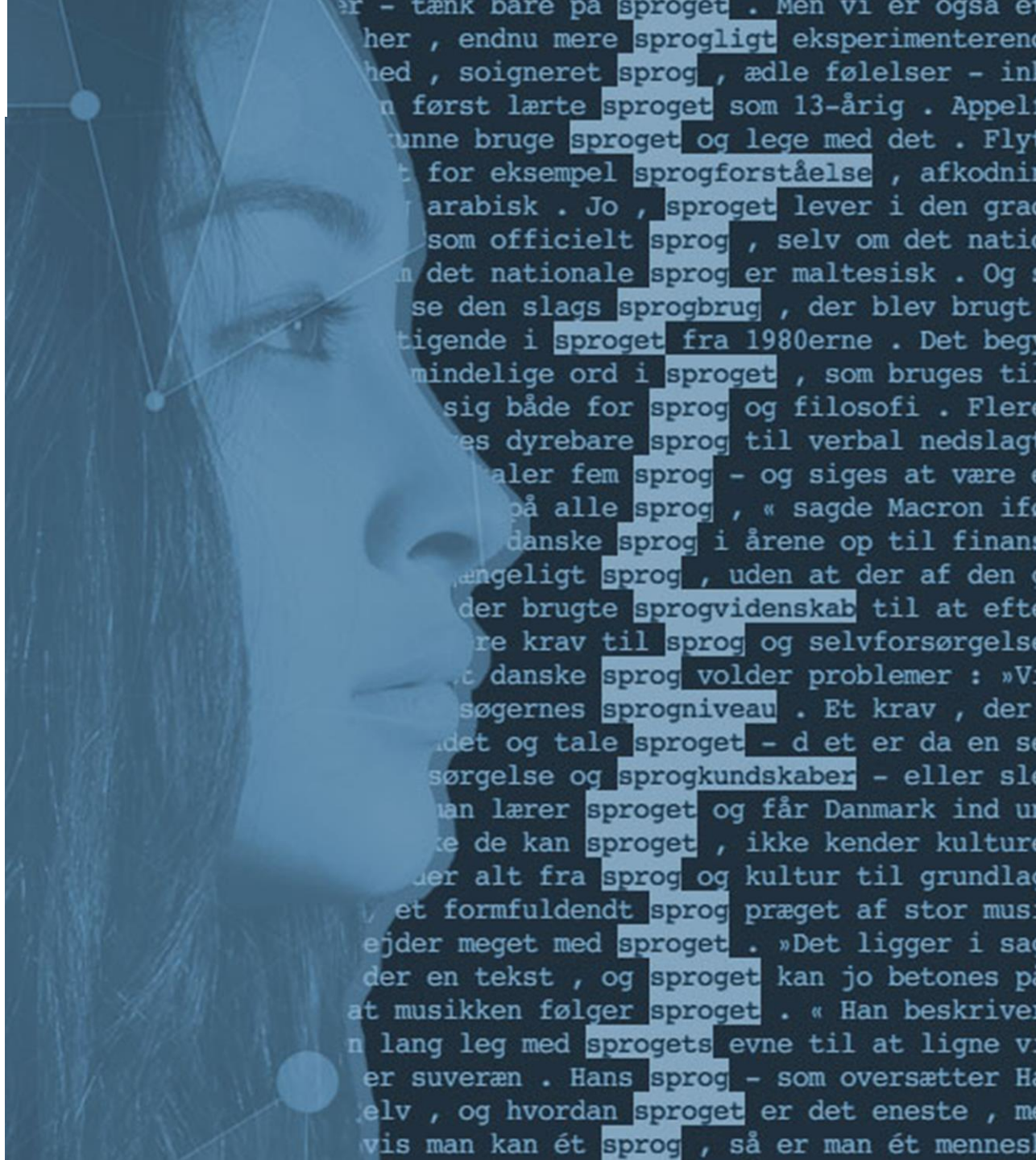
Lexical Conceptual Resources in the Era of Neural Language Models

Bolette Sandford Pedersen
Centre for Language Technology

Autumn Linguistics Seminars Sofia 2022

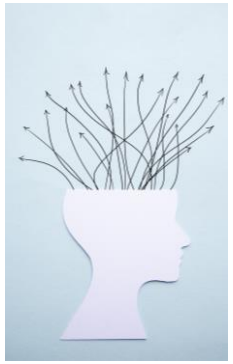
UNIVERSITY OF COPENHAGEN



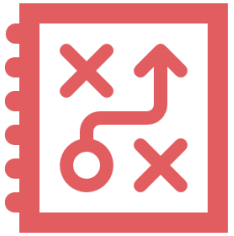
Contents of my talk

Limitations of current language models

Why lexical conceptual resources are still relevant



Lexical-conceptual resources as part of an infrastructure



Example from my own work on Danish lexical conceptual resources

COR
DET CENTRALE ORDREGISTER

Limitations of current language models

Neural language models have in many ways disrupted the field of NLP

- We now apply large pretrained language models and fine-tune them to specific tasks
- Pretrained models generally outperform previous models on these tasks
- However, language models **are not performing natural language understanding**!
- Synthetic text produced via language models is **not meaningful**!

Limitations of current language models

Models based on text only are biased

- towards these specific texts
- towards the 'general nature' of texts

Texts do not necessary contain the background knowledge which is necessary for **the interpretation** of texts

Limitations of current language models

A popular, simple, and illustrative example is the so-called *black sheep* problem

What is the colour of a sheep?

- The language model: *black*
- The lexical conceptual resource: *white/grey*

Limitations of current language models

In texts we typically describe/highlight:

- the extraordinary
- the spectacular
- The controversial
- the 'forefront'
- The 'interesting'

This is particularly true for **web-scraped** text corpora and newswire

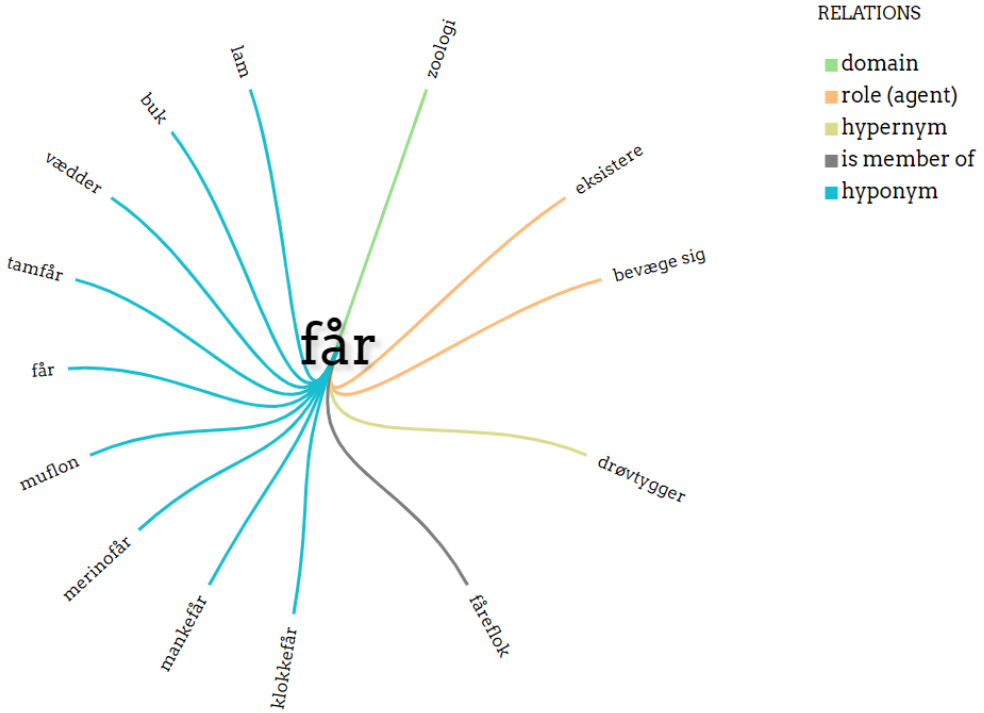
This approach leads to **damaging bias** (cf. Bender et al. 2021) if we are not very careful in our curation

Limitations of current language models

Definition in DanNet/The Danish Dictionary:

mellemstor drøvtygger med meget kraftig, oftest hvidlig uldpels

(a medium-sized ruminant with very thick, mostly whitish woolly fur)



Limitations of current language models

Other examples from the Danish resources:

What is the form of a roof?

Language model: *flat*

Lexical-conceptual resource: *A-formed*

What is the sex of a doctor or a priest?

Language model: *female*

Lexical conceptual resource: *a person who..*

Limitations of current language models

The lexical conceptual resources describe exactly

- the background
- the prototypical

As well as:

- the syntagmatic/relational information
- sentiment information at lexicon level

How to inform LMs with semantics and world knowledge?

Several studies investigate semantics in BERT, e.g.

Rogers et al. (2020). **A Primer in BERTology: What We Know About How BERT Works**. Transactions of the Association for Computational Linguistics, 8:842–866.

Reviews current state of the art and experiments with so-called probing tasks where BERT is tested on semantic tasks and tasks requiring world knowledge

How to inform LMs with semantics and world knowledge?

Peters et al. (2019). **Knowledge Enhanced Contextual Word Representations**. EMNLP-IJCNLP, Hong Kong, China.

Presents 'KnowBERT' which extends BERT with structural knowledge from wordnets and wikipedia via a 'Knowledge Attention and Recontextualisation Component'

Bevilacqua & Navigli (2020). **Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information.** ACL Proceedings

Why lexical conceptual resources are still relevant

- Lexical conceptual resources in terms of e.g. **wordnets**, **framenets**, and **ontologies** have been compiled for many languages during the last decades

Aim: to provide NLP systems with formally expressed information about the **semantics** of **words and phrases,** i.e.

- how they refer **internally to the text, externally to the world** and
- the **connotation** they evoke

Lexical-conceptual resources

Types of information captured in such resources:

- **Paradigmatic** info (wordnets, ontologies)
- **Syntagmatic** info (framenets, propbanks, verbnets)
- **Connotative** info (sentiment lexicons)

Paradigmatic descriptions

Word meaning is described via the words that can substitute it in a specific context corresponding to:

- Synonymy (*beautiful, wonderful*)
- Hyponymy (*horse, animal*)
- Meronymy (*hand, finger*)
- Antonymy (*cold, varm*)

Additional relations:

- Used_for
- Has_colour (cf. the SIMPLE lexicons)
- Located_in
- ..

Paradigmatic descriptions

Ontologies provide some of the same kinds of information as wordnets:

- the **is_a** relation which compares roughly to **hyponymy**

However

- Ontologies typically claim to be **language independent**
- Ontologies apply **classes, properties, attributes**, and **instances**
- **Relations** are established among **instances** (and sometimes among classes)

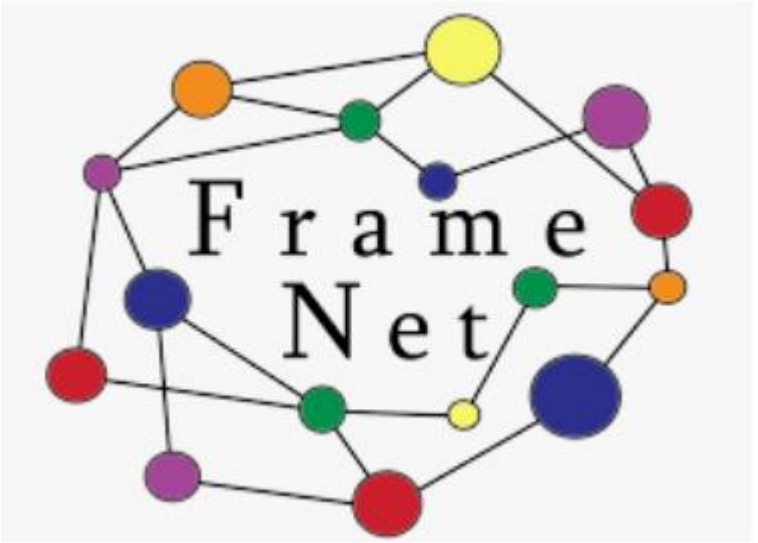
Instances

Instances are extensional references to e.g. persons, places, events etc.

These correspond to **proper names** in linguistics and are typically not included in dictionaries and wordnets but handled via gazatteers and NER applications

The ontology can be said to capture **encyclopedic** knowledge, i.e. information about the world

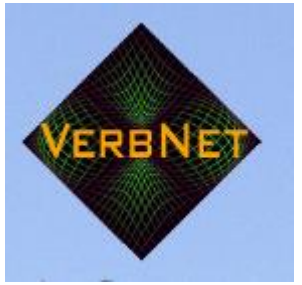
Syntagmatic descriptions



Who does what to whom, when and where?

In which way do the elements of a phrase relate semantically to each other? Which semantic roles?

These are the information types typically described in **framenets, propbanks** and **verbnets**



Communicator Addressee Reason

Den ungarske landstræner havde **talt med store bogstaver** til sine spillere i pausen

Jeg **skælder** hende **ud** for at være groft uansvarlig

I debatten **tordnes** der løs mod Det kgl. Teaters repertoire

Denotative and connotative descriptions

Previously mentioned resources aim to describe the **denotative** meaning, but many words have a **connotative** 'co-meaning', namely whether it is conceived as:

- **positive** or
- **negative**

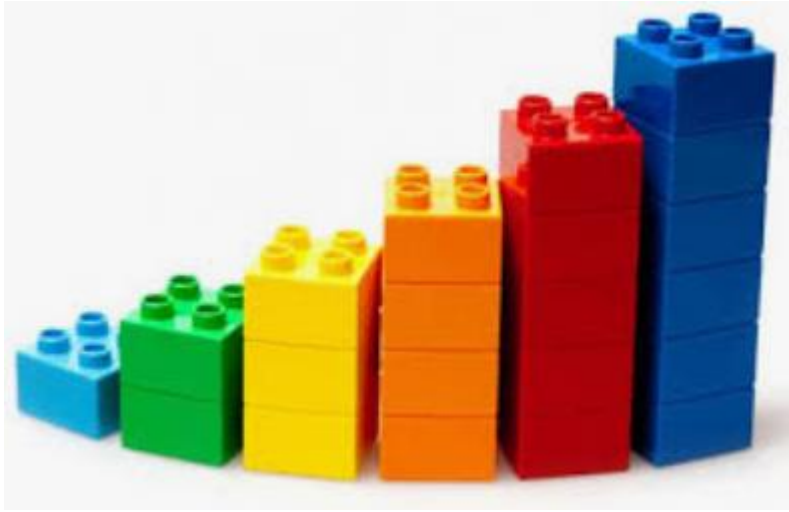
This information is typically described in **sentiment lexicons**

Lexical-conceptual resources as part of an infrastructure

Problems: NLP lexical resources are often fragmented and detached from other national lexical initiatives

They often struggle with:

- Scalability
- Coverage
- Maintenance (new words and meanings)
- Bias

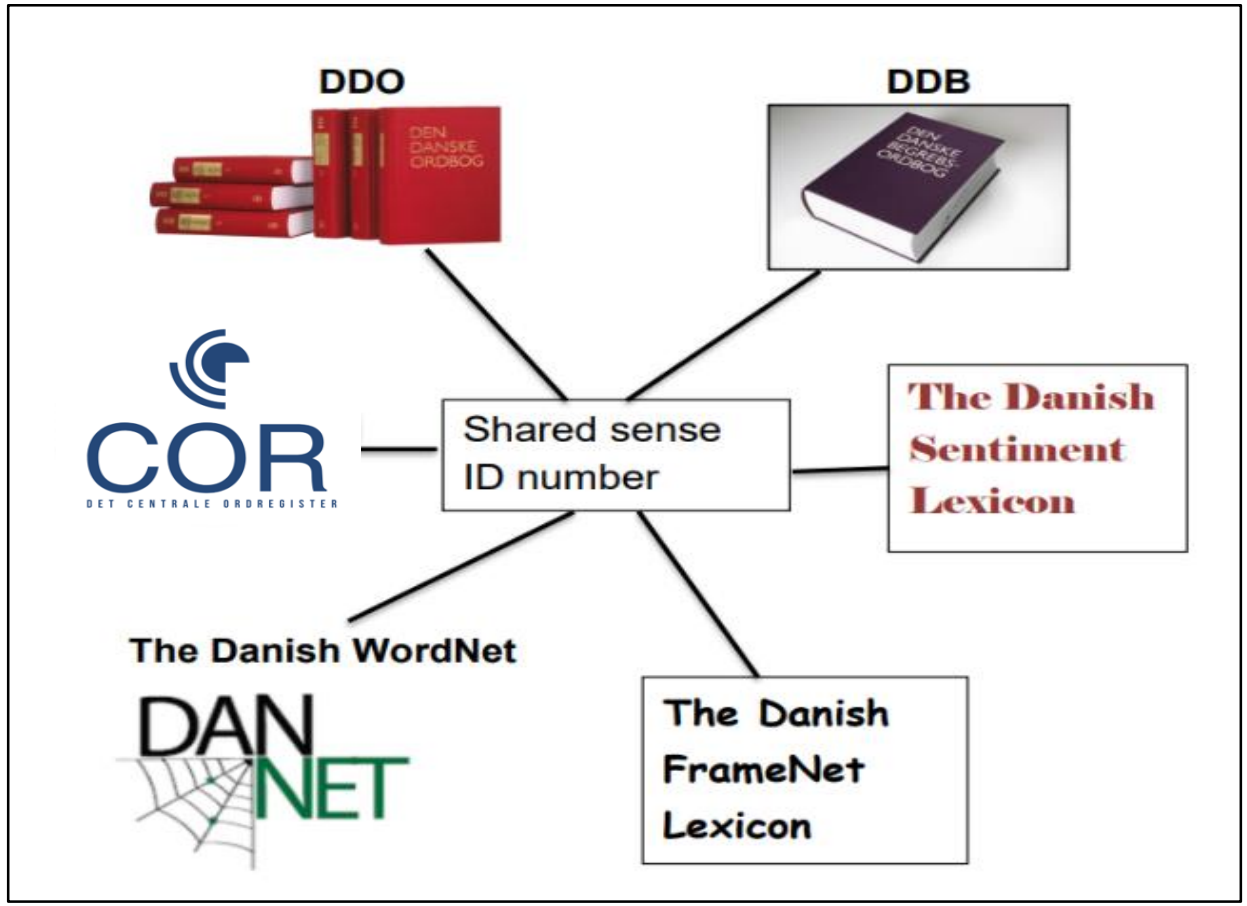


Lexical-conceptual resources as part of an infrastructure

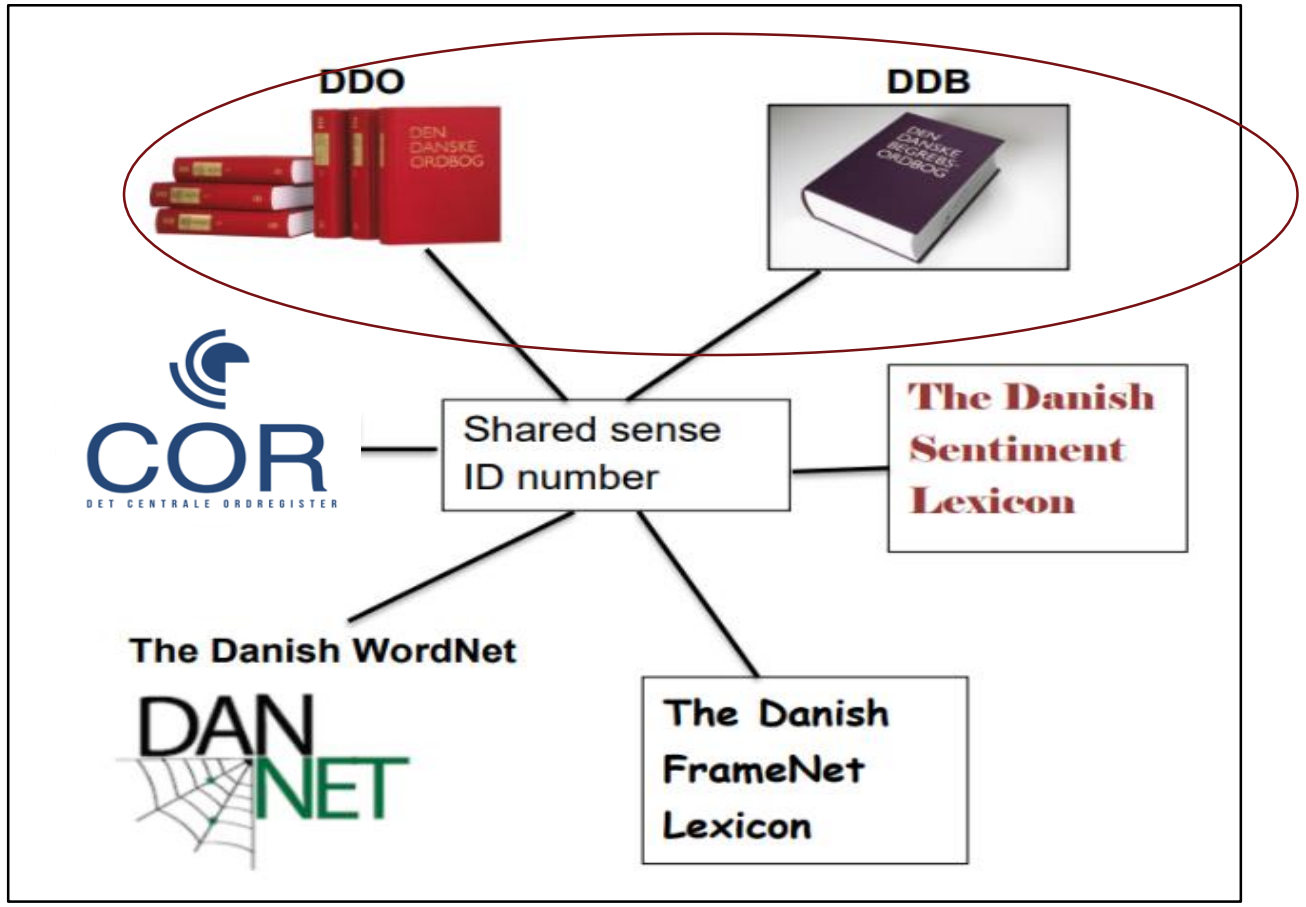
The **ELEXIS project** addressed exactly these problems by arguing for **better collaboration between NLP milieus and lexicographers** through improved infrastructure, tools and services

elexis european lexicographic infrastructure

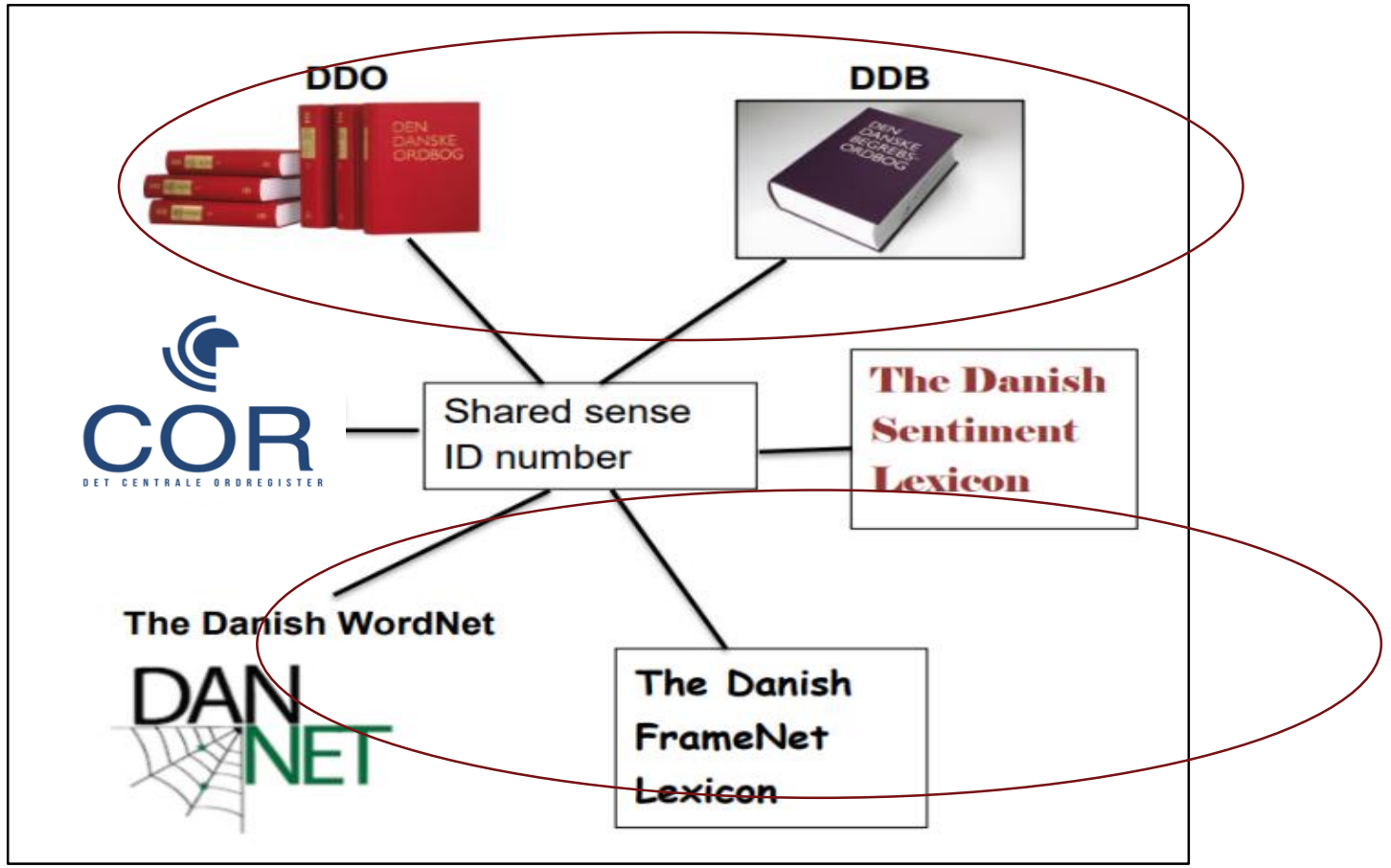
Examples from my own work on Danish conceptual resources



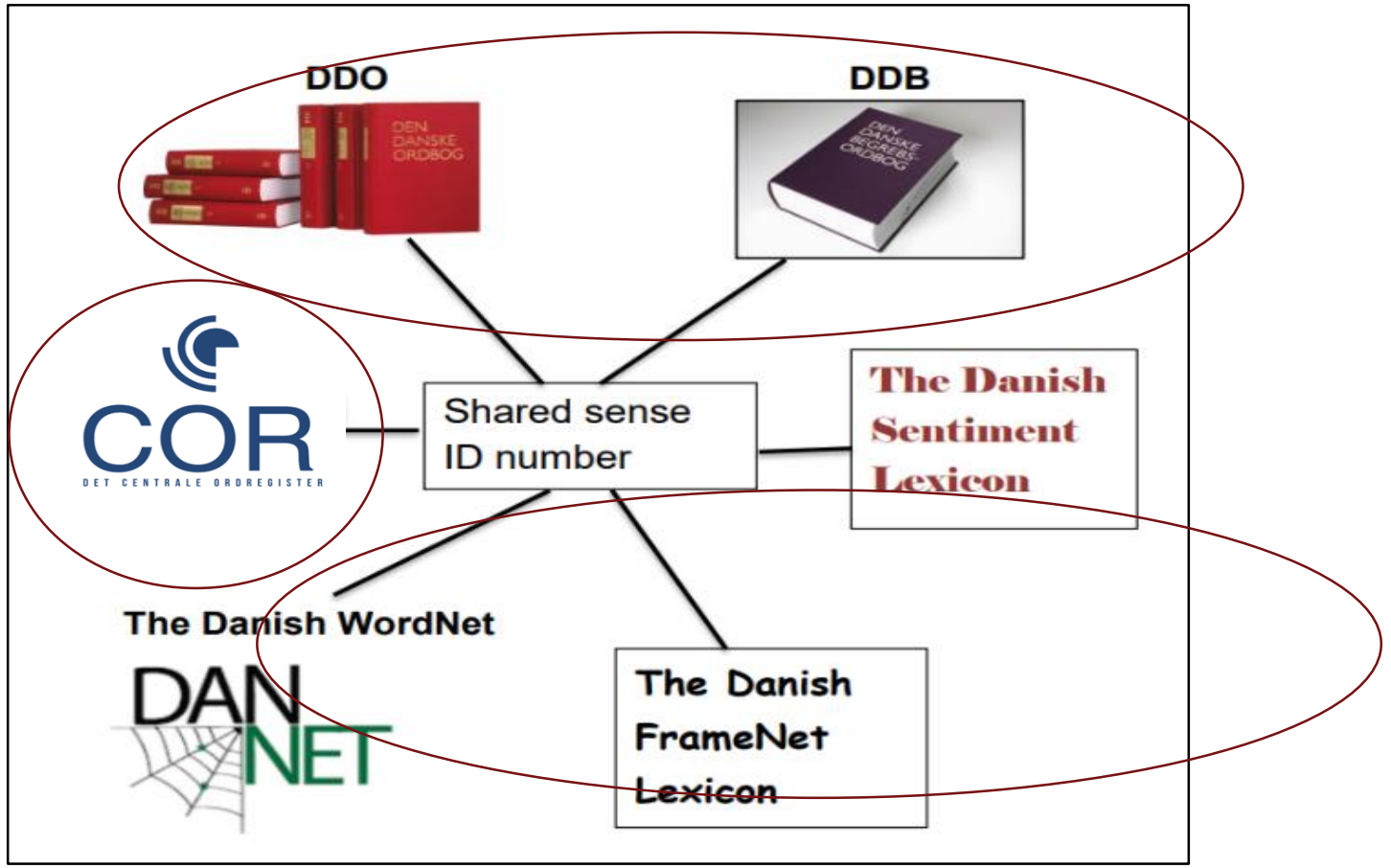
Examples from my own work on Danish conceptual resources



Examples from my own work on Danish conceptual resources



Examples from my own work on Danish conceptual resources



A few words about COR (Central Word Register for Danish)

Joint Danish project with

- The Danish Language Council
- The Danish Society for Language and Literature
- The Centre for Language Technology, UCPH

COR
DET CENTRALE ORDREGISTER

Financed by the Danish Agency for Digitalisation 2021-2023

Aim: To provide the Danish digital society with a large, high-quality lexical resource for NLP and AI

A few words about COR (Central Word Register for Danish)

- 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.)

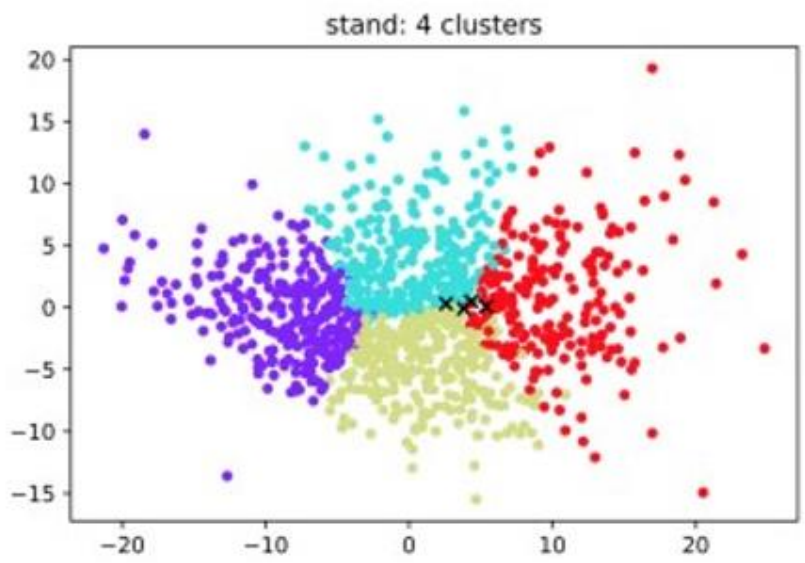
- The government has initiated a general effort to support AI in Denmark – COR is part of this initiative

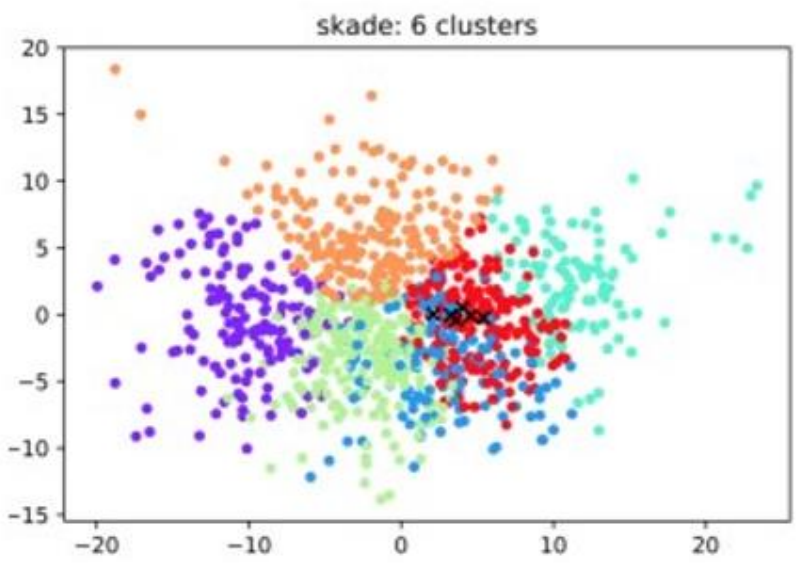
COR
DET CENTRALE ORDREGISTER

Language models inform our COR sense inventory

Questions asked:

- How well do LMs correlate with human judgments of senses?
- Which types of senses are distributionally confirmed in our language models?





(Olsen et al. 2020, Pedersen et al. 2022)

COR
DET CENTRALE ORDREGISTER

A gold standard for training on sense reduction

The gold standard consists of two parts:

- 3,500 highly polysemous lemmas
- 2,700 average polysemous lemma

Inter-annotator agreement: 0.82 (Cohen's *k*)

43% sense reduction from the 'classical' dictionary (DDO)

ML results are promising for the average polysemous part of the vocabulary (word2vec, BERT models) (Pedersen et al. 2022)

Summing up

Lexical conceptual resources:

- are important resources for defining the background knowledge implicit for language understanding
- describe the meaning of the vocabulary and how it refers to the world (paragdigmatic, syntagmatic, connotation)
- can inform LMs with semantic knowledge and vice versa: LMs can help curate them (e.g. sense inventory)
- should be part of a language's general lexical infrastructure in order to be **scalable**, **maintainable** and of **high quality**

Thank you! Selected references

Bender, Emily, Timnit Gebru, Angelina McMillan-Major, Schmargaret Schmitschell (2021) **On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?** FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610–623

Bevilacqua & Navigli, R. (2020). **Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information.** ACL Proceedings.

Krek, Simon, Iztok Kosem, John McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, & Tanja Wissik (2018). **European Lexicographic Infrastructure (ELEXIS)**. 881-892. Paper presented at the XVIII EURALEX International Congress, Ljubljana, Slovenia.

Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S., & Søgaard, A. (2017). **From Thesaurus to Framenet**. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic Lexicography in the 21st Century : Proceedings of eLex 2017 conference* (pp. 1-22). Lexical Computing CZ.

Nimb, S., Olsen, S., Pedersen, B. S., & Troelsgaard, T. (2022). **A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. I**n *Proceedings of the Language Resources and Evaluation Conference: LREC2022* (Vol. 2022, pp. 2826--2832). European Language Resources Association.

Olsen, I. R., Sayeed, A., & Pedersen, B. S. (2020). **Building Sense Representations in Danish by Combining Word Embeddings with Lexical Resources.** In *Globalex Workshop on Linked Lexicography: LREC 2020 Workshop Language Resources and Evaluation Conference* (pp. 45-52). European Language Resources Association.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., & Lorentzen, H. (2009). **DanNet - the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary**. *Language Resources and Evaluation*, *43*, 269-299.

Pedersen, B. S., Sørensen, N. C. H., Nimb, S., Flørke, I., Olsen, S., & Troelsgård, T. (2022). **Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR-Lexicon**. In *Proceedings of the 13th Language Resources and Evaluation Conference, Marseille, France* (pp. 51-60). European Language Resources Association

Peters, M. , M. Neumann, R. L. Logan, R. Schwartz, V. Joshi , S. Singh , and N. A. Smith (2019). **Knowledge Enhanced Contextual Word Representations**. EMNLP-IJCNLP, Hong Kong, China.

Rogers, A., Kolavera, O. & Rumsishki, A. (2020). **A Primer in BERTology: What We Know About How BERT Works**. Transactions of the Association for Computational Linguistics, 8:842–866.