Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes:

The Open Source COR Lexicon

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

UNIVERSITY OF COPENHAGEN

DET DANSKE SPROG- OG LITTERATURSELSKAB

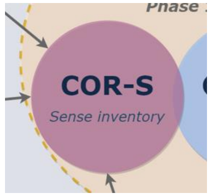
COR

DET CENTRALE ORDREGISTER

Table of Contents

Introducing COR

The semantic component: COR-S



Principles for defining core senses

Automatic sense clustering

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?

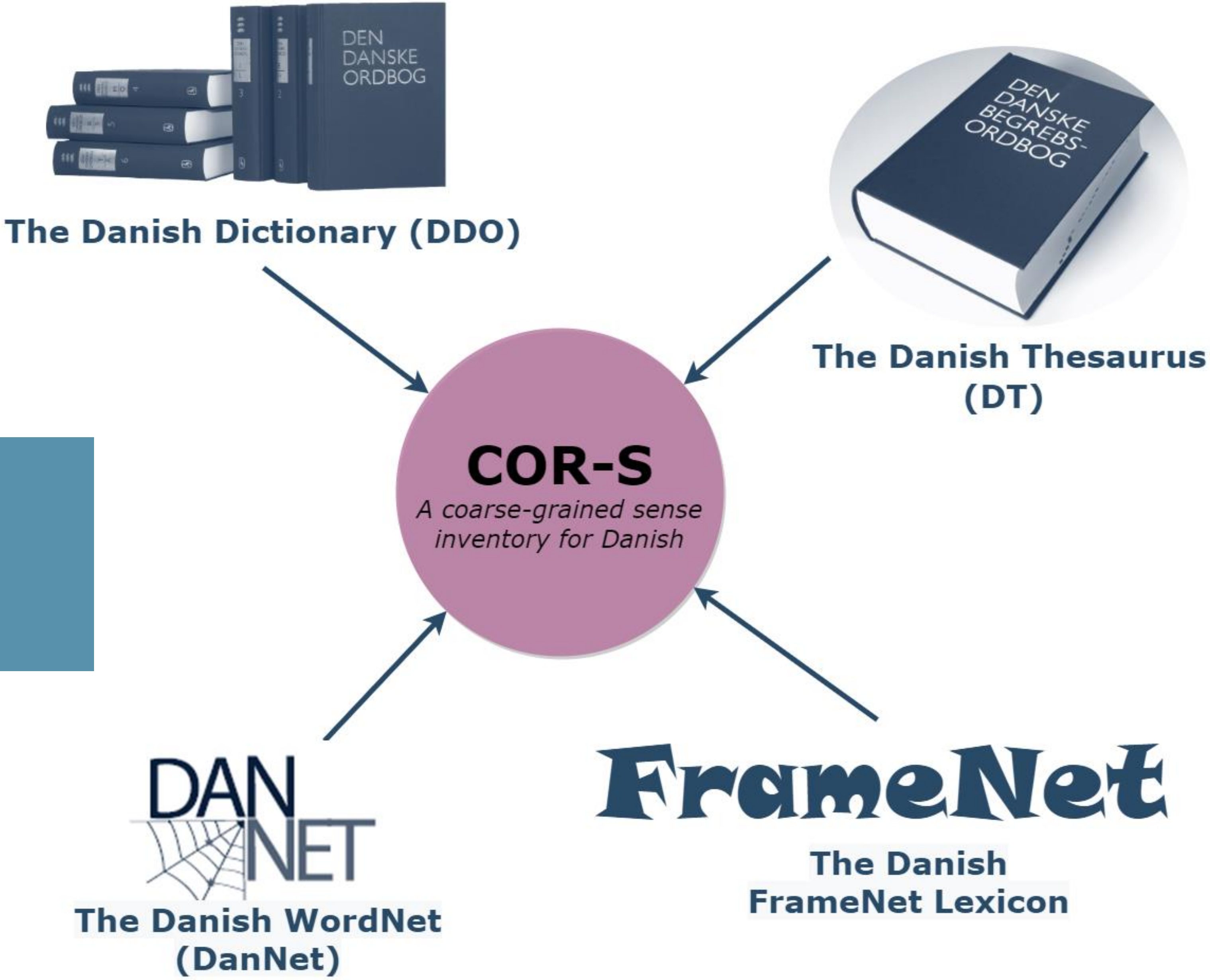
COR
DET CENTRALE ORDREGISTER

- 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 traditional lexica
- In other words on **high-quality, locally anchored knowledge about the Danish language and society!**

The Semantic Component: **COR-S**



The Danish Dictionary (DDO)

The Danish Thesaurus (DT)

COR-S

A coarse-grained sense inventory for Danish

The Danish WordNet (DanNet)

The Danish FrameNet Lexicon

How to achieve a suitable level of sense granularity?

COR
DET CENTRALE ORDREGISTER

Our aim:

- To establish generalized principles of **lexical semantic coreness**
- To reduce the DDO sense inventory accordingly
- Thereby achieve a **core sense inventory** which is potentially relevant for modern texts and **distinguishable on distributional grounds** and thus more suitable for NLP, still, however, capturing the central/relevant senses

Our method:

- To develop a hand-coded and extensive **gold standard** for highly polysemous and more average parts of the vocabulary
- To apply aumomatic methods for the rest of the vocabulary based on this standard

Related work on sense granularity (for factual references see the paper)

COR
DET CENTRALE ORDREGISTER

In **lexicography** and **lexical semantics**, the discussion of **sense granularity** has been ongoing for decades

A typical, slightly simplified, categorisation of lexicographers into being either **lumpers** or **splitters**

Where very rich sense descriptions seem to correspond well to the needs of human users, very subtle sense descriptions tend to cause **notorious problems for NLP and WSD**

In fact, this has been the case to an extent where traditional dictionaries have been **deemed somewhat useless** in relation to NLP

The **ELEXIS and COR projects** are trying to remedy this problem

Principles for sense structure in DDO

COR
DET CENTRALE ORDREGISTER

- A close semantic relationship between a main sense and its sub-senses
- While 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, in order 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 found to be described as main senses**
- In other words: idiosyncracies have to be taken into account

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

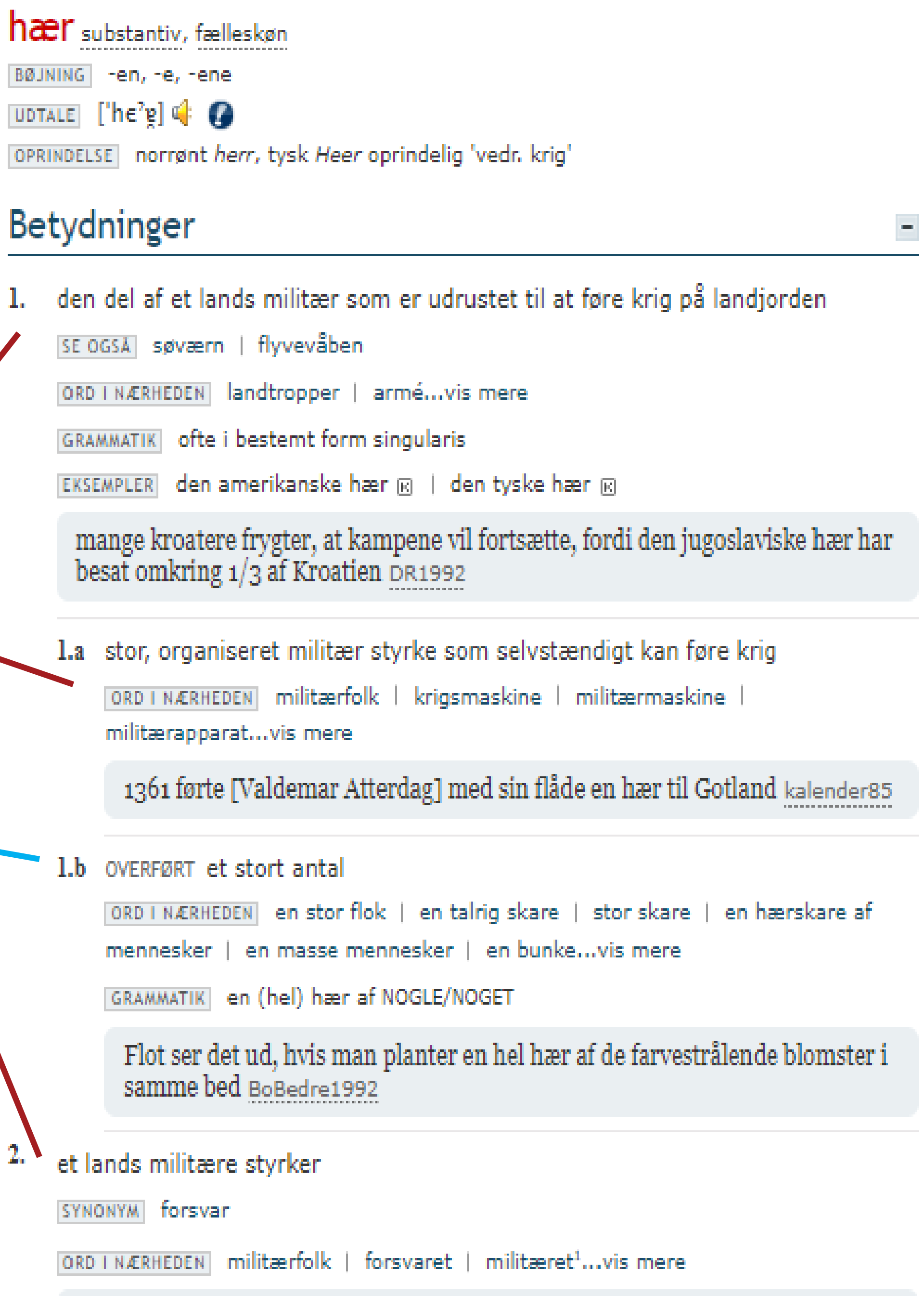
An example: *Hær* (army..)

COR senses for *hær:*

Sense 1 : Army/military forces (HUMAN_GROUP)

Sense 2 : A big quantity of something (ABSTRACT)

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

The gold standard

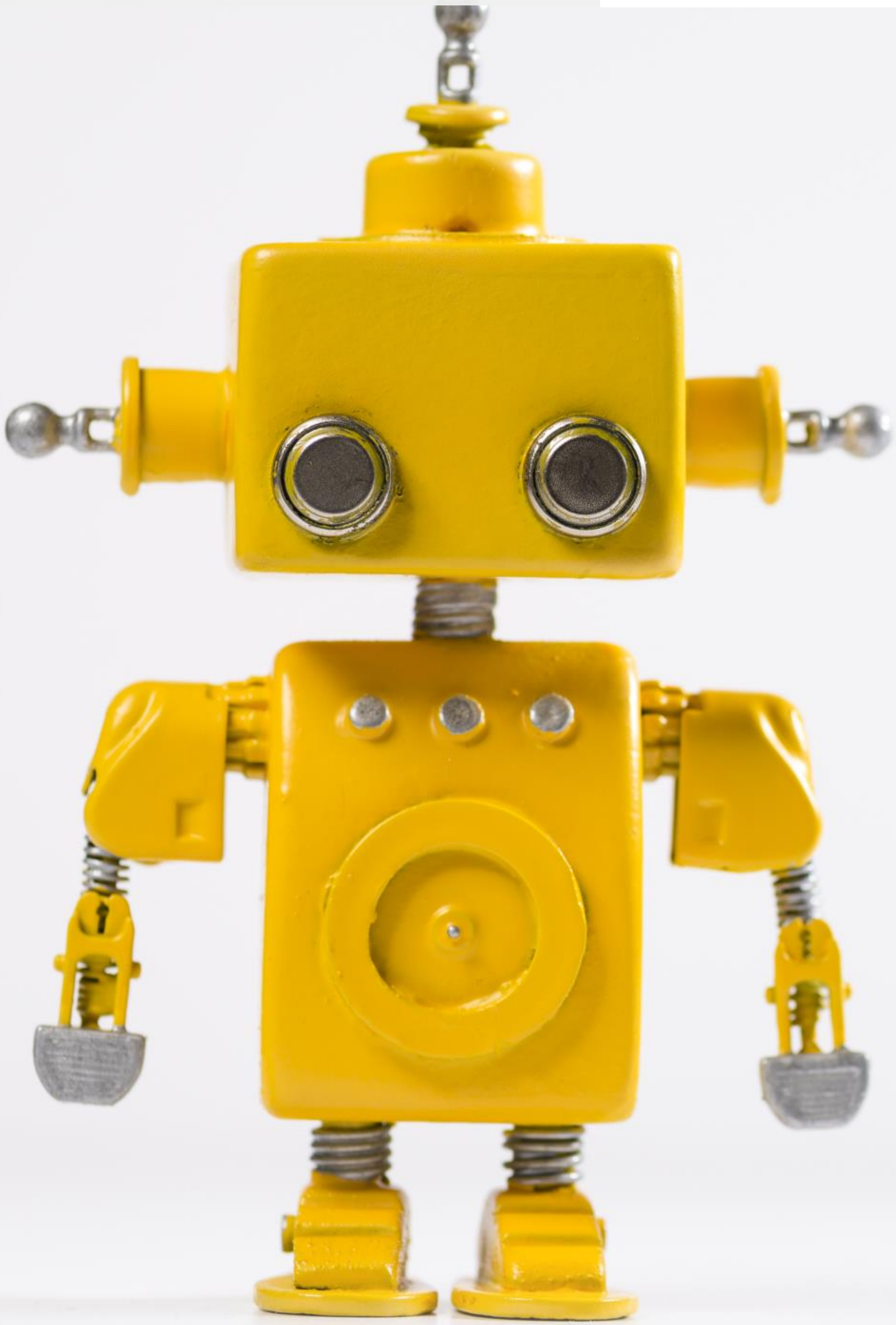
The gold standard consists of two parts:

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

Inter-annotator agreement

- We use Cohen's *k*
- The average agreement of **0.82**
- The principles are actually manageable

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



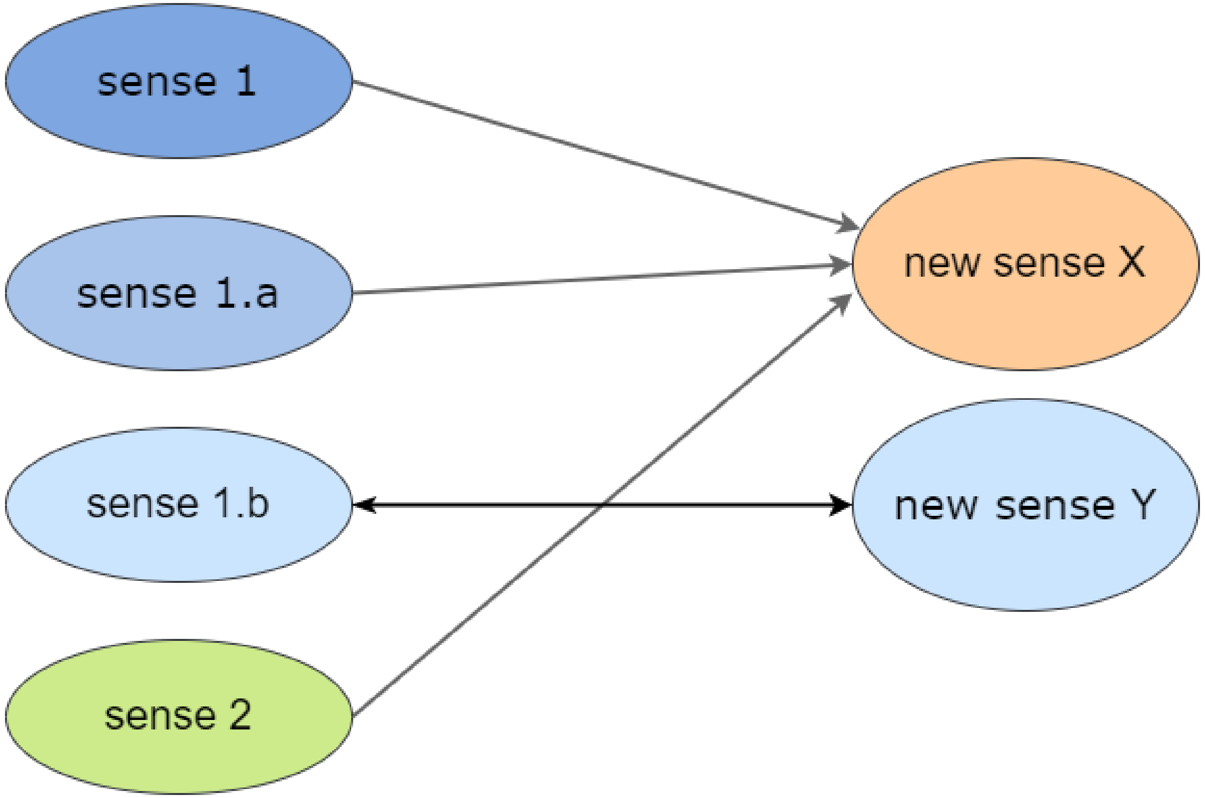
Experiments with automatic sense clustering

The task

Can we replicate the hand annotations with an automatic method?

Use dictionary and wordnet information to **partition** the **set** of a lemma’s non-deleted **senses** into ***k*** **clusters**.

k = number of senses in COR



Challenges

- **Varying** and **unknown** numbers of clusters for different lemmas
  - 2-step method inspired by
    - dataset alignment (McCrae & Buitelaar, 2018)
    - ELEXIS Clusty tool (Martelli et al, 2019)

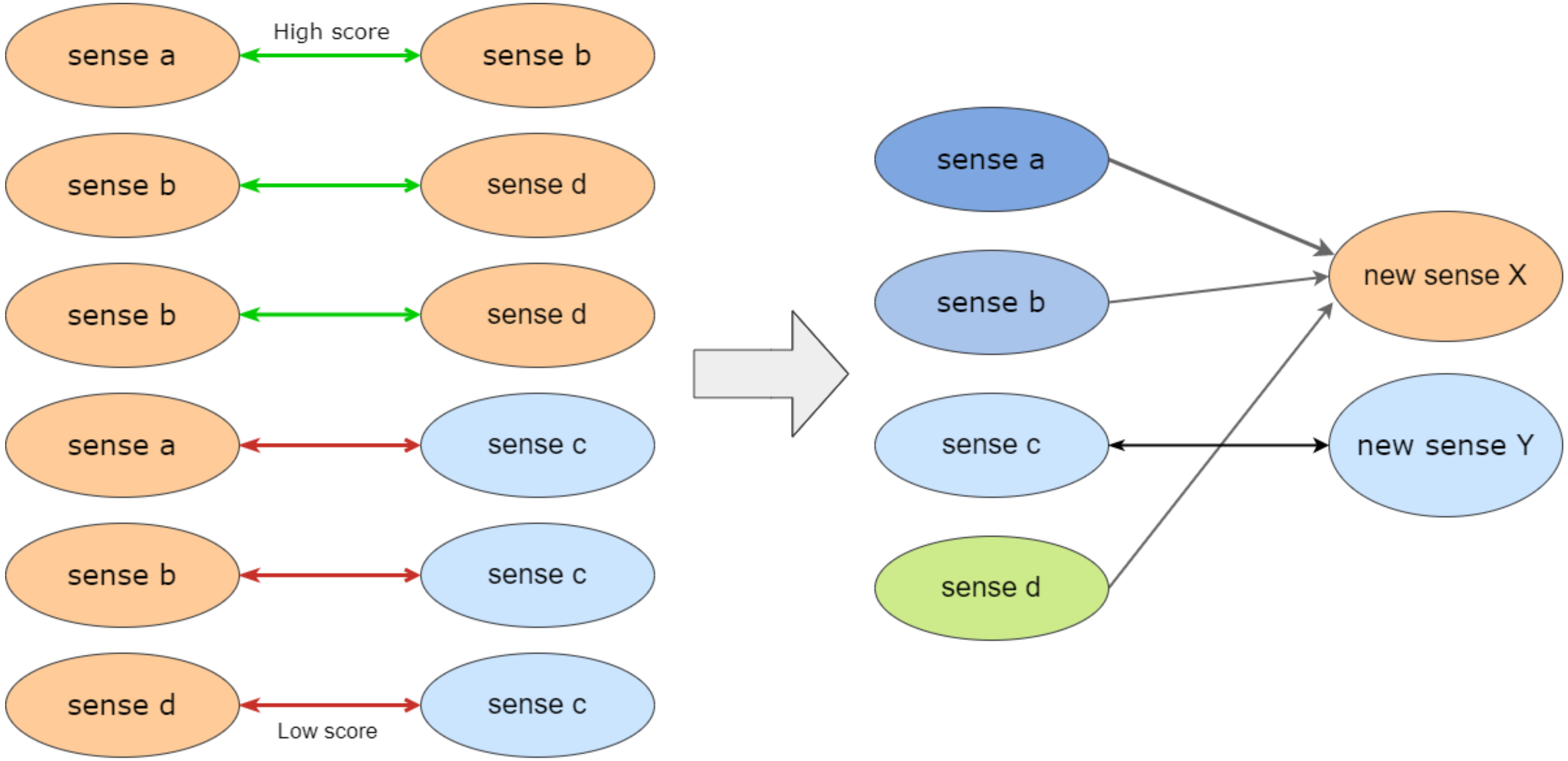
- How to **model** information from **dictionaries**?

  - Text-based: definition & quotes with a word embedding model (**word2vec, BERT**)

    +

  - Rule-based: Hand-selected features from **DDO** and **DanNet**

The 2 step method

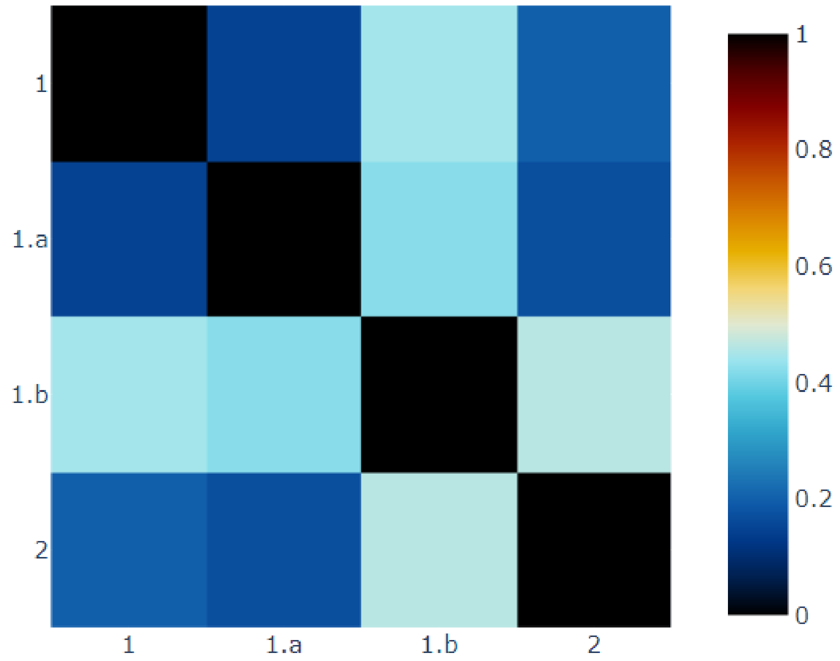


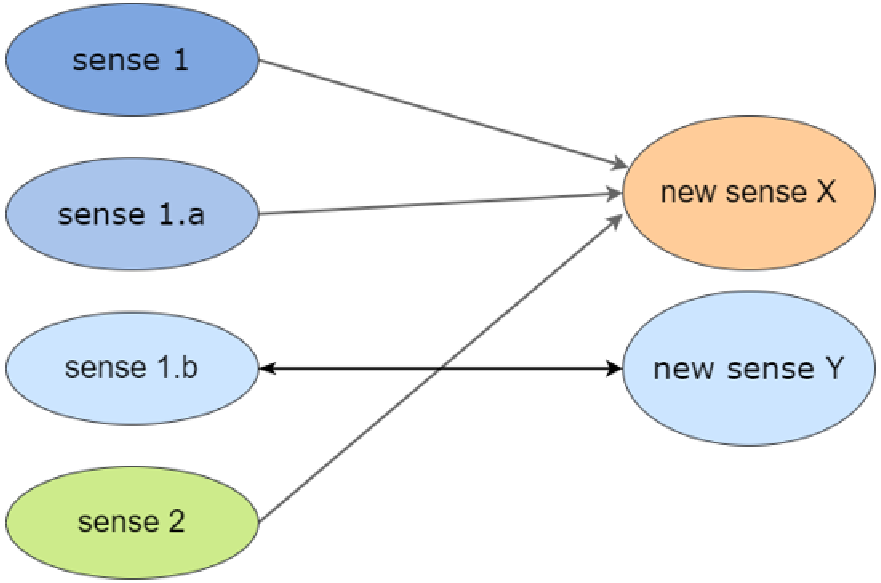
Step 1: Calculate pairwise sense proximity score using a model

Step 2: Clustering based on the sense proximity score

How similar are pairs of senses (sense proximity/similarity)?

1. Semantic Textual Similarity (STS) using BERT or word2vec
2. Use the principles (rule-based)





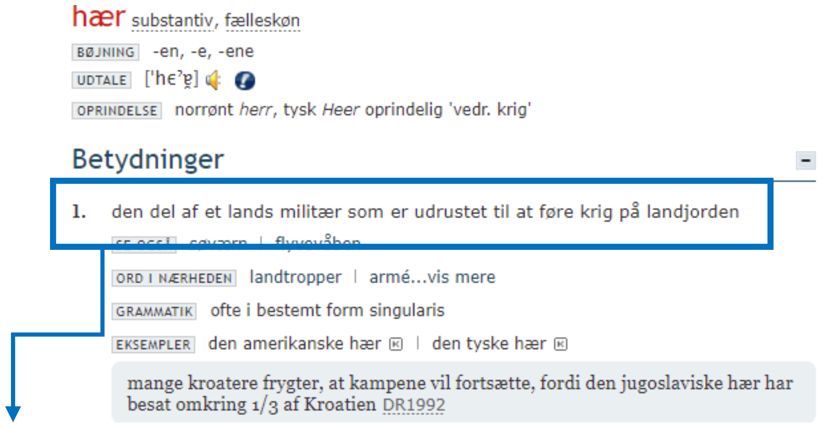
Text-based models

Word2vec

- Centroid embedding of the bag-of-words from definitions + quotes

  (punctuation + stopwords removed)

- Cosine distance as similarity measure



hær substantiv, fælleskøn

BØJNING -en, -e, -ene

UDTALE ['hɛˀɐ̯]

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

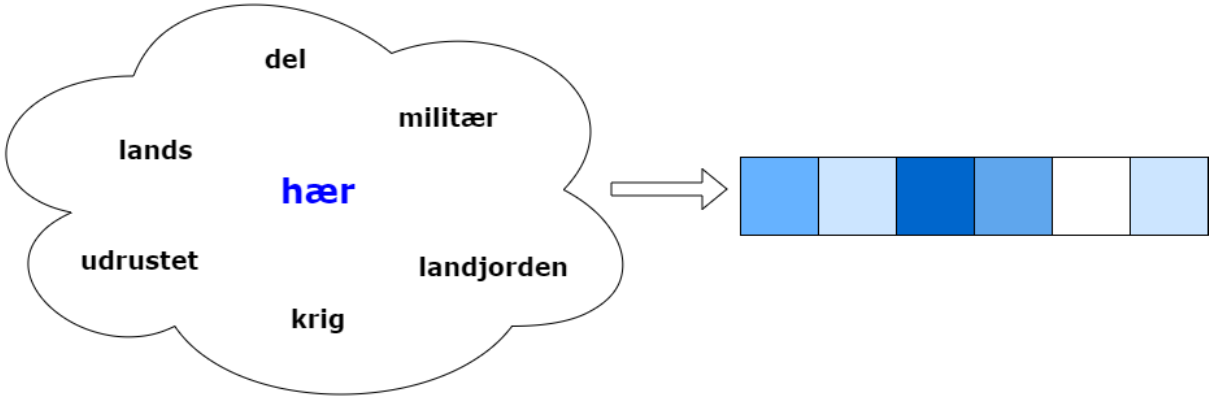
SE OGSÅ søværn | flyvevåben

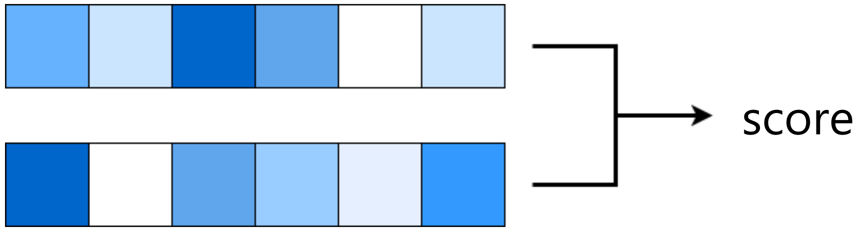
ORD I NÆRHEDEN landtropper | armé...vis mere

GRAMMATIK ofte i bestemt form singularis

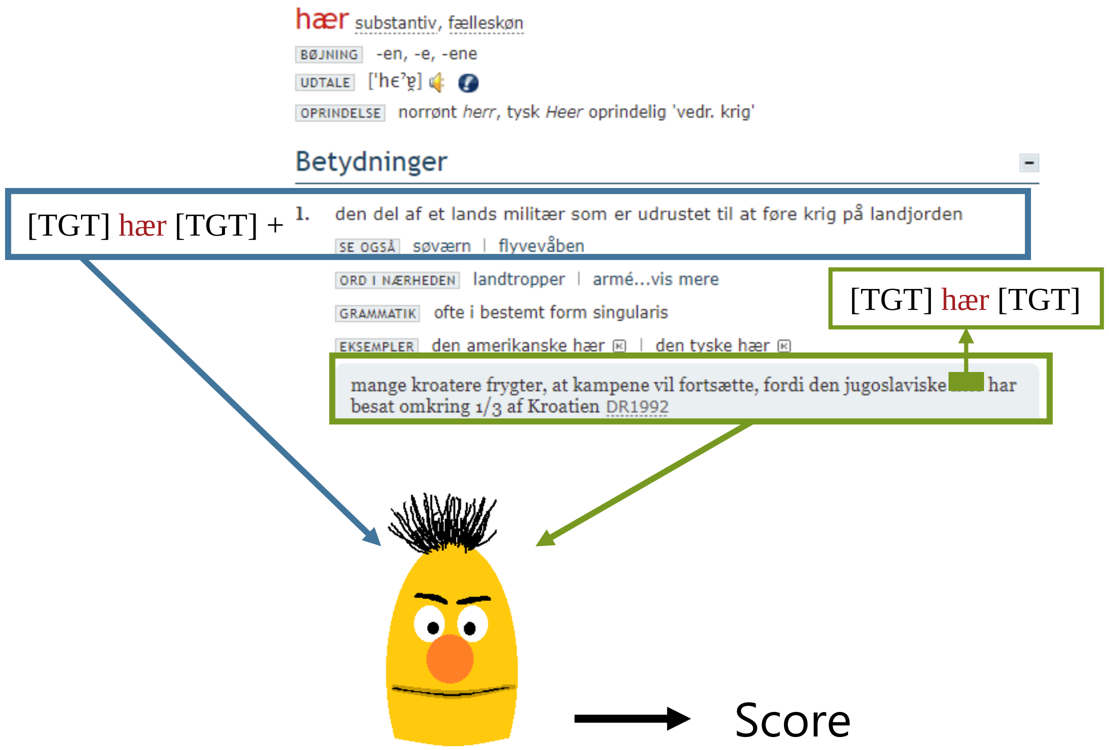
EKSEMPLER den amerikanske hær | den tyske hær

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





Text-based models



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

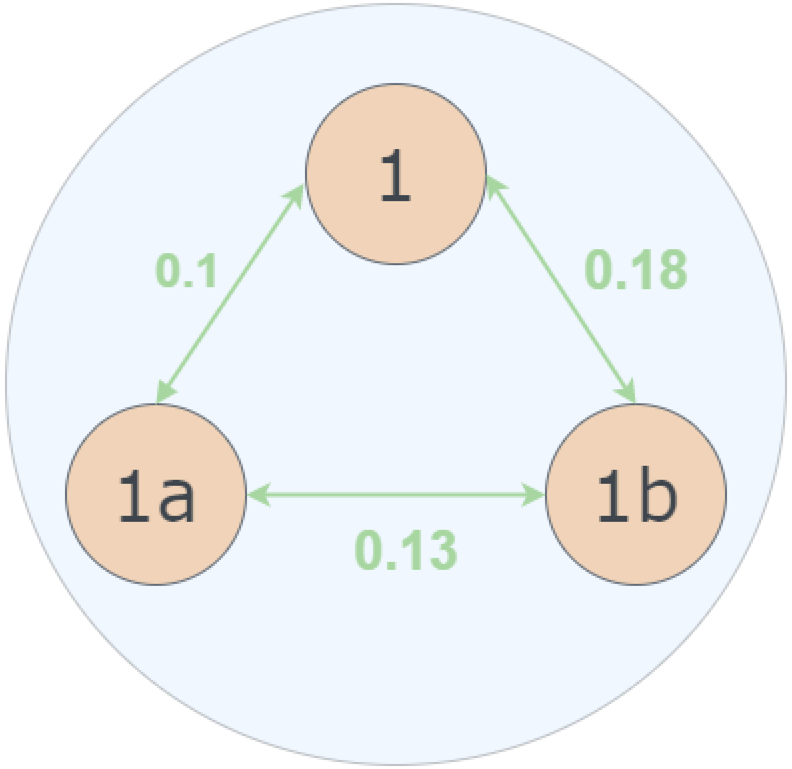
EKSEMPLER den amerikanske hær | den tyske hær

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

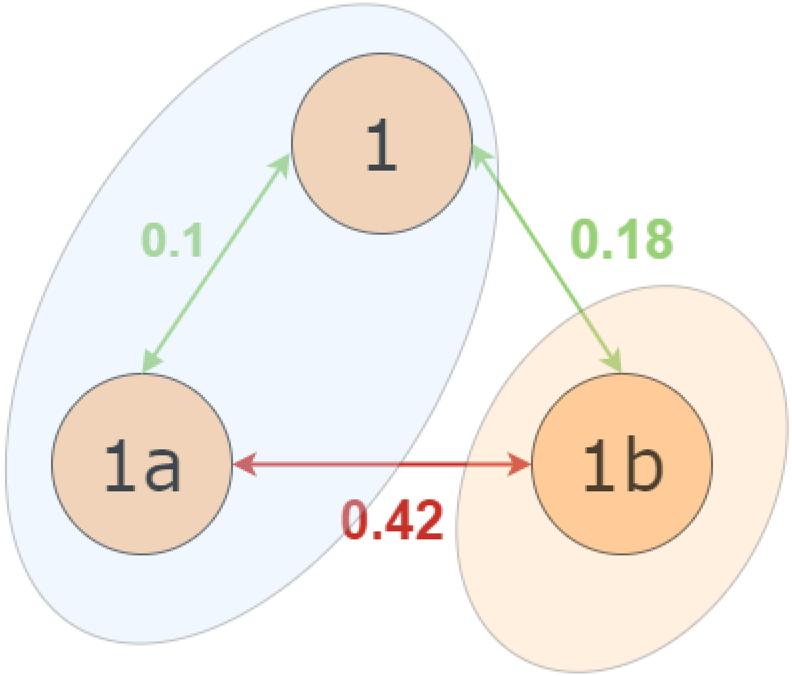
BERT

- Input: two sentences / contexts (quote and/or definition)
  - target lemma marked with [TGT] token
- Output: similarity score
- Fine-tune on 80% of the main annotation

Clustering based on similarity scores

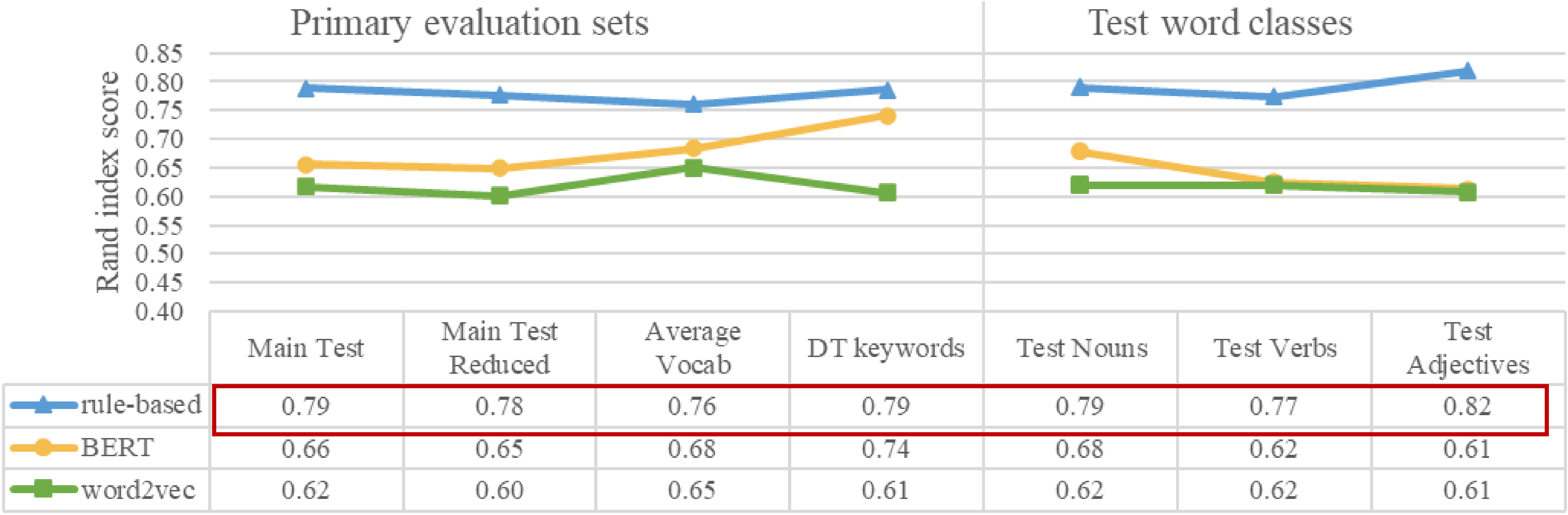


1 merged COR sense



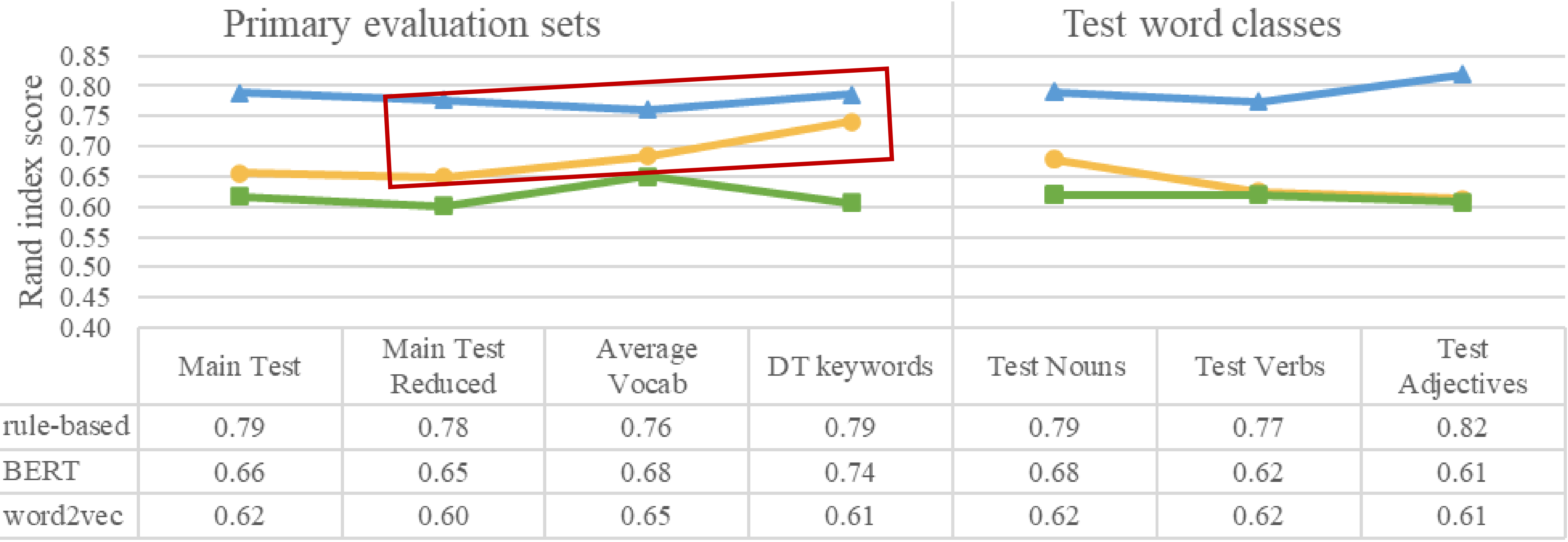
2 COR senses (blue, orange)

Results



	Main Test	Main Test Reduced	Average Vocab	DT keywords	Test Nouns	Test Verbs	Test Adjectives
▲ rule-based	0.79	0.78	0.76	0.79	0.79	0.77	0.82
● BERT	0.66	0.65	0.68	0.74	0.68	0.62	0.61
■ word2vec	0.62	0.60	0.65	0.61	0.62	0.62	0.61

Results



	Main Test	Main Test Reduced	Average Vocab	DT keywords	Test Nouns	Test Verbs	Test Adjectives
▲ rule-based	0.79	0.78	0.76	0.79	0.79	0.77	0.82
● BERT	0.66	0.65	0.68	0.74	0.68	0.62	0.61
■ word2vec	0.62	0.60	0.65	0.61	0.62	0.62	0.61

Conclusion

Conclusion

- With DDO as starting point we establish a notion of "coreness" and establish principles for merging senses
- Sense reduction of 43% from DDO to COR with an intercoder agreement of 0.82 -> in other words, the principles seem sound and manageable
- Rule-based model shows promise for automatic sense reduction of the remainder of the vocabulary
- Word classes should be treated differently
- Text-based approaches struggle with highly polysemous lemmas – why hand annotation are still necessary

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)