Reusing the Danish WordNet for a New Central Word Register for Danish

DANNET

Bolette S. Pedersen, Sanni Nimb, Nathalie C. H. Sørensen, Sussi Olsen, Ida Flörke, Thomas Troelsgård

UNIVERSITY OF COPENHAGEN

DET DANSKE SPROG- OG
LITTERATURSELSKAB

COR
DET CENTRALE ORDREGISTER

Table of Contents

Introducing COR and DanNet

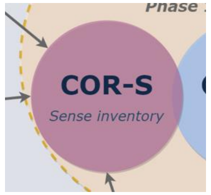
A Reduced Sense Inventory

Hyponymy Revisited

A Simplified Ontological Typing

COR Datastructure and Coverage

Feedback to DanNet



Introducing COR

COR
DET CENTRALE ORDREGISTER

- Companies in Denmark are right now entering the field of **language-centered AI** – and are therefore working intensively with Danish language data from an **NLP perspective**

- In this context, there is an increasing request for **a standardised machine usable lexicon of Danish** with basic morphology and semantics (core senses, sentiment etc.)

- The government has initiated a general effort to support AI in Denmark – COR is part of this initiative funded by the **Agency for Digitisation** under the Ministry of Finance

Which Partners and which Background?

COR
DET CENTRALE ORDREGISTER

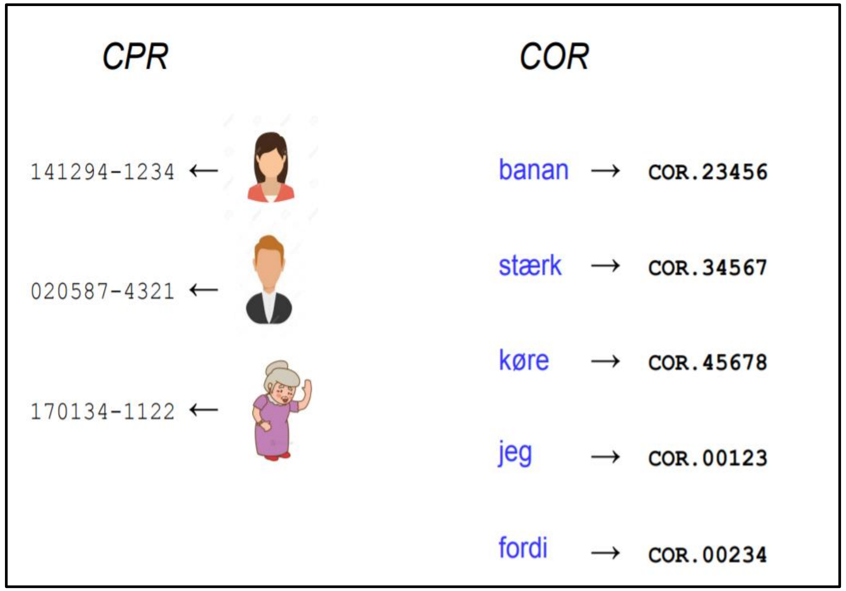
- Danish Language Council
- Society for Danish Language and Literature
- Centre for Language Technology (CST) at the University of Copenhagen

COR is based on **existing Danish dictionaries**

- We take advantage of the very rich and socially contextualised information on word meaning already described in DanNet and The Danish Dictionary (DDO)
- In other words, on **high-quality, locally anchored knowledge about the Danish language and society!**

All Danish Lemmas Linked to a Common Numerical Index

Dansk Sprognævn



CPR

141294-1234 ←

020587-4321 ←

170134-1122 ←

COR

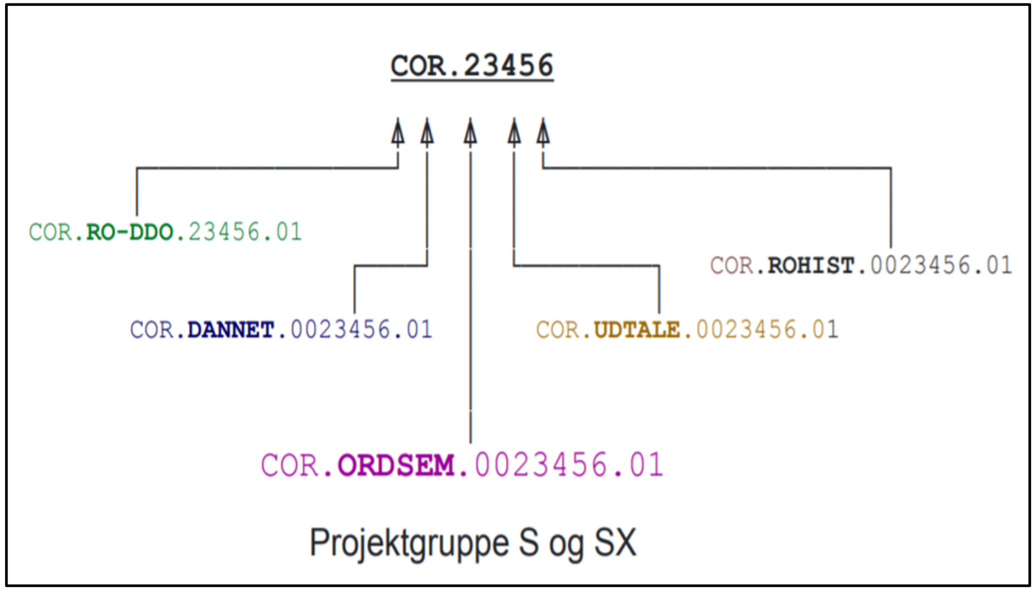
banan → COR.23456

stærk → COR.34567

køre → COR.45678

jeg → COR.00123

fordi → COR.00234



COR.23456

COR.RO-DDO.23456.01

COR.ROHIST.0023456.01

COR.DANNET.0023456.01

COR.UDTALE.0023456.01

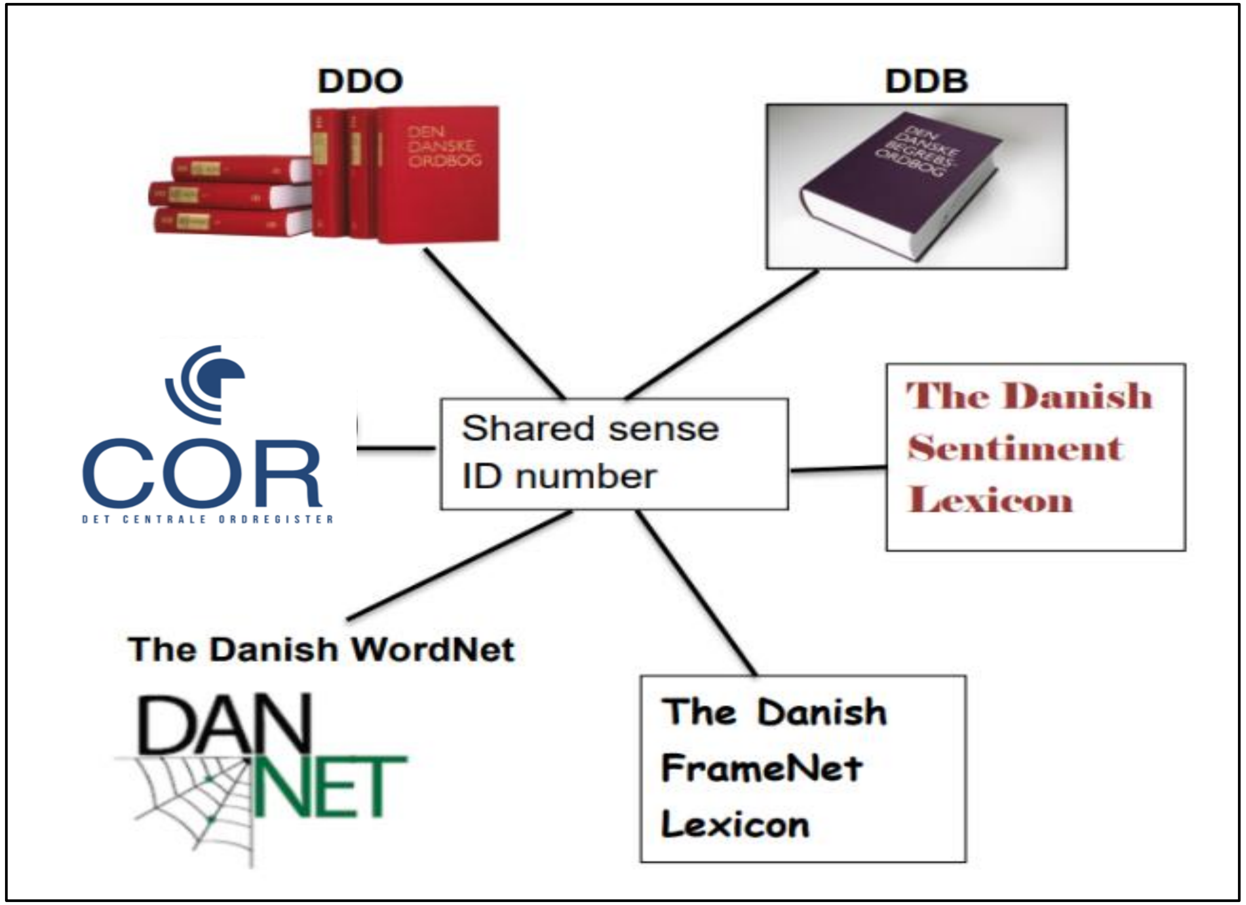
COR.ORDSEM.0023456.01

Projektgruppe S og SX

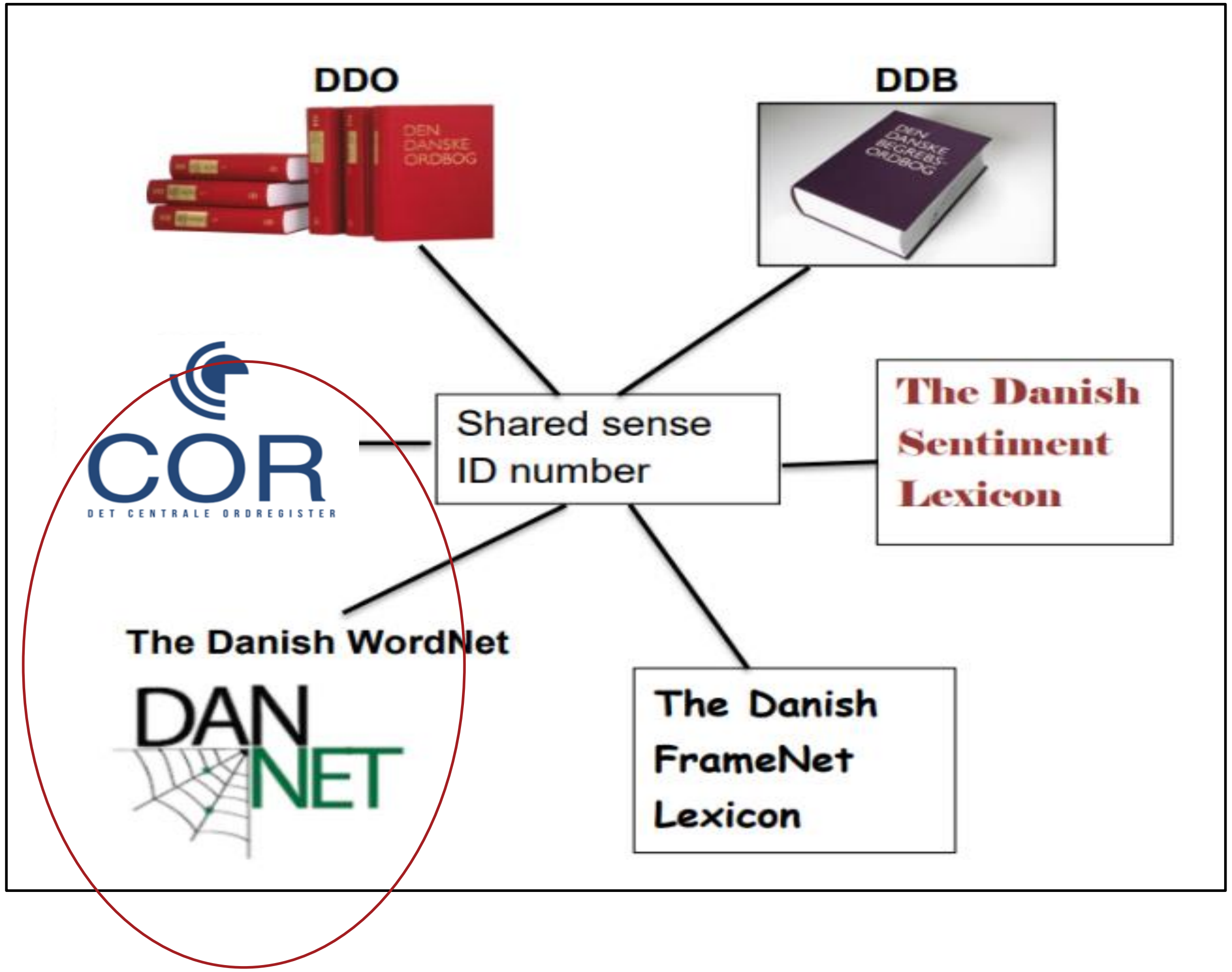
COR

DET CENTRALE ORDREGISTER

The Semantic Component: **COR-S**



The Semantic Component: **COR-S**



DanNet as one of the Background Resources

A Danish WordNet built in 2009 and extended in 2021:

- Semi-automatically compiled based on the DDO
- Approx 70,000 synsets with ontological types and approx. 300,000 semantic relations
- Approx 10,000 synsets linked to Princeton WordNet (merge approach)
- Roughly takes the sense inventory of DDO with some simplifications, however, **only covers approx. half of DDO**
- Main task wrt COR: **completion, curation** and **simplification**

DAN NET

A Reduced Sense Inventory in COR

Background: Sense structure in DDO:

- Semantic relationship between a **main sense** and its **sub-senses;** Sub-senses denote either **a broader**, **a narrower** or **a figurative** nuance of its main sense
- **Main senses are in principle semantically unrelated to each** other although etymologically deriving from the same lemma. However, to avoid deep sense structures in the printed dictionary, senses that in fact could have been classified as sub-senses from the above criteria, **are actually sometimes described as main senses**
- So: idiosyncracies have to be taken into account!

Sense structure in DanNet: took over (a subset of) DDO senses, but did not structure senses, nor reorganize/cluster in a principled way

A Reduced Sense Inventory in COR

(Pedersen et al. LREC 2022)

COR
DET CENTRALE ORDREGISTER

COR principles of ‘coreness’:

Delete a DDO main or sub-sense if it:

- is marked as rare, historic, very domain-specific, colloquial, or slang in DDO (and/or has a very *low sense weight*)

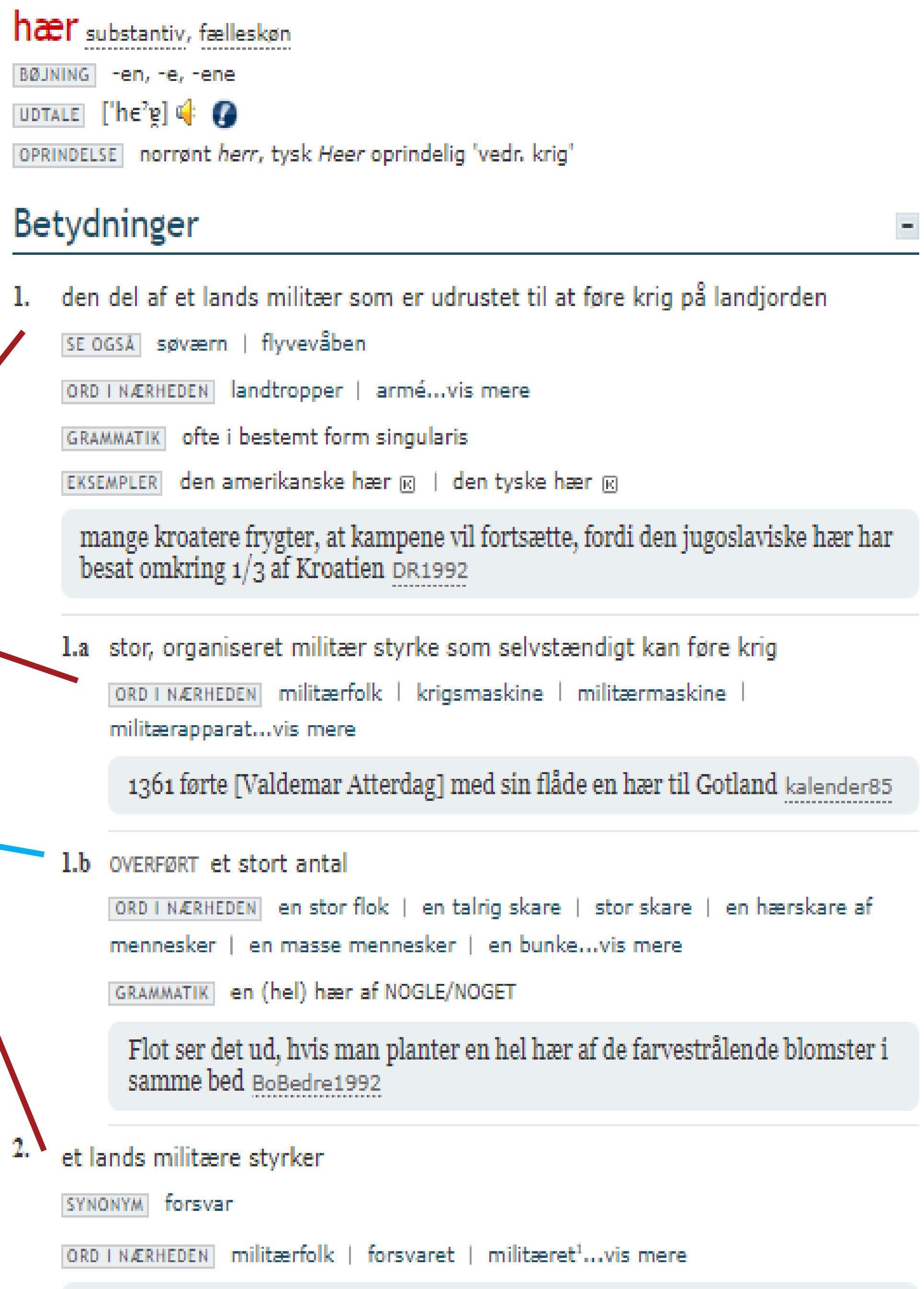
Merge a DDO sub-sense with its main sense, unless a sub-sense is:

- Marked with a different ontological type in the wordnet
- Marked as figurative sense in DDO

In some specific cases: **Merge** semantically close main senses

An Example: *Hær* (army..)

DDO senses:



hær substantiv, fælleskøn

BØJNING -en, -e, -ene

UDTALE [ˈhɛˀɐ̯]

OPRINDELSE norrønt *herr*, tysk *Heer* oprindelig 'vedr. krig'

Betydninger

1. den del af et lands militær som er udrustet til at føre krig på landjorden

SE OGSÅ søværn | flyvevåben

ORD I NÆRHEDEN landtropper | armé...vis mere

GRAMMATIK ofte i bestemt form singularis

EKSEMPLER den amerikanske hær | den tyske hær

mange kroatere frygter, at kampene vil fortsætte, fordi den jugoslaviske hær har besat omkring 1/3 af Kroatien DR1992

1.a stor, organiseret militær styrke som selvstændigt kan føre krig

ORD I NÆRHEDEN militærfolk | krigsmaskine | militærmaskine | militærapparat...vis mere

1361 førte [Valdemar Atterdag] med sin flåde en hær til Gotland kalender85

1.b OVERFØRT et stort antal

ORD I NÆRHEDEN en stor flok | en talrig skare | stor skare | en hærskare af mennesker | en masse mennesker | en bunke...vis mere

GRAMMATIK en (hel) hær af NOGLE/NOGET

Flot ser det ud, hvis man planter en hel hær af de farvestrålende blomster i samme bed BoBedre1992

2. et lands militære styrker

SYNONYM forsvar

ORD I NÆRHEDEN militærfolk | forsvaret | militæret¹...vis mere

COR senses for *hær:*

Sense 1 : Army/military forces (HUMAN_GROUP)

Sense 2 : A big quantity of something (ABSTRACT)

Some Manual and some Automatic Work

(Pedersen et al. LREC 2022)

- The reduction is done manually for the most complex (i.e. most polysemous) part of the vocabulary,
- whereas automatic methods are used for treating the least polysemous part of the vocabulary (2-4 senses per lemma), using however, the hand-coded examples as a gold standard.
- We apply a rule-based method, a word2vec model (Mikolov et al. 2013) and a BERT model (Devlin et al., 2019) for our automatic merges (cf. Pedersen et al. 2022: Section 4).
- All automatic reduction is manually curated

Some Manual and some Automatic Work

(Pedersen et al. LREC 2022)

A **gold standard** for reduction of senses:

- Part I contains 3,500 highly polysemous lemmas (~15,000 senses in DDO)
- Part II: 2,700 average polysemous lemma

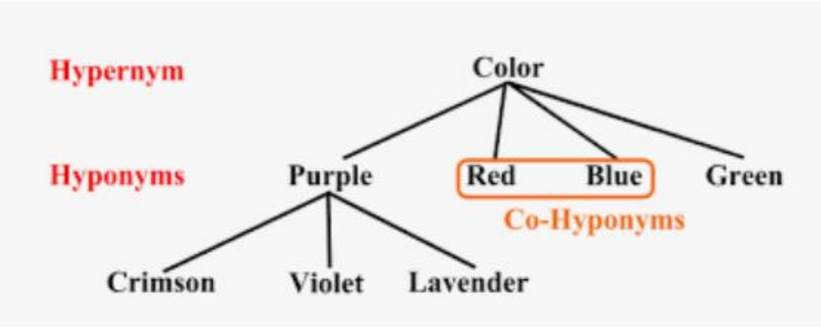
Inter-annotator agreement

- The average Cohen's *k* agreement is **0.82**
- The principles are actually manageable!

43% sense reduction (4.3 senses in DDO to 2.4 senses in COR on average)

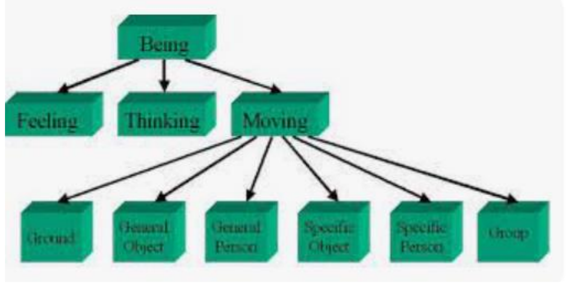
Hyponymy Revisited

- The skeleton of the wordnet in the sense of its hyponymy structure is essentially taken over in COR
- All senses in COR include a link to its most suited hypernym
- Some adjustment: very specialist taxonomies are simplified, reflecting in COR a layman's perspective to i.e., natural entities (e.g., plants and animals)



Hyponymy Revisited

- The hypernyms of **abstract** and **2nd Order Entities** (events, properties, states etc.) in DanNet often relate to synsets that are based on highly polysemous DDO lemmas
- Therefore these have now been adjusted
- In particular, the inventory of **verbal hypernyms** has been reduced to a limited set to ensure consistency among verbs



A Simplified Ontological Typing

Reduced with 36% from DanNet to now 130 types in COR

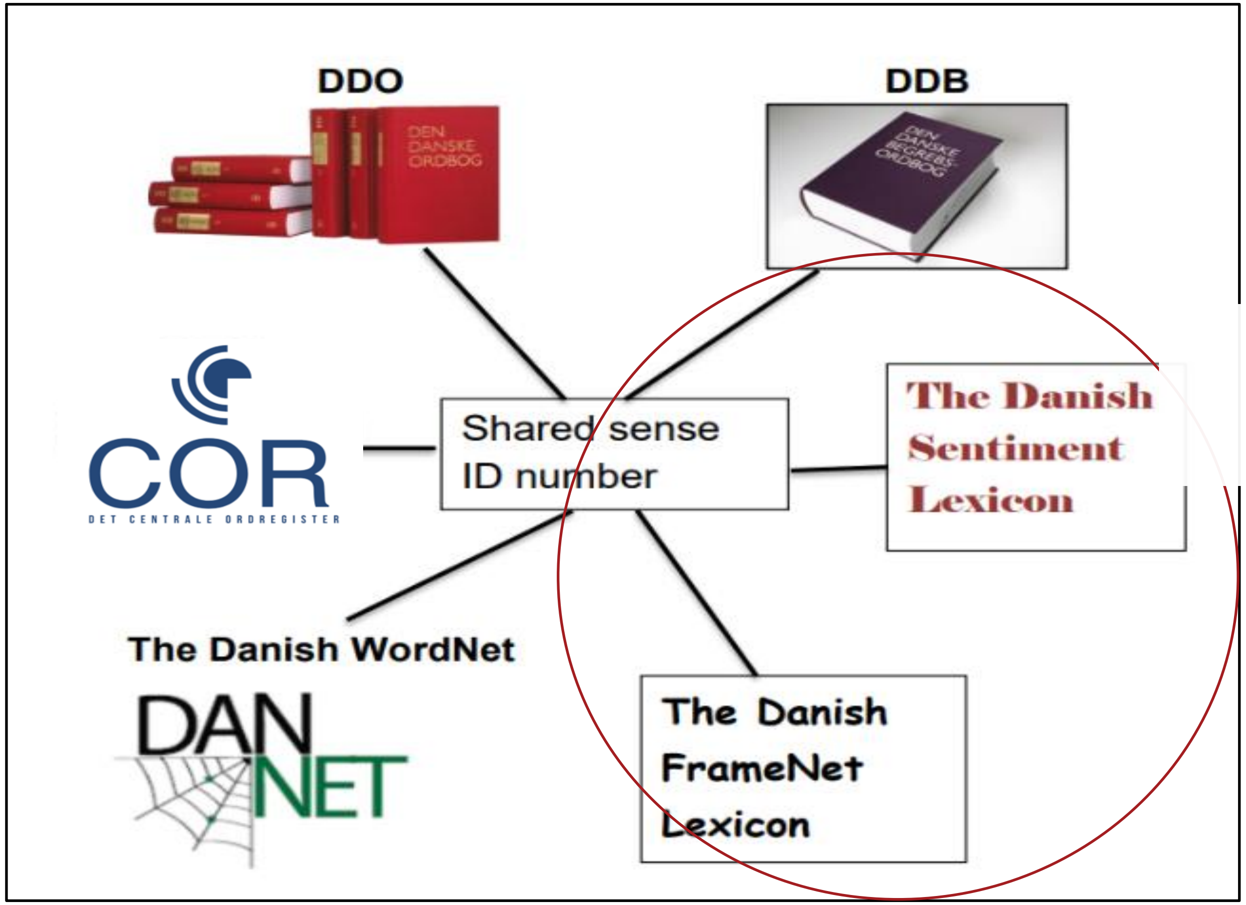
DanNet	COR-S
UnboundedEvent	Event
BoundedEvent	
UnboundedEvent+Agentive	Act
BoundedEvent+Agentive	
Dynamic+Agentive	
3rdOrderEntity+Mental+ Purpose	Abstract+Purpose
3rdOrderEntity+Mental+ Purpose+Manner	
BoundedEvent+Agentive+ Purpose+Possession	Act+Possession
BoundedEvent+Agentive+ Purpose+Possession+Social	

COR-S Datastructure

COR-S unites existing semantic lexical resources for Danish

***Bemærke* (verb)**	**COR sense 1 (PERCEIVE)**	**COR sense 2 (UTTER)**
Definition (short from **DDO/DanNet**)	**Become aware**	**Mention**
Synonyms, related words (**Thesaurus DDB**)	Observe	Talk about
Example (DDO)	*Flere naboer **bemærkede** en banken på et stuevindue .. (the neighbours **noticed** a knock on the window)*	*Ja, fru Nielsen er flink, bemærkede Linda adspredt (*Yes, Mrs Nielson is nice, Linda dispersedly noticed)
Hypernym (DanNet)	{opfatte_1} (**PERCIEVE**)	{ytre_1} (**UTTER**)
Ontological type (DanNet)	ACT+MENTAL	ACT+COMMUNICATION
Frame (Danish FrameNet Lexicon)	BECOMING_AWARE	MENTION
Sentiment (Danish Sentiment Lexicon)	NONE	NONE
Systematic polysemi	NONE	NONE

The Semantic Component: **COR-S**



17,000 DDO senses with polarity -3 to +3

20,000 DDO senses with frame value

COR-S : Coverage

Which ressource?	**DDO lemmas**		**DDO senses**
DanNet + **linked to PWN core concepts**	4.600 lemmas	13.000 **central lemmas**	29.600 DDO-senses (5900 polyseme lemmas with 22.500 DDO senses 7.100 monoseme lemmas)
DanNet + **keyword in The Danish Thesaurus**	11.500 lemmas		
DanNet + **monoseme in the DDO dictionary**	29.000 lemmas		29.000 DDO senses
DanNet + **polyseme in the DDO dictionary** Manually annotated training data Automatic sense reduction	 2.690 lemmas 4.841 lemmas (with 2,3 or 4 senses)		 6.225 DDO senses 11.276 DDO senses
TOTAL	**49.500 lemmas**		**76.150 DDO senses**

Feedback to DanNet?

COR provides curated feedback to DanNet

Completion

- DDO sense in COR, but not in DanNet

  → add synset, automatic transfer of hypernym and ontological type

Curation

- Validation of hypernyms → transferred to DanNet
- Validation of ontological type → transferred to DanNet

DanNet Sense inventory?

Information on rare senses/synsets in DanNet? Merging of senses/synsets?

Conclusion

- COR re-uses existing semantic lexical resources for Danish and unites them via a common index.
- DanNet is crucial to the project: information on ontological type and hyperonym
- Compared to DanNet, COR applies a simplified approach to information and sense structure
- COR: focus on lemma – provides information on all central senses of a lemma
- DanNet: focus on concepts – high number of semantic relations, synonyms, links to English (PWN)
- The two resources are linked

Thank You – and Acknowledgements

- The COR development project is funded by the **Danish Agency for Digitisation** as part of an AI initiative embarked by the Danish Government in 2020
- The research behind the COR project also relies on the **European Lexicographic Infrastructure (ELEXIS)** project under the European Union's Horizon 2020 research and innovation programme (grant agreement No 731015)