb**
2	**brændstof**			3	slagteri	2	**butik**
				3	**mad**	4	*kvalifikation*
				2	slagte	4	*certifikat*
				2	**selvbetjening**	3	**penge**
				2	**marked**	3	**kasseterminal**
						2	**reklame**

Figur 2.2

Godstransport		**Godschauffør**		**Taxichauffør**		**Buschauffør**	
Frq	**1. led**	**Frq**	**1. led**	**Frq**	**1. led**	**Frq**	**1. led**
37	*job*	16	*job*	41	*job*	27	*job*
20	**gods**	14	**gods**	15	**køre**	12	**køre**
18	**køre**	8	*arbejde*	14	*arbejde*	10	*arbejde*
17	*arbejde*	7	vej	11	**taxi**	8	*kvalifikation*
10	*kvalifikation*	3	**køre**	8	*kvalifikation*	4	**rute**
9	*certifikat*	3	*kvalifikation*	4	**befordring**	4	person
7	vej	3	**færdsel**	3	virksomhed	4	*certifikat*
5	**transport**	2	virksomhed	3	**færdsel**	3	**færdsel**
4	**færdsel**	2	myndighed	2	**taxa**	2	**kørsel**
4	**påhæng**	2	**lande**	2	service	2	Forsikring
4	lande	2	*certifikat*	2	forsikring	3	**bus**
3	**vogn**			6	*certifikat*	2	virksomhed
3	sættevogn			5	person	2	**kunde**
3	myndighed			3	navigation		
3	**kørsel**			2	**vogn**		
2	**vare**			2	**kunde**		
2	virksomhed			2	konkurrence		
2	sikkerhed			2	**fører**		
2	navigation						
2	**godstransport**						
2	entreprenør						

Figur 2.3

Det første der falder en i øjnene er at førsteleddene faktisk giver en pæn beskrivelse af kursernes emne. Det er dog ikke alle førsteled der er lige karakteriserende for dokumentets emne. De mindre beskrivende svarer groft set til dem der ikke har nogen markering i skemaet. Ordet *branche* siger fx ikke meget om emnet *godstransport*. For de emnemæssigt beslægtede kurser kan man se at der både er fælles ord og ord der adskiller dem. Fx er ordene *køre, færdsel, virksomhed* og *kunde* fælles for kurserne til

taxachauffør og buschauffør; mens *taxi, taxa, vogn* og *fører* er unikke for taxachaufførkurset, og *rute, kørsel* og *bus* er unikke for buschaufførkurset.

Hvis proceduren blev implementeret i et system til automatisk klassifikation af dokumenter, ville de mindre karakteriserende ord selvfølgelig give nogen støj og dermed sløre præcisionen; men det vil være minimalt i forhold til at betragte alle ordene eller bare alle nøgleordene i dokumentet. [2]

I det undersøgte domæne viser vores undersøgelse at hypotesen holder stik: førsteled i sammensætninger hjælper til indkredsning af dokumentets emne. Det interessante er nu om det også er tilfældet i andre type tekst fra andre domæner.

2.2.2 Forsøg III: Fra "førsteled" til dokumentemne i forskellige teksttyper

Vi har indsamlet ni forskellige tekster af forskellig type (dog er to af dem websider, men af forskellig karakter). Fælles for teksterne er at de alle har et fagligt indhold; men specialiseringsgraden, subjektiviteten samt dokumentlængden er forskellig. I skemaerne i figur 2.4, 2.5 og 2.6 ses de ti første nøgleord samt alle førsteled med frekvens over en for hver af teksterne. Ved hver figur gives en nærmere angivelse af teksternes emne og længde.

Advokattekst (1)		**Avisartikel (2)**		**Marketing (3)**	
Nøgleord	**"1. led"**	**Nøgleord**	**"1. led"**	**Nøgleord**	**"1. led"**
revisor	år	kvinde	Vold	Beocom	Telefon
oktober	erhverv	stening	Stening	telefon	Lyd
dagsorden		stat		trådløs	
udkast		mand		telefonbog	
godkendelse		hor		hånd	
bestyrelsesmøde		far		funktion	
årsrapport		dommer		design	
referat		dom		øre	
juni		død		stik	
firma		bevis		stemme	

Figur 2.4: **1**: Advokattekst, referat fra bestyrelsesmøde, 486 ord **2**: Avisartikel, om stening af en nigeriansk kvinde, 389 ord, **3**: Marketing, Bang og Olufsen trådløs telefon, 183 ord.

Kursusbeskrivelse (4)		**Sygejournal (5)**		**Rapport (6)**	
Nøgleord	**"1. led"**	**Nøgleord**	**"1. led"**	**Nøgleord**	**"1. led"**
medarbejder	job	indlæggelse	dag	ontologi	meta
jobområde	gods	samtale	fod	sprog	web
kompetence	køre	forældre	alkohol	dokument	internet
køretøj	arbejde	medicin	læge	metadata	tekst
arbejdsplads	kvalifikation	kontakt		værktøj	standard
godschauffør	certifikat	aftale		begreb	data

[2] Opløsning af sammensatte ord vil også kunne bruges til "automatisk trunkering". Selvom termen trunkering oftest bliver brugt i forhold til manuel informationssøgning, kunne man forestille sig at trunkerede nøgleord kunne bruges til partiel match på en søgestreng.

transport	vej	læge		figur	søge
gods	transport	afdeling		system	sprog
udførelse	færdsel	weekend		xml	ord
teknologi	lande	tilstand		rdf	udtryk

Figur 2.5: **4**: Beskrivelse af AMU-kursus, om godstransport ad landevej, 2445 ord, **5**: Sygejournal, fra psykiatrisk afdeling, 5462 ord, **6**: Rapport, VID-rapport om ontologier og metadata, 13805 ord

Hjemmeside I (7)		**Hjemmeside II (8)**		**Behandlingsbeskrivelse (9)**	
Nøgleord	**"1. led"**	**Nøgleord**	**"1. led"**	**Nøgleord**	**"1. led"**
udstilling	kultur	virksomhed	dokument	ansøgning	patent
kulturhistorie	natur	projekt	It	trin	ansøgning
kultur		viden	Sprog	skrivelse	indlevering
krop		informationssøgning		opfindelse	
bibliotek		dokumentproduktion		indlevering	
udgave		undersøgelse		ansøger	
træffetid		sprogteknologi		modtagelse	
trykken		sprog		besvarelse	
torsdag		metode		rapportering	
topografi		dokumentation		ansøgningstekst	

Figur 2.6: **7**: Hjemmeside, personlig side for en universitetsansat på Europæisk Etnologi, 311 ord, **8**: Hjemmeside, om VID-projektet, 327, **9**: Behandlingsbeskrivelse, om behandlingsforløb for patentansøgning, 1117 ord

Uden at gå i dybden med analysen af teksttype og -genre, kan vi se at forsøget ikke virker lige godt for alle teksterne. Specielt giver førsteleddene for advokatteksten, sygejournalen og den personlige hjemmeside ikke megen mening. De tekster der giver de bedste resultater, er tekster om meget specialiserede fagområder. Resultatet er ikke overraskende for det er ved specialiserede fagområder at man oftest har brug for specialiserende termer og disse dannes typisk ved hjælp af sammensætninger på dansk.

Vi kan derfor konkludere at man med held kan beregne dokumenters emne via sammensætningers førsteled *hvis* man arbejder med højt specialiserede tekster.

2.3 Konklusion

I dette kapitel har vi vist hvordan man med simple sprogteknologiske metoder kan generere nøgleord automatisk, og hvordan man ved at opløse sammensætninger kan nærme sig beskrivelsen af emner i tekst.

De næste udfordringer ligger i automatisk at komme fra konkret forekommende ord til beslægtede termer, og fra isolerede ord til relationer mellem ord. I de næste kapitler tager vi et skridt i den retning ved bl.a. at se på ontologier og tesaurusser, samt på forholdet mellem sammensætninger og navneordsfraser. Dette giver mulighed for yderligere at udvide listen af nøgleord med beslægtede termer.

3 Præproces 2: Sprogbaseret ontologiopbygning

3.1 Ekspertviden versus viden i tekster

At opbygge ontologier[3] er en tids- og resursekrævende proces, selvom ontologierne kun modellerer begrænsede domæner. Traditionelt opbygges ontologier på baggrund af ekspertviden, men i den seneste tid har man forsøgt at opbygge eller evaluere ontologier ved at anvende store tekstsamlinger (korpora) der tilhører de pågældende domæner (Buitelar, Olejnik, Hutanu, Schutz, Declerck & Sintek 2004, Pedersen, Navarretta & Henriksen 2004).

Fordelene ved at inddrage tekstkorpora i opbygning af ontologier er mange. Først og fremmest kan korpora støtte og supplere den menneskelige introspektion i samlingen af den grundlæggende domænevokabular (både viden om begreber og relationerne som holder mellem disse begreber). Brugen af korpora som videnkilde kan forbedre konsistensen og kvaliteten af de opbyggede ontologier. Processen med at opbygge ontologier kan blive mindre tids- og resursekrævende fordi uddragelsen af information fra tekster delvist kan automatiseres. Endelig afspejler teksterne den reelle brug af domænesproget. At afdække denne brug er især vigtig når man bygger ontologier der skal anvendes i tekstorienterede applikationer der fx tillader brugeren at anvende frit sprog i brugergrænsefladen.

I dette kapitel beskriver vi nogle af de sprogteknologiske og/eller sprogbaserede metoder som vi har anvendt til at støtte ontologiopbygningen. Som et supplement til disse kan ses de metoder vi beskriver i VID-rapport nr. 7, hvor der gives en redegørelse for de *statistiske* metoder til ontologiopbygning. Disse går ud på at gruppere ord semantisk efter i hvor høj grad de optræder i samme sproglige kontekst.

Det er vigtigt at understrege at de sproglige og statistiske metoder skal kombineres med ekspertviden for dermed at opnå det bedste resultat – på en mere effektiv måde. Der er nemlig visse begrænsninger ved at anvende tekster til ontologiopbygning: det er ikke alt den nødvendige domæneviden der er udtrykt i teksterne, og det kan derfor være svært for ontologiudviklerne ud fra teksterne alene at få overblik over domænet. På grund af disse begrænsninger bør tekster ikke betragtes som den eneste videnkilde til opbygning af ontologier, og domæneeksperter bør stadig deltage aktivt i ontologiprocessen.

I VID-rapport nr. 3 beskriver vi hvorledes vi har bearbejdet teksterne med sprogteknologiske værktøjer, og hvorledes vi semiautomatisk har uddraget termerne til ontologien fra teksterne. De automatisk genererede lister er blevet gennemgået og færdigredigeret af Zaccos egne termeksperter.

Den ontologi vi beskriver i de nedenstående afsnit bygger altså på disse termudtræk og er blevet udviklet i forbindelse med at virksomheden er i gang med at omstrukturere deres arbejdsgang i relation til deres standarddokumenter. Ontologien indeholder i alt

[3] For en definition på begrebet ontologi henviser vi til VID-rapport nr. 1.

ca. 400 begreber; dog er der mulighed for at det midterste lag i ontologien på længere sigt kan udbygges yderligere af virksomhedens termsekperter.

3.2 OWL i Protégé-2000

I VID-projektet har vi bestræbt os på at udvikle ressourcer der var nemme at udveksle imellem projektets partnere, og som overholdt gældende standarder sådan at chancerne for genbrug på intra- og internet blev sandsynliggjort. Vi har derfor valgt at anvende W3C Ontology Web Language) (http://www.w3.org/TR/owl-ref/) som det formelle sprog i ontologiopbygningen. Som kodningsværktøj har vi anvendt open source-værktøjet Protégé-2000 og det dertilhørende OWL-plugin, begge udviklet ved Stanford University (http://protege.stanford.edu/). I VID-rapport nr. 1 argumenterede vi for disse valg hvor vi sammenlignede med en række andre mulige formelle sprog og værktøjer.

3.3 Ontologiens top

Vi refererer til VID-ontologien som en *lingvistisk ontologi* af tre årsager: (i) den er sprogligt forankret i og med at den primært bygger på tekstkorpora, (ii) den er sprogspecifik i de nederste lag (dansk) selv om ontologiens øverste lag bygger på en sprogneutral topontologi, (iii) den tager højde for sproglige problemstillinger i form polysemi og synonymi. Ontologiens fundament udgøres af inklusionshierarkiet, altså tanken om at begreber grundlæggende forholder sig til hinanden som enten over-, under- eller søsterbegreber. Den er opbygget ved at kombinere 'bottom up'- og 'top down'-strategier.

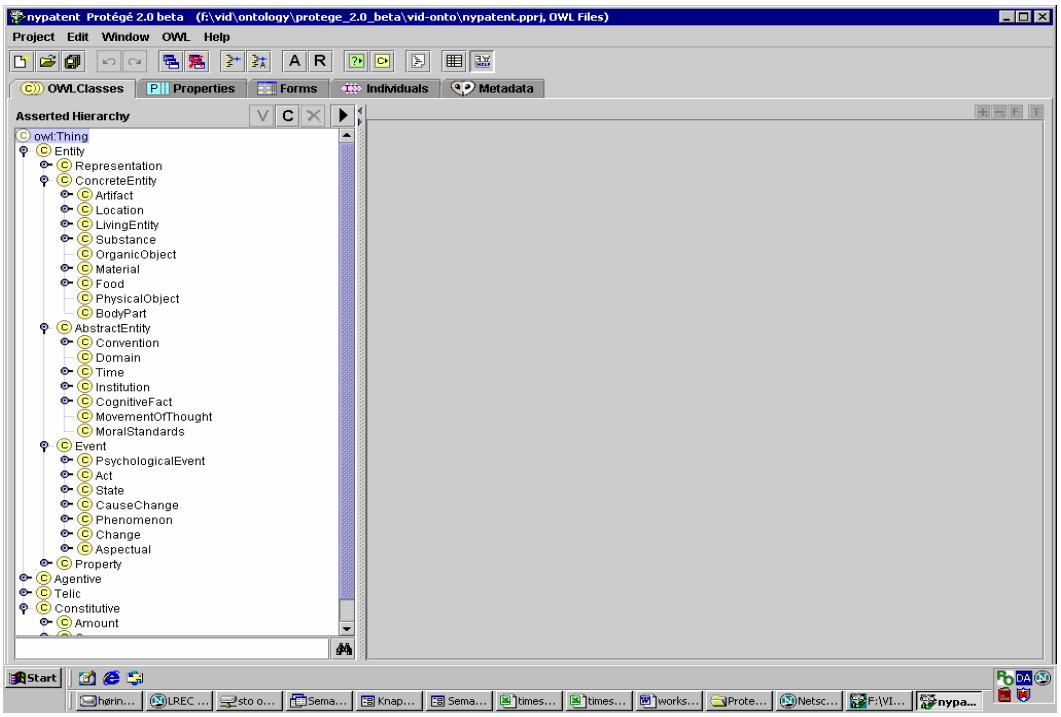
Som ontologiens top har vi anvendt SIMPLE-ontologien (Semantic Information for Plurilingual, Multifunctional LExica) som består af 135 top-begreber af typen *Human, Semiotic_artifact, Event* mv. (cf. Lenci et al. 2001, Pedersen & Paggio 2004). SIMPLE er oprindeligt bygget som et organiseringsværktøj for semantiske ordbøger for 12 europæiske sprog; således indgår den flersproglige dimension også som et vigtigt parameter – ontologien skal kunne anvendes for flere sprog. Dette kan blive relevant i forbindelse med et projekt som Zaccos som også på længere sigt skal relateres til virksomhedens norske, svenske og engelske tekster.

Som for de fleste andre formelle top-ontologier (som fx SUMO, DOLCE, BFO og andre, se VID-rapport nr. 1) er inklusionshierarkiet det bærende fundament i SIMPLE. SIMPLE-ontologien er imidlertid multidimensionel og forsøger at tage højde for begrebers forskellige kompleksitetsgrader. Den anvender ortogonal nedarvning ud fra Pustejovskys principper om de fire qualiaroller (Formal, Constitutive, Agentive og Telic), som bedst kan forklares som dækkende hhv. type ('is a'-relation) , intern struktur (fx del-helhedsstruktur), oprindelse (fx 'made by'-relation) og formål (fx 'used for'-relation). En af de fundamentale antagelser i SIMPLE er nemlig af begreber varierer mht til intern kompleksitet. Simple typer dækker såkaldte basiskategorier med rigide egenskaber (Guarino 2000:Sec.3.1) så som *Human* og *Vegetal*. Basiskategorier regnes

for endimensionale og nedarver derfor kun fra Formal role. Derimod anses komplekse typer som flerdimensionale, og de kan derfor arve fra flere qualiaroller. Det gælder fx *sagfører* som vil gå under den komplekse type AgentOfPersistentActivity i og med at sådan en person udfylder en bestemt funktion. Alle begreber hører dog til under en bestemt basiskategori; således hører *sagfører* naturligt under basiskategorien *Human*.
Til forskel fra de fleste andre formelle ontologier, anvendes der ikke aksiomatisk karakteristik i SIMPLE i form af formel-logiske udtryk til anvendelse i inferens. I stedet anvendes lingvistiske tests som brugeren kan anvende som guide til at placere begreberne korrekt i strukturen. Den lingvistiske test til bestemmelse af kategorien AgentOfPersistentActivity ser fx ud som følger:

- Klassen tæller individer, ex *fem sagførere* = 5 forskellige individer; (i modsætning til *fem kunder* = 5 eller færre individer)
- Klassen kombinerer dårligt med visse tidsadjektiver (**en hyppig sagfører* i modsætning til *en hyppig kunde*)

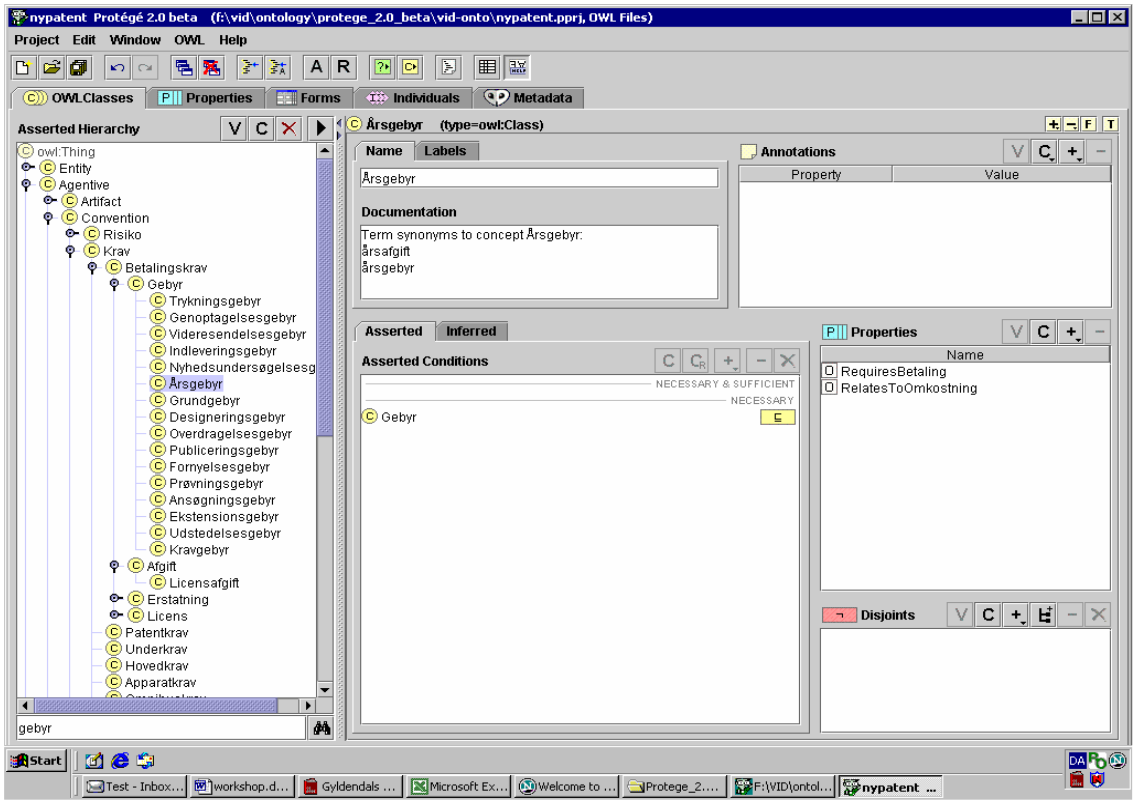
Vi har kodet SIMPLE-topontologien i OWL og et uddrag heraf kan ses i figur 3.1. Eksempler på multidimensionale kategorier som arver fra hhv. Constitutive, Telic og Agentive Role er BodyPart, Artifact og MovementofThought, hvorimod eksempler på simple typer er OrganicObject og Location (i figuren kan man dog kun se én moderknude ad gangen).



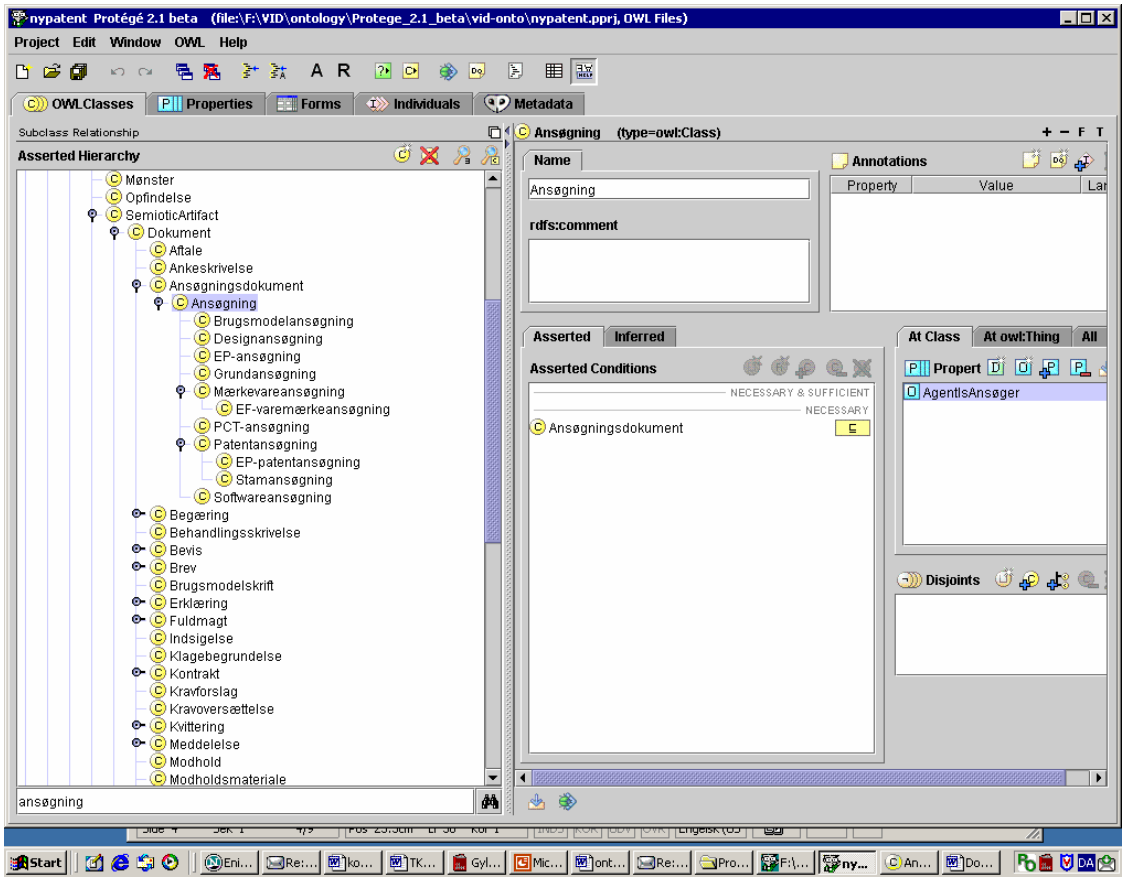
Figur 3.1: Uddrag af SIMPLE-topontologien

3.4 Ontologiens nederste lag

I ontologiens nederste lag anvendes der for de nederste knuder en 'bottom up'-strategi på basis af de termer der er blevet udtrukket. Ved specialiserede termer udtrykt i form af sammensætninger, genereres de nederste lag i første omgang automatisk, idet fx alle specialiseringer af *gebyr* hægtes under *gebyr* som det ses i figur 3.2 og for *ansøgning* i figur 3.3.



Figur 3.2: Uddrag af patentontologi med fokus på *gebyr* og dets underbegreber



Figur 3.3: Uddrag af patentontologi med fokus på *patentansøgning*

Alle øvrige begreber som der refereres til i termlisten, placeres i første omgang under de top-knuder hvor de ud fra de lingvistiske tests bedst hører hjemme.

3.5 Ontologiens mellemlag

Opbygningen af ontologiens mellemlag udgør langt den sværeste proces, og det er her ekspertviden spiller den vigtigste rolle uanset hvor mange sprogteknologiske værktøjer der er i spil. En del af denne viden kan være udtrykt i termordbøger eller lærebøger som man kan forsøge at udnytte, særligt hvis man har adgang til elektroniske versioner af dem. Vi har haft en scannet version af Zaccos egen patentordbog til rådighed, og hermed har vi som ikke-eksperter været i stand til at indsætte en del af mellemlaget i form af genus-delen i patentdefinitioner, fx defineres *justifikationssag* som en slags *retssag* hvorunder der sker en prøvning af, om et fogedforbud er nedlagt med rette; altså placeres *justifikationssag* under *retssag* i ontologien.

I den eksperimentelle ontologi udviklet i VID har termeksperterne fra Zacco indtil videre kun været involveret i mindre grad (under afholdelse af 2-3 fælles møder). Det er tanken at virksomheden kan vælge at udvide ontologiens mellemlag yderligere hvis de

skønner det relevant. Under møderne var det dog klart at nogle af de ontologiske kategorier kunne vinde ved en yderligere inddeling, fx indeholder kategorien *dokument* mere end 30 underbegreber som termeksperterne helt intuitivt foreslog underinddelt i interne og eksterne dokumenter, eksterne dokumenter med og uden retslig status osv. Termeksperterne skønnede med andre ord at en sådan understrukturering kunne lette det senere dokumenthåndteringsarbejde. Endvidere syntes en tidslig strukturering relevant; nogle dokumenter danner fx forudsætningen for andre dokumenter i sagsbehandlingen af patentsager. Disse aspekter er ikke på nuværende tidspunkt indarbejdet i ontologien.

3.6 Ontologiens tværgående relationer

Ontologier i form af inklusionshierarkier bygger på vertikale relationer i form af IS-A-relationer. Vi har også eksperimenteret med tværgående relationer i ontologien idet vi ville gøre det muligt at undersøge i hvor høj grad tværgående relationer kunne forbedre søgning.

Som angivet hos bl.a. Loukachevitch & Dobrov (2004) giver søgeekspansion på tværgående relationer - i endnu højere grad end ekspansion på vertikale relationer - en lav precision, men et højt recall. Tværgående relationer må derfor etableres med en vis forsigtighed og med direkte fokus på den konkrete applikation. Vi har på eksperimentel basis udarbejdet 18 tværgående relationer i ontologien hvor vi har udvidet den ontologiske beskrivelse i relation til nogle af domænets kernebegreber så som *patent*, *gebyr, patentansøgning*, og *sagsbehandler*. Fx er der, som det ses i figur 3.2, etableret en relation fra *gebyr* til *betaling* via RequiresBetaling. For *ansøgning* i figur 3.3 er der ligeledes etableret en tværgående relation til *ansøger* via AgentisAnsøger. Tanken er at hvis man har en forespørgsel af typen: 'betalinger i forbindelse med udvidelse af patent til et land uden for EU', så vil man også få hits i form af tekster der indeholder *gebyr* og *afgift* fordi disse er forbundet via en tværgående relation.

Det har af tidsmæssige grunde ikke været muligt at undersøge de tværgående relationers teoretiske egenskaber nærmere. Fx ville det være interessant at få belyst i hvor høj grad tværgående relationer kunne erstattes af rene aksiomatiske beskrivelser. I kapitel 5 om søgning og dokumenthåndtering beskrives nogle praktiske synsvinkler omkring spørgsmålet om søgeekspansion på relationer.

4 Syntaksbaseret søgefunktionalitet: synonyme udtryk i søgning med fokus på sammensætninger

4.1 Sprogbaseret søgning

Kapitel 4 og 5 i denne rapport omhandler sprogbaseret søgning der udnytter nogle af de elementer vi har beskrevet i de tidligere kapitler samt syntaktisk viden om dansk som kan udtrækkes fra STO-ordbasen. Eksperimenterne læner sig op ad mange tidligere eksperimenter foretaget primært for engelsk idet der for engelsk traditionelt har været flere sprogteknologiske ressourcer til rådighed. Mest kendte er eksperimenterne foretaget med WordNet (Voorhees 1994, Voorhees 1994, Smeaton & Quigley 1996) og EuroWordNet (Gonzales et al. 1998) hvor man har udnyttet ordnettenes ontologiske struktur til ekspansion på søgestrengen efter devisen om at begreber der ligger semantisk tæt på hinanden fx i og med at de er underbegreber til hinanden, ofte erstatter hinanden i tekst. Allen (2000) fremfører imidlertid at søgeekspansion på basis af ontologisk viden ikke forbedrer søgeresultaterne væsentligt, et standpunkt der dog tilbagevises senere i Loupy & El-Bèze (2002) hvor der rapporteres om vellykkede sprogteknologiske resultater fra TREC-6. Helt nye eksperimenter rapporterer også om interessante resultater i forbindelse med ekspansion på tværgående ontologiske relationer som fx beskrevet i Loukachevitch, N. & B. Dobrov (2004).

Det danske OntoQuery-projekt er det første forskningsprojekt der forsøger at overføre nogle af disse eksperimenter til dansk (cf. Andreasen et al. 2004) bl.a. ved hjælp af den danske SIMPLE-ordbase (Pedersen & Paggio 2004) samt en specialudviklet ontologi for ernæring. I OntoQuery-projektet eksperimenteres også med en kombination af syntaksberegning og beregning af semantiske relationer ud fra hypotesen om at en udregning af hvilke semantiske relationer der eksisterer mellem begreberne i et NP, kan forbedre beregningen af semantisk similaritet. Hvis man fx har beregnet at begreberne i NPet *øjets hornhinde* sammenholdes af relationen LOKATION, og at det samme er tilfældet med *hornhinden i øjet*, vil man kunne udlede at de to NPer er semantisk meget tæt forbundne, og dette kan udnyttes i søgning (se også Paggio, Pedersen & Haltrup 2003).

Eksperimenterne i VID adskiller sig fra de ovennævnte søgeeksperimenter ved dels at afprøve grænsen for hvor langt man kan komme med morfologi og syntaks alene (som det ses i kapitel 2 og i dette kapitel), dels med en kombination af flere videnkilder i form af ontologi og metadataoplysninger (kapitel 5).

4.2 Sammensætninger og synonyme udtryk

Ankiro er en teknologiudviklervirksomhed som bl.a. arbejder med at udvikle intelligente søgemaskiner. Et af deres mål som deltagere i VID-projektet er at undersøge nogle sprogteknologiske metoder i forbindelse med et par meget specifikke, ontologisk

relaterede problemer så som hvordan man automatisk kan identificere synonymer og synonyme udtryk; altså udtryk der refererer til samme begreb.

I dette kapitel fokuseres der på en meget specifik problemstilling i relation til dette, nemlig *sammensætninger* og søgefunktionaliteten i relation til dette. Det er et kendt problem at det kan være vanskeligt at søge med sammensatte ord fordi disse typisk har mange alternative udtryksformer; udtryksformer som har samme semantiske indhold. Om man siger *byrådsmedlem* eller *medlem af byrådet*, *husholdningsaffald* eller *affald fra husholdninger* er altså mere eller mindre underordnet; udtrykkene er parvis synonyme. Men søger man på *byrådsmedlem* vil man med en almindelig søgemaskine kun få fremfundet tekster hvor lige præcis det sammensatte ord forekommer, og det betyder at søgemaskinens recall bliver relativt lavt[4]; der er altså højst sandsynligt mange relevante tekster der ikke bliver fundet, se også forskningsresultater for sammensætninger i svensk (Chen & Gey 2003 og Dalianis 2005) som støtter denne antagelse. Virksomheden er derfor interesseret i at få belyst hvilke sprogteknologiske analyser der kan være med til at bestemme om hits med 'splittede'[5] sammensatte ord er gode eller dårlige svar på forespørgsler med sammensatte ord. Målet er på lang sigt at udvikle mere fleksible og sprogorienterede søgemaskiner der kan ekspandere på søgestrengen til synonyme udtryk uden at give for meget støj.

Opgaven kan opdeles i to delspørgsmål:

- Kan vi automatisk beregne hvilke sammensætninger der har hyppige alternative udtryksformer?

- For den gruppe der har hyppige alternative udtryksformer: kan vi på basis af automatiske, sprogteknologiske analyser frasortere dårlige hits?

Som baggrund for at svare på disse to spørgsmål har virksomheden leveret to korpora bestående af en række tekstudsnit i form af søgehits som er fremkommet ved at man har splittet sammensatte ord og søgt på disse. I stedet for at søge på *byrådsmedlem* har man altså søgt på *byråd* og *medlem*. En af de grundlæggende hypoteser for de sprogteknologiske analyser har været at hits hvor begge søgeord var at finde inden for den samme navnefrase (NP), sandsynligvis var gode hits som skulle prioriteres relativt højt.

[4] Hvis en given database antages at indeholde i alt 50 dokumenter, der kan karakteriseres som værende relevante i forhold til en forespørgsel, og samme forespørgsel fremfinder alle 50 dokumenter, så har den en *recall* på 1. Hvis der kun fremfindes 10 af de 50 relevante dokumenter, så er *recall* på 10/50=0,2, hvilket omvendt betyder 0,8 (80%) af de relevante dokumenter ikke blev fundet, og derfor stadigvæk er ukendte for brugeren (http://www.pce-web.dk/search/size.htm).

[5] *Medlem af byrådet* opfattes altså i denne sammenhæng som 'splittet'.

4.3 Tekstmaterialet og søgningerne

KommuneInformations tekstdatabase, der primært indeholder love, bekendtgørelser og cirkulærer, danner baggrund for det tekstmateriale der er blevet arbejdet med i dette eksperiment. Ankiro (v. Steen Bøhm Andersen) har automatisk fremfundet 1000 sammensatte ord som derefter er blevet manuelt gennemgået. Derefter er de sammensatte ord automatisk blevet splittet op i to søgeord, og der er blevet søgt i tekstdatabasen efter tekstudsnit hvori begge ord er fundet inden for afstand af max. 10 ord.

4.4 Manuel analyse: forskellige kategorier af sammensætninger

Morfologisk set er det udtrukne korpusmateriale meget homogent, idet der kun er blevet set på sammensætninger med *substantiver* i forskellige kombinationer[6]:

- substantiv + substantiv, fx *båd+designer*
- substantiv + sammensætning (bestående af substantiv + substantiv), fx *aktie+* (*beskatnings+regel*) eller
- sammensætning (bestående af substantiv + substantiv) + substantiv, fx. (*arbejds+giver*)+*medlem*

Sammensætninger med andre ordklasser behandles altså ikke, men kategorien af sammensætninger der består af substantiver, kan imidlertid yderligere inddeles bl.a. ud fra de ordklasser som substantiverne er afledt af. Hypotesen i dette eksperiment har nemlig været at de forskellige substantivsammensætninger nok ikke dannede synonyme udtryk på samme måde eller i samme udstrækning – og at det derfor ikke var lige relevant at splitte alle sammensætninger, men kun en del af dem. Derfor har vi foretaget en klassifikation af materialet i fire overordnede typer:

- Deverbale sammensætninger, altså sammensætninger hvor det sammensatte ords andet led er afledt af et verbum, som fx. sammensætningen *depotanbringelse*.
- Deadjektivale sammensætninger, hvor andet led er afledt af et adjektiv, som i *dækningsmulighed*.
- Sammensætninger som på anden måde er relationelle, altså hvor andet led er valensbærende[7] selv om det ikke stammer fra et verbum, som fx *bistandspligt*.

[6] Andre former for sammensætninger er bestemt ikke uinteressante, omend ikke særlig hyppige i forhold til substantiv+substantivsammensætninger. Vi har imidlertid skønnet at de muligvis var knap så systematiske i deres måde at danne synonyme udtryk på og derfor i første omgang ikke så relevante for vores undersøgelse.

[7] Vi har anvendt informationerne i den korpusbaserede STO-ordbasen (Braasch & Pedersen 2002) til at vurdere om et substantiv er valensbærende eller ej.

- Ikke-relationelle sammensætninger som fx *afdelingsregnskab*.

For at kunne analysere nuanceret på resultaterne, har vi foretaget endnu en underinddeling af først og fremmest de deverbale sammensætninger. Den rolle som første led udfylder i en deverbal sammensætning, er influeret af det verbum som andet led er afledt af. Verbets syntaktiske og semantiske valens bestemmer altså første leds rollepotentiale, men der er tilsyneladende yderligere restriktioner end dette på hvilken rolle første led kan have. Ifølge Grimshaw (1990:14-15) gælder der for engelske deverbale sammensætninger følgende restriktioner: "when the head takes more than one internal argument, the least prominent must be inside the compound, and the prominent outside", som det ses i

cookie-baking for children (*children bake cookies)* vs.

**child-baking of cookies*

hvor *cookie* (theme) altså regnes for mindre prominent end *child* (agens) ifølge et hierarki af semantiske roller (hvor agens er den mest prominente svarende til det logiske subjekt), og derfor er 'tilladt' som første led. Imidlertid er dansk tilsyneladende mindre restriktivt på dette område, som angivet i Ørsnes 1999:109 bl.a. med eksemplet:

mediadækningen af damefodbold

hvor agens ses at udfylde førsteled i sammensætningen: *medierne dækker damefodbold.*

I vores materiale forekommer der imidlertid kun en begrænset mængde af deverbale sammensætningstyper, og de synes alle at overholde Grimshaws hierarkiske princip; de udviser med andre ord systematik mht. hvilke verbaltyper der tillader hvilke roller som første led. Således ses rollen agens som første led kun ved intransitive verber som i:

tandlægeindgreb (*tandlægen griber ind*)

Transitive verber derimod forekommer hyppigst med første led som theme (svarende til logisk objekt) som i:

apoteksovertagelse (*nogen overtager et apotek*)

eller som i

båddesigner (*designeren designer både*)

Den sidstnævnte kaldes også nomenagentis i og med at agens er indeholdt i deverbalet.

Endelig ses en tredje rolleudfyldning i form af lokation (svarende til et stedsadverbial) ved både transitive og ditransitive verber som i:

gartneriproduktion (*produktionen foregår i gartneriet*)

De ikke-afledte relationelle sammensætninger kan også underinddeles semantisk ud fra hvilken rolle første led udfylder:

- første led er *patiens*, som i *børnetilskud* (*tilskud for børn*) og *chaufførsats* (*sat for chauffører*)

- første led udfylder en mere vag semantisk rolle, her kaldet *concerns* som i *erstatningsprincip* (*princip vedrørende erstatning*)

Gruppen af ikke-relationelle substantiver i materialet har vi valgt ikke at underinddele yderligere, selv om også de kunne underinddeles i en lang række undertyper afhængig af den rolle som 1. led spiller (hos Järborg 2003 fx 14 forskellige typer for svensk, for engelsk angives hos Mueller 1998 30 forskellige kategorier; se også Johnston & Busa 1999 der anvender Pustejovskys Qualia model til beregning af relationstyper i ikke-relationelle sammensætninger). Det lader nemlig til at de ikke- relationelle substantiver har en svagere tendens til at danne synonyme udtryk; derfor synes der ikke at være det store perspektiv i en underinddeling.

4.5 Automatisk analyse af tekstmaterialet med sprogteknologisk opmærkning

I VID-rapport nr. 3 beskrev vi den automatiske analyse som blev foretaget på CST af Ankiros tekstmateriale (Jongejan, Pedersen, Navarretta 2004). Som udgangspunkt ekspanderede Ankiros søgemaskine på begge søgeord ud fra bøjninger og afledninger, vægtet med en såkaldt T-værdi på hhv. 9 og 8. I modsætning hertil er ikke ekspanderede søgeord som er fundet i søgningerne, tilskrevet en værdi på 10. For et mindre delmateriale er der yderligere ekspanderet ud fra en tesaurus. Det maksimale antal hits for en forespørgsel er sat til 200.

Som nævnt i rapport 3 tokeniseres teksten hvorpå der foretages ordklasseopmærkning (Haltrup 2000); herefter analyseres teksten med en NP-parser skrevet i CASS-formalismen (Haltrup 2002). NP-genkenderen finder alle NPer med op til to præpositionsforbindelser (se Bilag A).

Alle søgeresultater er herefter vægtet med følgende algoritme hvori også T-værdien, som blev beskrevet ovenfor, indgår:

$$vægt = 10(I - N) + \overline{T}$$

hvor

$I =$ antal fundne søgeord (<Tn>-tags)

$N =$ antal NPer i et søgeresultat som indeholder <Tn>-tags

$\overline{T} =$ gennemsnittet af T-værdierne i søgeresultatet.

Hvis der fx er fundet 2 Tn-tags i et søgeresultat sættes *I* til 2. Hvis antallet af NPer med Tn-tags også er 2 sættes N til 2. I dette tilfælde bliver $I - O = 0$ og vægten på 10 udløses derfor ikke idet vi implicit har beregnet at søgeordene ikke er i samme NP. Hvis derimod antallet af NPer med Tn-tags er 1, hedder regnestykket $2 - 1 = 1$, og det betyder implicit at de to Tn-tags må være i samme NP, hvorfor vægten på 10 tilskrives.

Prioriteringen er foretaget således at resultater med de højeste vægte er de bedste (og derfor skal præsenteres først) ifølge hypotesen om at et resultat som samler flere søgeord ord i én NP, er semantisk tættere på det sammensatte ord end et resultat som spreder søgeordene over flere NPer, jf figur 4.1 og 4.2 nedenfor med søgninger der ser hhv. på synonyme udtræk til det sammensatte ord *hudreaktion* og *værdigodkendelse*.

wght	#	<WORD>hudreaktion reaktion hud</WORD><COUNT>14</COUNT>
		3 hits:1 x 9.0, 2 x 19.0
19.0	1368937	pågældende organer . Tilpasningsreaktioner (f. eks . indvandring af makrofager i lungevævet , leverhypertrofi og enzyminduktion , **[NP1 [NP [ADJ hyperplastiske] [N <T9>reaktioner</T9>]] [PRÆP på] [NP [ADJ irriterende] [N stoffer]]]**) . Lokale **[NP1 [NP [N <T9>reaktioner</T9>]] [PRÆP i] [NP [N <T9>huden</T9>]]]** på grund af gentagen dermal anvendelse af et stof , som bedre klassificeres med R38 »Irriterer **[NP [N <T9>huden</T9>]]**«. Hvor der
19.0	651711	pågældende organer . Tilpasningsreaktioner (f. eks . indvandring af makrofager i lungevævet , leverhypertrofi og enzyminduktion , **[NP1 [NP [ADJ hyperplastiske] [N <T9>reaktioner</T9>]] [PRÆP på] [NP [ADJ irriterende] [N stoffer]]]**) . Lokale **[NP1 [NP [N <T9>reaktioner</T9>]] [PRÆP i] [NP [N <T9>huden</T9>]]]** på grund af gentagen dermal anvendelse af et stof , som bedre klassificeres med R38 »Irriterer **[NP [N <T9>huden</T9>]]**«. Hvor der
9.0	626651	eller **[NP1 [NP [V_PARTC_PAST gentagen] [N berøring]] [PRÆP af] [NP [N <T9>huden</T9>]]]** eller slimhinderne . 8 . Sensibiliserende : Stoffer og produkter , som ved indånding eller **[NP1 [NP [N optagelse]] [PRÆP gennem] [NP [N <T9>huden</T9>]]]** kan fremkalde overfølsomheds-**[NP [N <T9>reaktion</T9>]]** , således at der ved yderligere eksponering af stoffet eller produktet fremkommer karakteristiske symptomer . 9 . Kræftfremkaldende : Stoffer og produkter

Figur 4.1: Søgehits der inkluderer *hud* og *reaktion* prioriteret på basis af sprogteknologiske analyser: begge søgeord i samme navnefrase=høj vægt

vægt		<WORD>værdigodkendelse godkendelse værdi</WORD><COUNT>200</COUNT>
19.0	549290	I **[NP2 [NP1 [NP [N forbindelse]] [PRÆP med] [NP [N <T9>godkendelse</T9>]]] [PRÆP af] [NP [V_PARTC_PAST forhøjede] [N <T9>værdier</T9>]]]** skal det fastlægges , hvor ofte der skal udtages prøver til kontrol af , at dispensationen ikke overskrides .
19.0	549291	I **[NP2 [NP1 [NP [N forbindelse]] [PRÆP med] [NP [N <T9>godkendelse</T9>]]] [PRÆP af] [NP [V_PARTC_PAST forhøjede] [N <T9>værdier</T9>]]]** træffer amtsrådet efter vandforsyningslovens § 62 , stk . 3 , beslutning om , hvilke foranstaltninger , der skal træffes i den periode , hvor
19.0	718589	Dispensation fra kvalitetskravene . Ikke-almene anlæg skal som udgangspunkt overholde de samme kvalitetskrav som almene anlæg , og **[NP1 [NP [N <T9>godkendelse</T9>]] [PRÆP af] [NP [V_PARTC_PAST forhøjede] [N <T9>værdier</T9>]]]** kan kun meddeles , når dette er anført i bekendtgørelsens bilag 1 , og når **[NP1 [NP [N forudsætningerne]] [PRÆP for] [NP [N_GEN amtsrådets] [N <T9>godkendelse</T9>]]]** er
19.0	549837	af 4. juni **[NP [NUM 1971] [N godkendte] [N <T9>værdi</T9>]]** . Landsskatteretten bemærkede , at skatterådet ved den i skrivelsen af 4. juni **[NP2 [NP1 [NP [NUM 1] [N meddelte] [N <T9>godkendelse</T9>]] [PRÆP af] [NP [N <T9>værdien</T9>]]] [PRÆP af] [NP [N B] [EGEN Hovedgård]]]** måtte anses for at have taget fornødent forbehold , og da skatterådet endvidere ikke kunne give bindende
19.0	549292	**[NP1 [NP [N <T9>Godkendelse</T9>]] [PRÆP af] [NP [V_PARTC_PAST forhøjede] [N <T9>værdier</T9>]]]** meddeles for et bestemt tidsrum , som højst kan være 5 år .
19.0	104281	til kl . 22 , en dag ugentlig dog til kl . 24 . Klageren gjorde over for landsskatteretten til støtte for **[NP1 [NP [N <T9>godkendelse</T9>]] [PRÆP af] [NP [PRON_DEMO den] [ADJ selvangivne] [N <T9>værdi</T9>]]]** af den fra arbejdsgiveren modtagne frie kost gældende , at han kun havde modtaget ét måltid om dagen i
19.0	77467	medføre uacceptable skattemæssige konsekvenser , hvis kurs 80 ikke blev godkendt , idet denne kurs havde været **[NP2 [NP1 [NP [PRON_UBST en] [N forudsætning]] [PRÆP for] [NP [N_GEN skattemyndighedernes] [N <T9>godkendelse</T9>]]] [PRÆP af] [NP [N <T9>værdierne</T9>]]]** . I pådømmelsen deltog dommerne Sigrid Ballund , Jochimsen og Jakob Jakobsen (kst.). Det måtte lægges til grund , at sagsøgerens omdannelse
9.0	996600	edb-rettigheder , ligesom der ikke af anklagemyndigheden er taget stilling til indholdet og **[NP [N <T9>værdien</T9>]]** af det omhandlede edb-program . " .Det ses heraf , at **[NP1 [NP [N_GEN politiets] [V_PARTC_PAST påståede] [N <T9>godkendelse</T9>]] [PRÆP af] [NP [N softwaren]]]** er forsynet med en række væsentlige forbehold . 2. 4 . Økonomiske transaktioner
9.0	594247	blev erlagt i form af en græsningsret for svigerfaderens kreaturer på klagerens ejendom . **[NP1 [NP [N <T9>Værdien</T9>]] [PRÆP af] [NP [PRON_DEMO de] [ADJ købte] [N kreaturer]]]** efter de af **[NP1 [NP [N_GEN statens] [N_GEN lignings] [N direktorat]] [PRÆP med] [NP [N_GEN ligningsrådets] [N <T9>godkendelse</T9>]]]** fastsatte normalpriser var på købstidspunktet 1. 100 kr . eller 700 kr . lavere end
9.0	569995	afsnit A. C. 1. 2 , note 5) , finder reglen om , at **[NP1 [NP [PRON_DEMO den] [ADJ samlede] [N <T9>værdi</T9>]] [PRÆP af] [NP [N ejerenskapitalpensionsordninger]]]** skal afgiftsberigtiges samtidig , kun anvendelse for ordninger , der omfattes af **[NP [N <T9>godkendelsen</T9>]]** . Hvis f. eks . en skatteyder med en godkendt pensionsalder på 55 år har
9.0	1362834	3 dele : Beløbsmæssig begrænsning , begrænsning af den kreds af foreninger mv . , der kan opnå **[NP [N**

		<T9>godkendelse</T9>]] til at modtage fradragsberettigede ydelser , samt [NP2 [NP1 [NP [N begrænsning]] [PRÆP af] [NP [PRON_DEMO den] [ADJ skattemæssige] [N <T9>værdi</T9>]]] [PRÆP af] [NP [N fradraget]]] . Af bevismæssige grunde må det , for at der kan gives fradrag , kræves

Figur 4.2: Søgehits der inkluderer *godkendelse* og *værdi* prioriteret på basis af sprogteknologiske analyser: begge søgeord i samme navnefrase=høj vægt

Figur 4.1 og 4.2 illustrerer eksempler på at de hits hvor begge søgeord er i samme NP, generelt ligger semantisk tættere på forespørgslen end de øvrige hits.

4.6 Evaluering

4.6.1 Metode

For at få nogle tal for hvor godt hypotesen holder for det output som den sprogteknologiske beregning genererer, har vi ønsket at beregne precision og recall på materialet. Precision er et mål for mange af de fundne synonymfraser der er korrekte, mens recall angiver hvor mange af de korrekte synonymfraser der er fundet. I og med at alle tekstudsnittene på en vis måde er 'fundne' i og med at de er genereret af Ankiros søgemaskine, har vi fastlagt en tærskel for hvad vi opfatter som 'fundet' i den specifikke, sprogteknologiske sammenhæng. Værdien 10 er sat som tærskel i og med at hits med værdi under 10 ikke har begge søgeord inden for samme NP. Hits med værdi over 10 er altså registreret som fundne, svarende til kolonne E i figur 4.3.

Ønskede resultater er identificeret manuelt ud fra hvorvidt vi har ment at et givent tekstudsnit indeholdt en frase der var synonym til det sammensatte ord. I overensstemmelse med gængs praksis for beregning udregnes de to tal på følgende måde hvor Ø står for ønskede og F for fundne:

$$\text{Recall} = |\, Ø \cap F \,| \,/\, |\, Ø \,|$$

Recall er altså beregnet ud fra fællesmængden af de fundne og de ønskede divideret med antallet af ønskede.

$$\text{Precision} = |\, Ø \cap F \,| \,/\, |\, F \,|$$

Precision derimod beregnes ud fra fællesmængden af de fundne og de ønskede divideret med antallet af fundne.

I figur 4.3 svarer dette til en udregning på følgende kolonner:

$$\text{Recall} = (E\text{-}G) \,/\, (E+H\text{-}G)$$

$$\text{Precision} = (E\text{-}G) \,/\, E$$

A: Sammensat ord	**B: 1. led er afledning el. valensbærende**	**C: semantisk rolle på 2. led**	**D: antal tekstudsnit i alt**	**E: antal fundne (score over 10)**	**F: antal ikke fundne (score under 10)**	**G: antal ikke ønskede i fundne**	**H: antal ønskede i ikke fundne**	**I: sproglige karak. ved fundne**	**J: årsag til ikke ønskede i fundne**	**K: årsag til ønskede i ikke fundne**	**L: precision ((E-G) / E)**	**M: recall (E-G) / (E+H-G)**
acceptinstrument			0	0	0	0	0				-	-
adgangsproblem	*Dn1tCL (STO-kode)*	CONCERNS theme	39	3	36	0	2	omkring		kompleks NP perfektum participium	1	0,6
administrationsopgørelse	*Dn2GPnw-over* *Dn2GPn-af*		89	1	88	0	2	over		i samme VP	1	0,3
adsorptionskapacitet	*Dn1Pn-påDn1Pni-til: Dn1G*		0	0	0	0	0				-	-
advarselsplan			0	0	0	0	0				-	-
aerosoldannelse	deverb *Dn1GDn1Pn-af*	theme/agent	6	3	3	0	3	af		koordineret NP	0,5	0,5

Figure 4.3: Uddrag af kolonneoversigt over resultater

4.6.2 Sammensatte ord hvor der ingen hits er

Som nævnt er der foretaget 1000 søgninger med forskellige sammensatte ord i splittet form. Af disse er der for 330s vedkommende ingen hits, dvs. søgning med det sammensatte ord i splittet form har ikke givet nogle søgeresultater (noteret som 0 under tekstudsnit i kolonne D). Dette kan have to forklaringer, hvoraf den første lader til at være gældende for størstedelen af materialet:

- det sammensatte ord er leksikaliseret i en udstrækning så det ikke forekommer i splittet form; dette er den mest plausible forklaring i tilfælde som *aftrækskanal*, *autovaskehal*, *beboelseskompleks*, *blodbudding* og *bjerglandskab*, altså typisk sammensætninger hvor andet led er ikke-relationelt og refererer til konkrete genstande.

- det sammensatte ord er temmelig sjældent evt. fordi det er produktivt konstrueret, dvs. opstået i en ganske bestemt og meget begrænset kontekst; som en følge heraf ses synonyme udtryk stort set ikke. Dette er den mest plausible forklaring i tilfælde som *bevistema*, *blandingshonorar*, *farveindskrænkning*, og *feriegirosystem*.

4.6.3 Evaluering af beregningerne på synonymer til deverbale sammensætninger

Ud af de resterende 670 søgninger vedrører de 32 deverbale sammensætninger (noteret som 'deverbal' i kolonne B). Heraf falder 18 inden for gruppen hvor første led udfylder rollen *theme* (kolonne C). Vores hypotese var at især disse meget systematisk forekom med en synonym søsterkonstruktion i form af et NP med præpositionsforbindelse indledt med *af*, som i *apoteksovertagelse – overtagelse af apotek*. Dette viste sig at holde stik; i materialet er disse typer synonymer generelt *meget* hyppigt forekommende hvorfor man synes at kunne tale om en helt systematisk og hyppigt forekommende omskrivning for disse ords vedkommende. Ligeledes viser omskrivning med genitiv sig at danne et klart mønster; *biblioteksdrift – bibliotekernes drift*.

For hele gruppen af deverbale sammensætninger får vi følgende tal for precision og recall:

Precision: 0,9 - Recall: 0,6

Som det ses er resultatet for precision meget tilfredsstillende, hvorimod recall lægger op til en nærmere undersøgelse af hvorfor så relativt mange ønskede hits tilsyneladende ikke har opnået den forventede score. Her viser der sig to primære årsager, hvoraf den første dog er langt den hyppigste:

- Analysefejl i NP-genkenderen: NP-genkenderen er ikke tilstrækkelig udbygget til at kunne håndtere de relativt komplekse NPer der forekommer i teksttypen. Især skaber koordinerede NPer problemer.

- Der optræder en del *sætninger* som vi har skønnet er synonyme til det sammensatte ord; altså relevante hits som modsiger hypotesen om at kun hits

med begge søgeord i samme NP er synonyme, nemlig eksempler som *badevandskvalitet – badevandet er af en bestemt kvalitet*.

Eksempler på mangelfuld NP-analyse gives nedenfor; først og fremmest er det som nævnt koordinerede NPer der giver problemer:

> *gartneriproduktion*: [NP2 [NP1 [NP [N Bygninger]] [PRÆP til] [NP [ADJ erhvervsmæssig] [N <T9>produktion</T9>]]] [PRÆP vedrørende] [NP [N landbrug]]] , skovbrug , [NP [N <T9>gartneri</T9>]] ,

> *arveerhvervelse*: [NP [N <T9>Erhvervelse</T9>]] og afståelse af en fordring eller en kontrakt ved gave , [NP [N <T9>arv</T9>]] eller arveforskud sidestilles i denne lov med køb henholdsvis salg .

> *apoteksovertagelse*: overføres endvidere bl .a. [NP1 [NP [N udgifter]] i forbindelse [PRÆP med] [NP [N <T9>overtagelse</T9>]]] , nyanlæg eller [NP1 [NP [N flytning]] [PRÆP af] [NP [PRON_UBST et] [N <T9>apotek</T9>]]] , såsom flytteudgifter for apotekerens husstand og bohave , honorar for sagførerbistand m. m. Landsskatteretten fandt ikke , at der

Men også appositionslignende NPer giver anledning til fejl:

> *arveerhvervelse*: Hovedreglen om afgiftsfrihed i § 8 , stk . 1 , omhandler de i henhold til lovens foregående bestemmelser [NP [ADJ afgiftspligtige] [N <T9>erhvervelser</T9>]] , herunder foruden [NP [ADJ egentlig] [N <T9>arv</T9>]] eksempelvis de i §§ 4 og 5 nævnte dødsgaver og gaver , hvoraf giveren har forbeholdt sig indtægten for

Ligesom der forekommer relevante NPer med mere end to præpositionsforbindelser.

Som det kan ses af Bilag A, er NP-genkenderen først og fremmest designet til at kunne identificere simple NPer i og med at kompleksitetsgraden kun indbefatter tre niveauer i analysen: NP, NP1 og NP2 refererende til NPer med hhv. ingen, en eller to præpositionsfraser. Det må overvejes at udvide NP-genkenderen til at kunne håndtere en større kompleksitetsgrad; problemet er blot at fx området koordination er meget komplekst og åbner op for utallige muligheder for hvilke slags syntaktiske størrelser der kan koordineres. I praksis kan kun en meget raffineret semantik helt entydigt afklare hvilke led der er sideordnede med hvilke (jf. beskrivelsen af koordination på dansk i Fersøe & Kirchmeier-Anderesen 1989). Det analyserede tekstmateriale, der som nævnt indeholder mange bekendtgørelser o. lign., er generelt karakteriseret ved at indeholde meget varierende brug af koordination.

Vi har derfor manuelt gennemgået resultaterne i de tilfælde hvor NP-genkenderen overser strukturerer, for at genberegne recall på en korrekt NP-genkendelse. Med en sådan korrektion får vi en recall på 0.7 hvilket må siges at være en mærkbar forbedring. Dog må man regne med at precision til en vis grad kan blive påvirket negativt såfremt

NP-genkenderen udvides. Dvs. irrelevante hits hvor NPet ikke er synonymt med det sammensatte ord, vil forekomme, som det ses i eksemplet nedenfor hvor tekstudsnittet med et koordineret NP ikke har noget med *behandlingsanvisning* at gøre (i den medicinske forstand):

> *behandlingsanvisning*: [NP [N <T9>anvisning</T9>]] og [NP1 [NP [N <T9>behandling</T9>]] [PRÆP af] [NP [N affald]]]

Konkluderende for deverbale sammensætninger må man dog sige at hypotesen holder: NPer med begge søgeord i samme NP er generelt synonyme med det sammensatte ord. Testresultaterne er stort set tilfredsstillende (i hvert fald med en udvidet NP-genkender), og der synes således at være god mening i at ekspandere deverbale sammensætninger til NPer med samme søgeord, idet søgningen vil forbedres markant. Automatisk registrering af hvorvidt et sammensat ord er deverbalt eller ej, kan udledes af fx STO-ordbasen (Braasch & Pedersen 2002).

4.6.4 Evaluering af beregningerne på synonymer til andre relationelle sammensætninger

Også deadjektivale og andre relationelle sammensætninger følger i materialet et forudsigeligt mønster som kan udnyttes i søgning. For disse ord alene (15 undersøgte i alt) er der for søgninger på NPer som synonym til sammensat ord følgende resultat:

> Precision: 0,9 - Recall på 0,8

NP-parallellen til disse relationelle sammensætninger har typisk en leksikalsk styret præposition, som i *pligt til bistand* som synonym til *bistandspligt* og *risiko for tyveri* som synonym til *tyveririsiko*. De specifikke præpositioner der anvendes følger med andre ord ikke et gennemskueligt grammatisk mønster, men synes primært at være leksikalsk motiveret, hvorfor de kan udledes automatisk ud fra STO-basens syntaktiske oplysninger (for *pligt* er fx angivet koden *Dn2GPni-til* som står for optionel relationel genitiv, samt optionel præpositionsforbindelse med præpositionen *til* efterfulgt af et NP eller infinitivsætning).

4.7 Konkluderende bemærkninger

Synonymer til ikke-relationelle sammensætninger er generelt langt sværere at identificere med sprogteknologiske midler end de øvrige typer. Generelt kan man konkludere at ikke-relationelle sammensætninger har en svagere tendens til at have den type parallelle synonymer som vi har set for de andre to kategorier, og at når de har det, er de generelt mindre forudsigelige. Når vi ikke har udregnet precision og recall specifikt for denne gruppe, skyldes det at en stor andel af den slet ikke har genereret nogle hits (jf. afsnit 5.5.3).

For hele materialet til sammen, dvs. både relationelle og ikke-relationelle sammensætninger, er resultatet som følger:

Precision: 0,8 - Recall: 0,5

Som det kan udledes er det primært de ikke-relationelle sammensætninger der påvirker resultatet negativt. Konklusionen må derfor være at det ikke uden videre kan betale sig at søge på 'splittede' sammensatte ord af denne type; i hvert fald må man i så fald regne med nogen støj. For deverbale og andre relationelle sammensætninger er der derimod gode perspektiver i at søge efter NPer med begge dele af det sammensatte ord, og disse kan som nævnt identificeres automatisk via STO-ordbasen.

Et interessant aspekt, som vi kun kan berøre sporadisk her, er ekspansion af søgestrengen til at fange eksempler på ordsynonymer og specialisering. I ganske mange tilfælde ser vi i materialet at der optræder specialiserede termer for et af søgeordene. Havde vi udnyttet en tesaurus eller en ontologi og ekspanderet på søgestrengen til synonymer og underbegreber, ville man have høstet en lang række relevante hits: *valutaomregning - omregning af lånevaluta*, *biblioteksdrift - drift af folkeskolebiblioteker, besætningsudskiftning - udskiftning af svinebesætning*. Muligvis ville en sådan ekspansion også forbedre resultaterne for de ikke-relationelle sammensætninger. Det virker sandsynligt at de sammensætninger der synes leksikaliserede (fx *blodbudding*), netop kun ses i splittet form når man vil specificere 2. led yderligere som fx i *budding af lammeblod*.

5 Prototype til søgning og dokumenthåndtering med ontologi og metadata

For at afprøve søgning på domænespecifikke dokumenter ved hjælp af nogle af de lingvistiske og ontologiske oplysninger og metadata som er blevet beskrevet i de foregående kapitler, har vi i samarbejde med Ankiro opbygget en prototype til søgning i Zaccos standarddokumenter. Ankiro har implementeret prototypens søgemaskine. Prototypen køres gennem internettet og anvender browseren Netscape som brugergrænseflade.

Systemet søger efter indholdsord i tekster og i teksternes XML-metadata og ved at ekspandere de søgte ord med lingvistiske og semantiske oplysninger. Følgende oplysninger udnyttes i søgning:

- Oplysninger om ordenes bøjningsformer. I søgning anvendes den omvendte metode af lemmatisering i det der tages udgangspunkt i en fuldformsordbog.

- Oplysninger om ordenes synonymer og hyponymer, samt oplysninger om andre ord som er relateret til søgeordene via nogle mindre stærke relationer, som fx nærsynonymi og nogle af de tværgående relationer som er diskuteret i afsnit 3.6.

- Oplysninger om led i sammensætninger, således at første led bindes til sammensætningen via en tværgående relation, mens andet led bindes til sammensætningen som dennes overbegreb. Fx er *patentansøgning* relateret til *ansøgning* via en IS-A relation, mens den er bundet til *patent* via en tværgående og derfor mindre stærk relation.

Prototypen returnerer tekster hvor søgeordene, eller ord relateret til søgeordene, er blevet fundet. Systemet søger på indholdsord, og ikke, som almindelige søgemaskiner, på strenge. Søgeresultaterne bliver tilknyttet forskellig vægt afhængigt af de oplysninger der ligger til grund for søgningen. Højest vægt tildeles resultater opnået ved ekspansion på morfologiske oplysninger (bøjningsformene), samt resultater fundet ved at følge synonymiske og hyponymiske relationer. Resultater opnået ved ekspansion på indirekte (tværgående) relationer tilskrives en lavere vægt. Afstanden mellem de relevante indholdsord i de returnerede tekster påvirker også den vægt resultaterne får. I prototypen returneres resultaterne i prioriteret rækkefølge, således at de bedste resultater, dvs. resultaterne tilknyttet den højeste vægt, står forrest.

Søgemaskinen er forbundet til en database hvor oplysningerne om ordenes morfologi og semantiske relationer bliver kodet. Databasen er også udviklet af Ankiro.

Prototypen søger på Zaccos standarddokumenter indenfor patentdomænet. Udover disse danske dokumenter indeholder prototypen seks engelske og fire norske dokumenter inden for samme domæne. Standarddokumenterne, som er i Word, er automatisk konverteret til simpelt tekstformat, se Jongejan et al. (2004). Dernæst har vi automatisk konverteret disse tekster til XML-dokumenter som indeholder metadata med oplysninger om teksternes dato, forfatter, emneord mm. Som metadata har vi valgt

Dublin Core's metadata. XML-skabelonen som er blevet anvendt i dokumenterne kan ses i figur 5.1. Elementerne, hvis navn starter med "dc:" er Dublin Core-metadata.

```
<document xmlns:dc="http://purl.org/dc.elements/1.1">
    <dc:Title></dc:Title>
    <dc:Creator></dc:Creator>
    <dc:Subject></dc:Subject>
    <dc:Description></dc:Description>
    <dc:Publisher></dc:Publisher>
    <dc:Contributor></dc:Contributor>
    <dc:Date></dc:Date>
    <dc:Type></dc:Type>
    <dc:Format></dc:Format>
    <dc:Identifier> </dc:Identifier>
    <dc:Language></dc:Language>
    <dc:Relation></dc:Relation>
    <dc:Coverage></dc:Coverage>
    <dc:Rights></dc:Rights>
    <body></body>
</document>
```

Figur 5.1: XML-skabelon med Dublin Core-metadata

Følgende Dublin-Core metadata er blevet udfyldt i dokumenterne: *dc:Title*, *dc:Creator*, *dc:Subject*, *dc:Date*, *dc:Identifier*, *dc:Language*.

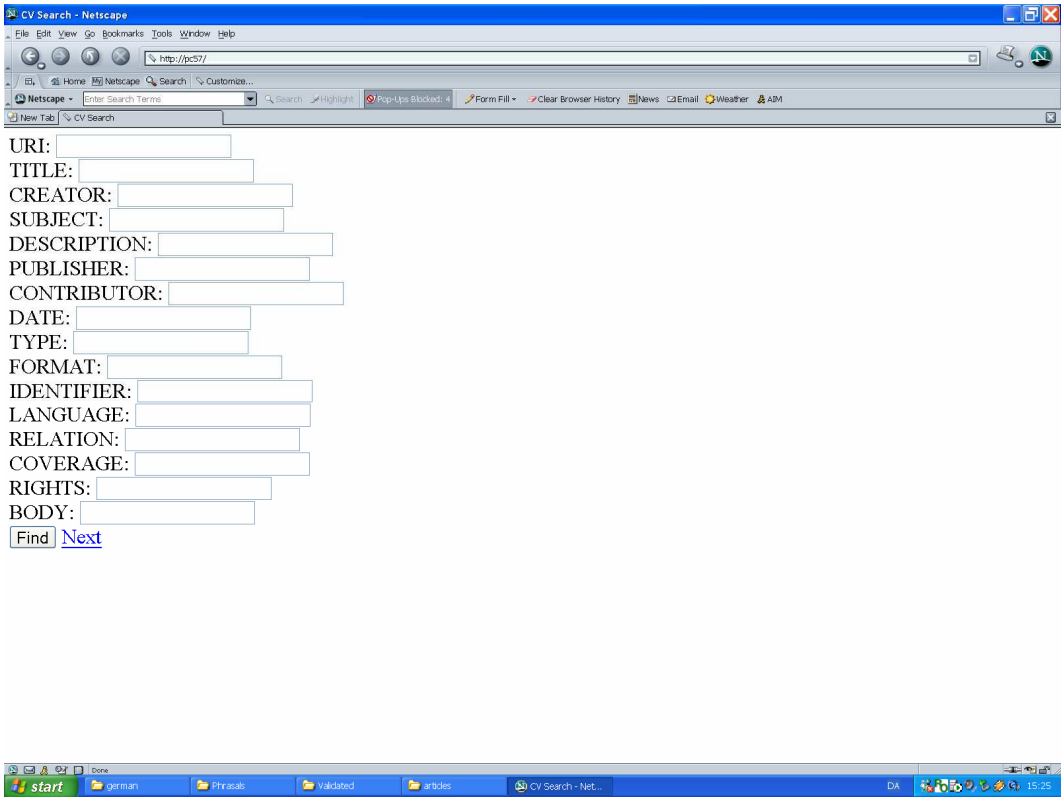
dc:Title indeholder dokumentets titel, *dc:Creator* angiver den der skrev dokumentet, *dc:Date* indeholder dato for skabelsen af dokumentet, *dc:Identifier* indeholder internetadressen for selve dokumentet på internettet (i det aktuelle tilfælde på projektets intranet), *dc:Language* angiver det sprog dokumentet er skrevet i. *dc:Subject* indeholder dokumentets nøgle/emneord. Disse nøgleord blev automatisk uddraget fra dokumenterne med nogle af de teknikker beskrevet i kapitlet 2 om indeksering. Formålet med at opmærke dokumenter med nøgleord er at begrænse fritekstsøgning og give de ansatte mulighed for at søge i dokumenterne gennem emneord.

Indholdet af XML-elementet *body* udgøres af teksten i det aktuelle dokument.

Ankiros database indeholder oplysninger om almensproglige ord, mens vi har tilføjet oplysninger om domænespecifikke termer. Disse ord og termer er blevet semiautomatisk uddraget af Zaccos dokumenter som beskrevet i Jongejan et al. (2004). For hvert ord/term er der indsat morfologiske oplysninger, samt semantiske oplysninger uddraget fra domæneontologien beskrevet i kapitel 3. Alle hyponymer og synonymer er blevet indsat i databasen, og oplysninger om alle led i sammensætninger blev kodet. For at begrænse antallet af tværgående semantiske relationer i databasen, har vi brugt et antal prototypiske forespørgsler. Nogle af disse forespørgsler har vi selv produceret,

andre har vi fået af Zacco. Kun de ontologiske tværgående relationer der var nødvendige til at opfylde disse forespørgsler, er blevet kodet.

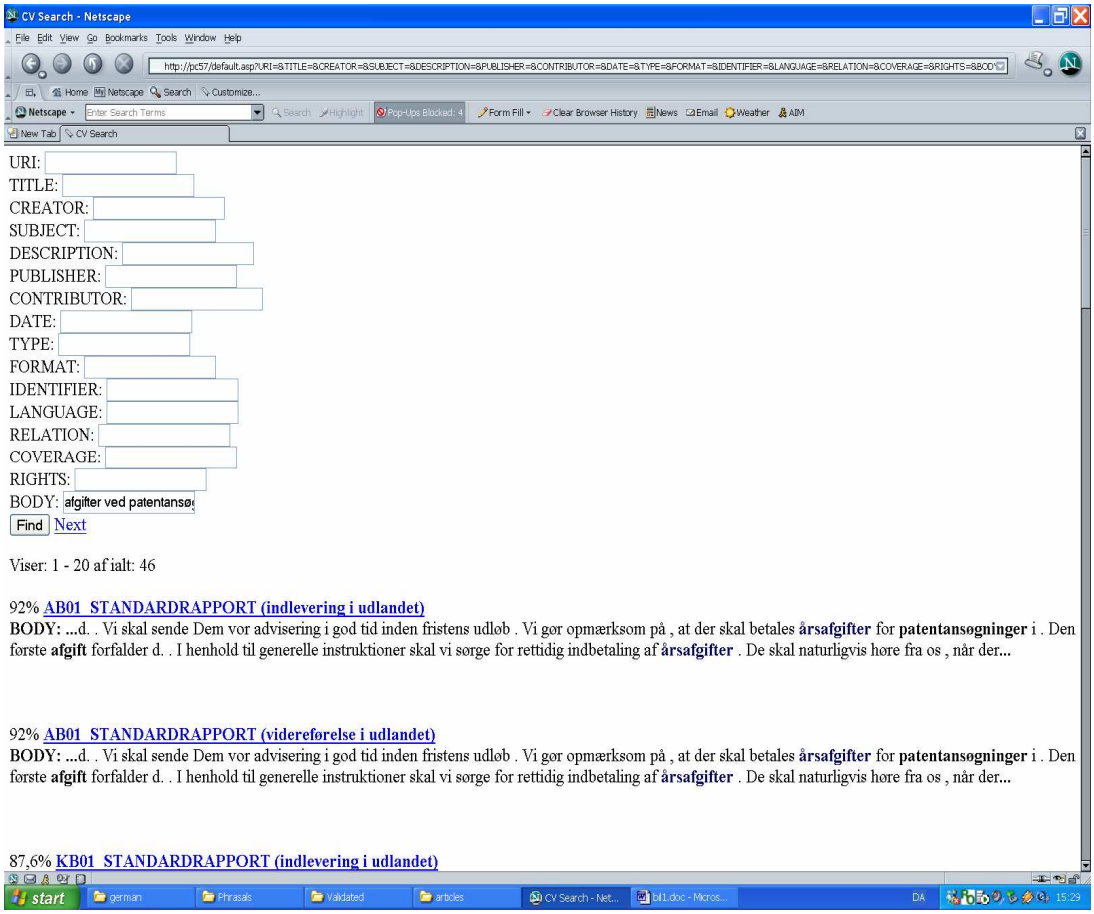
I figur 5.2 vises prototypens grænseflade i Netscape.



Figur 5.2: prototypens brugergrænseflade

Felterne i grænsefladen svarer til XML-elementerne i teksterne. URI-feltet indeholder samme oplysninger som dc:Identifier, dvs. internetadressen for dokumenterne. Brugeren kan søge i en eller flere felter ved at indtaste søgeordene efter felternes navne. Søgemaskinen anvender kun ekspansionsmekanismen på ord i BODY-feltet.

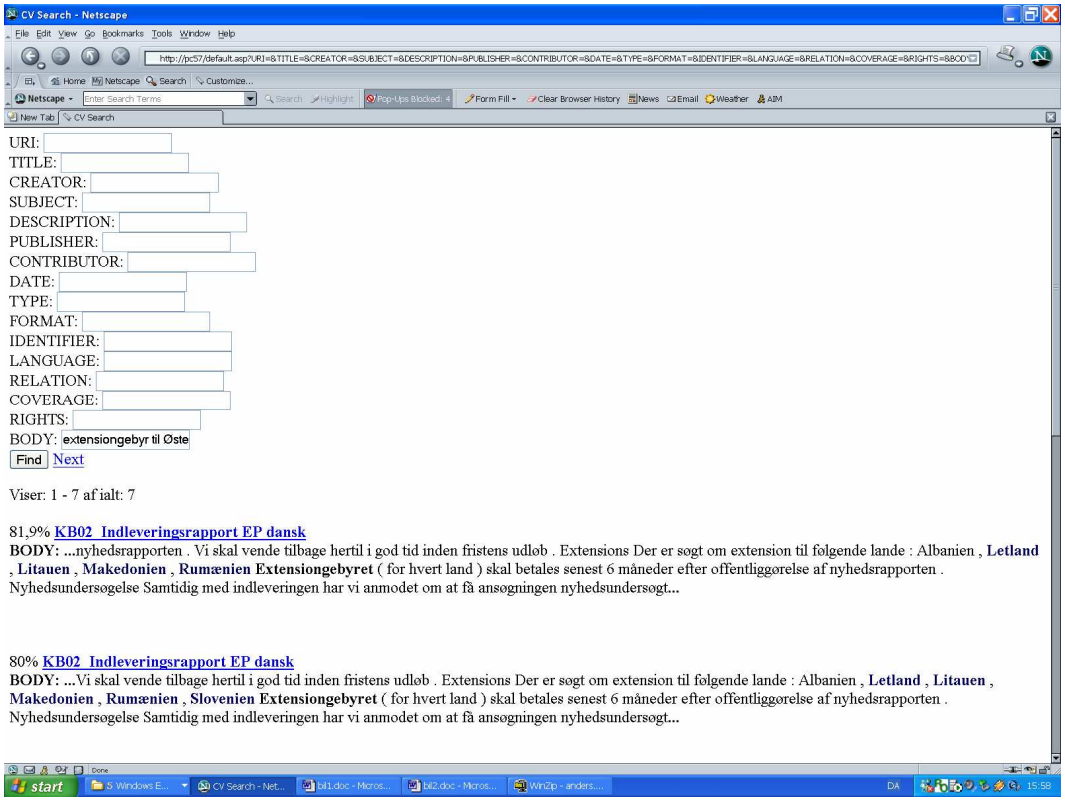
I det følgende vises resultaterne for forskellige typer forespørgsler i prototypen. I figur 5.3 vises skærmbilledet af de højest vægtede resultater af forespørgslen ”afgifter ved patentansøgninger” i dokumenter skrevet på dansk (felt Language=da).



Figur 5.3: resultater af forespørgslen ”*afgifter ved patentansøgninger*”

Søgemaskinen har ikke fundet tekster hvor begge indholdsord *afgifter* og *patentansøgninger*, (i alle mulige bøjningsformer), optræder i samme sætning. Derfor er de bedste resultater opnået ved at ekspandere via indholdsordenes semantiske relationer. I de første (og bedste) to søgeresultater, vist i figur 5.3, returneres tekster hvor *patentansøgninger* og en hyponym for det første indholdsord *afgifter*, i dette tilfælde *årsafgifter*, optræder sammen. Begge resultater har fået tildelt en vægt på 92%.

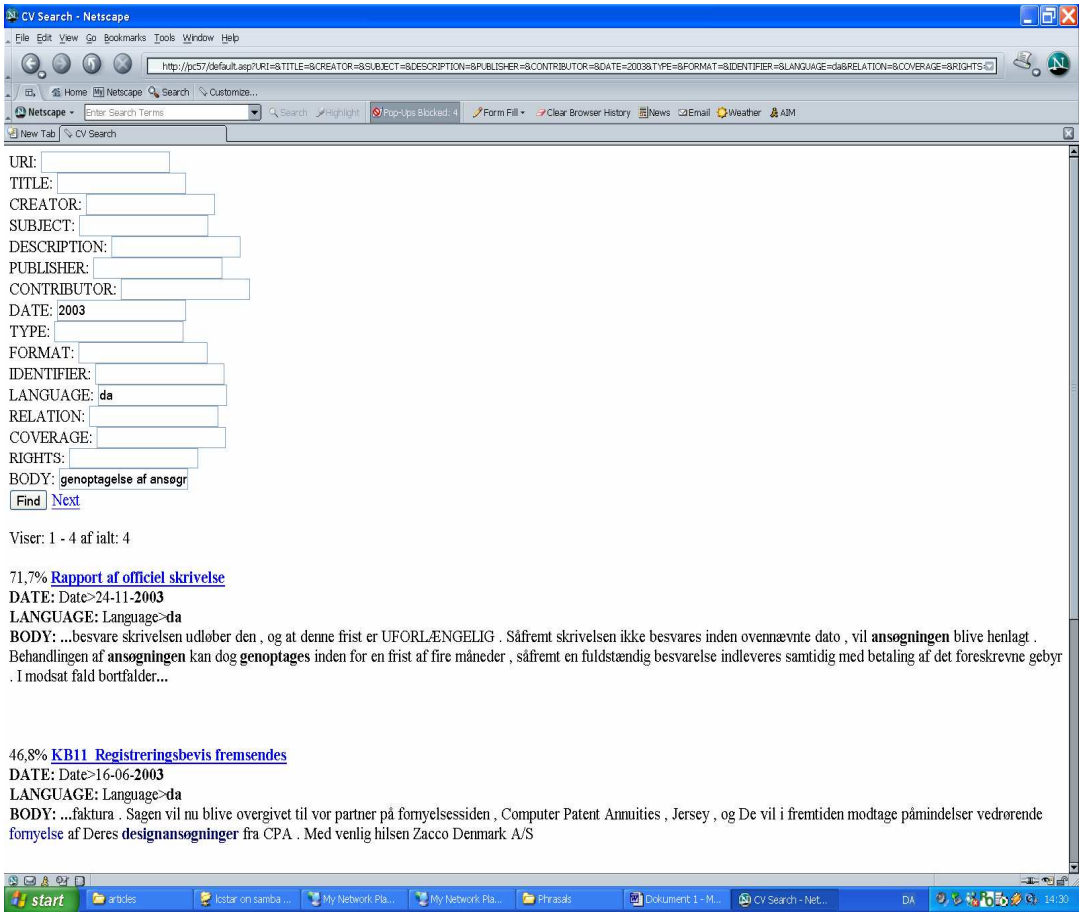
I figur 5.4 vises de bedste resultater fra forespørgslen ”extensiongebyr til Østeeuropa”.



Figur 5.4: resultater af forespørgslen ”*extensiongebyr til Østeuropa*”

Heller ikke i 5.4. fandt systemet nogen tekster med begge indholdsord fra forespørgslen i samme sætning. Derfor returneres som bedste resultater to tekster hvor det første søgeord, *extensiongebyr*, i bøjningsformen *extensiongebyret*, optræder sammen med hyponymerne for det andet indholdsord, *Østeuropa*. I det aktuelle tilfælde er hyponymerne navne på østeuropeiske lande. Disse svarer til instanser af klassen Østeuropa i ontologien.

I skærmbilledet i figur 5.5 vises resultater fra forespørgslen ”genoptagelse af ansøgning i danske dokumenter skrevet i 2003 ” (Language=da, Date=2003).

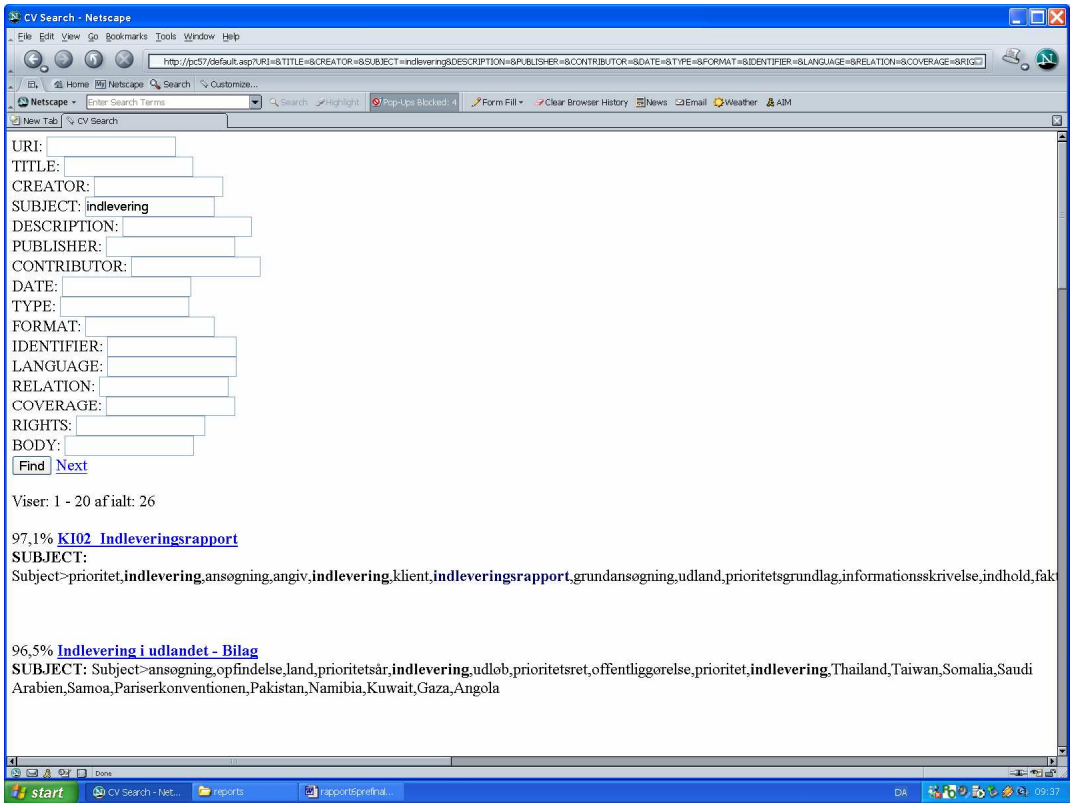


Figur 5.5: resultater af forespørgslen "*genoptagelse af ansøgning*"

I det første søgeresultat i 5.5 (vægt på 71 %) bliver der returneret en tekst som indeholder verbet *genoptage*, som er relateret til substantivet *genoptagelse*. Det andet indholdsord i forespørgslen er ekspanderet morfologisk (*ansøgningen*).

I det andet søgeresultat (vægt 46.8%) har systemet ekspanderet på overbegrebet til det andet led i det sammensatte søgeord *designansøgninger*, samt ekspanderet fra *genoptagelse* til *fornyelse* via en tværgående semantisk relation. Dette relevante resultat ville aldrig være blevet fundet af en søgemaskine baseret på simpel strengeanalyse.

I skærmbilledet i figur 5.6 vises resultaterne fra en søgning på emneord i SUBJECT-feltet. Ordet der søges på i eksemplet er *indlevering*.



Figur 5.6: søgning på emneord indlevering

5.1 En første evaluering af prototypen og nogle konkluderende bemærkninger

Vi testede prototypen med en repræsentant fra Zacco. Den første test bestod i at lade hende analysere søgeresultaterne fra de forespørgsler som Zacco tidligere havde sendt til os. Hun fandt at resultaterne var både præcise og relevante. I den anden test skulle Zaccos repræsentant afprøve prototypen med nye forespørgsler. Også resultaterne af denne test blev fundet præcise og relevante. I den tredje test afprøvede vi selv prototypen med tilfældige forespørgsler. I denne sidste test fandt vi enkelte tilfælde hvor de lavest vægtede søgeresultater ikke var relevante efter vores mening. Dette skyldtes ekspansionen på ord som tilhører det almensproglige domæne, men som også anvendes i det specifikke patentdomæne og i andre domæneområder som Ankiro har arbejdet med. For at løse dette problem, kan man markere de forskellige domæner som ord anvendes i og de tilsvarende domænespecifikke relationer, men der vil altid være tilfælde som er svære at behandle korrekt uden at tage den aktuelle kontekst i betragtning.

Det er indlysende at VID-prototypen baseret på en søgemaskine der ekspanderer på både lingvistiske og semantiske oplysninger, giver mange flere og mere nuancerede resultater end almindelige strengbaserede søgemaskiner. Det fremgår også klart, fra

vores første evaluering af systemet, at præcisionen af de opnåede resultater er meget høj, og at ontologisk og lingvistisk viden bør derfor spille en vigtig rolle i et domænespecifikt dokumenthåndteringssystem. Det er dog stadig uklart hvorvidt søgeresultaterne afhænger af domænestørrelsen og typen. Det bør også undersøges videre, hvor vidt søgeekspansion via tværgående semantiske relationer påvirker søgeresultaterne i større og mere komplekse domæner. I vores test fandt vi enkelte irrelevante resultater opnået gennem semantisk ekspansion af ord som tilhører både det almensproglige vokabular og patentdomænet. Fremtidigt arbejde bør derfor også bestå i afprøvning af ontologibaseret søgning på større domæner og evaluering af forholdet mellem præcision og relevans af resultater opnået gennem ekspansion via tværgående semantiske relationer. Et andet emne som vi ikke nåede at tage fat i arbejdet med prototypen, er hvorvidt man kan forbedre søgeresultater i tilfælde af tvetydige søgeord, og om det er muligt at udnytte ontologisk viden om relaterede begreber til at disambiguere konteksten i tekster.

6 Konklusion

I denne rapport har vi præsenteret de resultater vi har opnået i den del af VID-projektet der omhandler indholdsbaseret søgning og dokumenthåndtering med sprogteknologi. Der er tale om en række eksperimenter der er igangsat dels af de involverede teknologivirksomheder Navigo og Ankiro, dels af Zaccos 'case' vedrørende et fleksibelt og sprogbaseret dokumenthåndteringssystem.

I kapitel 2 undersøgte vi en sprogteknologibaseret tilgang til automatisk generering af nøgleord. Vi viste hvordan man ved at raffinere eksisterende metoder med mere sprogteknologisk viden, kunne opnå langt bedre resultater for dansk.

I kapitel 3 beskrev vi de sprogteknologiske aspekter ved ontologiopbygning, og vi gjorde rede for hvorfor en tekstbaseret tilgang udgør et meget væsentligt supplement til top-down-baserede metoder baseret primært på eksperters viden. Ikke overraskende tilføjede de automatiske metoder mest viden til ontologiens nederste lag, hvorimod ontologiens mellemlag stadig primært må bygge på ekspertviden, dog støttet af viden repræsenteret i eventuelle termordbøger. I kapitlet beskrev vi også et eksperiment med kodning af tværgående ontologiske relationer.

De to sidste kapitler omhandlede søgeekspansion på basis af dels syntaksviden, dels ontologisk viden. I kapitel 4 så vi på sammensætninger og deres synonyme paralleller i form af fraser. Vi viste hvordan man ud fra den sprogteknologiske ordbase med syntaksoplysninger, kan identificere de sammensætninger der hyppigt har synonymparalleller: nemlig sammensætninger hvor kernen er relationelt, altså valensbærende. Vi præsenterede også en algoritme baseret på NP-analyse som var i stand til at vægte gode hits højere end dårlige hits. Endelig præsenterede vi i kapitel 5 VID-prototypen som søger på Zaccos dokumenter på basis af sprogteknologisk viden.

Et af de overordnede mål med VID-projektet har været at udnytte den dynamiske 'trekant' i projektsammensætningen hvori indgår både forskere, teknologiudviklere og brugere. I de eksperimenter vi har beskrevet i denne rapport, har der således været 1 bruger, 2 teknologiudviklere og 1 forskningsinstitution. Denne konstellation har generelt været krævende – men også givende. Krævende fordi inputtene og interesseområderne har været mange og forskelligartede og ikke altid har peget i den samme retning. Givende fordi det har tvunget os som forskere ud i nogle afkroge hvor vi ikke ellers ville være kommet. Den store udfordring i projektet har været at sætte de enkelte eksperimenter ind i en større sammenhæng, og her har prototypen i vid udstrækning fungeret som det samlende element. Selv om prototypen på flere punkter er rudimentær i sin nuværende form, så har det her været muligt at afprøve flere af de udførte eksperimenter i en samlet kontekst og dermed få et samlet billede af de forbedringer som integreret sprogteknologi kan give til ontologibaseret teksthåndtering.

Referencer

Andreasen, T. P.A. Jensen, J.F. Nilsson, P. Paggio, B.S. Pedersen, H.E.Thomsen (2004) 'Content-based text querying with ontological descriptiors', in: *Database and Knowledge Engineering Journal no. 48: pp. 199-219*, Elsevier Science B.V., Holland.

Allan, J. (2000). Natural Language Processing for Information Retrieval, *Tutorial notes presented at the NAACL/ANLP language technology joint conference*, Washington.

Braasch, A. B.S. Pedersen (2002). Recent Work in the Danish Computational Lexicon Project „STO", in *EURALEX Proceedings 2002,* Center for Sprogteknologi, København.

Buitelar, P., Olejnik, D., Hutanu, M., Schutz, A., Declerck, T. & Sintek, M.(2004), Towards ontology engineering based on linguistic analysis, in *Proceedings of LREC-2004*, Lisboa, Portugal, pp. 7-10.

Chen, A. and F. Gey (2003). Combining Query Translation and Document Translation in Cross Language Retrieval CLEF 2003 http://clef.iei.pi.cnr.it/2003/WN_web/05.pdf

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

de Loupy, C. & M. El-Bèze (2002). Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet, in *Proceedings from the workshop on Using Semantics for Information retrieval and Filtering*, 20-27. LREC 2002, Las Palmas de Gran Canaria , Spain

Fersøe, H. & S. Kirchmeier-Andersen (1989). Maskinoversættelse af simple paratagmer. I: *Skrifter om anvendt og matematisk lingvistik nr. 14*. Københavns Universitet.

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

Johnston, M. and F. Busa (1999). Qualia structure and the compositional interpretation of compounds. In E. Viegas (ed.), *Breadth and Depth of Semantics Lexicons*. Dordrecht: Kluwer Academic Publishers p. 167-187.

Jongejan, B. & Haltrup, D. (2001). *The CST Lemmatiser*, Teknisk Rapport, Center for Sprogteknologi, Københavns Universitet.

Järborg, J. (2003). Semantisk uppmärkning. Metoder, problem og resultat. *Research Reports from the Department of Swedish,* Göteborg Universitet.

Mueller, E.T. (1998) Evaluating Representations of Common Sense, Talk for *AAAI-1998 Panel*, tilgængelig på http://www.signiform.com/compoundnoun/classes.htm

Navarretta, C., B. S. Pedersen, D. Haltrup Hansen. (2004). Human Language Technology Elements in a Knowledge Organisation System -The VID project. *In Proceedings of LREC-2004, vol.1:75-78.* Lissabon.

Paggio, P., B. S. Pedersen, D. Haltrup (2003) Applying Language Technology to Ontology-based Querying - The OntoQuery Project. Applied Artificial Intelligence Journal. Artificial Intelligence for Cultural Heritage and Digital Libraries, Vol. 17 Numbers 8-9:817-833.

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

Pedersen, B. S., C. Navarretta, L. Henriksen. (2004). Building Business Ontologies with Language Technology Techniques - The VID project. In *Proceeding of ONTOLEX Workshop in conjunction with LREC 2004*. 30-35. Lissabon.

Porter, M. (1980). An algorithm for suffix stripping. *Program*, vol. 14, no. 3, 130-137 http://www.tartarus.org/~martin/PorterStemmer/def.txt

Loukachevitch, N. & B. Dobrov (2004). Ontological Types of Associative Relations in Information-Retrieval Thesauri anmd Automatic Query Expansion. In: *Proceedings from Ontolex 2004- Ontologies and Lexical Resources in Distributed Environments* pp. 24-29. Lissabon.

Lovins, J. B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1-2):22-31.

Luhn, H. P (1957). 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development,* **1**, 309-317

van Rijsbergen, C. J. (1999). *Information Retrieval*, second edition. http://www.dcs.gla.ac.uk/~iain/keith/

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

Voorhees, E. (1993). Using WordNet to disambiguate word senses for text retrieval. In: Korfhage, Robert, Edie Rasmussen and Peter Willett, eds., *Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, 171–180, Pittsburgh.

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.

Ørsnes, B. (1995). *The Derviation and Compounding Event Nominals in Modern Danish – an HPSG Approach with an Implementation in Prolog.* PhD dissertation, University of Copenhagen.

Bilag A Cass-grammatik

Dette er np_pp_pp_gram.reg

```
:np
  determ = PRON_DEMO | PRON_UBST | PRON_INTER_REL ;
  attr = ADJ | FORK | UL | FORK SYMBOL | FORK ADJ | NUM | NUM_ORD |
V_PARTC_PAST | V_PARTC_PRES;
  gen = N_GEN | EGEN_GEN | ADJ_GEN | PRON_POSS |
           PRON_DEMO_GEN | PRON_UBST_GEN | NUM_GEN |PRON_REC_GEN |
PRON_INTER_REL_GEN;

  DP =  FORK? determ+ | determ SKONJ determ | gen SKONJ* gen* ;
  AP =  attr+|ADV? ADJ+ |ADV? attr TEGN ADV? attr|ADV? attr TEGN? ADV?
attr? SKONJ ADV? attr ADV? attr?;
  NP_G = DP AP* gen | DP* AP gen | EGEN gen;


  kerne = N | EGEN+ | NUM | UL | FORK| XX SKONJ N ;
  NP ->  ADJ? DP? AP? kerne | ADJ? NP_G? AP? kerne| PRON_DEMO attr |
NUM N N | PRON_UBST ADJ? N N | N EGEN+;


:np1
  PP = PRÆP NP;

  NP1 -> NP PP;


:np2
  PP = PRÆP NP;

  NP2 -> NP1 PP;
```