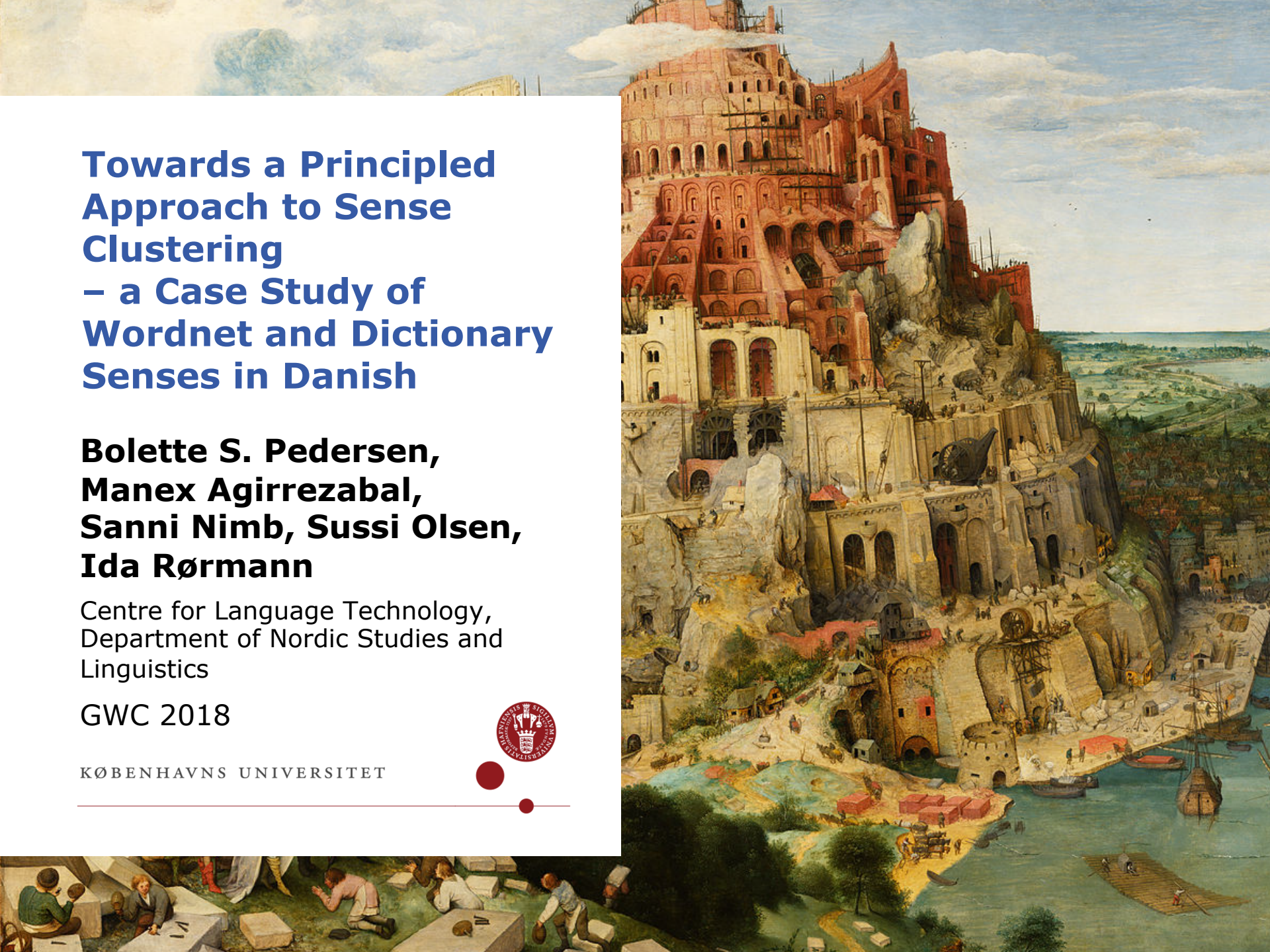


Towards a Principled Approach to Sense Clustering – a Case Study of Wordnet and Dictionary Senses in Danish

Bolette S. Pedersen, Manex Agirrezabal, Sanni Nimb, Sussi Olsen, Ida Rørmann

Centre for Language Technology, Department of Nordic Studies and Linguistics

GWC 2018

KØBENHAVNS UNIVERSITET

Overall goal

- To make existing lexical resources and their sense inventories more practically useful in NLP
  - not too fine-grained to be operational
  - yet fine-grained enough to be worth the trouble
- Questions asked:
  - which senses/sense clusters are manageable for human annotators
  - which senses/sense clusters work in WSD
- Data examined:
  - 10 of the most polysemous nouns in Danish
  - Senses as described in DanNet and DDO compared to occurrence in a corpus

Contents

1. Introduction, what's the problem
2. Sense organization in DDO and DanNet
3. Principled establishment of clusters
4. Corpus and annotation
5. Annotation results
6. Word sense disambiguation using the LibLINEAR package
7. Concluding remarks

Introduction, what's the problem

- Dealing with finegrained lexical sense inventories in NLP is a challenging task, selecting the correct sense in a specific context is incredibly hard when word meaning is richly described with subtle and detailed sense distinctions as found in most wordnets and lexica

- Conventional dictionaries have a highly structured sense inventory typically describing the vocabulary by means of *main- and subsenses*

- Wordnets are generally fine-grained and *unstructured,* in some cases ontologically tagged

Approaches

Coarse-grained word-sense disambiguation has become a well-established discipline over the years.

- Approach 1: Supersense tagging using for instance WordNet's *first beginners* as a cross-lingual sense inventory (comparable to the categories used in Named Entity Recognition)

- Approach 2: cluster existing inventories from dictionaries
  - manually or
  - automatically

Approaches

Coarse-grained word-sense disambiguation has become a well-established discipline over the years.

- Approach 1: Supersense tagging using for instance WordNet's *first beginners* as a cross-lingual sense inventory (comparable to the categories used in Named Entity Recognition)

- Approach 2: cluster existing inventories from dictionaries
  - manually or
  - automatically

Approaches

Informativeness ↓ **Coarse-grained**

Supersense tagging

Reduced clusters of DDO/DanNet

Clusters of DDO/DanNet

Full sense inventory from DDO/DanNet (”regular”)

Fine-grained

Cross-linguality ↑ **Language independent**

Language specific

Approaches

Informativeness ↓ **Cross-linguality** ↑

Coarse-grained — **Language independent**

Supersense tagging

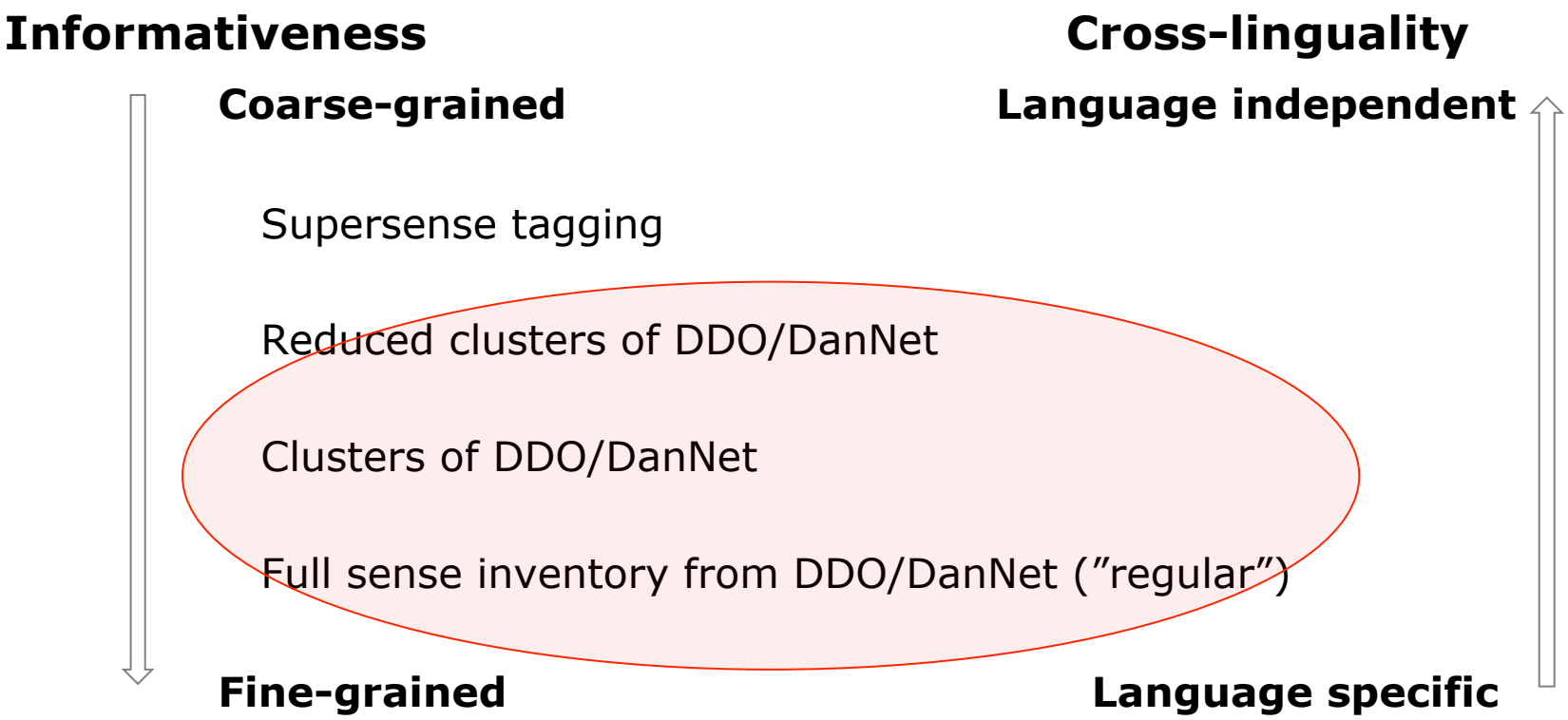
Reduced clusters of DDO/DanNet

Clusters of DDO/DanNet

Full sense inventory from DDO/DanNet (”regular”)

Fine-grained — **Language specific**

Approaches


Informativeness
Cross-linguality
Coarse-grained
Language independent
Supersense tagging
Reduced clusters of DDO/DanNet
Clusters of DDO/DanNet
Full sense inventory from DDO/DanNet (”regular”)
Fine-grained
Language specific

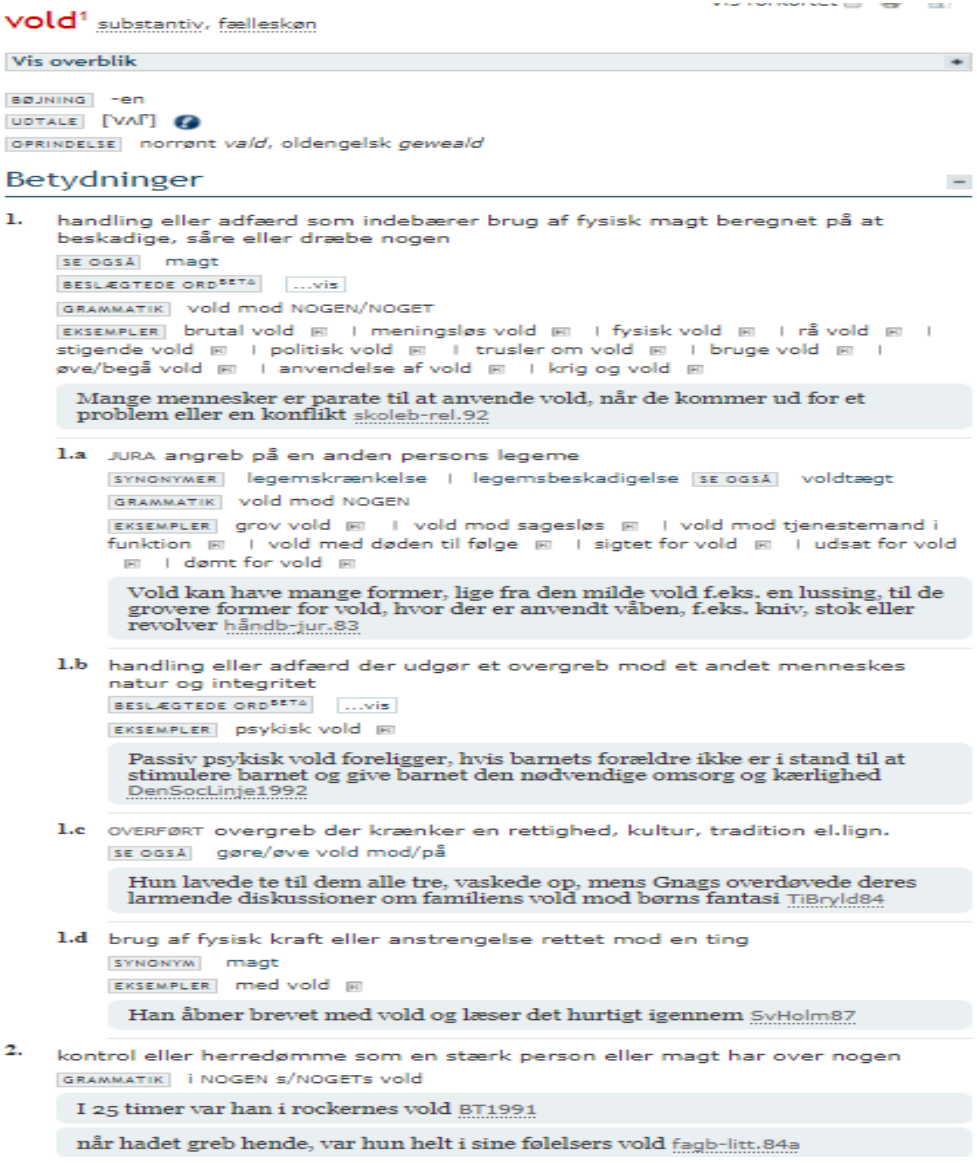
Sense organization in DDO and DanNet

Den Danske Ordbog (DDO)

The Danish Wordnet, DanNet

DANNET

Sense organization in DDO



vold[1] substantiv, fælleskøn

Vis overblik

BØJNING -en
UDTALE ['vʌl']
OPRINDELSE norrønt *vald*, oldengelsk *gewald*

Betydninger

1. handling eller adfærd som indebærer brug af fysisk magt beregnet på at beskadige, såre eller dræbe nogen
SE OGSÅ magt
BESLÆGTEDE ORD BETA ...vis
GRAMMATIK vold mod NOGEN/NOGET
EKSEMPLER brutal vold | meningsløs vold | fysisk vold | rå vold | stigende vold | politisk vold | trusler om vold | bruge vold | øve/begå vold | anvendelse af vold | krig og vold

> Mange mennesker er parate til at anvende vold, når de kommer ud for et problem eller en konflikt skoleb-rel.92

1.a JURA angreb på en anden persons legeme
SYNONYMER legemskrænkelse | legemsbeskadigelse SE OGSÅ voldtægt
GRAMMATIK vold mod NOGEN
EKSEMPLER grov vold | vold mod sagesløs | vold mod tjenestemand i funktion | vold med døden til følge | sigtet for vold | udsat for vold | dømt for vold

> Vold kan have mange former, lige fra den milde vold f.eks. en lussing, til de grovere former for vold, hvor der er anvendt våben, f.eks. kniv, stok eller revolver håndb-jur.83

1.b handling eller adfærd der udgør et overgreb mod et andet menneskes natur og integritet
BESLÆGTEDE ORD BETA ...vis
EKSEMPLER psykisk vold

> Passiv psykisk vold foreligger, hvis barnets forældre ikke er i stand til at stimulere barnet og give barnet den nødvendige omsorg og kærlighed DenSocLinje1992

1.c OVERFØRT overgreb der krænker en rettighed, kultur, tradition el.lign.
SE OGSÅ gøre/øve vold mod/på

> Hun lavede te til dem alle tre, vaskede op, mens Gnags overdøvede deres larmende diskussioner om familiens vold mod børns fantasi TiBryld84

1.d brug af fysisk kraft eller anstrengelse rettet mod en ting
SYNONYM magt
EKSEMPLER med vold

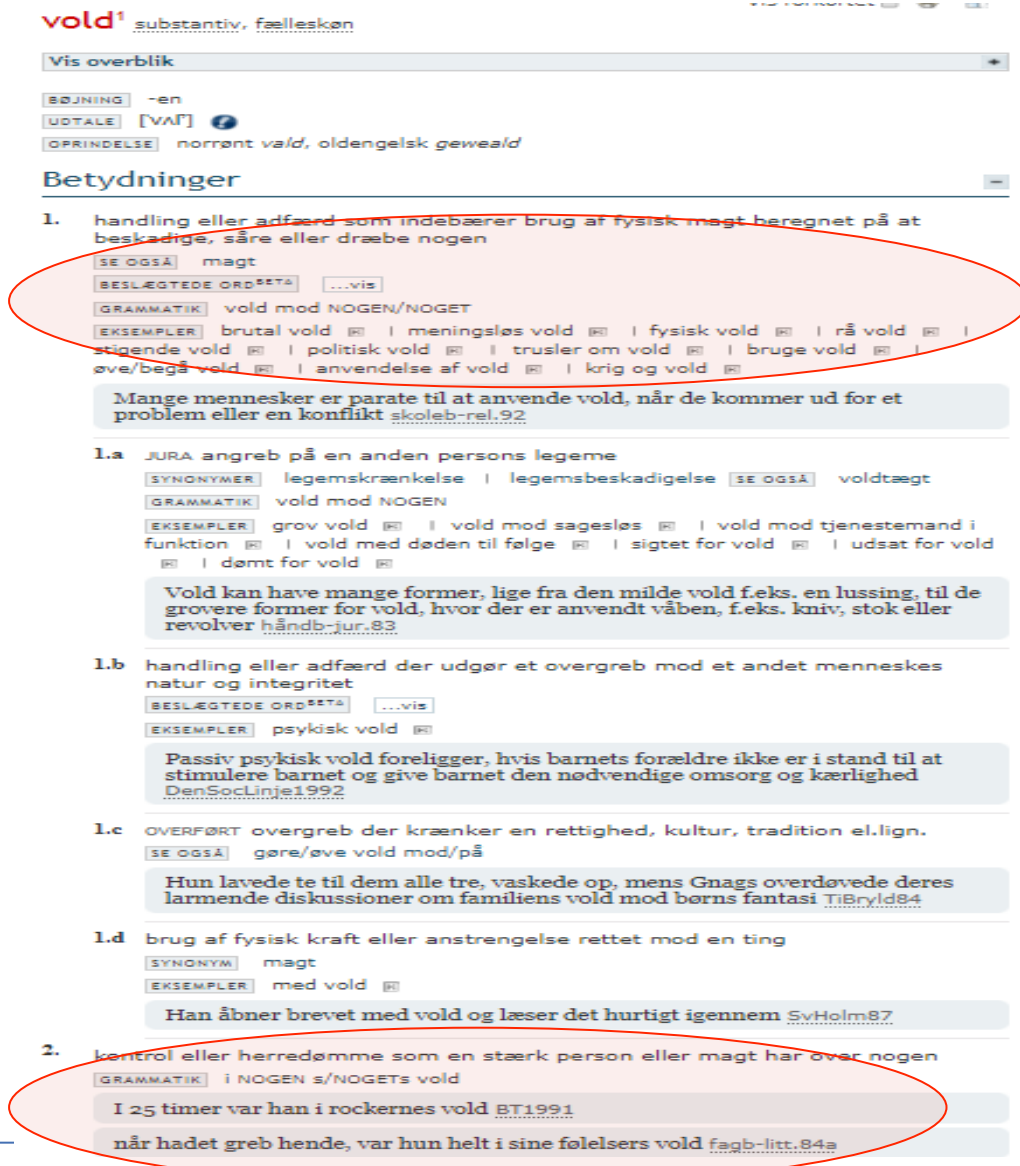
> Han åbner brevet med vold og læser det hurtigt igennem SvHolm87

2. kontrol eller herredømme som en stærk person eller magt har over nogen
GRAMMATIK i NOGEN s/NOGETs vold

> I 25 timer var han i rockernes vold BT1991

> når hadet greb hende, var hun helt i sine følelsers vold fagb-litt.84a

Sense organization in DDO



vold[1] substantiv, fælleskøn

Vis overblik

BØJNING -en
UDTALE ['vʌl']
OPRINDELSE norrønt *vald*, oldengelsk *gewealdd*

Betydninger

1. handling eller adfærd som indebærer brug af fysisk magt beregnet på at beskadige, såre eller dræbe nogen
SE OGSÅ magt
BESLÆGTEDE ORD BETA ...vis
GRAMMATIK vold mod NOGEN/NOGET
EKSEMPLER brutal vold | meningsløs vold | fysisk vold | rå vold | stigende vold | politisk vold | trusler om vold | bruge vold | øve/begå vold | anvendelse af vold | krig og vold

> Mange mennesker er parate til at anvende vold, når de kommer ud for et problem eller en konflikt skoleb-rel.92

1.a JURA angreb på en anden persons legeme
SYNONYMER legemskrænkelse | legemsbeskadigelse SE OGSÅ voldtægt
GRAMMATIK vold mod NOGEN
EKSEMPLER grov vold | vold mod sagesløs | vold mod tjenestemand i funktion | vold med døden til følge | sigtet for vold | udsat for vold | dømt for vold

> Vold kan have mange former, lige fra den milde vold f.eks. en lussing, til de grovere former for vold, hvor der er anvendt våben, f.eks. kniv, stok eller revolver håndb-jur.83

1.b handling eller adfærd der udgør et overgreb mod et andet menneskes natur og integritet
BESLÆGTEDE ORD BETA ...vis
EKSEMPLER psykisk vold

> Passiv psykisk vold foreligger, hvis barnets forældre ikke er i stand til at stimulere barnet og give barnet den nødvendige omsorg og kærlighed DenSocLinje1992

1.c OVERFØRT overgreb der krænker en rettighed, kultur, tradition el.lign.
SE OGSÅ gøre/øve vold mod/på

> Hun lavede te til dem alle tre, vaskede op, mens Gnags overdøvede deres larmende diskussioner om familiens vold mod børns fantasi TiBryld84

1.d brug af fysisk kraft eller anstrengelse rettet mod en ting
SYNONYM magt
EKSEMPLER med vold

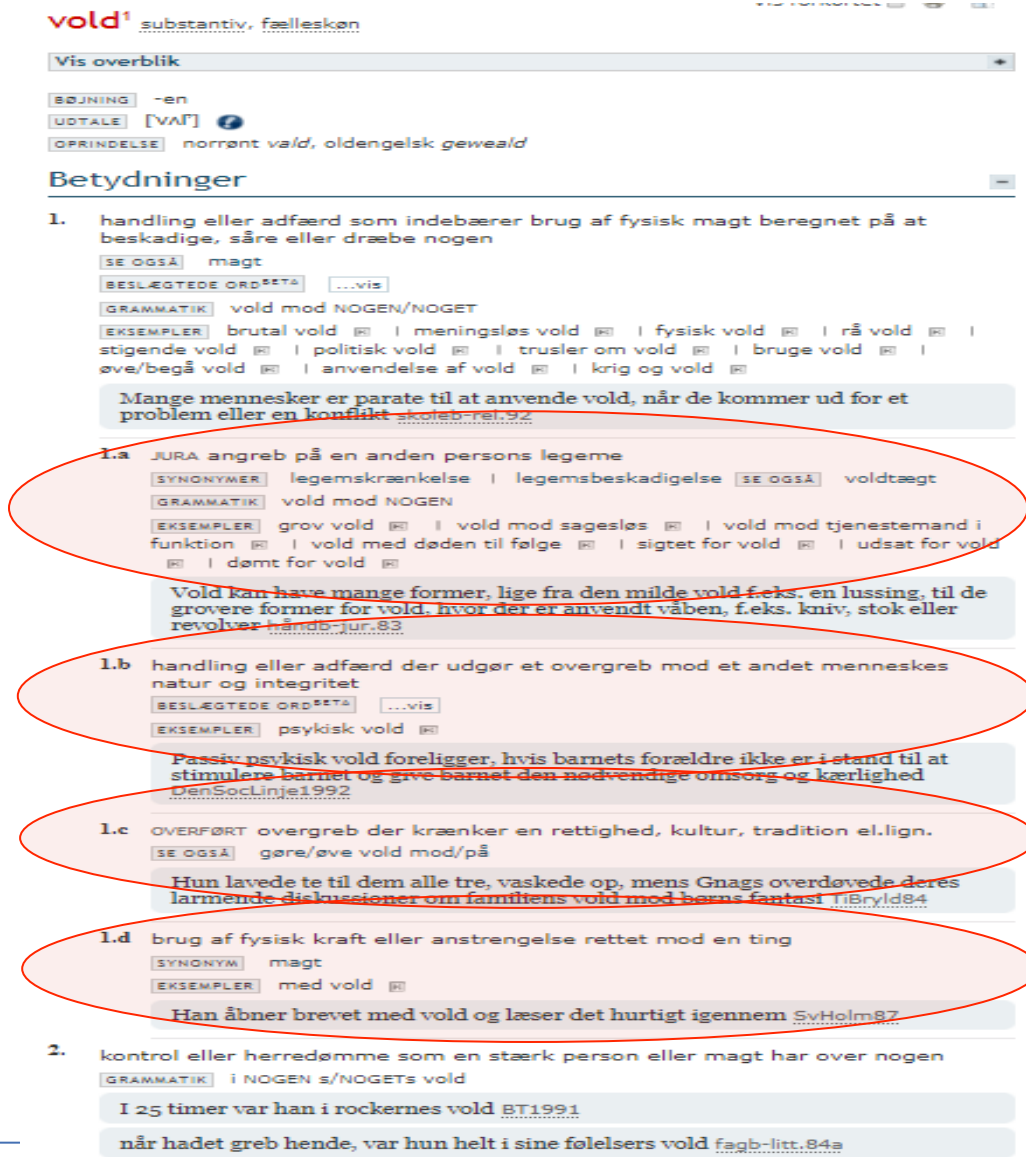
> Han åbner brevet med vold og læser det hurtigt igennem SvHolm87

2. kontrol eller herredømme som en stærk person eller magt har over nogen
GRAMMATIK i NOGEN s/NOGETs vold

> I 25 timer var han i rockernes vold BT1991

> når hadet greb hende, var hun helt i sine følelsers vold fagb-litt.84a

Sense organization in DDO



vold[1] substantiv, fælleskøn

Vis overblik

BØJNING -en
UDTALE ['vʌl']
OPRINDELSE norrønt *vald*, oldengelsk *gewealdd*

Betydninger

1. handling eller adfærd som indebærer brug af fysisk magt beregnet på at beskadige, såre eller dræbe nogen
   SE OGSÅ magt
   BESLÆGTEDE ORD BETA ...vis
   GRAMMATIK vold mod NOGEN/NOGET
   EKSEMPLER brutal vold | meningsløs vold | fysisk vold | rå vold | stigende vold | politisk vold | trusler om vold | bruge vold | øve/begå vold | anvendelse af vold | krig og vold

   Mange mennesker er parate til at anvende vold, når de kommer ud for et problem eller en konflikt skoleb-rel.92

   1.a JURA angreb på en anden persons legeme
   SYNONYMER legemskrænkelse | legemsbeskadigelse SE OGSÅ voldtægt
   GRAMMATIK vold mod NOGEN
   EKSEMPLER grov vold | vold mod sagesløs | vold mod tjenestemand i funktion | vold med døden til følge | sigtet for vold | udsat for vold | dømt for vold

   Vold kan have mange former, lige fra den milde vold f.eks. en lussing, til de grovere former for vold, hvor der er anvendt våben, f.eks. kniv, stok eller revolver håndb-jur.83

   1.b handling eller adfærd der udgør et overgreb mod et andet menneskes natur og integritet
   BESLÆGTEDE ORD BETA ...vis
   EKSEMPLER psykisk vold

   Passiv psykisk vold foreligger, hvis barnets forældre ikke er i stand til at stimulere barnet og give barnet den nødvendige omsorg og kærlighed DenSocLinje1992

   1.c OVERFØRT overgreb der krænker en rettighed, kultur, tradition el.lign.
   SE OGSÅ gøre/øve vold mod/på

   Hun lavede te til dem alle tre, vaskede op, mens Gnags overdøvede deres larmende diskussioner om familiens vold mod børns fantasi TiBryld84

   1.d brug af fysisk kraft eller anstrengelse rettet mod en ting
   SYNONYM magt
   EKSEMPLER med vold

   Han åbner brevet med vold og læser det hurtigt igennem SvHolm87

2. kontrol eller herredømme som en stærk person eller magt har over nogen
   GRAMMATIK i NOGEN s/NOGETs vold

   I 25 timer var han i rockernes vold BT1991

   når hadet greb hende, var hun helt i sine følelsers vold fagb-litt.84a

Sense organization in DDO

- **Auto-hyponymy**: narrowed meaning with same hypernym, as in *to drink alcohol* as a subsense to *to drink*
- **Auto-superordination**: extended meaning as in *man* (person) vs *man* (male)
- **Auto-meronymy**: a part instead of the whole as in *door* meaning a piece of wood, metal or the like in contrast to *door* in the broader opening sense (as in *the door was made of wood* vs. *he closed the door*).
- **Auto-holonymy**: a whole instead of the part as in *body* meaning the whole body in contrast to *body* in the sense of the torso only.
- **Figurative**: sense where only part of the meaning is derived from the core sense but used in a figurative/metaphorical context as in *window* in the sense *a window to the world*.

Sense organization in DDO

Factors that overrule these principles:

- **Frequency of the senses** "big words" tend to establish main senses where they should actually have been subsenses according to Cruse

- **Communicative factor** of the structure: overall goal was to compile an 'easy to read' printed dictionary, especially by avoiding very deep sense structures

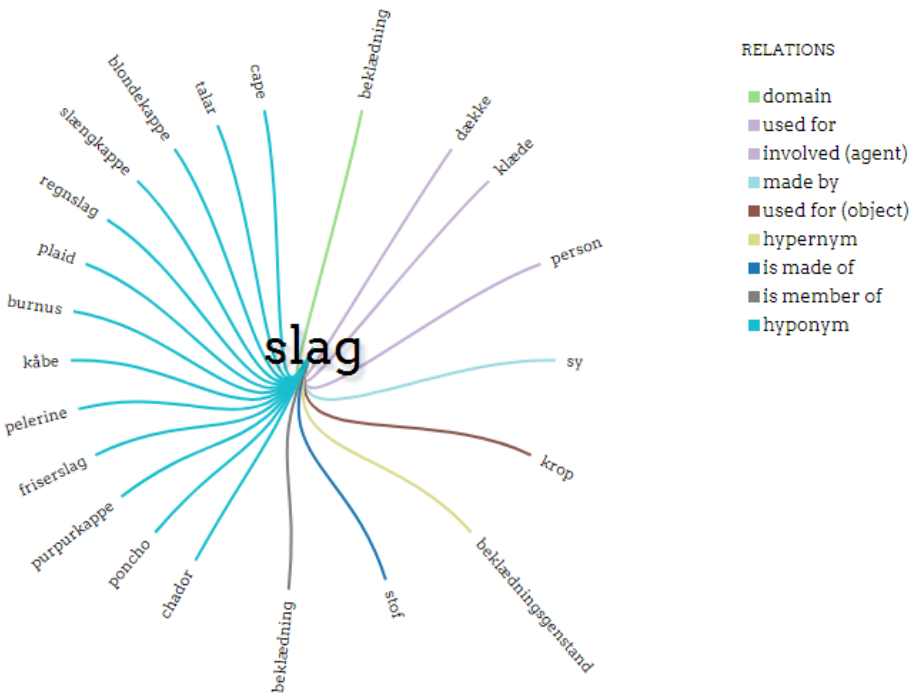
Sense organization in DanNet

- Senses in DanNet are organized in terms of synsets
- Each synset is assigned an ontological type based on EuroWordNets' top ontology
- All synsets all have equal status, i.e. no main and sub-senses
- Further, each synset is inter-related to other synsets via semantic relations

DanNet relations

slag 7 N

(lang) vid beklædningsgenstand som ikke har ærmer, ...



DanNet: Ontological types

(EuroWordnet topontology)

- Origin
  - Natural
    - Living
      - Plant
      - Human
      - Creature
      - Animal
  - Artefact
- Form
  - Substance
    - Solid
    - Liquid
    - Gas
  - Object
- Composition
  - Part
  - Group
- Function
  - Vehicle
  - Representation
    - MoneyRepresentation
    - LanguageRepresentation
    - ImageRepresentation
  - Software
  - Place
  - Occupation
  - Instrument
  - Garment
  - Furniture
  - Covering
  - Container
  - Comestible
  - Building

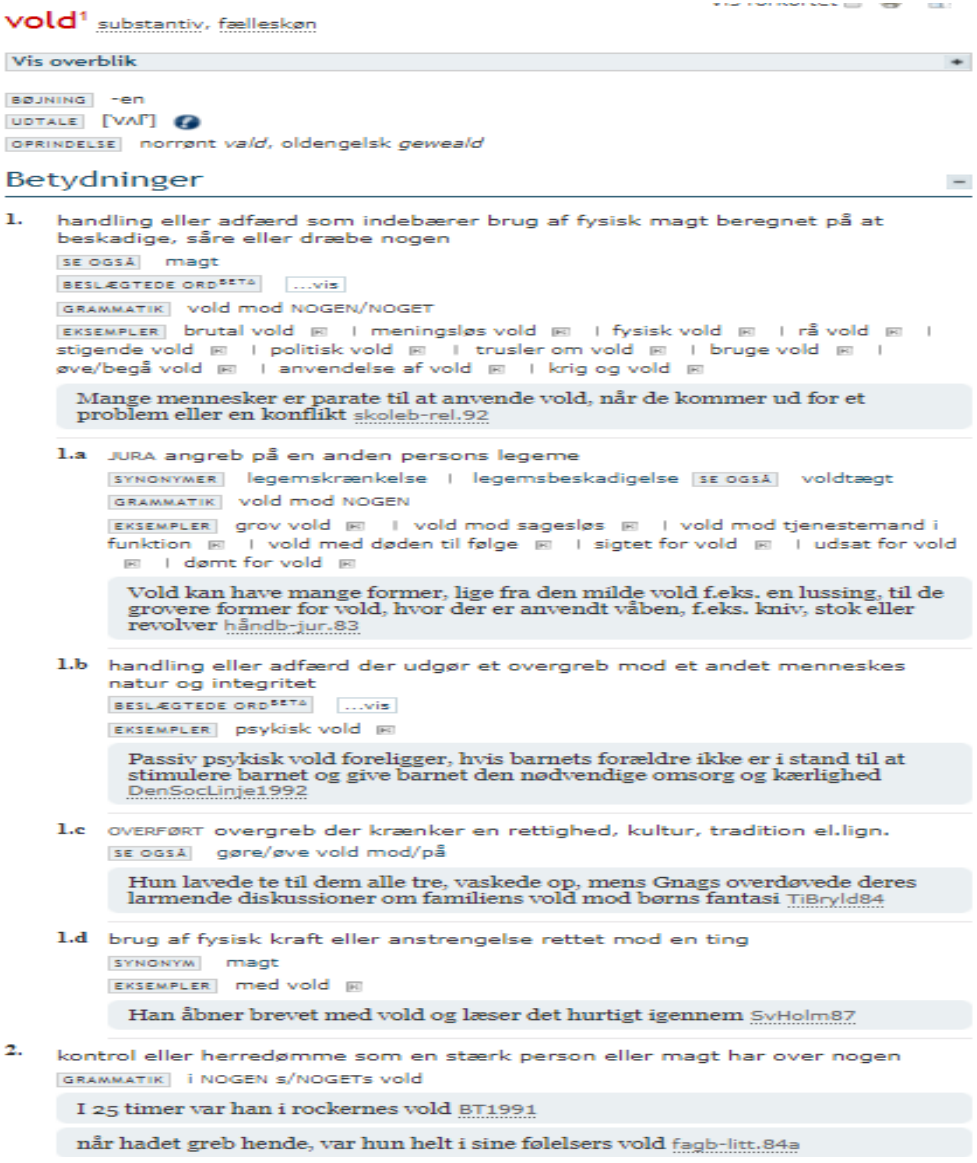
- SituationType
  - Dynamic
    - BoundedEvent
    - UnboundedEvent
  - Static
    - Property
    - Relation
- SituationComponent
  - Cause
    - Agentive
    - Phenomenal
    - Stimulating
  - Communication
  - Condition
  - Existence
  - Experience
  - Location
  - Manner
  - Mental
  - Modal
  - Physical
  - Possession
  - Purpose
  - Quantity
  - Social
  - Time
  - Usage

Establishment of clusters

Exploiting semantic info from both sources

- **Experiment 1** ('regular') where all main and subsenses are maintained
- **Experiment 2** ('clustered') where subsenses are clustered if they are of the same ontological type
- **Experiment 3** ('clustered reduced') where also main senses are clustered if they are of the same ontological type.

Establishment of clusters



vold[1] substantiv, fælleskøn

Vis overblik

BØJNING -en
UDTALE ['vʌl']
OPRINDELSE norrønt *vald*, oldengelsk *gewealdd*

Betydninger

1. handling eller adfærd som indebærer brug af fysisk magt beregnet på at beskadige, såre eller dræbe nogen

SE OGSÅ magt
BESLÆGTEDE ORD BETA ...vis
GRAMMATIK vold mod NOGEN/NOGET
EKSEMPLER brutal vold | meningsløs vold | fysisk vold | rå vold | stigende vold | politisk vold | trusler om vold | bruge vold | øve/begå vold | anvendelse af vold | krig og vold

Mange mennesker er parate til at anvende vold, når de kommer ud for et problem eller en konflikt skoleb-rel.92

1.a JURA angreb på en anden persons legeme

SYNONYMER legemskrænkelse | legemsbeskadigelse SE OGSÅ voldtægt
GRAMMATIK vold mod NOGEN
EKSEMPLER grov vold | vold mod sagesløs | vold mod tjenestemand i funktion | vold med døden til følge | sigtet for vold | udsat for vold | dømt for vold

Vold kan have mange former, lige fra den milde vold f.eks. en lussing, til de grovere former for vold, hvor der er anvendt våben, f.eks. kniv, stok eller revolver håndb-jur.83

1.b handling eller adfærd der udgør et overgreb mod et andet menneskes natur og integritet

BESLÆGTEDE ORD BETA ...vis
EKSEMPLER psykisk vold

Passiv psykisk vold foreligger, hvis barnets forældre ikke er i stand til at stimulere barnet og give barnet den nødvendige omsorg og kærlighed DenSocLinje1992

1.c OVERFØRT overgreb der krænker en rettighed, kultur, tradition el.lign.

SE OGSÅ gøre/øve vold mod/på

Hun lavede te til dem alle tre, vaskede op, mens Gnags overdøvede deres larmende diskussioner om familiens vold mod børns fantasi TiBryld84

1.d brug af fysisk kraft eller anstrengelse rettet mod en ting

SYNONYM magt
EKSEMPLER med vold

Han åbner brevet med vold og læser det hurtigt igennem SvHolm87

2. kontrol eller herredømme som en stærk person eller magt har over nogen

GRAMMATIK i NOGEN s/NOGETs vold

I 25 timer var han i rockernes vold BT1991

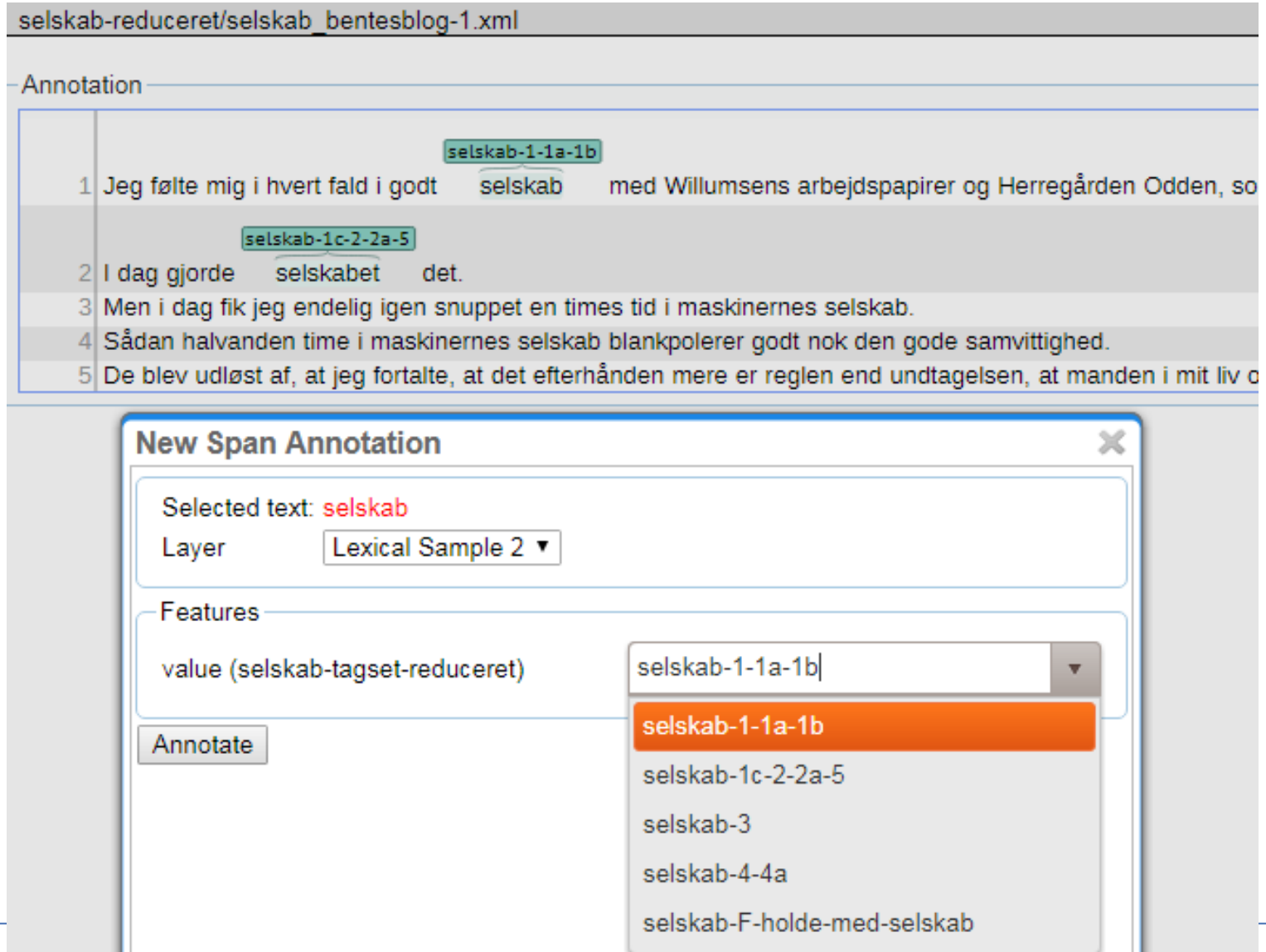
når hadet greb hende, var hun helt i sine følelsers vold fagb-litt.84a

Corpus and annotation

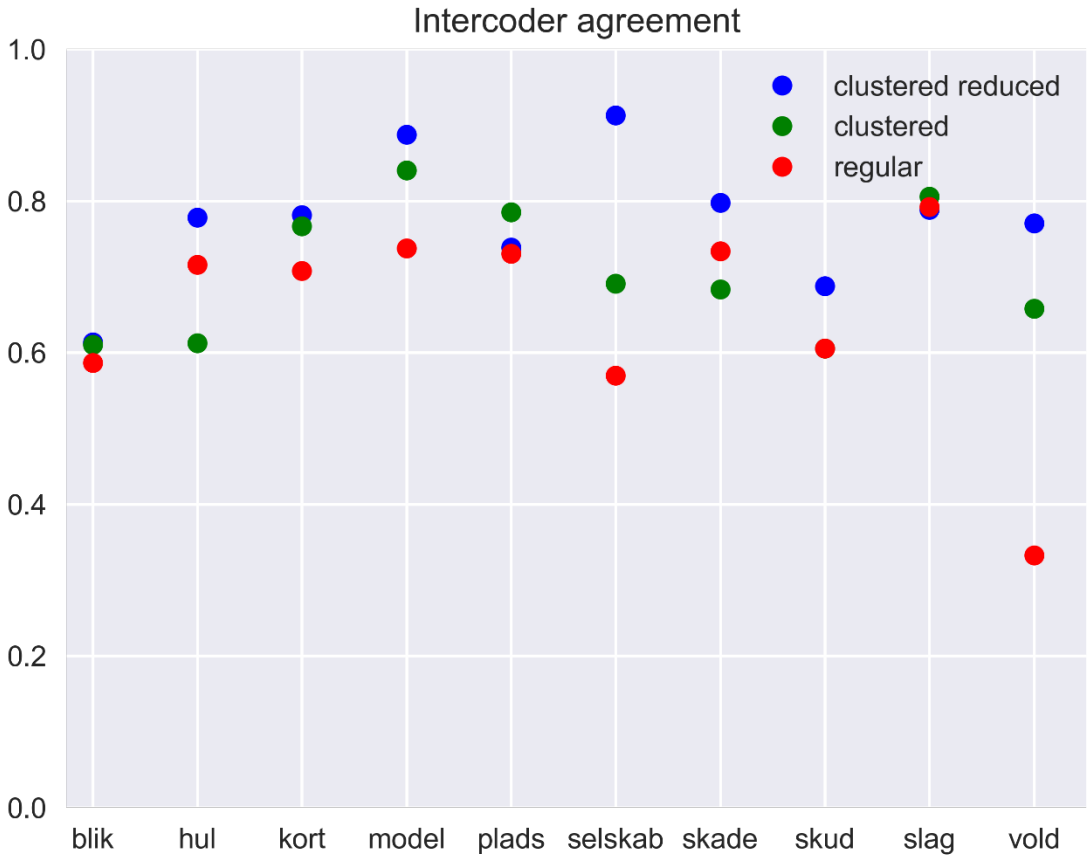
- The texts selected for annotation have been extracted from the 45 million words CLARIN Reference Corpus.
- The corpus contains a wide variety of text types and domains: blog, chat, forum, magazine, Parliament debates, and newswire.
- The number of annotated sentences for each noun varies according to the number of DDO senses of the noun (100 + 15*no. of senses), resulting in from 175 to 535 sentences per noun.

Corpus and annotation

WebAnno tool:



Intercoder agreement using Krippendorffs α



Intercoder divergences

Divergence types identified (when curating 2% of the material)

- **Underspecified examples**: Diverging annotations where the precise word sense could not be deduced from the isolated example (most divergences).

- **Incomplete or unclear tag set***:* Diverging annotations in cases where a new/unconventional sense of the word was not covered by the tag set, or where the lexical description of a tag was unclear or blurred.

- **Plain errors**: Diverging annotations due to wrong POS tags or because the annotator had erroneously skipped a word, for instance in cases with more than one lexical occurrence per sentence.

Intercoders' report

- Annotation tasks are generally reported to be very hard! In particular with the full sense inventory where the distinctions are often very subtle.
- In contrast, they report that the generated clusters are somewhat more intuitive for them to work with, but still hard
- One example is *selskab* where groups of people doing things together is described by many senses in the fine-grained experiment (party, group)– but in only one temporary cluster in the cluster experiments; a fact which increased agreement quite a lot
- In some cases, clusters are reported to be too coarse *kort* where two very different kinds of artifacts are clustered (playing cards and maps) due to same ontological type: Image Representation)
- **Special challenges**: metaphors and the digital universe – concrete or not?

WSD using the LibLINEAR package

A corresponding automatic disambiguation task using empirical methods (LibLINEAR package included in *scikit-learn* from Python).

- Disambiguate the ambiguous words in context (lexical sample task)
- See if there is any significant improvement of the prediction accuracies when using clustered word senses.

The features:

- Bag of lemmas of the whole sentence.
- Next and previous four lemmas (primarily devised to disambiguate idiomatic expressions whose structure is mostly fixed).

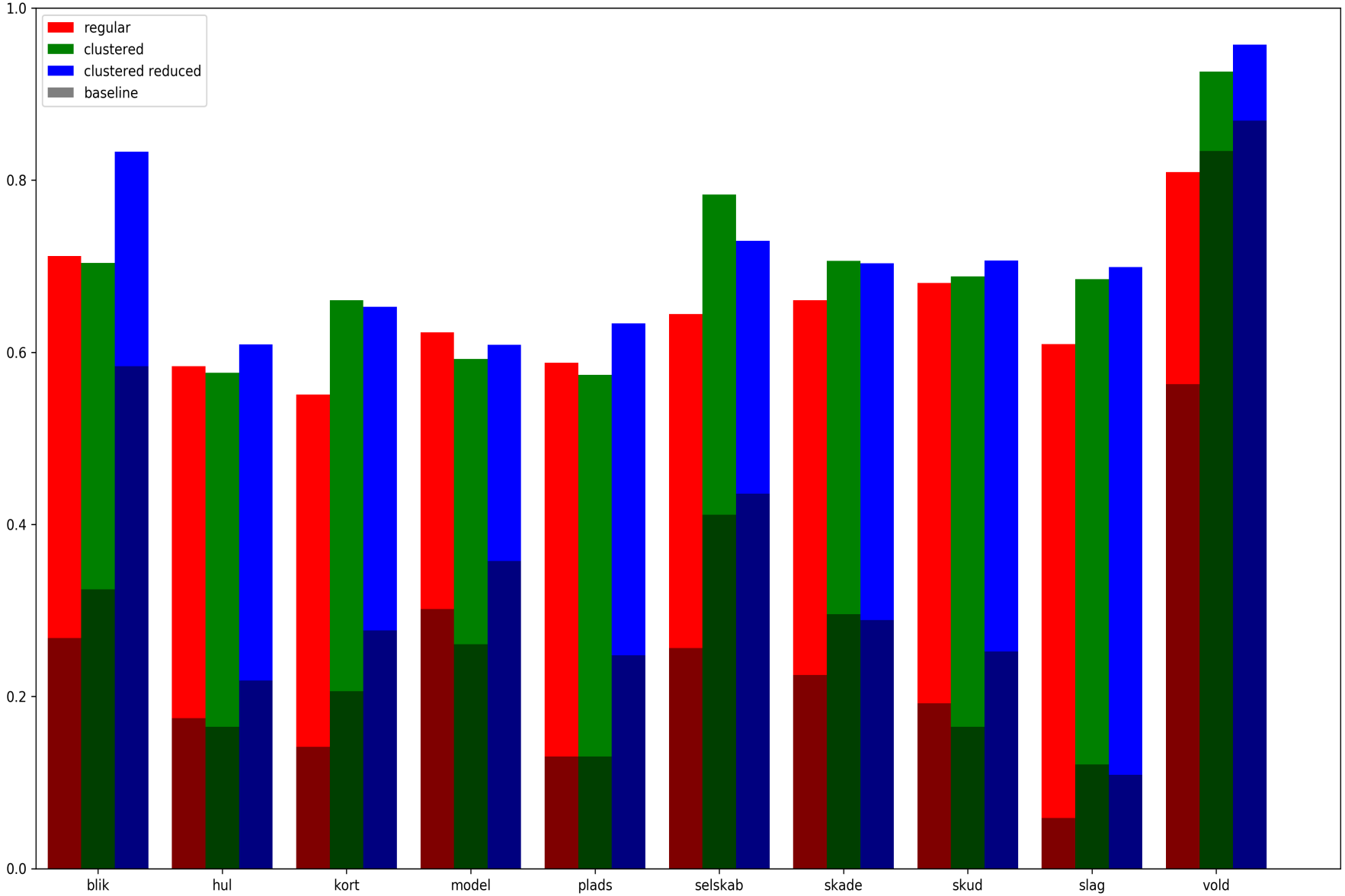
WSD using the LibLINEAR package

Evaluation of a model

- If two annotators have tagged a word in a sentence with diverging sense cluster tags, we consider it correct if an ML classifier classifies that instance as one of those sense clusters (either of them).
- This corresponds well to the fact that most divergences are caused by underspecified corpus examples.

For learning

- if two different annotators have tagged an instance, we consider it to be two different instances, resulting in some cases where we can have two instances with the same attributes, but with different outputs.


regular
clustered
clustered reduced
baseline
1.0
0.8
0.6
0.4
0.2
0.0
blik
hul
kort
model
plads
selskab
skade
skud
slag
vold

Concluding remarks

The task:

- How to cluster noun senses in a principled way based on existing semantic info (main and sub-senses *and* ontological typing) in order to obtain more convenient sense inventories
- Focus on some of the hardest and most polysemous nouns in Danish
- Examine how clusters influence inter-annotator agreement and automatic word sense disambiguation

Conclusion:

- Reduced clusters provides a more manageable inventory for both human annotators and the automatic disambiguation system.

Concluding remarks

Questions to be addressed in future work:

- How would random clustered have performed?
- How relevant are the sense clusters established for a specific NLP task (i.e. question/answering?)
- How do clusters based on lexicons and wordnets compare to the word profiles that appear with word embeddings and sense induction methods?
- How well will our method scale up to include verbs and adjectives?

Intercoder agreement

- Krippendorffs α calculates chance corrected agreement coefficients, i.e. sets off the fact (to some degree) that it is easier to agree on few tags than on many.
- An α value of 1 represents perfect agreement and a value of 0 indicates absence of agreement.
- It is customary to require $\alpha \geq .80$ in most annotations tasks, however, for sense annotation where more tentative conclusions are still acceptable, we consider $\alpha \geq .67$ reasonable and useful