



Linguistic Challenges in DanNet

Sanni Nimb

Danish Society for Language and Literature



DET DANSKE
SPROG- OG
LITTERATURSELSKAB

Linguistic Challenges in DanNet

- Introduction

1. The hyponymy hierarchy in DanNet

- Reusing data from Den Danske Ordbog (DDO): advantages and problems
- Information added in DanNet

2. Other semantic relations in DanNet

- Reusing data from DDO: advantages and problems
- Information on relations added in DanNet

- Conclusions

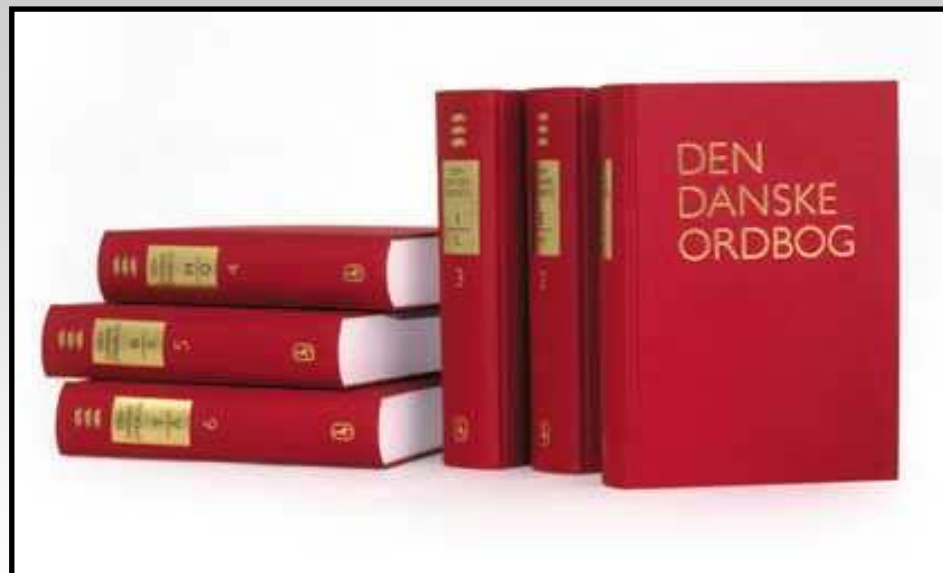
Introduction

- Method: monolingual (**not** translation of synonym sets from Princeton WordNet)
- Reasons:
 - **DDO** (The Danish Dictionary) completed in 2005
100,000 sense definitions
accessible in a **machine-readable** form
corpus-based: frequent senses in Danish general language text material
-> aim of DanNet: resource for computational processing of the same kind of text material
 - **loyal picture** of conceptualization in Danish

1. The hyponymy hierarchy in DanNet

- Reusing data from DDO: advantages and problems

DDO

**ladcykel** sb. fk.

ladcyklen (el. -en), ladcykler, ladcyklerne, [ˈlað-]

tra cykel med lad foran til transport af varer m.m.; □ Felix Ask
svingede fløjtende ladcyklen ind over fortovet KnHolt92.

```

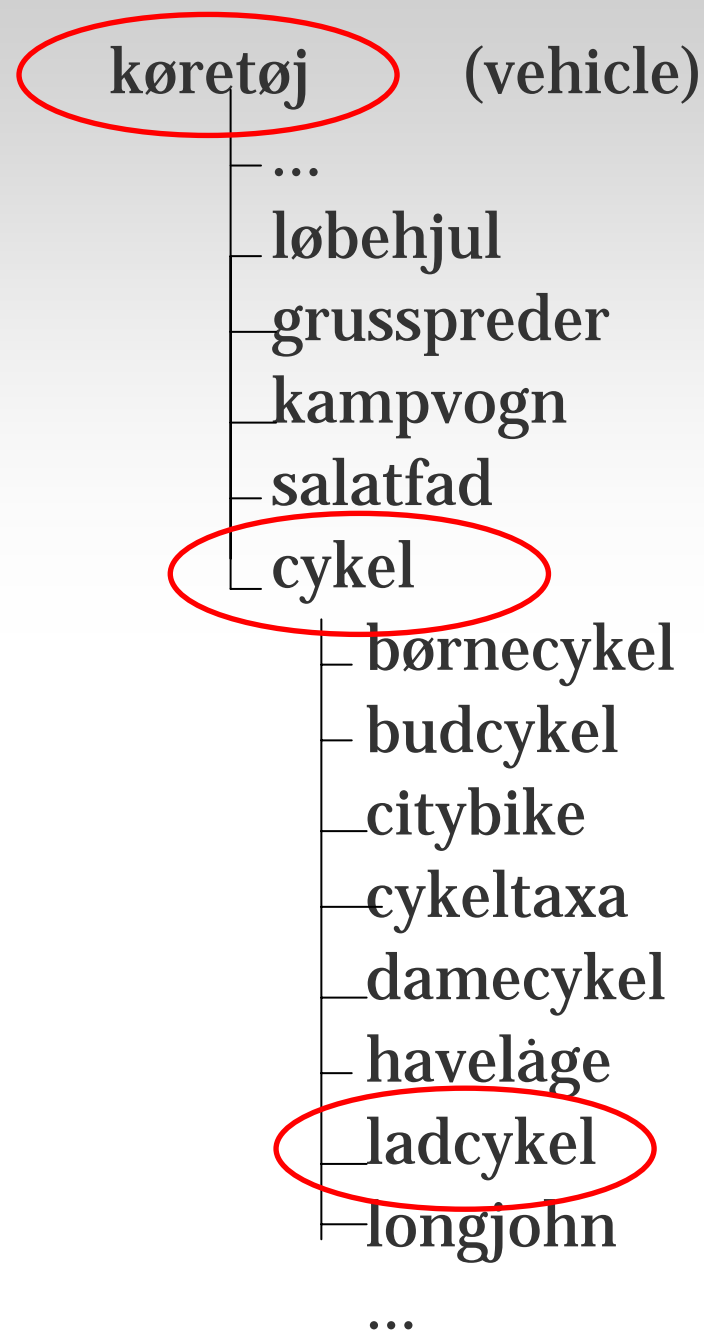
< Semdel >
  < Semem >
    < Restspec >
      < Sysfag > tra
    < Denbet DanNetSemID="21046023"
      DanNetSemType="Semem" > cykel med lad foran til
      transport af varer m.m.
    < Genprox > cykel
    < Dok_DokStatus="a" >
    < Citat >
      < txt > Felix Ask svingede fløjtende ladcyklen
      ind over fortovet
    < Kilde >
      < DDOkilde > KnHolt92
      < Kildeid > LcXI
  
```

Reusing data from DDO

Genus proximum
specified for each sense
in DDO



automatic extraction of a first
version of a DanNet
hyponymy hierarchy



Reusing data from DDO

Problems: Hyponymy hierarchy

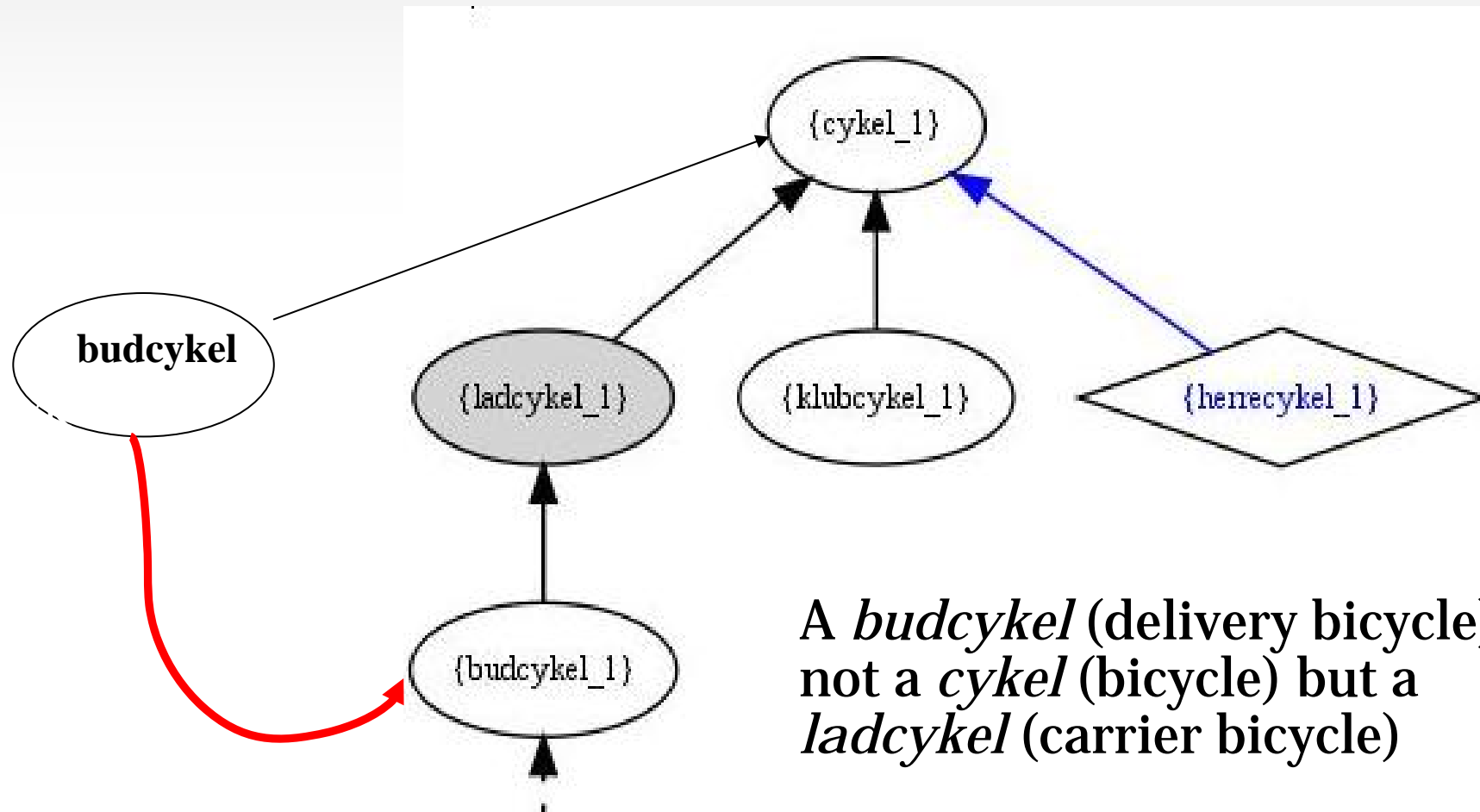
1. Manual disambiguation in cases of polysemous genus expressions in DDO (*celle* (cell) 1: room, 2: organism)
2. Choice of 'top' hypernym (superordinate, 'overbegreb') when definitions in DDO are circular

sted: område el. plads
(spot: area or place)

område: sted
(area: spot)

plads: område
(place: area)

3. Moving of concepts when DDO genus expression refers to a concept at a higher level in the hierarchy than the nearest one from a structural point of view



4. Choice between different genus expressions in DDO in cases of concepts of the same kind

slips: dekorativt, ret smalt **stykke stof** e.l. der bindes rundt om halsen...

(tie: decorative, narrow **piece of fabric/cloth** which is tied around the neck)

Focus on material

krave: **beklædningsgenstand** der anbringes rundt om halsen, især som pynt

(collar: **garment** which is placed around the neck, especially for decoration)

Focus on function



■ **DanNet:** hypernym with most semantics to inherit:

slips and **krave** is a **beklædningsgenstand** (garment)

.... and thereby inherits:

made_by: *sy* (to sew)

made_of: *stof* (fabric)

used_for: *klæde* (to dress)

involved_agent: *person*

- Information on hyponym groups added in DanNet

Hyponyms of cykel (bicycle): one group in DDO

BMX-cykel	(bmx bike)
ladcykel	(carrier bicycle)
bycykel	(community bicycle)
børnecykel	(children's bike)
herrecykel	(men's bike)
havelåge	(slang, 'old bike')
jernhest	(slang, 'old bike')
klubcykel	(standard bicycle)
mountainbike	(mountainbike)
racercykel	(racer bicycle)
cykellig	(bicycle wreck)
turistcykel	(tourist bicycle)
minicykel	(minibike)
damecykel	(women's bike)
budcykel	(delivery bicycle)

...but two groups in DanNet

1. Taxonomical



2. Non-taxonomical

Mutually exclusive

(‘Orthogonal’)

Linguistic test: X is a kind of Y (Cruse 2002)

what is left...

Which of the bicycles are **kinds** of

bicycles?:

BMX-cykel	(bmx bike)
ladcykel	(carrier cycle)
bycykel	(community bicycle)
klubcykel	(standard bicycle)
mountainbike	(mountainbike)
racercykel	(racer bicycle)
turistcykel	(standard bicycle)
minicykel	(mini bike)

børnecykel	(children’s bike)
herrecykel	(men’s bike)
havelåge	(slang, old bike)
jernhest	(slang, old bike)
cykellig	(bicycle wreck)
damecykel	(women’s bike)

⊖ {ret,2_1; madret_1} (Comestible): mad som er tilberedt el. a

⊖ orthogonal hyperonymy

... {anretning_1} (Comestible): ret el. måltid som er pænt

... {dag_11; dagens ret} (Comestible): en (varm) madret

... {egnsret_1} (Comestible): madret der er karakteristisk

... {forret,1_1} (Comestible): mindre madret der serveres

... {frokostret_1} (Comestible): let, kold el. lun madret der

... {færdigret_1} (Comestible): madret der købes færdiglavet

... {herreret_1} (Comestible): lækker og mættende (tradit

... {hors d'oeuvre_1} (Comestible): lille madret der serveres

... {hovedret_1} (Comestible): den største og mest mættende

... {livret_1} (Comestible): madret som man sætter særlig

... {middagsret_1} (Comestible): (varm) ret der indtages s

... {nationalret_1} (Comestible): madret der er meget udbredt

⊖ taxonomic hyperonymy

... {bankekød_1} (Comestible+Artifact+Substance): varm

... {biksemad_1} (Comestible+Artifact+Substance): varm

... {brunkål_1} (Comestible+Artifact+Substance): madret

... {burger_1} (Comestible+Artifact+Object): madret best

... {carpaccio_1} (Comestible+Artifact+Object): madret a

11

97

Synset

⊖ {træ_1} (Plant+Object): høj, mangeårig plante bestående af en tyk,

⊖ orthogonal hyperonymy

8

... {ammetræ_1} (Plant+Object): hurtigtvoksende træ der plantes

... {bonsai_1_1} (Plant+Object): træ el. busk dyrket på denne måde

... {hængetræ_1} (Plant+Object): træ med nedhængende grene

... {klatretræ_1} (Plant+Object): træ der er godt at klatre i, fx for

... {skovtræ_1} (Plant+Object): træ el. træart der oftest vokser i skov

... {stamtræ_2} (Plant+Object): træ (art el. individ) som et andet

... {vejtræ_1} (Plant+Object): træ der står i vejkanten, plantet som

... {vindfælde_1} (Plant+Object): træ der er væltet omkuld af vind

⊖ taxonomic hyperonymy

50

... {akacie_1} (Plant+Object): tæt træ el. busk med små, smalle r

... {ask_1} (Plant+Object): op til 40 m højt løvtræ med grålig bark

... {asp,1_2} (Plant+Object): træ inden for poppelslægten der har

... {balsatræ_1} (Plant+Object): hurtigtvoksende tropisk træ med

... {baobabtræ_1} (Plant+Object): løvfældende træ med meget ty

... {birk,1_1} (Plant+Object): løvfældende træ el. busk med hvid, g

... {bævreasp_1} (Plant+Object): op til 20 m højt, løvfældende træ

... {drageblodstræ_1} (Plant+Object): tropisk træ hvorfra man ud

... {el,2_1} (Plant+Object): løvfældende træ med ægformede blade

Why?

- Useful distinction in formal ontologies
- Taxonomic hyponyms are mutually exclusive:
a bicycle cannot be both a racer bicycle and a mountainbike
- Non-taxonomic hyponyms are not:
a bicycle can be
 - a men's bike
 - a bicycle wreck and
 - a racer bicycleat the same time



Investigations on the data

Example: 6.700 hyponyms of *genstand* (object),

- 160 direct hyponyms
- 50% taxonomic: *møbel* (furniture) *redskab* (tool), *bog* (book), *beholder* (container) etc.
- 50% non-taxonomic. Many words in Danish meaning any kind of object which is copied, invented, new, valuable, strange, small etc.
- Many non-taxonomic synsets with several members:

{*børge*, *dims*, *dimsedut*, *dingenot*, *dippedut*}

{*klods*, *kolos*, *mastodont*, *monstrum*, *skrummel*}

Some non-taxonomical hyponyms of 'genstand' (object):

ejendom (property)

blikfang (eye catcher)

eksemplar (specimen)

kopi (copy)

vare (article, product)

værdigenstand (article of value)

minde (souvenir)

motiv (motif)

mærkværdighed (curiosity)

nyhed (novelty)

offer (sacrifice)

opfindelse (invention)

original (original)

statussymbol (status symbol)

fund (find)

drøm, *herlighed* (dream)

Few words meaning a specific object (a shoe, a bicycle, a shirt) which is old, new, invented, copied, owned, valuable etc.

It seems that

- The more general a concept (e.g. *genstand* (object)), the more non-taxonomical hyponyms
- The more specific a concept (e.g. *cookery book*, *labrador*, *oak tree*, *burger*), the less non-taxonomical sisters

	Taxonomic hyponyms	Non-taxonomic hyponyms
Genstand (object)	80	80
bog (book)	28	14
hund (dog)	30	14
cykel (bicycle)	21	6
dør (door)	14	7
træ (tree)	50	8
madret (dish)	97	11
suppe (soup)	22	0
sko (shoe)	28	5
bukser (trousers)	16	0

2. Other semantic relations in DanNet

- Reusing data from DDO: advantages and problems

Reusing data from DDO

Advantages:

- Same aim of DDO and DanNet:
the native speaker's lexical knowledge about a concept
- Most definitions in DDO are 'true' definitions (Svensén 1993) and describe the semantic relations we want to bring in DanNet:



'ladcykel': bicycle with a platform in front for the transport of goods

Reusing data from DDO

But:

- DDO: semantics described 'bottom up' -
 - no schematic specifications for groups of words
 - no systematic coverage of all semantic aspects
- DanNet: semantics described 'top down'.
The top hypernyms function like specifications
 - systematic coverage needed

Reusing data from DDO

Problem:

- DDO definition: one well-formed, not too grammatically complex phrase aiming to capture typical meaning

bog: trykte el. beskrevne blade af papir indbundet el. på anden måde sammenhæftet i rækkefølge så de danner en helhed, ofte en sammenhængende tekst, beregnet på at blive læst

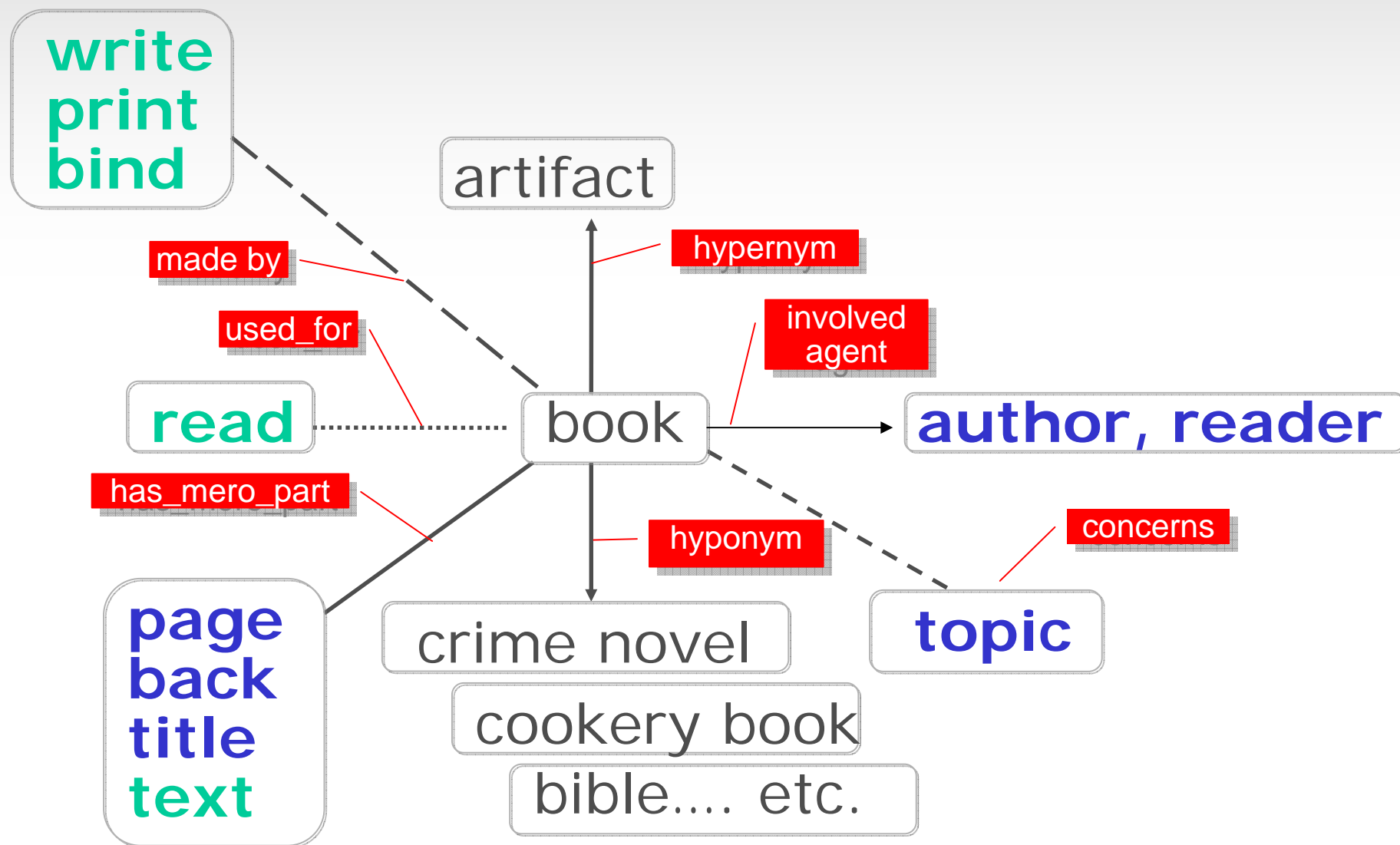
(book: printed or written sheets of paper, bound or in another way stuck together in order so that they form a whole, often a coherent text, meant to be read)

Nothing is said about a book having pages, a back, a title and a topic, an author and a reader

bog (a book)

Relations described in DDO: ●

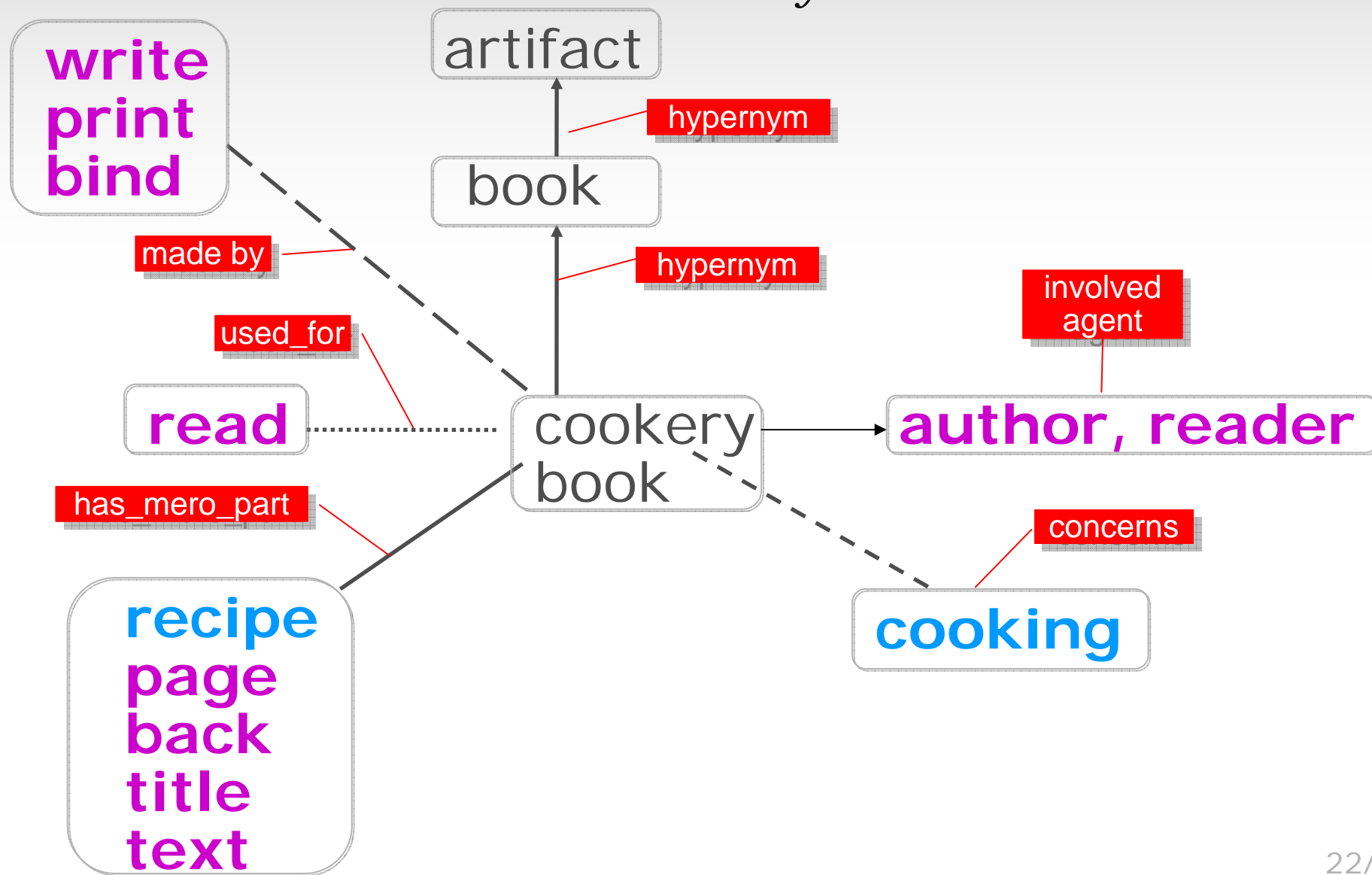
Relations added in DanNet : ●



Inheritance mechanism, *kogebog* (cookery book)

Relations inherited from *book*: ●

Relations added or restricted for *cookery book*: ●



Reusing data from DDO

Problem:

- Human reader →
a dictionary definition leans heavily on the reader's ability to make assumptions without any explicit statements in the text (Svensén 1993)
- in wordnets lexical knowledge must be made explicit by either features or links to other synsets

DDO:

budcykel: cykel med lad beregnet til at bringe varer e.l. ud på. *Jeg afventede min fars ordre til at samle varerne sammen og springe på budcyklen*

(‘budcykel’: carrier bicycle meant for bringing out goods.
I waited for my dad’s order to collect the goods and jump on the carrier cycle)

jeg = bud? (*I* = delivery boy?)

DDO:

Indlæggesseddél: dokument der er udstedt af en læge, og som foreskriver indlæggelse og behandling. *Doktoren gav Marie en indlæggesseddél til et sanatorium*

(hospital ticket: document issued by a doctor prescribing hospitalization. *The doctor gave Marie a hospital ticket for the sanatorium*)

Marie = patient?

butik: (mindre) lokale el. bygning hvor en handlende udstiller og sælger varer.

'se på butikker'

Vi gik ned ad Strøget. Mona standsede ved næsten alle butikkerne. Hun elskede at se på tøj.

(Shop: room or building where a shopkeeper exposes and sells goods.

'do window shopping'

We walked down the pedestrian street 'Strøget'. Mona stopped at almost all the shops. She loved looking at clothes.)

Underspecified: shops have show windows

Most used relations in 6800
 [artifact + object (+ *)] synsets
 (inherited relations not included)

Manually assigned relation	% of the 6800 'artifact + object' synsets
used_for	28 % (<i>bog/læse</i>)
has_mero_part	14 % (<i>bog/side</i>)
concerns	9 % (<i>julepynt/jul</i>)
made_by	6 % (<i>tøj/sy</i>)
involved_agent	6 % (<i>guitar/guitarist</i>)
has_holo_part (is a part of)	5 % (<i>side/bog</i>)
has_holo_location	3 % (<i>gulvtæppe/gulv</i>)
rest of relations	< 3 % each

In general:

- often found in DDO definition:
used_for
concerns
made_by
- rarely found in DDO definition:
involved_agent
all parts of complex artifacts
...but described in DanNet

Added information in DanNet, compared to DDO, examples of artifacts:

- *flyvecertificat* (pilot license): involved agent *pilot*
- *læbestift* (lipstick): involved agent *kvinde* (woman)
- *vielsesattest* (marriage certificate): involved agent *ægtepar* (married couple)
- *bageri* (bakery): involved agent *bager* (baker)
- *apotek* (pharmacy): involved agent *apoteker* (pharmacist)
- *bronkoskop* (bronchoscope): involved agent *læge* (doctor)
- *indlæggelsesseddel* (hospital ticket): involved agent *patient* (patient)
- *letlæsningsbog* (easy reader): involved agent *skoleelev* (pupil)
- *budcykel* (delivery bicycle): involved agent *bud* (delivery boy)
- *salmebog* (hymn book): involved agent *kirkegænger* (church goer)
- *butik* (shop): has part *butiksvindue* (shop window)

What can be studied in the data?

The semantic relations in DanNet make it possible to:

- study lexical relations in general at a large scale (e.g. between artifact and involved_agent)
- study relations between elements in compounds and the semantic relations:
 - bud-cykel* (delivery bike): involved_agent + hypernym
 - fugle-bog*: (bird book): concerns + hypernym
 - skov-træ*: (forest tree): location+hypernym
- group synsets depending on their relations etc.

Conclusions

- DDO indispensable basis for DanNet, but
- Semantics in DDO systematically structured in DanNet and underspecified data made explicit
- New data in DanNet on
 - hyponymy groups
 - users of artifacts
 - parts of 'complex' artifacts
- Linguistic resource for computational processing of Danish text material
- .. but also a resource for investigations on the Danish lexicon at a large scale
- and for onomasiological queries in the online version of DDO