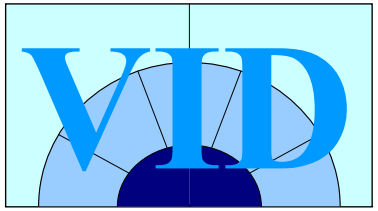


Sprogteknologiske komponenter i ontologi og søgning

Bolette Sandford Pedersen,

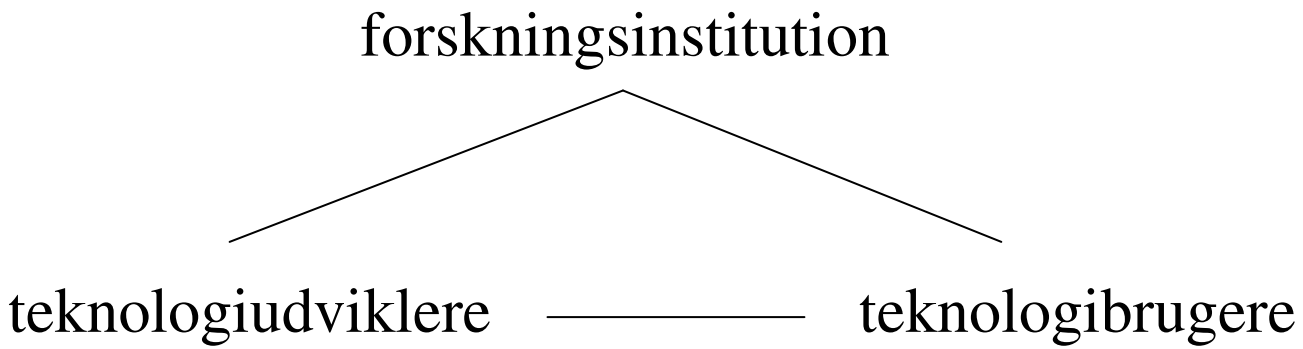
Costanza Navarretta,

Dorte Haltrup Hansen, Bart Jongejan

Center for Sprogteknologi, KU

VID-projektets mission

at foretage en række sprogteknologiske eksperimenter i et dynamisk trekantsmiljø:



Delprojekt 1: Ontologi og søgning med sprogteknologi

Hvorfor og hvor skal vi have (dansk) sprogteknologi i ontologi og søgning?

- **Præproces**: for at kunne automatisere indekseringsprocesser (nøgleord), lette opbygningen af ontologi/tesaurus
- **Søgning**: for at tilnærme os mere indholdsbaserede og fleksible søgeværktøjer
- **Dansk sprog**: for generelt at få søge- og navigeringsværkstøjer der virker ordentligt på dansk

Sprogteknologi ind på flere planer

I VID-projektets delprojekt om 'ontologi og søgning med sprogteknologi' har vi foretaget følgende fem eksperimenter:

1) Identifikation af termer fra brugervirksomhedernes tekster

2) Opstilling af fagontologi

3) Identifikation af nøgleord

4) Søgning med lingvistisk viden, ontologi og Dublin Core metadata (eksp.1-3)

5) Samme indhold i forskellig forklædning: identifikation af synonymfraser

Zacco-case (eksp. 1-4)

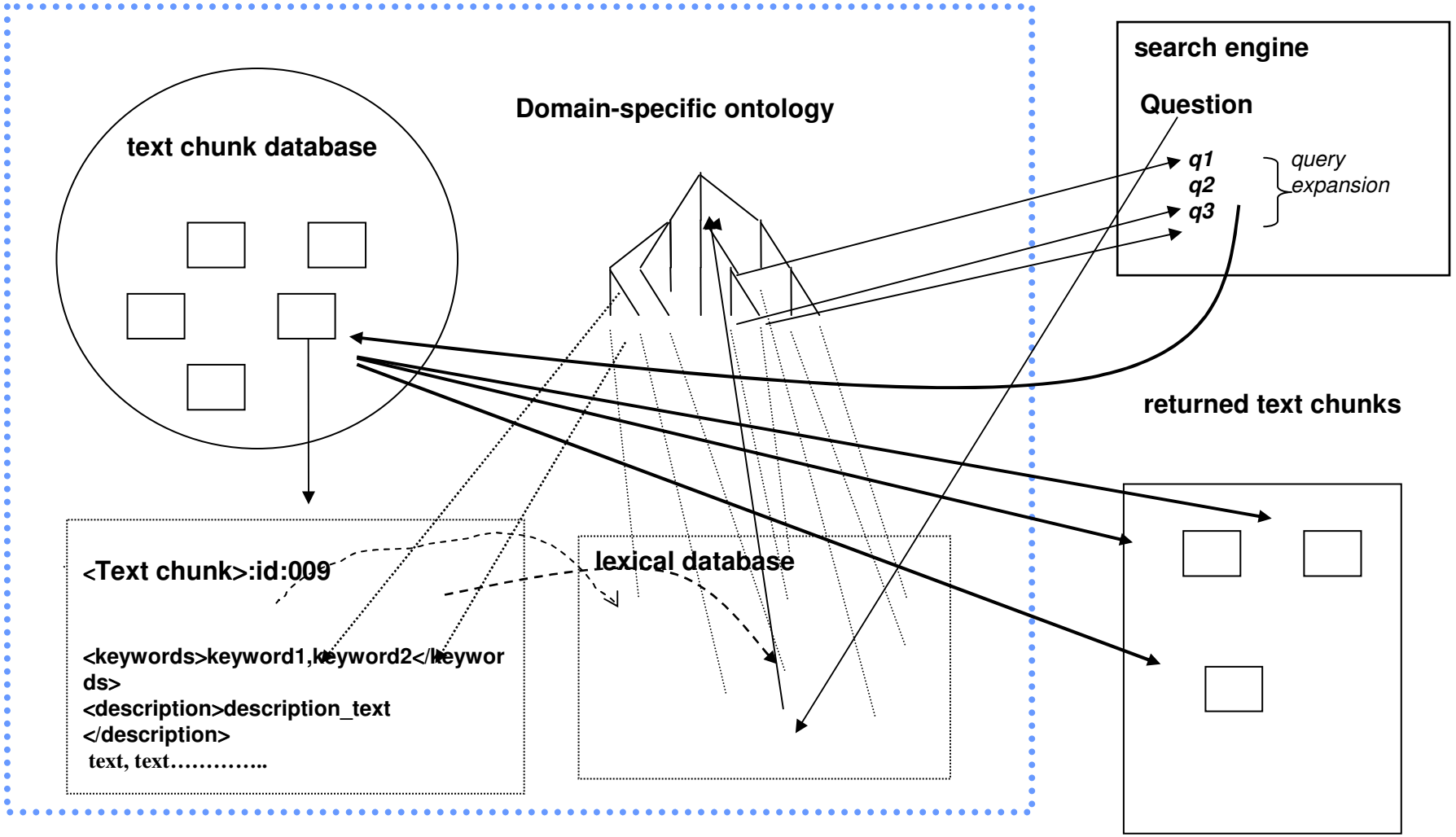
Scenarie: Zacco (IPR-virksomhed med afdelinger i flere nordiske lande) er i gang med at opbygge et videnorganiseringssystem således at de kan

- systematisere og effektivisere produktionen og ikke mindst vedligeholdelsen af standarddokumenter - også på tværs af sproggrænser

Dertil skal bruges:

- en videnmodel som bl.a. gør det muligt at søge fleksibelt efter dele af dokumenter som fx skal justeres i forhold til en ændret lovgivning

Zacco-case


search engine
Question
q1
q2
q3
query expansion
Domain-specific ontology
text chunk database
returned text chunks
<Text chunk>:id:009
<keywords>keyword1,keyword2</keywords>
<description>description_text</description>
text, text……..
lexical database

Termidentifikation

Opbygning af videnmodel kræver overblik over virksomhedens **terminologi** - dette kan skabes semiautoamtisk ud fra virksomhedens tekster

- Ordene identificeres (tokenising) og ordklasseopmærkes (POS tagging)
- Ordene neutraliseres til grundform (lemmatising)
- Navneord, tillægsord og udsagnsord udvælges
- Ordlisten sammenholdes med STO-ordbasen på 65.000 almensproglige ord, og ord der ikke er i STO, udvælges som termkandidater
- Listen suppleres med andet led fra udvalgte sammensatte ord også selv om ordet forekommer i STO, (fx. anses *gebyr* ud fra *extensionsgebyr*)

Resultater for termudtræk

Vi har stillet to spørgsmål til Zaccos termeksperterne:

- Er de fundne termkandidater rent faktisk termer ? (precision)
- Hvor mange af termerne har vi fundet ? (recall)

Evaluering af termeksperternes svar:

Precision:

71 % af de fundne termkandidater er rent faktisk virksomhedsrelevante termer

Recall:

77 % af termerne blev fundet af systemet

Resultater for termudtræk

Forklaring:

Precision (71 %) : analysefejl pga. af uidentificerede forkortelser, udenlandske ord eller stavemåder, slåfejl;

de sprogteknologiske værktøjer var ikke skræddersyet til teksttypen (skabeloner og delvist flersprogede udtryk)

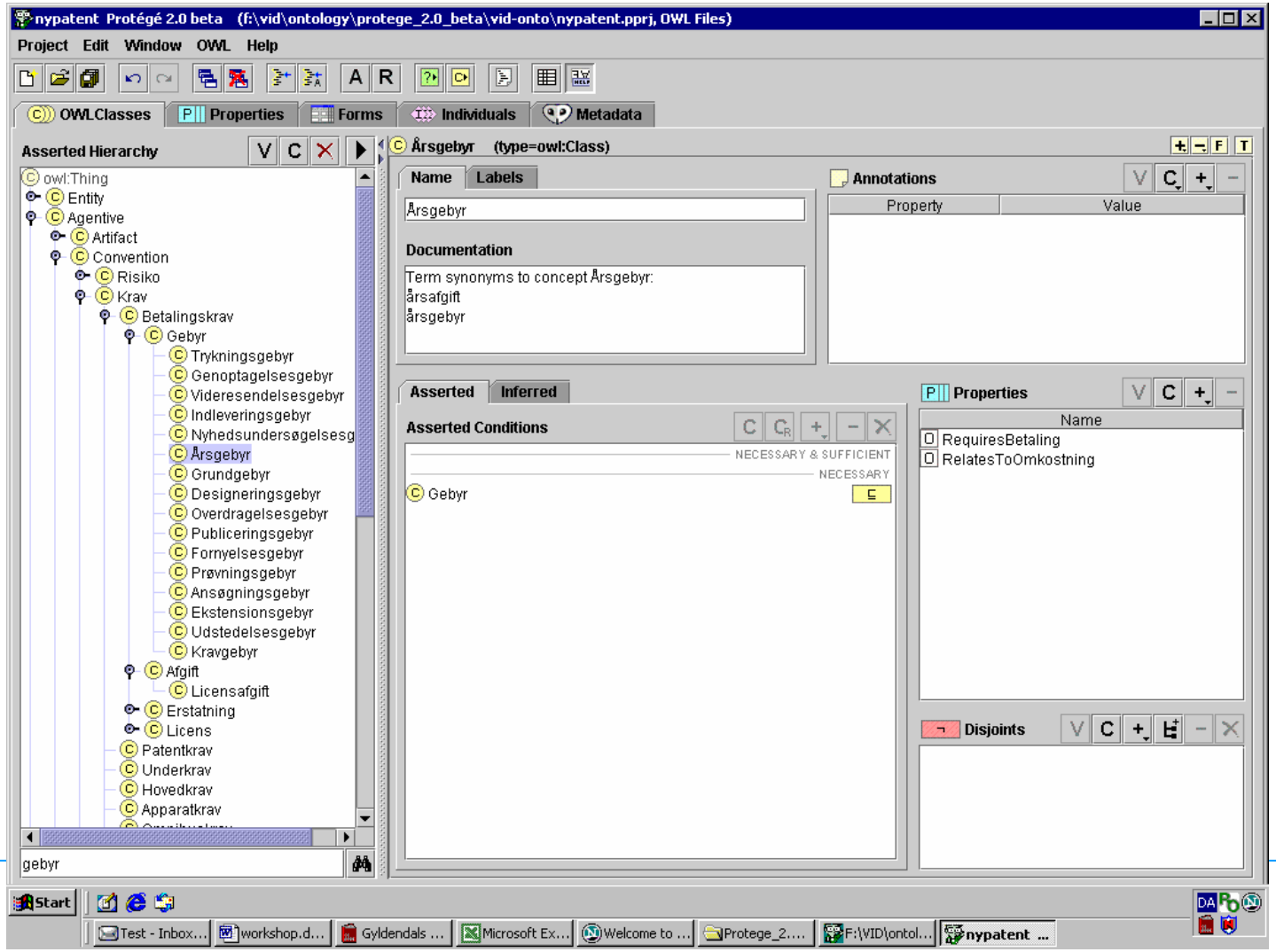
Recall:

- gråzoneord som optræder i STO gives ikke som termkandidater (*ret, domstol*)
- nogle termer optræder ikke i teksterne, men angives af termeksperterne som relevante termer

Ontologi som del af videnmodel

- Vi anvender standarder for at give mulighed for udveksling og genanvendelse (på intranet og internet), og mulighed for integration med metadatastandarder som Dublin Core
- Vi anvender standard W3C Ontology Web Language (OWL), (også i RDFS-format)
- Vi anvender Protégé-2000 som kodningsværktøj med OWL-plugins

Ontologiens nederste niveau



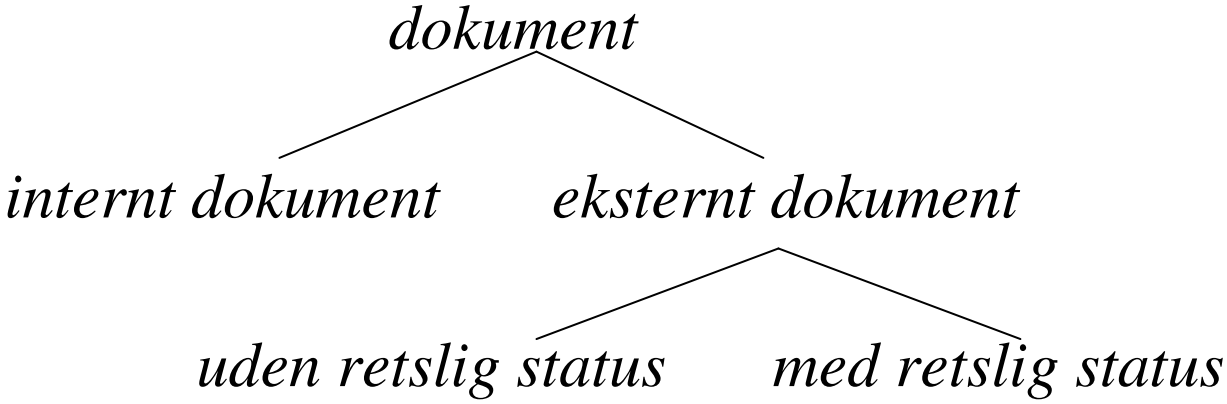
CST
CENTER FOR SPROGTEKNOLOGI

Ontologiens mellemlag

Vanskeligt at bygge, ekstralingvistisk viden, formålet med ontologien spiller ind

Kilder:

- En virksomhedsspecifik patentordbog
- Termeksperter definerer mellemklasserne



Ontologiens øverste lag

SIMPLE som top-ontologi (Semantic Information for Multifunctional, Plurilingual Lexica)

- 136 top-ontologiske klasser
- Hver ontologisk klasse har tests til at definere disse

Eksempel:

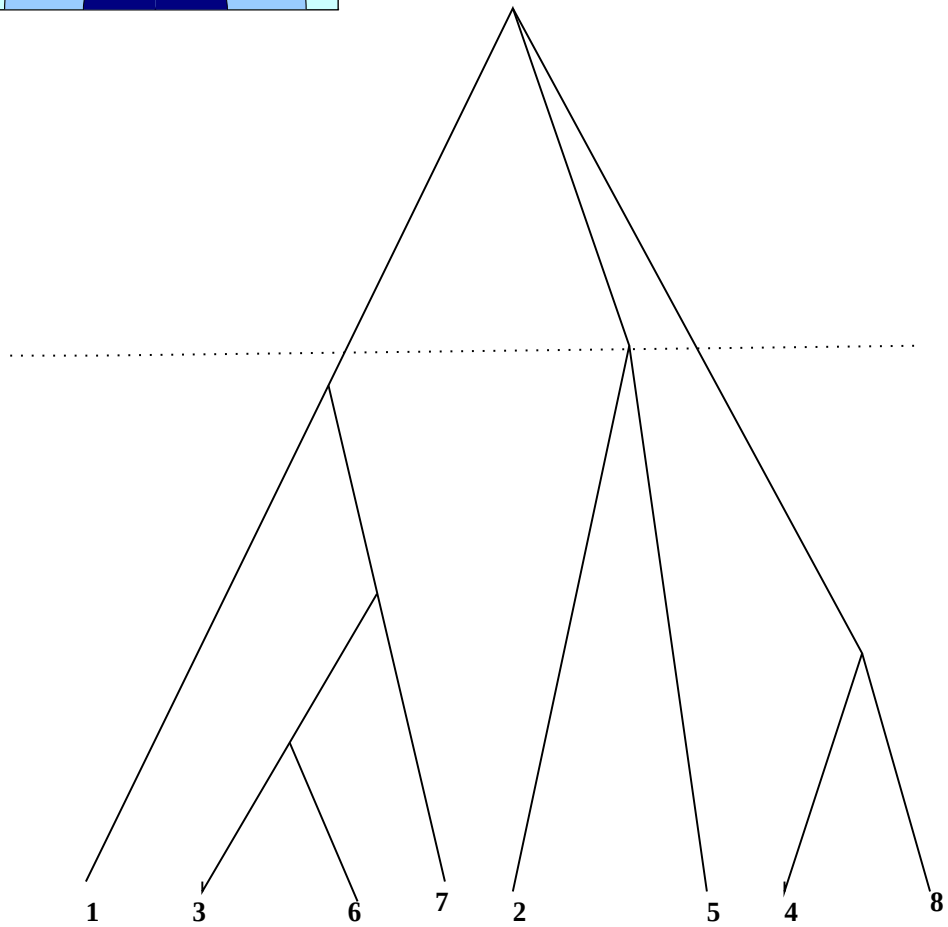
AgentOfPersistentActivity: (*sagfører*)

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

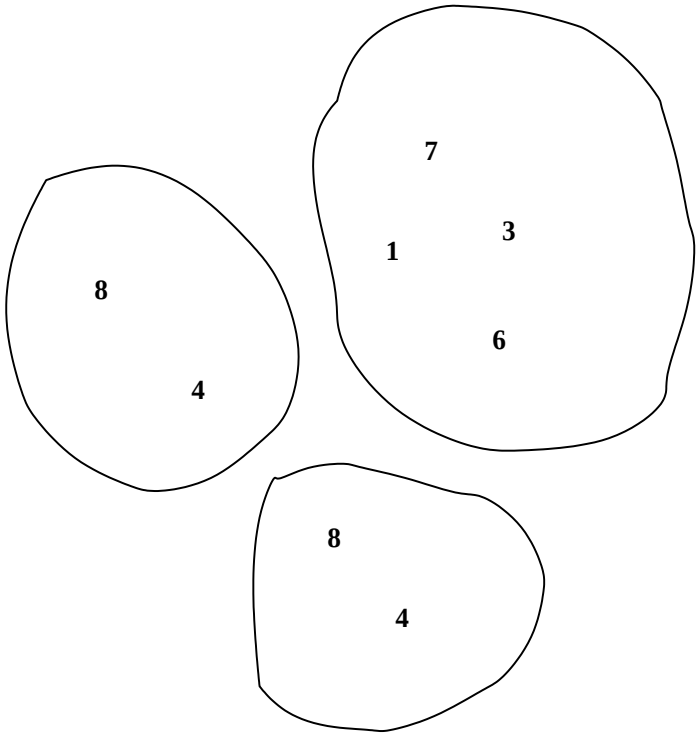
Statistiske tilgange til ontologi

- Semantisk clustering: ord som semantisk ligner hinanden mest, indsættes i samme gruppe (cluster), mens ord der er forskellige indsættes i separate grupper

- *Semantisk lighed* defineres som graden hvorpå ord kan erstatte hinanden i samme kontekst



Hierarkisk clustering



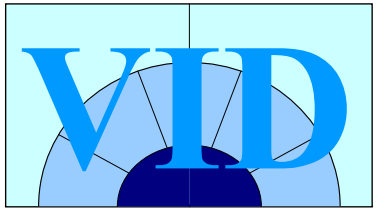
Ikke−hierarkisk clustering

Eksperimenter med Zaccos tekster

CMU-statistikpakken og Lnknet-systemet udviklet af MIT Lincoln Laboratory

Automatisk uddragede clusters med *K-means*- og *EM*-clustering (ikke-hierarkiske):

- patentansøgning, grundansøgning, ansøgning, oversættelse, patent
- gebyr, afgift, årsafgift, årsgebyr, fornyelsesafgift, kravgebyr
- rapport, indleveringsrapport, besvarelse, faktura
- konceptkopi, bilag, skrift, kopi



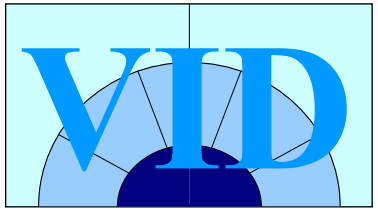
Evaluering af clustering

- Giver de bedste resultater på store mængder data (fx ignoreres lavfrekvente ord)
- Hjælper ontologidesigneren til at foretage de første grupperinger af data
- Understøtter ikke kodning af relationer mellem ord i de forskellige grupper og mellem grupperne

Nøgleord til indeksering (eksp. 3 med Navigo)

- **Indeksering:** den proces at udvælge et antal ord som tilsammen beskriver emnet i et dokument
- **Eksperimentets formål:** kan vi med simple sprogteknologiske metoder lave brugbar automatisk nøgleordsidentifikation for dansk ?

Med simpel metode mener vi uden adgang til viden om det domæne vi behandler, eller til resurser om ordbetydning



Nøgleord til indeksering

Traditionel metode:

- fjernelse af meget frekvente ord vha. stopordsliste
- stemming (hugge endelser af ordene)
- manuel udvælgelse af nøgleord OG/ELLER vægtning af ordforekomster hvorved nøgleordene fremfindes

Nøgleord til indeksering

Sprogteknologisk metode:

- Normalisering til grundform (lemmatisering, dvs. ordets grundform beregnes ud fra dets endelse og en række bøjningsregler fremkommet vha træning på en ordbase (STO))
- Udvælgelse af betydningsbærende ord (navneord)
- Vægtning af ordforekomster ud fra forholdet mellem ordets frekvens og antallet af dokumenter hvori ordet forekommer
- Resultat= nøgleordskandidater

Eksempel

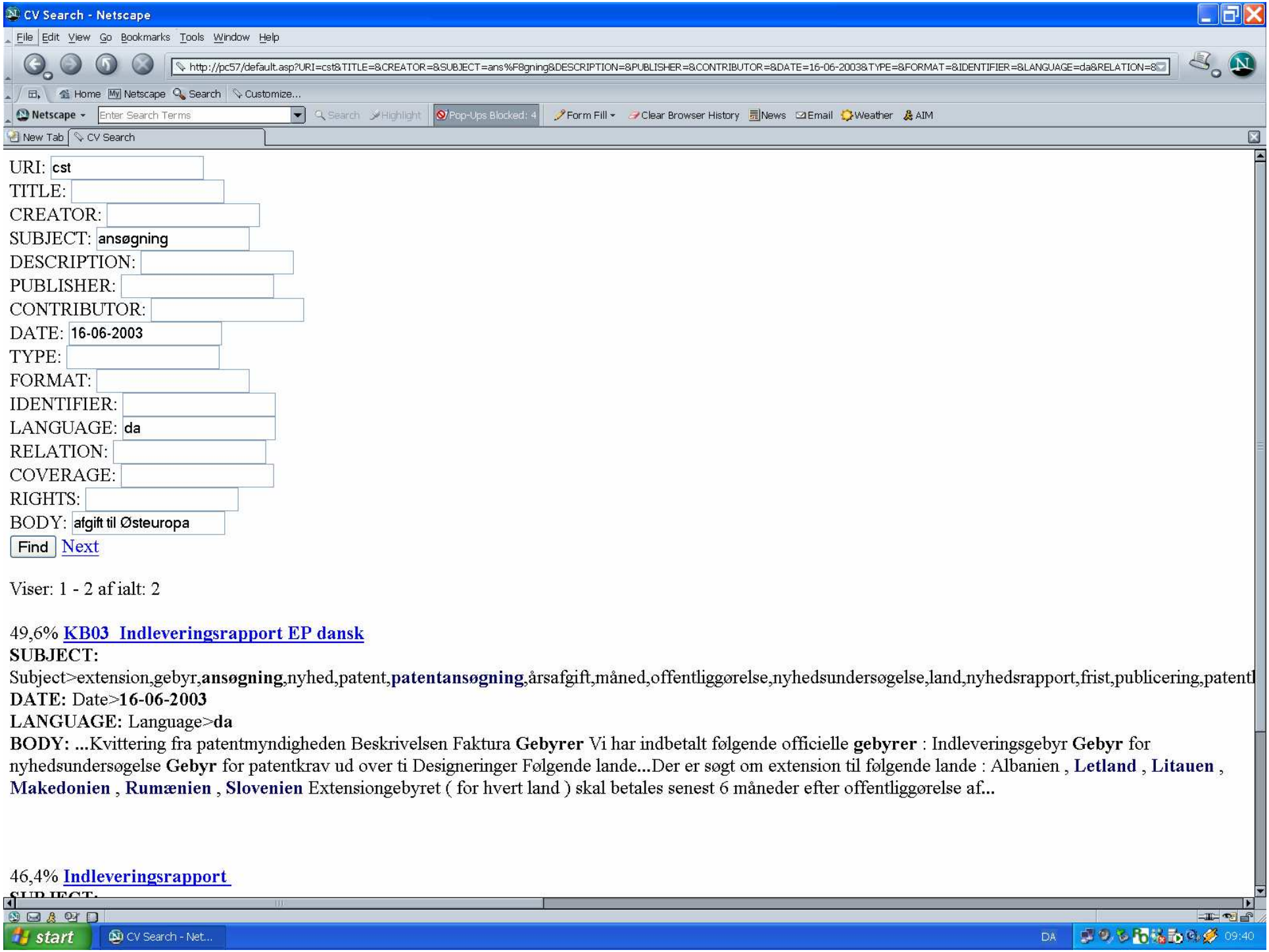
Tekst: Hjemmeside om VID	
Nøgleord	**1. led på sammensatte ord**
virksomhed	dokument
projekt	it
viden	sprog
informationssøgning	
dokumentproduktion	
undersøgelse	
sprogteknologi	
sprog	
metode	
dokumentation	

Prototype med Ankiro: en søgemaskine

Prototype på en søgemaskine der søger i Zaccos tekstmateriale

- prototypen kombinerer søgning i Dublin-Core metadata og søgning i tekst med ekspansion af søgestrengen på basis af sproglig og nogen ontologisk viden inden for patentområdet

Demo i kaffepausen!


CV Search - Netscape
File Edit View Go Bookmarks Tools Window Help
http://pc57/default.asp?URI=cst&TITLE=&CREATOR=&SUBJECT=ans%F8gning&DESCRIPTION=&PUBLISHER=&CONTRIBUTOR=&DATE=16-06-2003&TYPE=&FORMAT=&IDENTIFIER=&LANGUAGE=da&RELATION=&
Home Netscape Search Customize...
Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 4 Form Fill Clear Browser History News Email Weather AIM
New Tab CV Search
URI: cst
TITLE:
CREATOR:
SUBJECT: ansøgning
DESCRIPTION:
PUBLISHER:
CONTRIBUTOR:
DATE: 16-06-2003
TYPE:
FORMAT:
IDENTIFIER:
LANGUAGE: da
RELATION:
COVERAGE:
RIGHTS:
BODY: afgift til Østeuropa
Find Next
Viser: 1 - 2 af ialt: 2
49,6% KB03_Indleveringsrapport EP dansk
SUBJECT:
Subject>extension,gebyr,ansøgning,nyhed,patent,patentansøgning,årsafgift,måned,offentliggørelse,nyhedsundersøgelse,land,nyhedsrapport,frist,publicering,patentl
DATE: Date>16-06-2003
LANGUAGE: Language>da
BODY: ...Kvittering fra patentmyndigheden Beskrivelsen Faktura Gebyrer Vi har indbetalt følgende officielle gebyrer : Indleveringsgebyr Gebyr for nyhedsundersøgelse Gebyr for patentkrav ud over ti Designeringer Følgende lande...Der er søgt om extension til følgende lande : Albanien , Letland , Litauen , Makedonien , Rumænien , Slovenien Extensiongebyret (for hvert land) skal betales senest 6 måneder efter offentliggørelse af...
46,4% Indleveringsrapport
start CV Search - Net... DA 09:40

Synonymfraser (eksp. 5 med Ankiro)

- **Generelt problem ved søgning**: samme eller lignende semantisk indhold kan have mange forklædninger i tekst;
- **Specifikt undersøgelsesområde:** hvordan kan man automatisk identificere fraser som er synonyme med sammensatte ord ?

Ex: *byrådsmedlem*

Frase som er synonymt med *byrådsmedlem*:
medlem af byrådet

Sammensatte ord og synonymfraser

Eksperimentets formål:

- Kan vi på forhånd afgøre hvilke sammensatte ord som har synonymer i form af fraser?
- For dem der har parallelsynonymer; kan vi ud fra en sprogteknologisk beregning skelne gode hits fra dårlige og dermed automatisk frasortere støj?

Undersøgelsesmateriale leveret af Ankiro:

1000 forespørgsler på 'splittede' sammensatte ord i KommuneInformations database, op til 200 hits pr. forespørgsel

Typer af sammensatte ord

Klassifikation af sammensatte ord

- deverbale sammensætninger, *depotanbringelse*, *hudreaktion*
- andre relationelle sammensætninger, *dækningsmulighed*, *bistandspligt*
- ikke-relationelle sammensætninger som fx *autovaskehal, afløbspumpe, planteprodukt*

Typer af sammensatte ord

Hypotese:

- hits hvor begge søgeord er i samme navnefrase (NP) er nok betydningsmæssigt meget tæt på det sammensatte ord

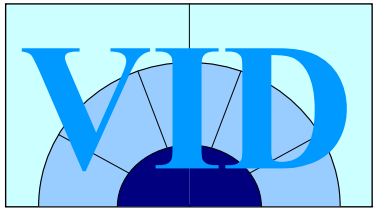
Sandsynligvis synonym til *apoteksovertagelse:*

- NP[Overtagelsen af apoteket] kan ske pr. 1. november

- Staten kan yde garanti for lån til NP[overtagelse og etablering samt til flytning og ombygning af apotek]

Sandsynligvis ikke synonym til *apoteksovertagelse:*

- Skatteyderen havde erhvervet et apotek og havde opnået statsgaranteret lån til finansiering af overtagelse af varelager, inventar m. v.



vægt		*<WORD>hudreaktion reaktion hud</WORD><COUNT>14</COUNT>*
		3 hits:1 x 9.0, 2 x 19.0
19.0		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		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		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

Resultater

Deverbale sammensætninger

- Precision: 90% af de fundne hits (fundet = vægt over 10) indeholder rent faktisk nærsynonymer til det sammensatte ord
- Recall: 60% af nærsynonymerne blev rent faktisk fundet af systemet

Andre relationelle sammensætninger
Precision: 90% - Recall: 80%

Ikke-relationelle sammensætninger (Precision: 80% - Recall: 50%)

Forklaring

Precision:

samme NP, men ikke synonym:
biblioteksdrift: drift af internetadgangen til bibliotekerne

Recall: begrænset NP-genkender som bl.a. ikke er gearet til at klare relativsætning, mange præpositionsforbindelser og koordinerede NPer (som er hyppig i bekendtgørelser o. lign.)

Resultater

Konklusion:

Med en udvidet NP-genkender er der god mening i at ekspandere søgestrengen ved **deverbale og andre relationelle sammensætninger**

(disse oplysninger kan bl.a. udledes fra STO-ordbasen)

Ved ikke-relationelle sammensætninger giver denne strategi en del støj

Konklusion

VID-projektet har sat vores forskning i perspektiv:

- Virkelige scenarier fra virksomhedsbrugere har påvirket vores forskning positivt - skræddersyning nødvendig!
- Dialogen med teknologiudviklerne har været meget frugtbar
- Sprogteknologi har en meget vigtig rolle at spille i behandlingen af digital information - ikke mindst for at få systemer der virker ordentlig på dansk