

Semantik, ontologi, tesaurus mv.

- Hvad er det, og hvilke fordele kan opnås ved fælles løsninger ?

Seniorforsker Bolette Sandford Pedersen,
Center for Sprogteknologi,
Københavns Universitet

Indhold

- Hvorfor er semantik relevant for informationsøgning ?
- Begrebsafklaring: semantik, ordnet, thesaurus, ontologi
- Hvordan opbygges sådanne videnbaser ?
Genbrug af eksisterende data, samt manuelle vs. halvautomatiske metoder
- To eksperimentelle projekter med ontologi og opmærkning med metadata

1. Hvorfor er semantik relevant ?

Vi får for mange informationer hvoraf meget af det er irrelevant

Vi får for få informationer i forhold til hvad der rent faktisk er tilgængeligt fordi vi ikke har 'ramt' den rigtige formulering i forespørgslen

Fremtidsvisionen: Vi vil gerne kunne spørge og få et svar - ikke bare i form af tekster

Hvorfor er semantik relevant ?

For mange informationer fordi:

ord kan betyde flere ting; de er flertydige

Ca. 25 % af alle ord der søges på er flertydige; ca. 10 % af alle navne der søges på er flertydige

homonymi: *pande* - hoveddel, køkkenredskab

polysemi: *mus* - dyr, styreredskab til computer

proprier: *Java*: ø, programmeringssprog

Entydiggørelse vha af sprogteknologiske metoder:

- domæneafgrænsning
- sprogteknologisk beregning på kontekst

Hvorfor er semantik relevant ?

**For få informationer fordi samme indhold kan have
forskellig forklædning:**

ordformen genkendes ikke af systemet:

trøffel - trøfler

synonymer og synonyme udtryk:

børnepasningsorlov - forældreorlov

byrådsmedlem - medlem af byrådet

Europæisk Union - EU

vi kan tale om ting på forskellige specificeringsniveauer:

mangelsygdom - beriberi

Hvorfor er semantik relevant ?

Samme udtryk i forskellig forklædning:

- lemmatisering på basis af dansk sprogteknologi
(stemming dur ikke for dansk)
- videnmodel der kortlægger domænets udtryksvariationer og begrebsmæssige struktur således at der kan foretages beregning på semantik 'nærhed'

2. Begrebsafklaring: semantik

Semantik: studiet af ords og sætningers betydning (Nudansk Ordbog)

- **Ordsemantik**

hvad betyder ordene ?

- **Sætningssemantik**

hvad betyder sætningen ?

- **Tekstsemantik**

hvad er indholdet i teksterne ?

Ordsemantik

Uformel tilgang: i almindelig ordbøger eller
tesaurusser vha *definition* og *brugseksempler*

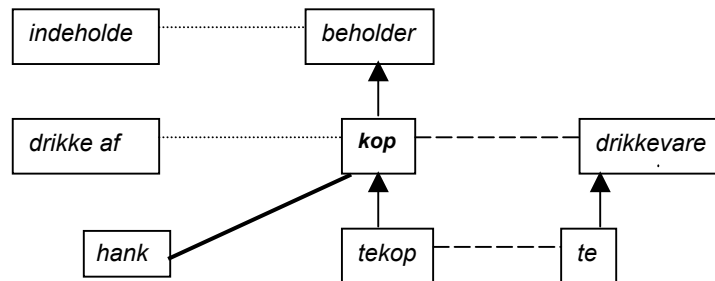
definition på *kop*: *lille skåleformet el.
cylindrisk beholder til at drikke af, typisk
med hank og brugt sammen med en dertil
hørende underkop til varme drikke*

(Den Danske Ordbog)

brugseksempler: *kopper og tallerkner,
kvinden stiller to kopper på bordet
forsvinder ind i stuens mørke og kommer
tilbage med kaffe*

Ordsemantik (light)

Formel tilgang: delmængde af de samme oplysningstyper som i en almindelig ordbog, blot formaliseret ud fra en sprogteknologisk anskuelsesvinkel, fx i et **ordnet** vha. etablerede relationer: *is_a*, *part_of*, *used_for*, *contains*



Ordsemantik

Formel tilgang kan indeholde langt mere formaliseret viden om betydning, fx

- aksiomatisk beskrivelse: *kæledyr*: 'der eksisterer en Person x og Dyr y sådan at x har Kæledyr y ' :

? $x, y : instance(x, Person) ? instance(y, Animal) ?$
 $pet(x, y)$

tekstsemantik (ekstra light)

Uformel tilgang:

Resumé, genfortælling

Formel tilgang:

- Nøgleord brugt som indeksord
- Opmærkning med metadata fx i form af ontologiske kategorier og relationer således at de faktuelle informationer registreres i et formelt sprog

Tekstsemantik er naturligvis meget mere end dette !

Ordnet

- Et ordnet er en database der som grundlæggende enheder indeholder betydningsdefinitioner, hvortil der knyttes de ord og vendinger som kan udtrykke pågældende betydning, såkaldte synsets
- Synsets forbindes med hinanden vha. semantiske relationer, fx modsætninger; *sort - hvid*, over- og underbegreber; *møbel - stol*, del-helhed; *finger - hånd*

WordNet

a lexical database for the English language

cognitive science laboratory | princeton university | 221 nassau st. | princeton, nj 08542

[About WordNet](#)

[Use WordNet online](#)

[Download](#)

[Changes in version 2.0](#)

[Frequently asked questions](#)

[WordNet manuals](#)

[WordNet statistics](#)

[Current events](#)

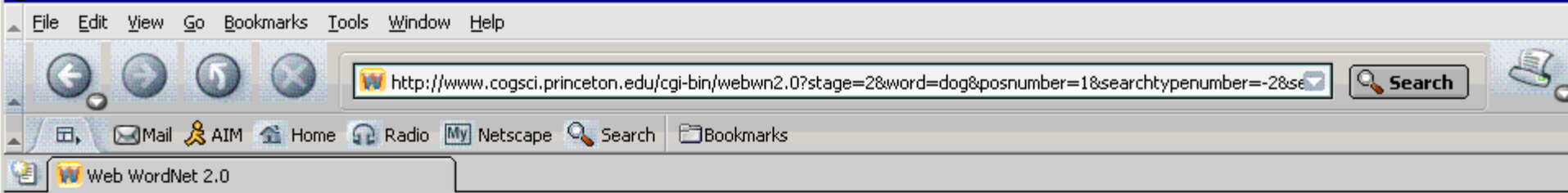
[Publications](#)

WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

WordNet was developed by the [Cognitive Science Laboratory](#) at [Princeton University](#) under the direction of [Professor George A. Miller](#) (Principal Investigator).

Over the years, many people have contributed to the success of WordNet. At the present time, the following individuals at Princeton work on the continuing development of WordNet and applying it to research:

- ◊ [Professor George A. Miller](#)
- ◊ [Dr. Christiane Fellbaum](#)
- ◊ [Ranee Tengj](#)
- ◊ [Susanne Wolff](#)
- ◊ [Pamela Wakefield](#)
- ◊ [Helen Langone](#)
- ◊ [Benjamin Haskell](#)



WordNet 2.0 Search

Search word: Find senses

Results for "Hypernyms (this is a kind of...)" search of noun "dog"

7 senses of dog

Sense 1

- dog, domestic dog, *Canis familiaris* -- (a member of the genus *Canis* (probably descended from the common wolf) that has been domesticated by man since prehistoric times)
- => canine, canid -- (any of various fissiped mammals with nonretractile claws and typically long muzzles)
- => carnivore -- (terrestrial or aquatic flesh-eating mammal; terrestrial carnivores have four or five clawed digits on each limb)
- => placental, placental mammal, eutherian, eutherian mammal -- (mammals having a placenta; all mammals except monotremes and marsupials)
- => mammal -- (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes)
- => vertebrate, craniate -- (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull)
- => chordate -- (any animal of the phylum Chordata having a notochord or spinal column)
- => animal, animate being, beast, brute, creature, fauna -- (a living organism characterized by voluntary movement)
- => organism, being -- (a living thing that has (or can develop) the ability to act or function independently)
- => living thing, animate thing -- (a living (or once living) entity)
- => object, physical object -- (a tangible and visible entity, an entity that can cast a shadow, "it was full of rackets, balls and other objects")
- => entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Sense 2

frump, dog -- (a dull unattractive unpleasant girl or woman: "she got a reputation as a frump": "she's a real dog")



Tesaurus

- En **tesaurus** er en specialordbog eller ordsamling, der angiver konceptuel, semantisk information om termer i form af definitioner, relationer, synonymer samt evt. anden inf.
- Termen anvendes især inden for specifikke fagdomæner
- I søgesammenhæng anvendes begrebet ofte om oversigter over søgeord til brug for indeksering



The Cook's Thesaurus

[home](#)

SEARCH RESULTS 1 - 5 of 5 total results for **zuchini**

Cook's Thesaurus: Summer Squash

...take the seed out--it's edible and tasty. Cooked chayotes make good low-fat substitutes for avocados. Substitutes: **zucchini** (stronger flavor, cooks more quickly) OR kohlrabi OR other summer squash OR carrots OR bell peppers (for...
<http://www.foodsubs.com/Squashsum.html>

Cook's Thesaurus: Cucumbers

...served raw in salads, sandwiches, drinks, sushi, and hors d'oeuvres to add crunch, but they can also be cooked like **zucchini**. Pickling cucumbers are usually smaller than slicing cucumbers, and often have thick, warty skins. They're...
<http://www.foodsubs.com/Squacuke.html>

Cook's Thesaurus: Asian Squash

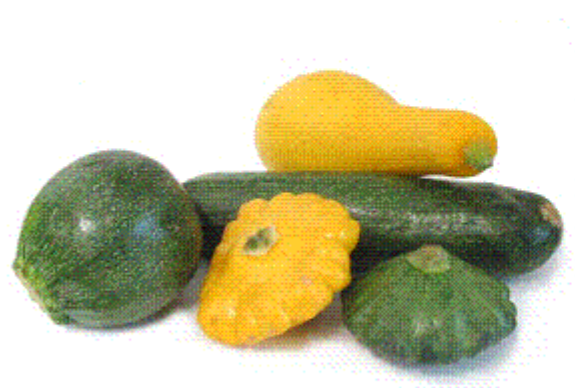
...angled loofah. Substitutes: **zucchini** Chinese winter melon foo gwa fuzzy melon = hairy melon = hairy cucumber = moqua Notes: This sweet and mild squash has a fuzzy feel to it. Substitutes: **zucchini** hairy...
<http://www.foodsubs.com/Squashasian.html>

The Cook's Thesaurus

 Search

[home](#) > [vegetables](#) > [fruit vegetables](#) > summer squash

Summer Squash



Unlike winter squash, summer squash can be eaten rind, seeds, and all. The different varieties vary in size, shape, and color, but they can be used interchangeably in recipes. Select summer squash that's small and firm.

Substitutes: eggplant (this must be cooked) OR bok choy (in stir-fries) OR cucumbers (if served raw) OR winter squash

Varieties:

bottle gourd

calabash 1. **spaghetti squash** 2. **cucuzza**

Ontologi

Aristoteles' undersøgelse af 'tingenes væsen':
identifikation af centrale begreber og
relationerne i mellem dem

mange forskellige faglige traditioner:

- filosofisk
- logisk
- datalogisk
- lingvistisk..

Ontologi

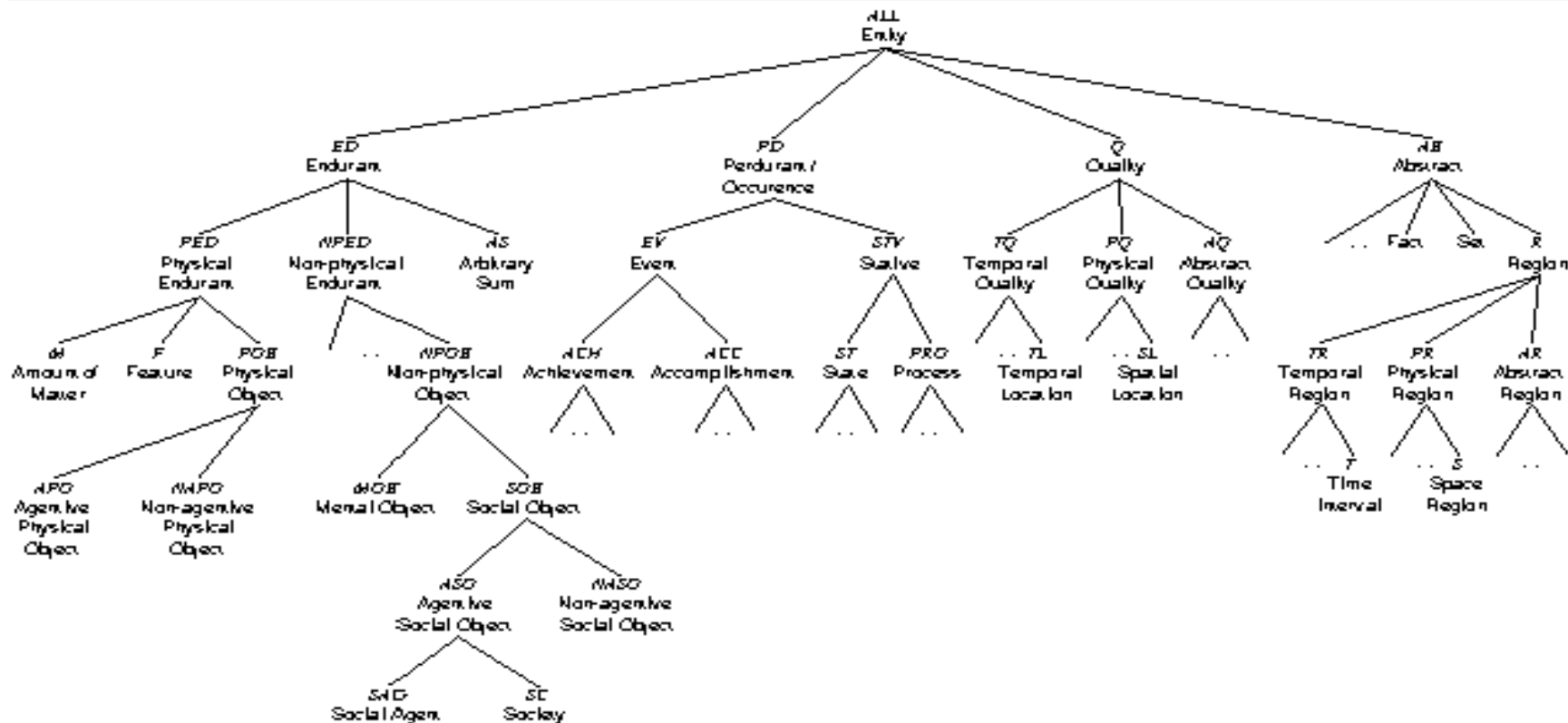
Karakteristika som adskiller ontologi fra de andre tilgange:

- I ontologi taler man om begreber eller klasser; dvs. man abstraherer over ordniveau
- En ontologi har ofte en **aksiomatisk karakteristik** af begreberne som muliggør **inferens** i informationssystemer

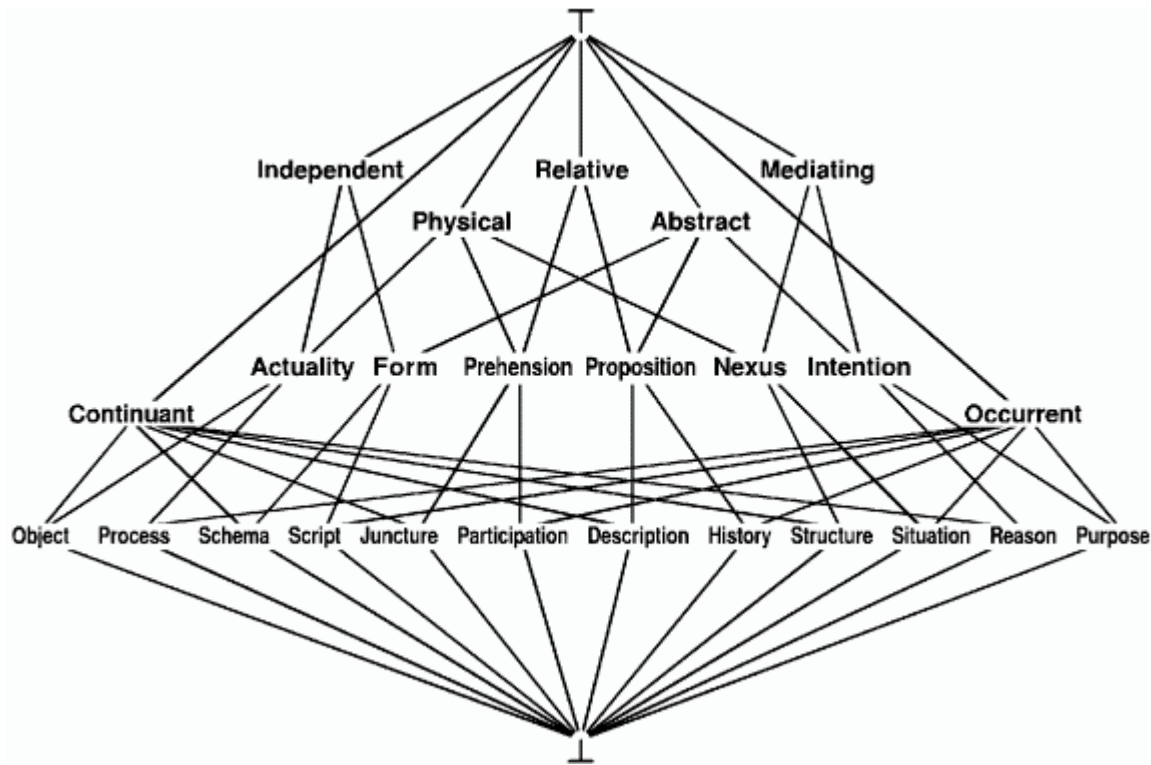
? $x, y : instance(x, Person) ? instance(y, Animal) ? pet(x, y)$

- En ontologi kan antage mange forskellige former (inklusionshierarki, gitterstruktur, graf, topic map)

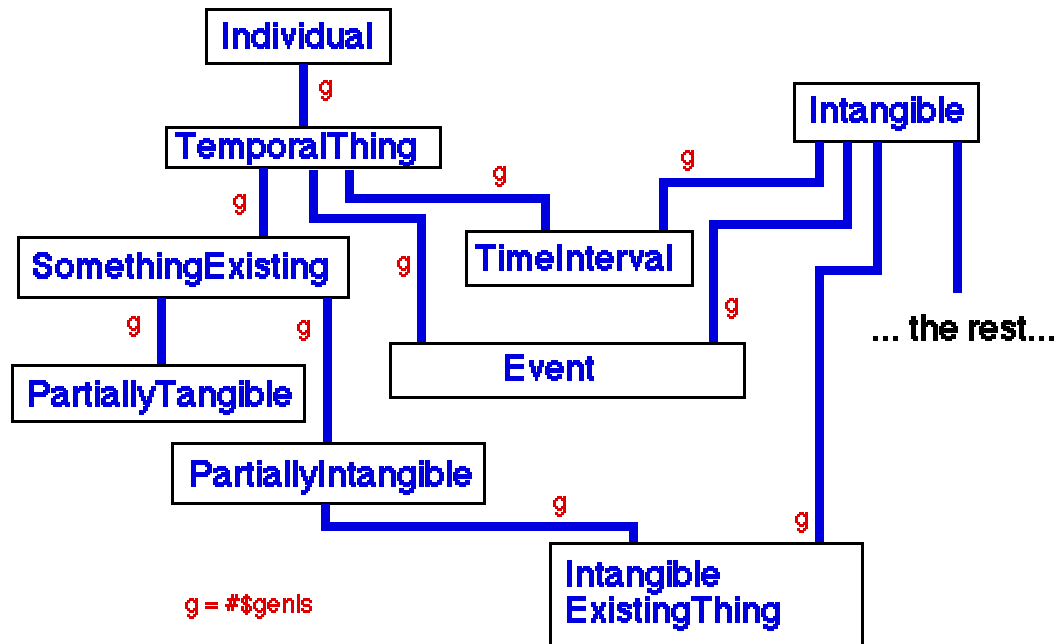
DOLCE Ontology



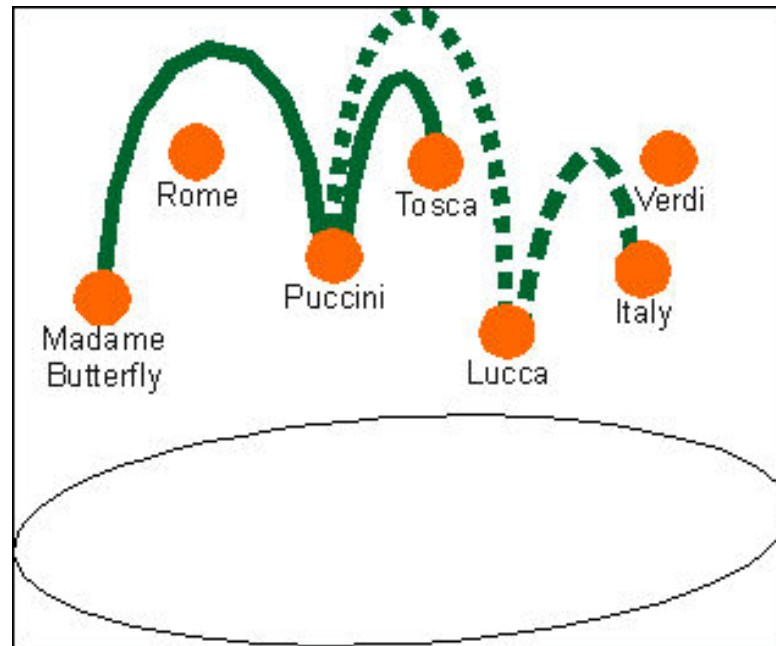
Sowas KR-ontology



Upper Cyc Ontology



Topic map



3. Opbygning af videnbaser

Opbygning af videnmodel kræver overblik over domænets **terminologi**

Kilder:

- Eksisterende ressourcer: egne ordlister, definitioner, tesaurusser mv.
- Uddragede termer fra tekster inden for domænet
- Termeksperternes viden, associationsmetoder

Termidentifikation ud fra tekster

Semiautomatisk termidentifikation, mulig metode:

- Ordene identificeres og ordklasseopmærkes Ordene neutraliseres til grundform
- Navneord, tillægsord og udsagnsord udvælges
- Ordlisten sammenholdes med evt. eksisterende termlister inden for domænet (+liste)
- Ordlisten sammenholdes med en almen ordbog (-liste)
- Finpudsning (flerordstermer, sammensatte ord mv.)

Resultater for termudtræk

Verifikation hos termeksperter:

- Er de fundne termkandidater rent faktisk termer ? (precision)
- Hvor mange af termerne er fundet med den automatiske metode ? (recall)

Afhængig af teksttype og af hvor stort et forarbejde man har gjort med at 'rense' teksterne for støj, bør man få precision og recall på over 80 %.

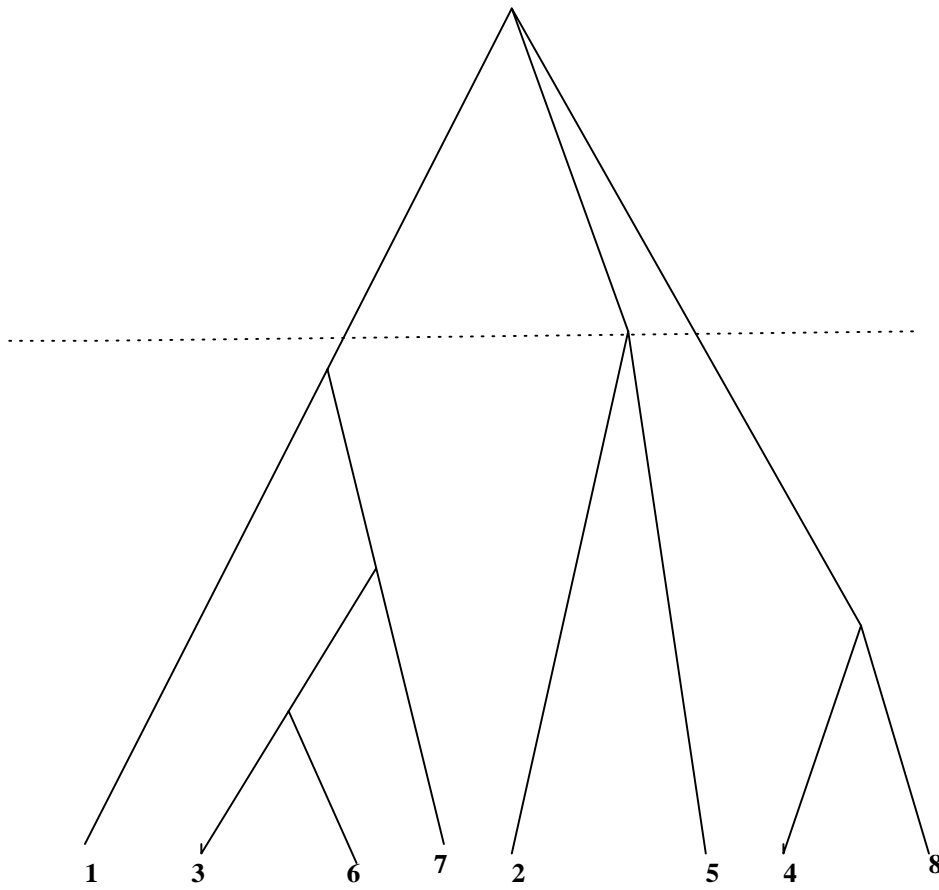
Fra termer til ontologi

Tre-lags model:

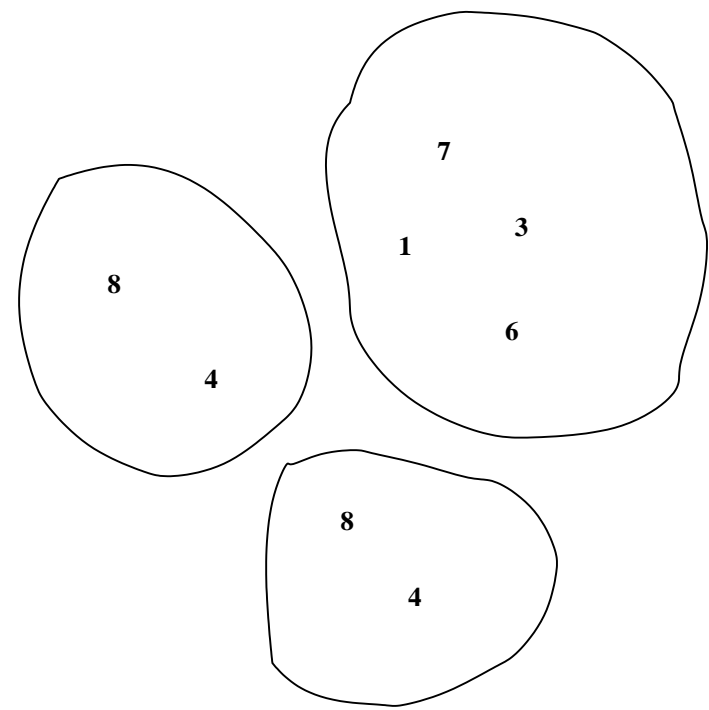
- **Nederste lag:** kan bygges semiautomatisk ud fra termlisten (især sammensatte ord), clustering-metoder
- **mellemlag:** kræver ekstralingvistisk viden om domænet; termeksperternes velvilje central, formålet skal stå klart
- **Øverste lag:** anvend i så høj grad som muligt ontologi- og metadatastandarder
- **Formelle sprog og værktøjer:** vi har eksperimenteret med Ontology Web Language (OWL), RDFS, Topic Maps, værktøj: Protégé

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

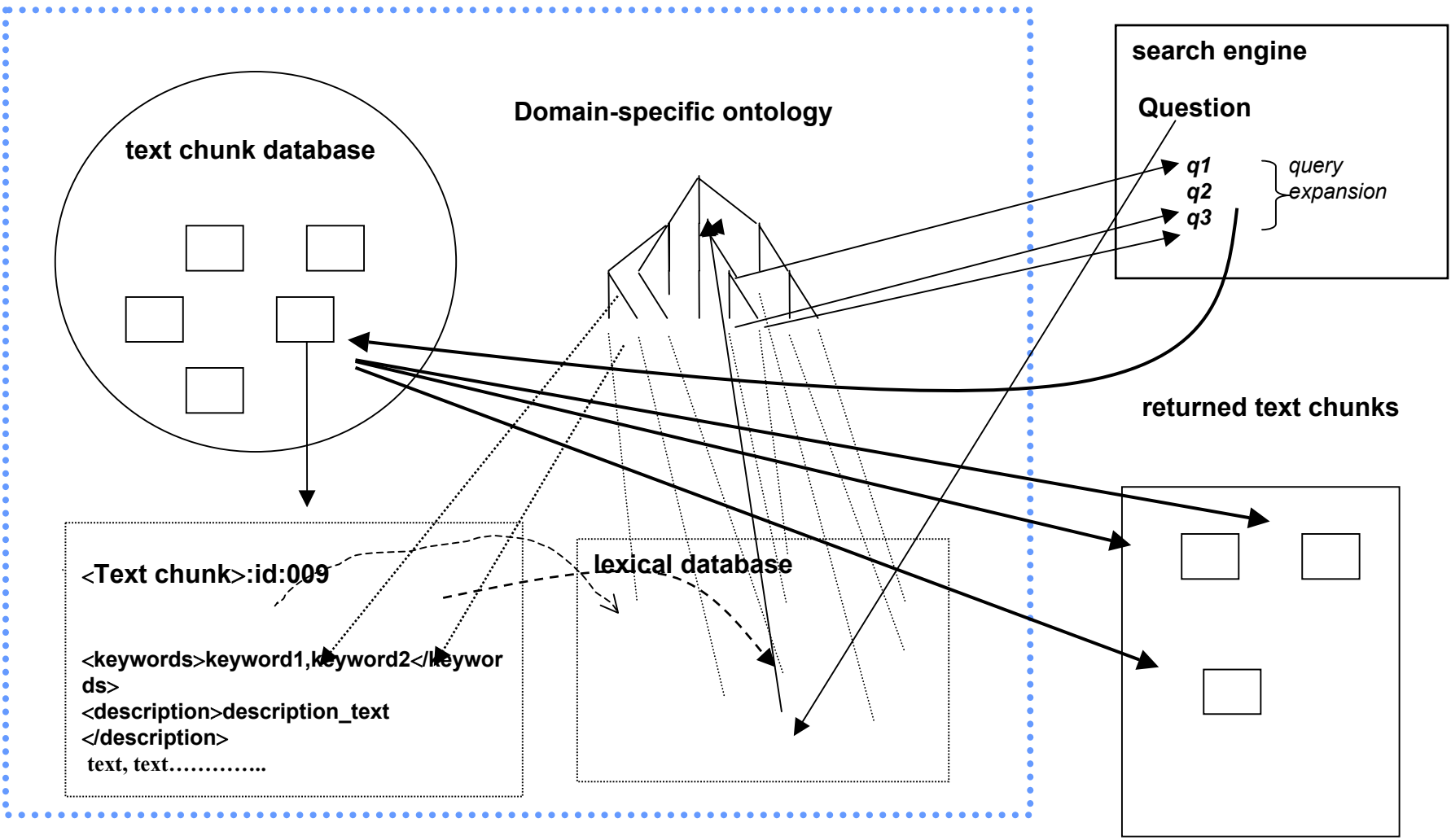
Vores erfaringer med clustering

- Vi har eksperimenteret med CMU-statistikpakken og Lnknet-systemet udviklet af MIT Lincoln Laboratory
- Giver de bedste resultater på store mængder data (fx ignoreres lavfrekvente ord)
- Hjælper ontologidesignereren til at foretage de første grupperinger af data
- Understøtter ikke kodning af relationer mellem ord i de forskellige grupper og mellem grupperne

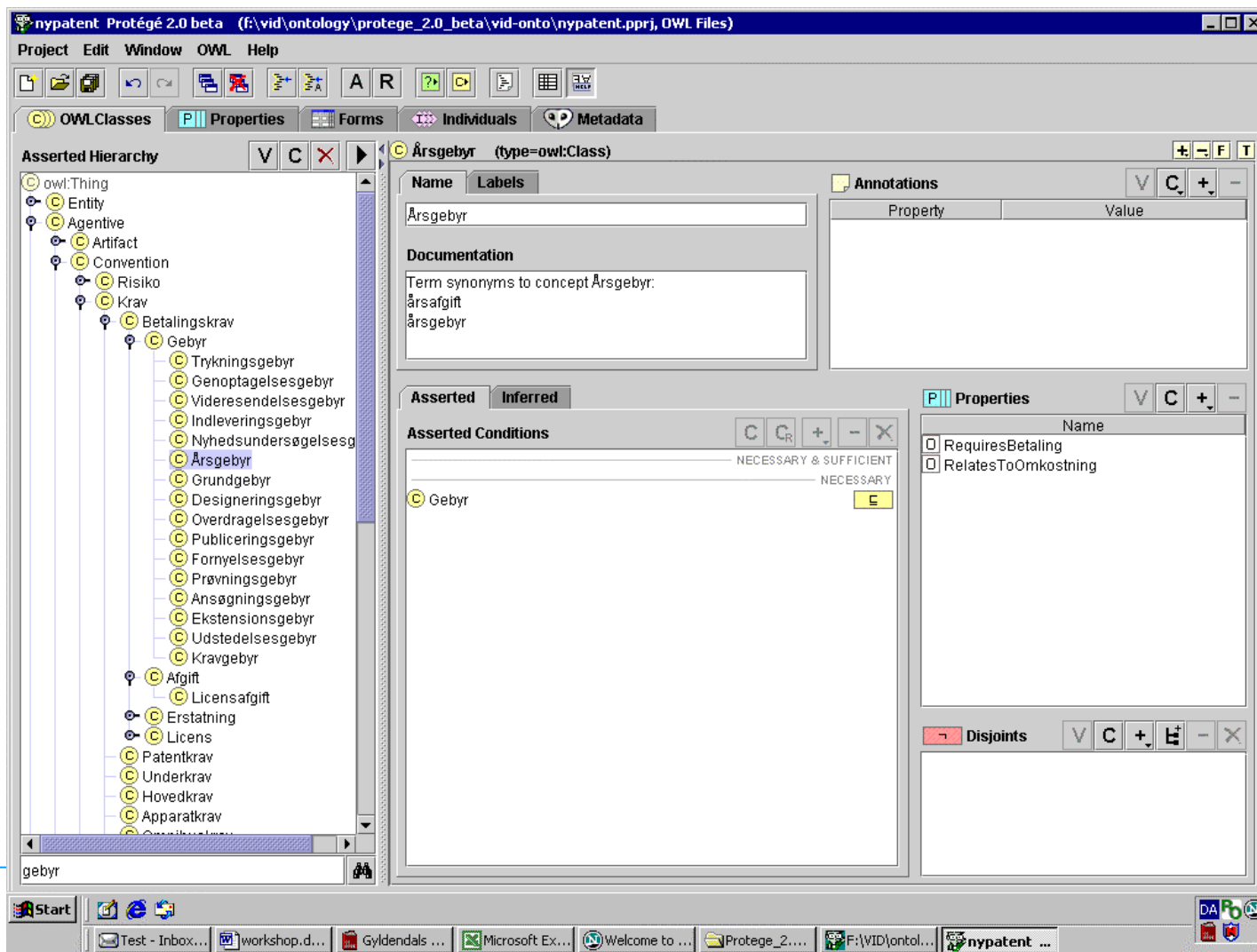
4. To eksperimenter med ontologi og metadata

- **VID:** Viden og Dokumenthåndtering med sprogteknologi
CST + 5 danske/nordiske virksomheder (Ankiro, Navigo, Nordea, B&O, Zacco)
formål: at udforske de forskellige muligheder som sprogteknologi frembyder inden for informationsøgning og dokumentproduktion
- **Moses:** **MO**dular and **Scalable Environment** for the **Semantic** web: Europæiske forskningsinstitutioner og virksomheder.
Danske partnere: CST og det Humanistiske Fakultet
formål: det semantiske web - også flersprogligt

VID: søgning og vedligeholdelse af standarddokumenter, IPR-domænet



IPR (OWL plugins i Protégé)



The screenshot shows the Protégé 2.0 beta interface. The 'OWLClasses' tab is active, displaying an 'Asserted Hierarchy' on the left. The hierarchy includes classes such as `owl:Thing`, `Entity`, `Agentive`, `Artifact`, `Convention`, `Risiko`, `Krav`, `Betalingskrav`, `Gebyr`, and various specific fees like `Trykningsgebyr`, `Genoptagelsesgebyr`, `Videresendelsesgebyr`, `Indleveringsgebyr`, `Nyhedsundersøgelsesgebyr`, `Årsgebyr`, `Grundgebyr`, `Designeringsgebyr`, `Overdragelsesgebyr`, `Publiceringsgebyr`, `Fornylesgebyr`, `Prøvningsgebyr`, `Ansøgningsgebyr`, `Ekstensionsgebyr`, `Udstedelsesgebyr`, `Kravgebyr`, `Afgift`, `Licensafgift`, `Erstatning`, `Licens`, `Patentkrav`, `Underkrav`, `Hovedkrav`, and `Apparatkrav`.

The main workspace shows the details for the selected class, `Årsgebyr (type=owl:Class)`. It includes a 'Name' field with the value 'Årsgebyr', a 'Labels' field, and an 'Annotations' table with columns for 'Property' and 'Value'. The 'Documentation' field contains the text: 'Term synonyms to concept Årsgebyr: årsafgift, årsgebyr'. Below this, the 'Asserted' tab shows 'Asserted Conditions' with a list containing `Gebyr`. The 'Properties' tab shows a list of properties: `RequiresBetinging` and `RelatesToOmkostning`. At the bottom, there is a 'Disjoints' section.

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,patent

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 Designinger 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](#)

SUBJECT:

Mål i Moses

At udvikle metoder til:

- Opbygning og vedligeholdelse af hjemmesiders indhold via semiautomatisk opmærkning
- Tværspørglig søgning i opmærkede ressourcer via brugerforespørgsler i naturligt sprog

Prototype: søgning i universiteters web-sites (KU, Roma III).





MOSES-ontologier

Udgangspunkt: DAML+OIL universitetsontologi

En italiensk og en dansk version defineret i ‘Topic Maps’- formalisme

Dansk ontologi: 200 begreber, 50 relationer

Ontologien bruges til strukturering/opmærkning af data og til sproglig analyse.

Messages  file:/oradisk03/GATE2/New_Danish_Annotated_corpus/  DKann12.xml_0004D  DKann112.xml_001DDText Annotations Annotation Sets Print 

Engelhardt, Juliane

JULIANE ENGELHARDT

Ph.d.-stipendiat

Kontor: 16.1.109.

Telefon: **35 32 94 15**

Personlige data:

Født i 1968 i Qaqortoq (Julianehåb), Grønland. Studentermedhjælp på Nationalmuseet i Brede. **Cand.mag. i historie og filosofi i 1998 fra Københavns Universitet.** Underviser i historie og filosofi på Askov højskole 1998-1999. Fagredaktør på Danmarks Nationalleksikon, historieredaktionen 1999-2001. Ph.d.-stipendiat (finansieret af forskningsstyrelsen) fra 2001.

Væsentligste produktion:

National identitet: konstruktion eller vækkelse? Askov højskoles årsskrift 1998.**"Ueberhaupt glauben wir uns als dänische Bürger". Slesvig-holstensk helstatspatriotisme 1784 - 1814. Meddelelser fra Thorvaldsens Museum 2001.****Adel er arvelig, men Dyd maaerhverves. Den patriotiske bevægelse 1780-1799. Fortid og Nutid, 2002, nr. 3.**

Nuværende forskning:

Projektets overordnede tema er statspatriotisme i det danske monarki 1780-1814. Mere konkret vil jeg kortlægge de utallige patriotiske selskaber, der opstod i Danmark, Norge og de tyske hertugdømmer Slesvig og Holsten i denne periode. Hvor opstod de, hvem stiftede dem, hvad ville medlemmerne og hvad var deres forhold til statsmagten? Ved at inddrage selskaber, der ikke var danske i den nationale forståelse af ordet, kan man samtidig afprøve spændevidden i den helstatspatriotiske ideologi.

Det er min hypotese, at selskaberne var konkrete resultater af den ideologiske udvikling, og dermed eksponenter for nogle af periodens centrale tanker. Medlemmerne var filantroper, men argumenterede samtidig for, at fattigdom skulle afhjælpes ved arbejdstvang. De ønskede bondens frihed og emancipation, men bondestanden skulle opdrages og disciplineres til arbejdsomhed og stræbsomhed. De var rationalister og mente, at fornuftens dannelse kunne højne menneskets erkendelse. Samtidig talte de i et højstemt sprog og brugte metaforer, der var pietistisk

Annotations Editor Features Editor Initialisation Parameters

Default annotations

- Dato
- Fakultet
- Institut
- Lookup
- PersonligHjem
- PhDStuderend
- Publikation
- Universitet
- Universitetsgra
- proporgphone
- proptelefonnum
- propwebadres
- zAffilieret
- zPersonHjem
- zSubOrganisati
- zUnivGrad

Original markups anno

- a
- b

Forespørgslerne

Hvilke ph.d.-studerende er der på det humanistiske fakultet på KU?

På hvilke universiteter kan man læse tysk filologi?

Applikationsdomæner:

Ansatte/Kurser/Forskning

Afslutning

- Samlede løsninger, harmonisering og brug af fælles standarder er absolut efterstræbelsesværdigt
- Men: vigtigt at genbruge eksisterende klassifikationer fra de enkelte delområder; større chance for gennemslagskraft;
- Sprogpolitisk mission:
Vælg sprogteknologiske løsninger som er udviklet for dansk!
- FOVITS-kurser: Forskningsbaseret Videreuddannelse i IT og Sprog
(KU, HHK, AAU)