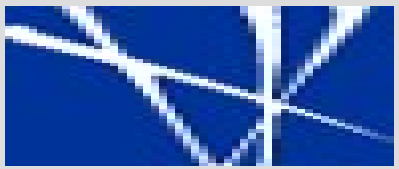
# Vidensekstraktion fra Internettet - en encyklopædisk browser?

DATALIngvistisk SEminar
Copenhagen Business School
29. marts, 2006

Jakob Halskov
Ph.d.-studerende
jh.id@cbs.dk



# Program

- Viden – hvad og hvorfor?
- Automatisk vidensfremfinding
- Den mønsterbaserede tilgang
- To problemer: precision og recall (**domæneafhængighed**)
- Inkrementel vidensfremfinding og -strukturering: En implementering
- Kommende arbejde: evaluering

# Motivationserklæring

Omverdensviden og domænespecifik viden
= vigtig i mange NLP-applikationer (fx. MT, dialogsystemer, IR etc.)

**MEN: "Knowledge Acquisition Bottleneck"**

Tre faser

1. **Vidensfremfinding** (*Knowledge Extraction*)
2. Vidensstrukturering (*Knowledge Engineering*)
3. Vidensdeling (*Knowledge Sharing*)

- Hvorfor bruge Internettet?
dynamisk, flersproget, tværfagligt, allestedsnærværende, gratis ...

# Hvad er viden?

Filosofisk set: holdning + bevismateriale ("justified true belief")

Terminologisk set: **begrebers trækspecifikationer** herunder
semantiske relationer mellem begreber

1. **implicit** viden (personlig, skjult i vores hoveder eller i termer)
2. **eksplicit** viden (frit tilgængelig, artikuleret og forklaret på skrift)

TO ANTAGELSER:

- Semantiske relationer realiseres i sproget som vidensmønstre
- Mange vidensmønstre er universelle (domæneuafhængige)

# Vidensmønstre og videnrige kontekster

- *knowledge probes* (Ahmad, 1992)
- *explicit relation markers* (Bowden, 1996)
- *knowledge patterns* (Meyer, 2001)

*Knowledge-rich context* (KRC) = "ontologiske byggeklodser"

"a context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation." (Meyer, 2001)

**NP + vidensmønster + NP** (fx. indexing **improves** retrieval speeds)

# Hvordan fremfindes viden automatisk?

- **Conceptual clustering** (Michalski 1983, Fisher 1987)
  + Induktiv/empirisk ("Distributional hypothesis": kontekstuel lighed = semantisk lighed)
  - uklassificerede "begrebsklynger", uklassificerede relationer

- **Mønsterbaserede tilgange** (Ahmad 1992, Hearst 1992)
  - Deduktiv/introspektiv (som udgangspunkt)
  + Klassificerede relationer

- **Hybride tilgange** (Cimiano & Staab 2005)

Begrænsning: fokus på taksonomier (generiske relationer)

# Hvilke mønstre og hvilke semantiske relationer?

Hvilke semantiske relationstyper er terminologisk interessante?

Prinicipielt uendeligt mange!

**Begrebsrelationer** contra **leksikalske relationer**

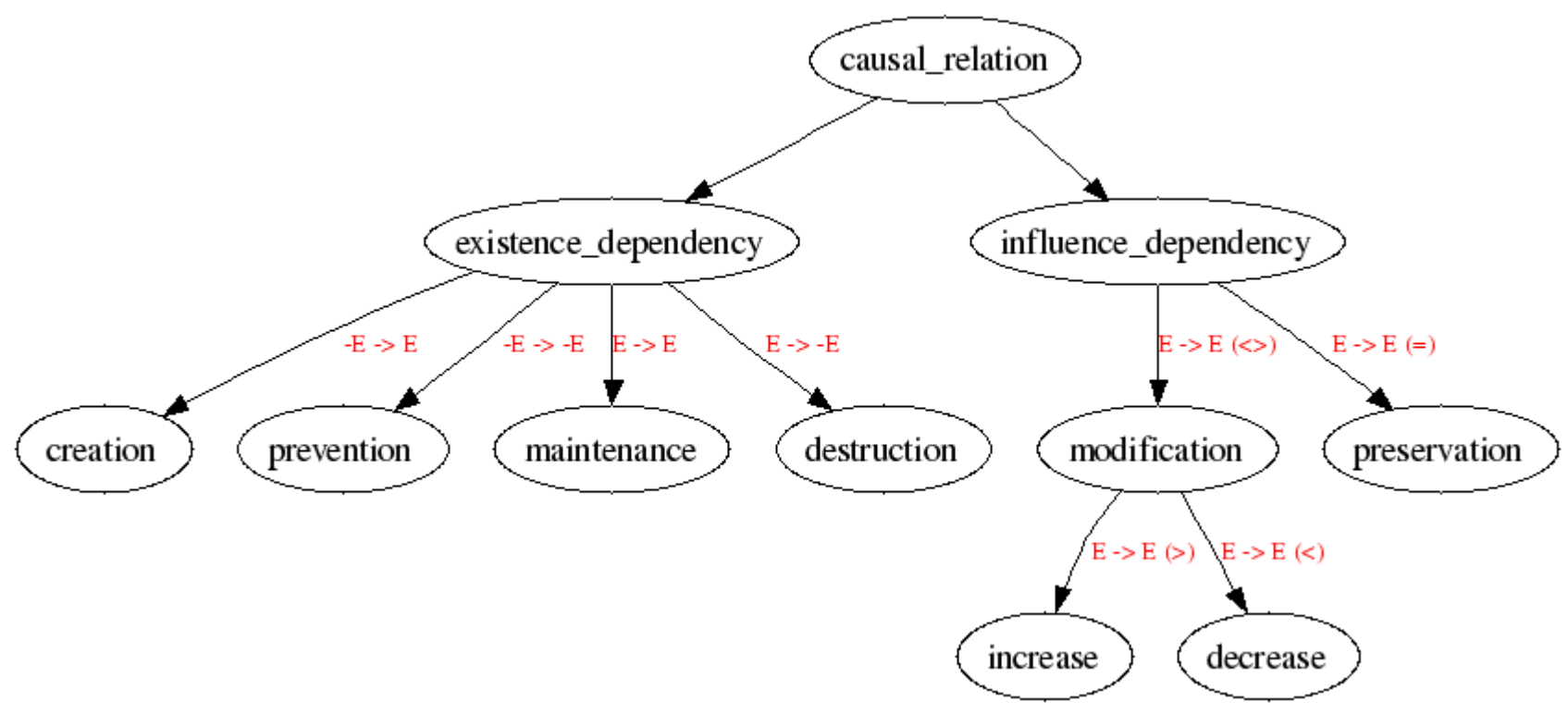
Fx. er synonymi ikke nogen begrebsrelation (Madsen, 2001)

Fem grupper af vidensmønstre (fra "TerminoWeb" - Barrière, 2005):
meronymi (7), hyponymi (8), synonymi (3), kausalitet (39), funktion (7)

Mulige tilføjelser: temporal, lokativ, rolle ..... (Madsen, 2001)

# Detaljeringsgrad

Underinddeling af den kausale relation (Barrière, 2002)



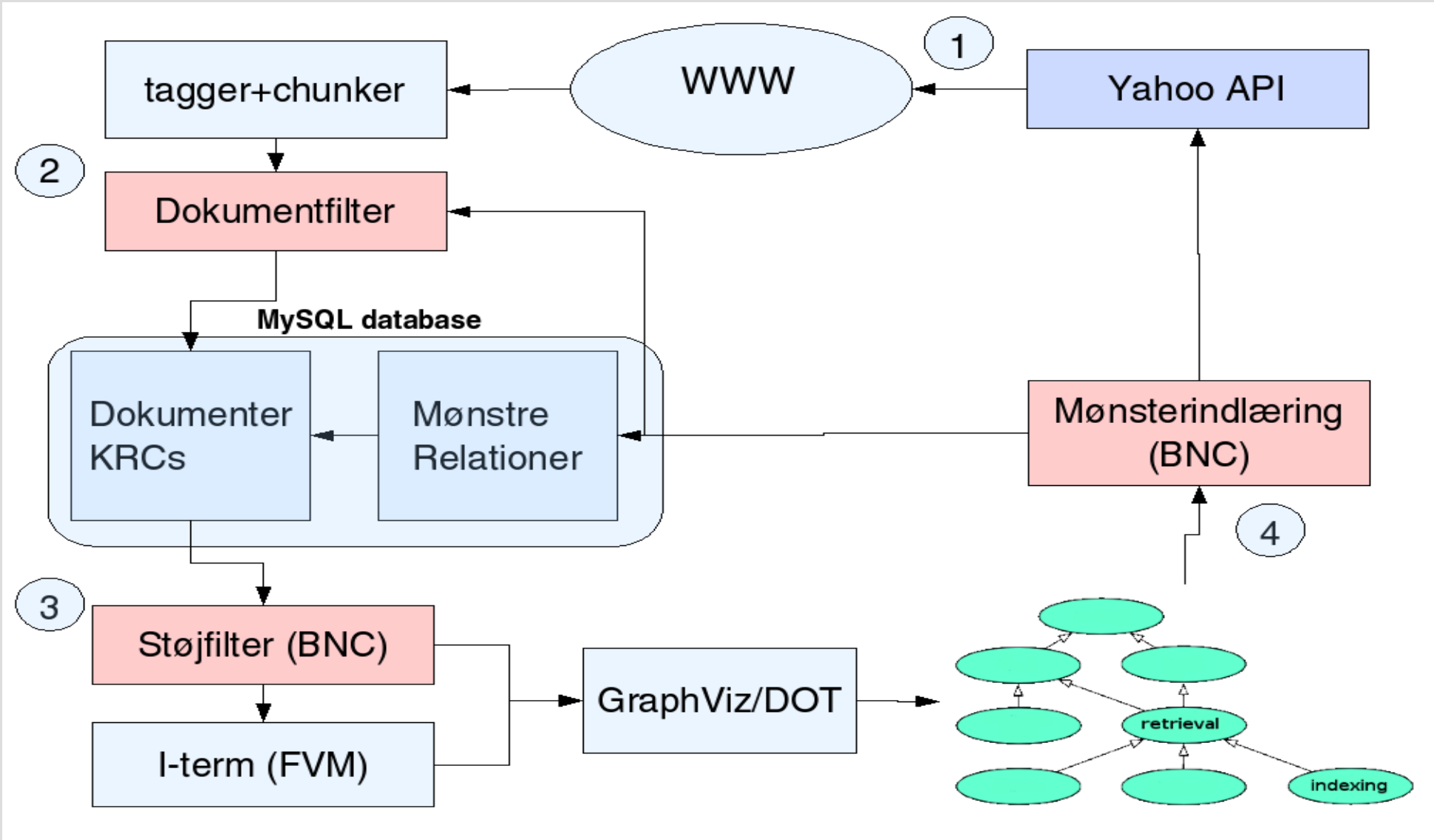

eksempler: CREATION: produce, generate, create ...
INCREASE: promote, aid, enhance, encourage ...

# 4 problemer med mulige løsninger

1. Meget viden er **implicit** (recall)
2. Internettet er fuld af underlødig information (precision)
3. Vidensmønstre er **støjende** (precision)
4. Mange vidensmønstre er **domæneafhængige** (recall)

    1. Brug WWW til at indhente tilstrækkelige data
    2. Implementér et dokumentfilter (tæthed af “akademiske ord”)
    3. Træn et støjfilter på et almensprogligt korpus
    4. Lær hvert domænes unikke mønstre vha. frekvenssammenligninger med et almensprogligt korpus

# Et vidensekstraktionssystem



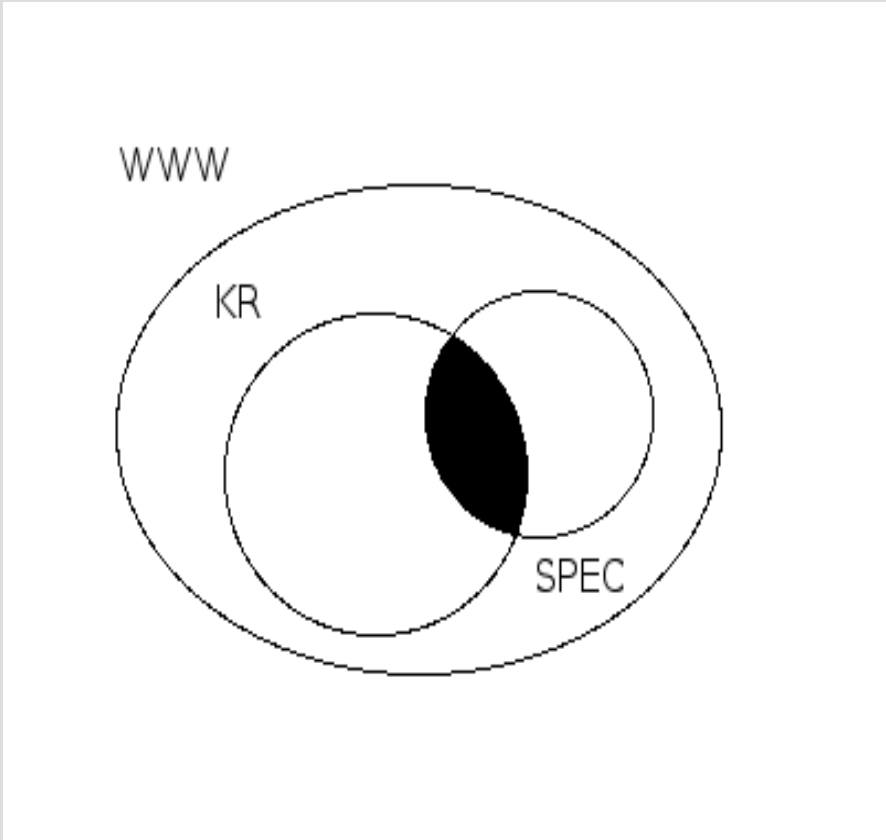
# Precision 1: et dokumentfilter

WWW: ustrukturerede, uklassificerede dokumenter => **Dokumentfilter**

Fællesmængden af dokumenter rige på **eksplicit viden** (KR) og **specialiserede** dokumenter (SPEC).

KR: høj tæthed af vidensmønstre (Barrière, 2005)

SPEC: høj tæthed af “akademiske ord” (Coxby, 2000)



# Akademiske ord m.v.

"Academic Word List" (Coxby, 2000)

**570 lemmaer** ("word families" - fx. *analyse*, *concept*, *data*, *research* ...)
ekstraheret fra et akademisk korpus på ca. 3,5 mio. ord

4 subkorpora (Arts, Commerce, Law, Science) og 28 fagområder

<u>Tre indlemmelseskriterier:</u>

- Må ikke forekomme blandt de 2000 hyppigste engelske ord
- Skal forekomme mindst 10 gange i hvert subkorpus && i mindst 15 fagområder
- Skal forekomme mindst 100 gange i hele korpusset

# Precision 2: et støjfilter

## To slags støj

- Vidensmønstret realiserer ikke nogen semantisk relation mellem de to argumenter

F.eks. “There **is a** sign on the door”, “What **prevents** an infinite loop?”

- Vidensmønstret realiserer en semantisk relation mellem to almensproglige ord (= meget overordnede kategorier)

F.eks. “The universe **is a** cold, dark place” (BNC)

# To måder at lære støj på

Støjkilde: **British National Corpus** (ca. 100 mio. ord)

- Active Learning: gentagen **manuel opmærkning** af minimalt antal grænsetilfælde
- Naïve Bayes ("Bag -of-Words"): **ingen opmærkning**, men antagelse som er delvist ukorrekt

Antagelse: kontekstuel lighed med BNC = støj!

$P(BNC|KP,features) = P(BNC|KP,f_{-n})*...*P(BNC|KP,f_{+n})*P(BNC|KP)$

$$\approx \frac{C(f_{-n}|KP,BNC)}{C(BNC)} * \ldots * \frac{C(f_{+n}|KP,BNC)}{C(BNC)} * \frac{C(KP|BNC)}{C(BNC)}$$

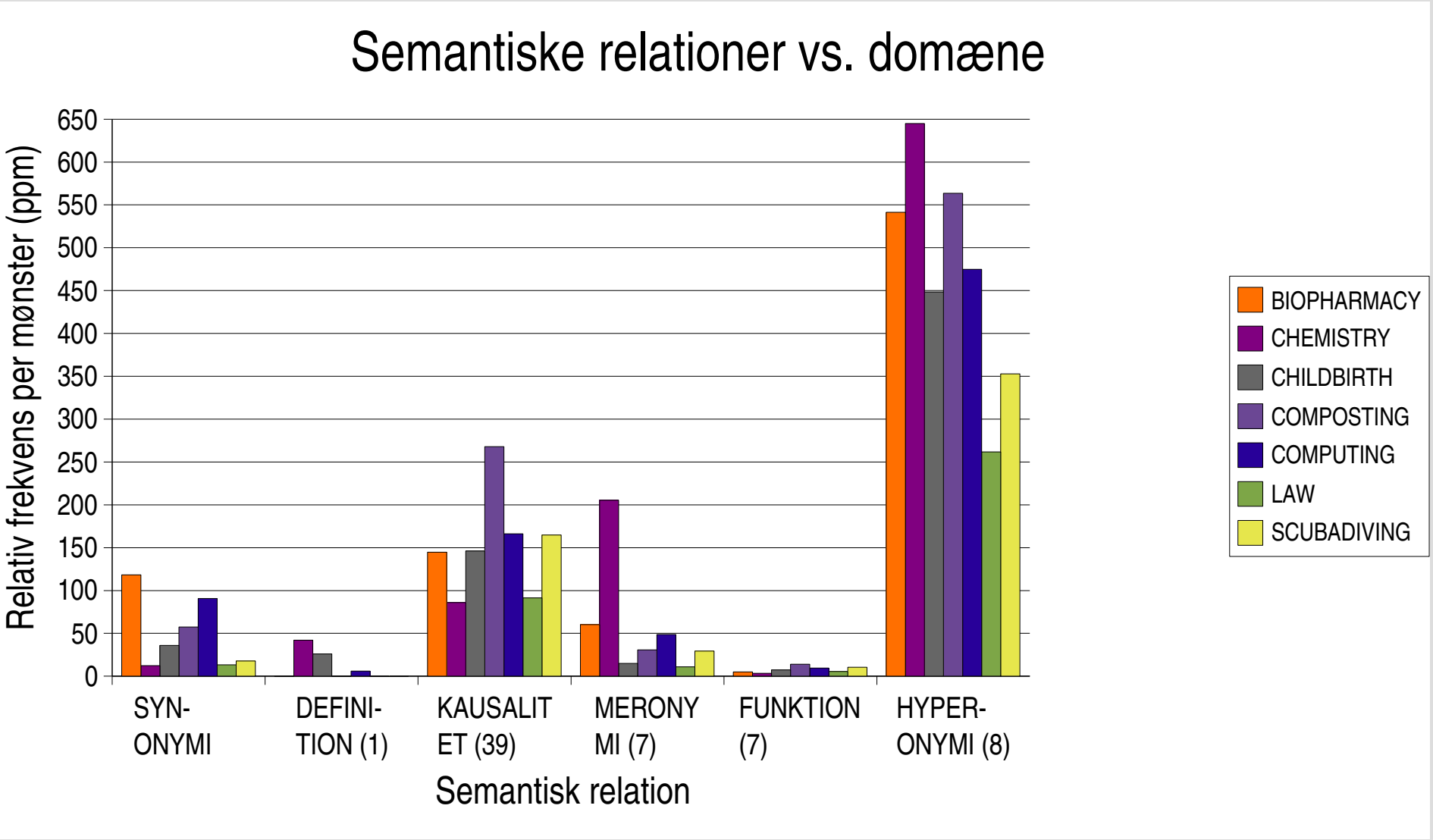
# Domæneafhængighed: Empiri

| | BIOPHARMACY | CHEMISTRY | CHILDBIRTH | COMPOSTING | COMPUTING | LAW | SCUBADIVING |
|---|---|---|---|---|---|---|---|
| Tokens | 166K | 216K | 546K | 93K | 650K | 1046K | 151K |

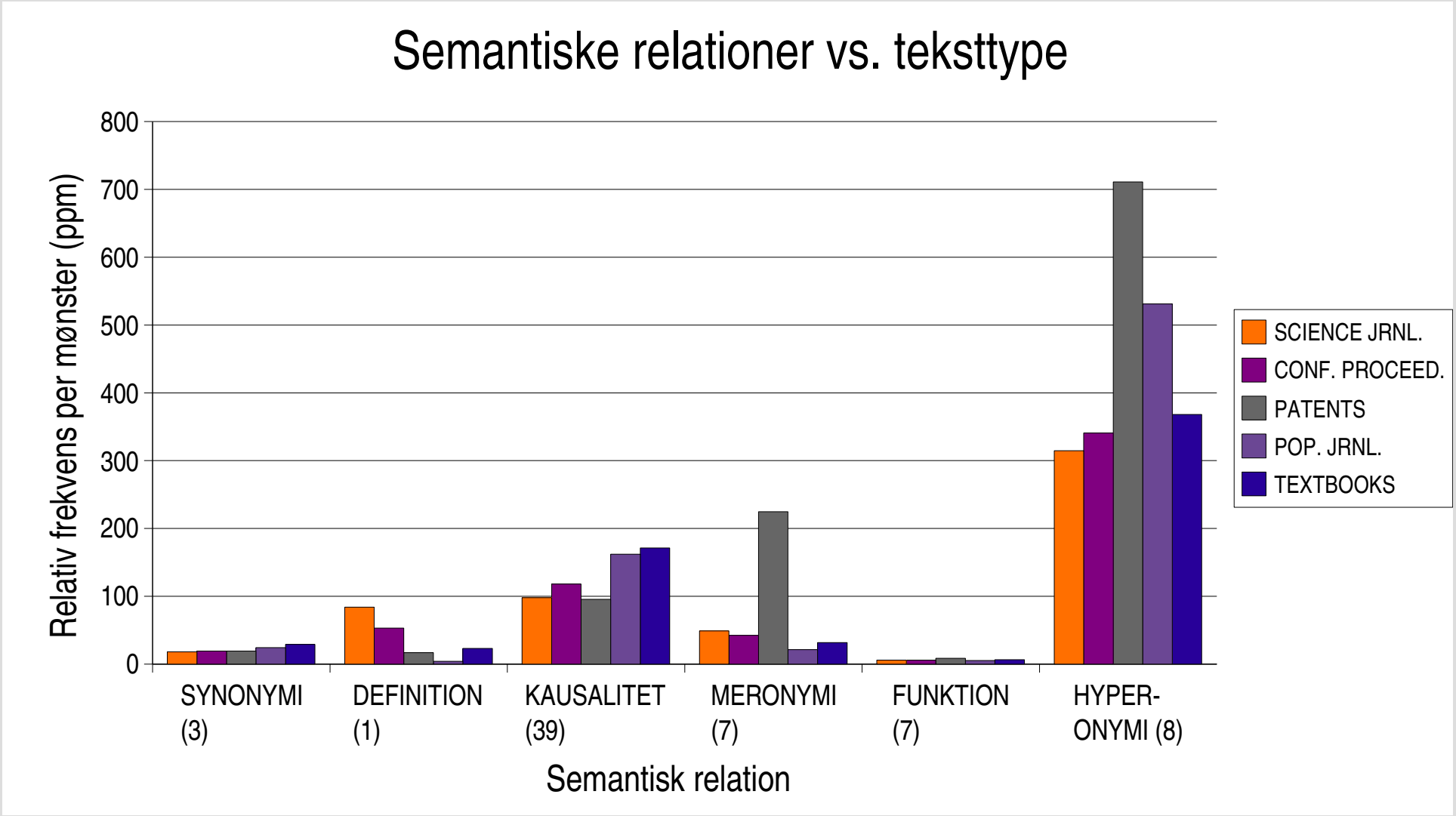
| | IT PATENTS | IT CONF. PROCEED. | IT SCIENCE JRNL. | IT TEXTBOOKS | IT POP. JRNL. |
|---|---|---|---|---|---|
| Enheder | 162 ansøgn. | 964 artikler | 508 artikler | 27 bøger | 65 numre |
| Tokens | 2028K | 5451K | 4447K | 5692K | 8536K |

| | BNC |
|---|---|
| Tokens | 110M |

# Fem relationstyper i syv domæner



# Fem relationstyper i fem teksttyper (IT)



# Recall: Nye vidensmønstre

Hvilke vidensmønstre er karakteristiske for et domæne?

| | korpus = LSP | korpus ≠ LSP |
|---|---|---|
| token = Vtrans | a | b |
| token ≠ Vtrans | c | d |

$$\log - odds\, ratio = \log\left(\frac{a * d}{b * c}\right)$$

Opmærkede korpora (IMS Treetagger + Yamcha):

TOKEN POS LEMMA CHUNK

Transitive verber (ikke-datid):

[chunk='.-NP'] [chunk='.-VP']* **[chunk='.-VP' & pos='VV[ZPNG]?'] [pos='IN]?** [chunk='B-NP']

# Eksempel 1: “Biopharmacy”

**De 10 verber med størst “overrepræsentation” i korpusset BIOPHARM contra BNC**

| Log-odds ratio | BIO RF* | BNC RF | verbum | TW | relation/klasse (Levin,1993) |
|---|---|---|---|---|---|
| -4.37 | 517.30 | 6.58 | **treat with** | nej | **CAUS_modification / ?** |
| -4.09 | 252.63 | 4.21 | **target** | nej | **? / ?** |
| -3.89 | 270.68 | 5.56 | **derive** | nej | **CAUS_creation / C&T** |
| -3.84 | 240.60 | 5.17 | **activate** | nej | **CAUS_creation / ?** |
| -3.23 | 276.69 | 10.94 | **stimulate** | nej | **CAUS_increase / Psych.** |
| -3.22 | 234.59 | 9.37 | **induce** | nej | **CAUS_creation / ?** |
| -3.10 | 156.39 | 7.06 | **evaluate** | nej | **? / Assessment verbs** |
| -2.60 | 300.75 | 22.38 | **demonstrate** | nej | **? / Communication verbs** |
| -2.51 | 162.41 | 13.19 | **enhance** | ja | **CAUS_increase / ?** |
| -2.51 | 842.11 | 68.67 | **develop** | nej | **CAUS_creation / C&T** |
| **...** | **...** | **...** | **...** | **....** | |

*** Mindst 50 ppm**

**C&T: Creation & Transformation**

# Eksempel 2: “Computing”

| Log-odds ratio | LSP_RF* | LGP_RF | verbum | TW | relation/klasse (Levin, 1993) |
|---|---|---|---|---|---|
| -4.56 | 53.81 | 0.56 | **disable** | nej | CAUS_destruction / ? |
| -4.13 | 69.18 | 1.11 | **read from** | nej | LOC_DYN_SRC / Comm. |
| -3.44 | 55.35 | 1.77 | **load** | nej | CAUS_creation / Putting |
| -3.35 | 129.14 | 4.51 | **run at** | nej | CAUS_maint. / Existence |
| -3.35 | 127.61 | 4.48 | **access** | nej | LOC_DYN_TAR / ? |
| -3.28 | 109.16 | 4.09 | **execute** | nej | CAUS_creation / ? |
| -3.19 | 112.23 | 4.62 | **store in** | nej | LOC_STA_ACT / H&K |
| -3.16 | 116.84 | 4.94 | **store** | nej | LOC_STA_ACT / H&K |
| -3.16 | 55.35 | 2.34 | **operate at** | nej | CAUS_maint. / Existence |
| -3.14 | 159.89 | 6.91 | **connect** | nej | LOC_DYN_TAR / C&A |
| ... | ... | ... | ... | ... | |

*Mindst 50 ppm
H&K: Hold & Keep Verbs
C&A: Verbs of Combining and Attaching

# Eksempel 3: “Law”

| Log-odds ratio | LSP_RF* | LGP_RF | verbum | TW | klasse iflg. Levin (1993) |
|---|---|---|---|---|---|
| -4.51 | 68.84 | 0.76 | **indemnify** | **nej** | CAUS_prevention / ? |
| -3.93 | 50.68 | 0.99 | **repudiate** | **nej** | ? |
| -3.82 | 51.63 | 1.13 | **bind(ing) on** | **nej** | patient-result? |
| -3.69 | 117.61 | 2.93 | **terminate** | **nej** | CAUS_destruction / Asp |
| -3.59 | 55.46 | 1.54 | **procure** | **nej** | CAUS_creation / C of P |
| -3.36 | 270.60 | 9.43 | **enter into** | **nej** | patient-result? |
| -3.34 | 123.35 | 4.36 | **specify in** | **nej** | LOC_STA_ENTITY |
| -3.32 | 257.21 | 9.32 | **recover** | **nej** | CAUS_creation/ CofP |
| -3.30 | 98.49 | 3.64 | **discharge** | **nej** | CAUS_destruction / VoR |
| -3.15 | 79.36 | 3.39 | **notify** | **nej** | ? / Communication verbs |
| ... | ... | ... | ... | | |

*Mindst 50 ppm
C of P: Change of Possession Verbs
VoR: Verbs of removing

# Universelle vidensmønstre?

Transitive verber med RF > 50 ppm og LO < -0.5 i mindst 3 ud af de 7 forskellige domæner:

| verbum (lemma) | antal forsk. domæner |
|---|---|
| **require** | 7 |
| **use**, provide, produce, include, contain, cause | 5 |
| **use in**, reduce, prevent, **determine**, allow | 4 |
| **use for**, **treat**, result in, **perform**, **involve**, increase, improve **enable**, depend on, **deliver**, consist of, **base on**, **avoid** **associate with**, **add** | 3 |

*Ikke i TerminoWeb: 14 / 27

# Fremfinding -> strukturering

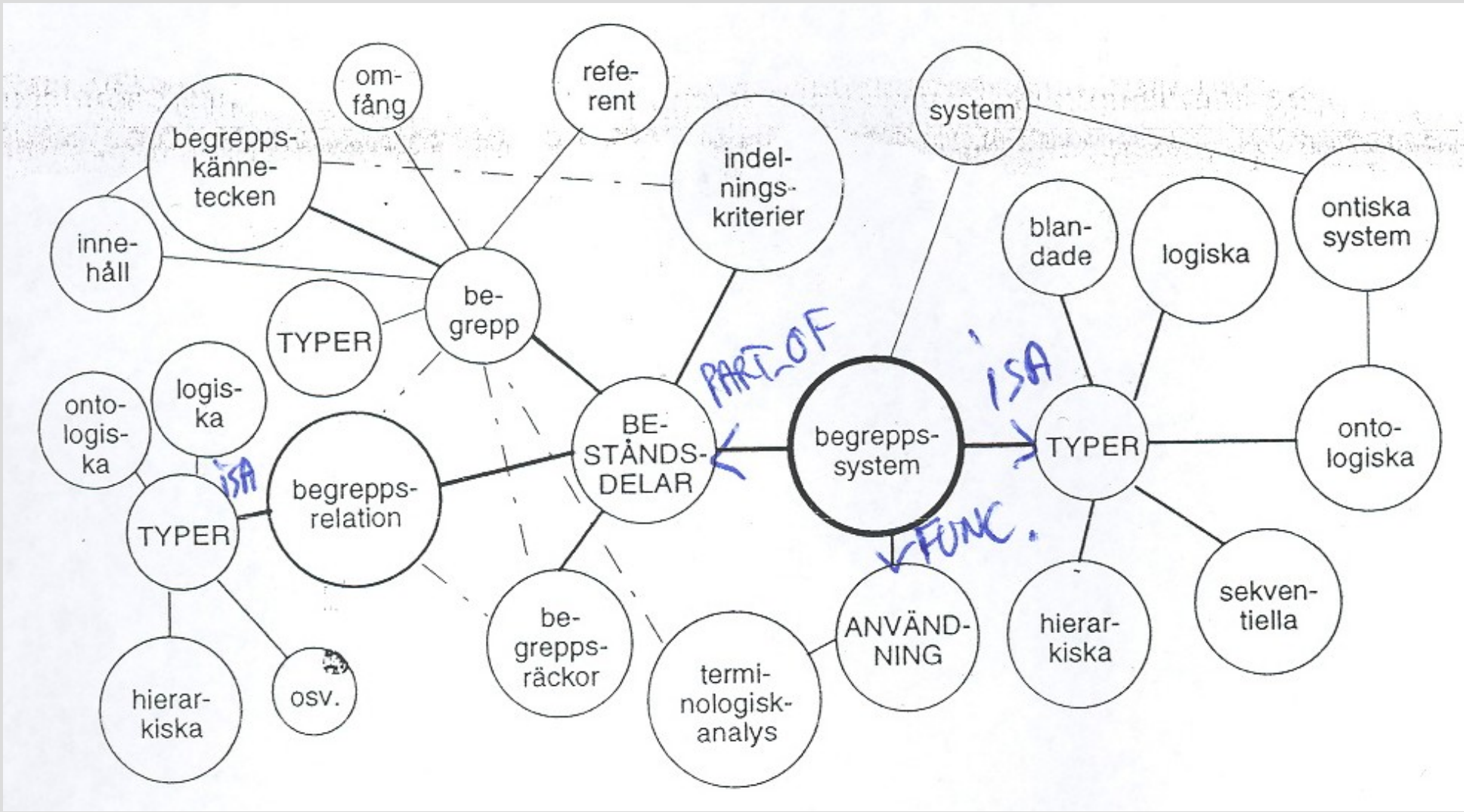
- Hierarkiske contra ikke-hierarkiske begrebssystemer
- Logiske contra ontologiske begrebsrelationer
- Intensionel (indhold) contra ekstensionel (omfang) begrebsanalyse

Visualisering:

Nuopponen (1994): **satellitsystem** (makrosystem, blandet)

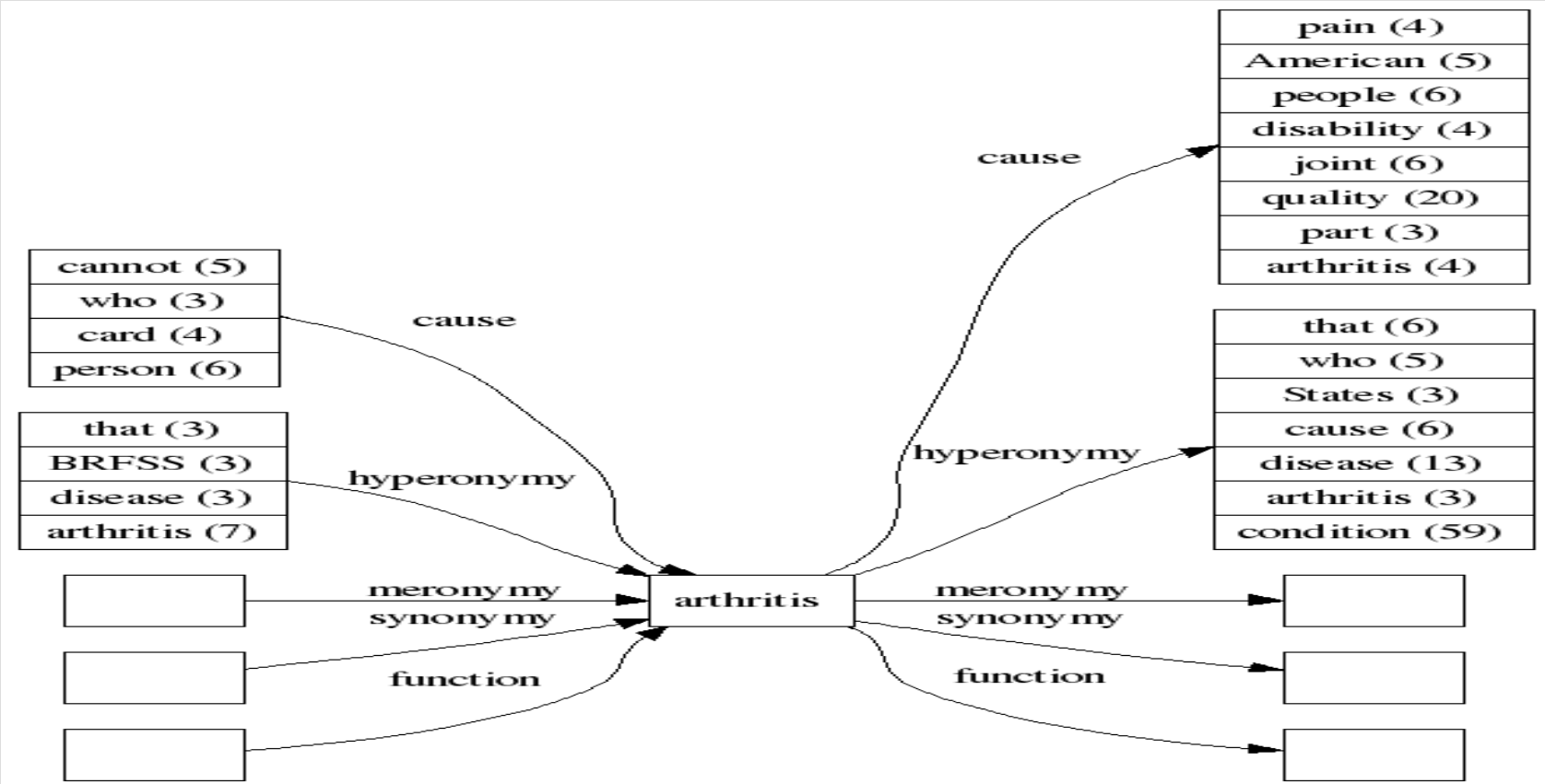
- Udgangspunkt: et kernebegreb
- Udforskningsmuligheder: 5+ relationstyper
- Mål: **inkrementel opbygning** af (blandede) begrebssystemer.

# Nuopponens (1994) satellitsystemer

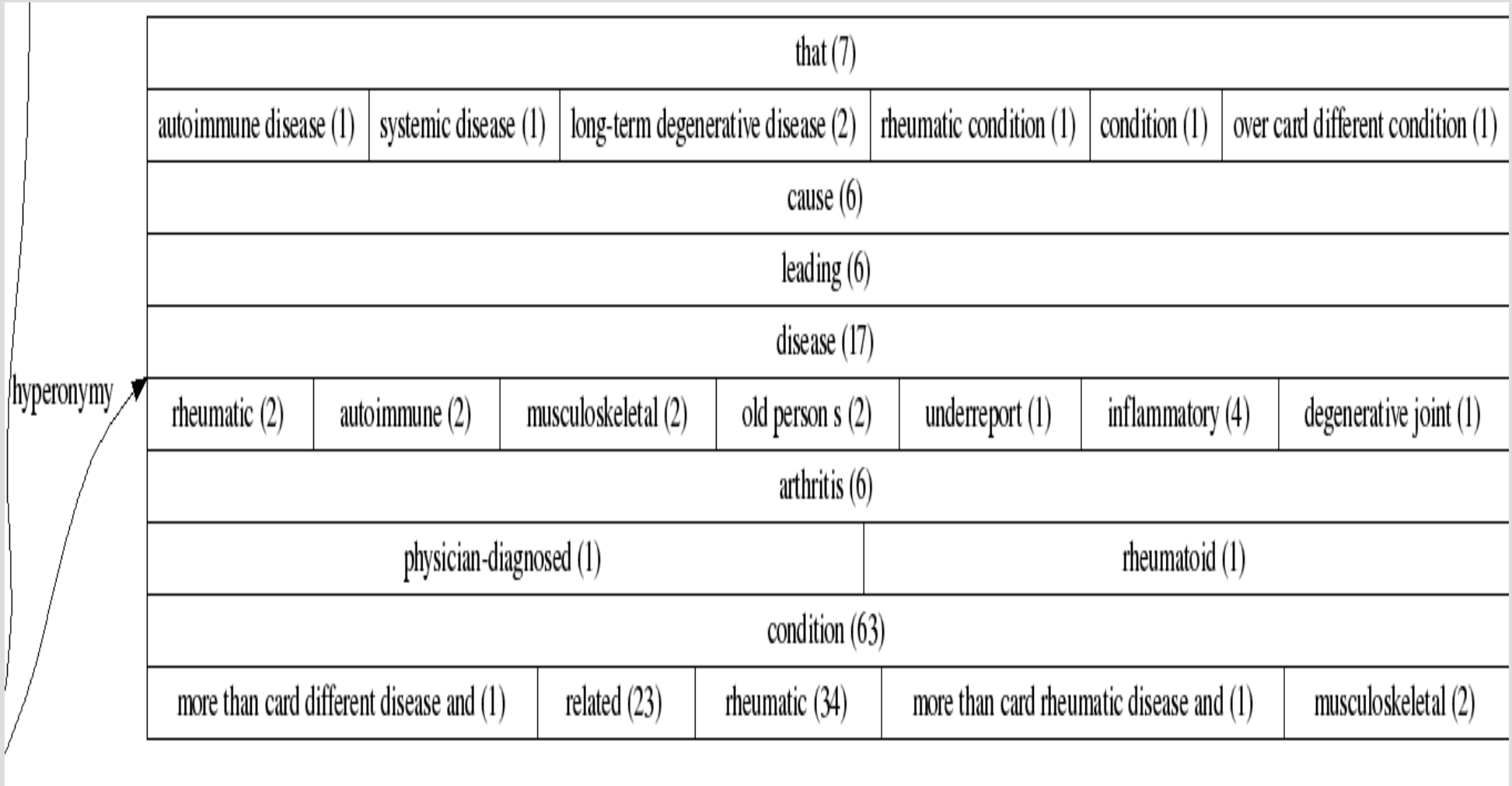


# Den Encyklopædiske Browser

Korpusstørrelse = 378.000 (50 dok.), forespørgsel = "**arthritis**", intet dokumentfilter, intet støjfilter, ingen mønsterindlæring, minimum frekvens af NP-kerne = 3

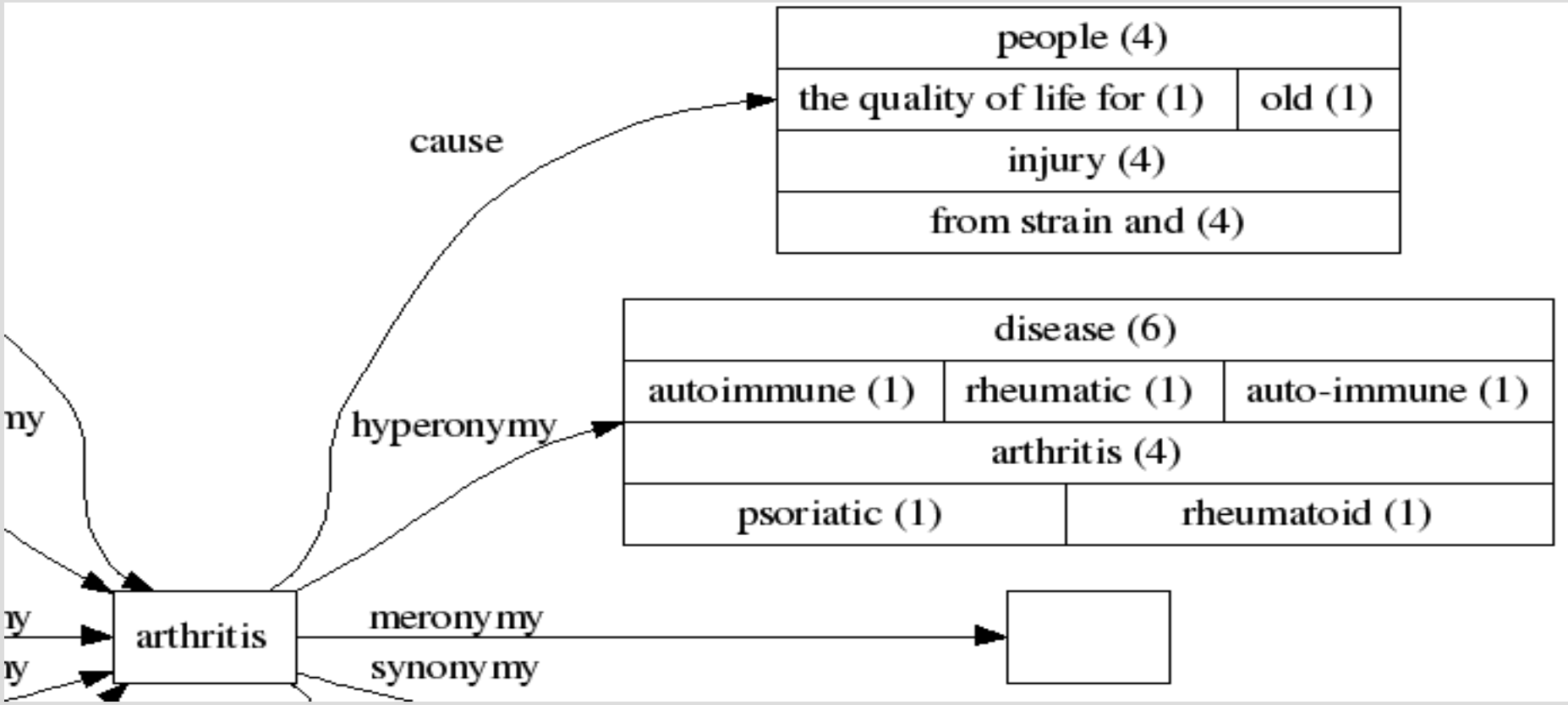


# Opløsning af komposita



# PDF eller HTML?

Korpusstørrelse = 76.000 (50 dok.), forespørgsel = "**arthritis**", intet dokumentfilter, intet støjfilter, ingen mønsterindlæring, minimum frekvens af NP-kerne = 3



# Evaluering af filtre og domæneafhg.

- Køre systemet for samme domæne

+/- dokumentfilter
+/- støjfilter (Active Learning contra Naïve Bayes, BOW)

og bede domæneeksperter **måle precision**

- Køre systemet for forskellige domæner

+/- mønsterindlæringsmodul

og bede domæneeksperter **vurdere recall**

# References


Ahmad, Khurshid (1992) “Semantic Relations an their Use in Elaborating Terminology” Tech. Report, University of Surrey

Barrière, Caroline; Akakpo Agbago (2005) “Corpus construction for terminology” I: *Proceedings of Corpus Linguistics 2005*, Birmingham, UK

Barrière, Caroline (2002) “Hierarchical refinement and representation of the causal relation” I: *Terminology*, vol. 8, pp. 91-111, John Benjamins

Bowden, P. R. (1996) “Extracting Conceptual Knowledge from Text Using Explicit Relation Markers” I: *Proceedings of the 9th EKAW Workshop on Advances in KA*, pp. 147-162

Coxby, Averil (2000) “A New Academic WordList” I: *TESOL Quarterly*, vol. 34, pp. 213-238

Cimiano, Philipp; Steffen Staab (2005) “Learning Concept Hierarchies from Text with a Guided Hierarchical Clustering Algorithm” I: *Proceedings of ICML* 2005

Fisher, D. H. (1987) “Knowledge Acquisition Via Conceptual Clustering” I: *Machine Learning*, vol. 2, pp. 139-172

Francis, Gill; Susan Hunston; Elizabeth Manning (1996) Collins Cobuild *Grammar Patterns 1: Verbs* Harper Collins Publishers

Hearst, Marti (1992) “Automatic Acquisition of Hyponyms from Large Text Corpora” I: *Procedings of COLING-92* pp. 539-545

Levin, Beth (1993) *English Verb Classes and Alternations* The University of Chicago Press

Madsen, Bodil Nistrup (2001) “Defining Semantic Relations for XXX

Meyer, Ingrid (2001) “Extracting knowledge-rich contexts for terminography” I: *Recent Advances in Computational Terminology* John Benjamins, Amsterdam

Michalski, R. S.; R. Stepp (1983) “Learning from observation: Conceptual clustering” I: *Machine Learning, An Artificial Intelligence Approach*, vol. 2 pp. 331-363

Nuopponen, Anita (1994) *Begreppssystem för terminologisk analys* Acta Wasaensia, WasaUniversity