Lexicography in natural language processing – is it useful?

Bolette Sandford Pedersen

Centre for Language Technology,
Department of Nordic Studies and Linguistics,
University of Copenhagen

bspedersen@hum.ku.dk

Contents

- Why lexicography in natural language processing?
- A Danish use case: Basing *a wordnet*, *a framenet* and *a semantically annotated corpus* on The Danish Dictionary (DDO) and The Danish Thesaurus (DT)
- The ELEXIS project: connecting the disconnected

Why lexicography in natural language processing

- NLP and Language Technology are *central building blocks in upcoming intelligent systems*
- Intelligent systems are undergoing an extreme fast development right now and are being integrated *everywhere in society*
- The closer intelligent systems get to our every day lives, the more need there is for *high-quality language technology based on our mother tongue*
- Important to base our language technology systems on locally anchored knowledge of *language* and *culture*

Why lexicography in natural language processing

- Dictionaries are not just systematic collections of words with information about morphology and syntax
- They are *cultural testimonies* in the sense that they describe the society and culture in which they are being compiled
- Therefore they constitute a central source for natural language processing in intelligent systems

Research focus

- Research and development within the field of *computational linguistics* and *lexicography* at the Centre for Language Technology at UCPH
- Main focus: to provide the HLT field with methodologies for reusing lexicographical resources and converting high quality lexicographical resources to formal lexica suitable for HLT
- Special focus on lexical semantics, sense inventories, sense clusters etc.
- Close collaboration with the Danish Society for Language and Literature

The special case of machine translation

- Machine translation used to be a good case for demonstrating the lexicon's role in natural language processing
- With rule-based MT the lexicon (mono- and multilingual) was indispensable
- With statistical and neural approaches the lexicon has become superfluous
- Will the same not happen with other AI systems when we get more data?

The special case of machine translation

- My claim is that MT is a very special case of natural language processing
- The human-translated parallel texts used for training constitute the *cultural clues*
- We don't have such *high-quality, culturally anchored clues* for other intelligent systems

Reusing and linking lexical resources

DSL

The Danish Dictionary

The Danish Thesaurus

Common sense id number

SemDax

Semantic corpus of Danish

Danish FrameNet

DAN NET

Danish wordnet

Reusing and linking lexical resources

DSL

The Danish Dictionary

The Danish Thesaurus

Common sense id number

SemDax

Semantic corpus of Danish

Danish FrameNet

DANNET

Danish wordnet

DSL + UCPH

Reusing and linking lexical resources

The Danish Dictionary

The Danish Thesaurus

Common sense id number

SemDax

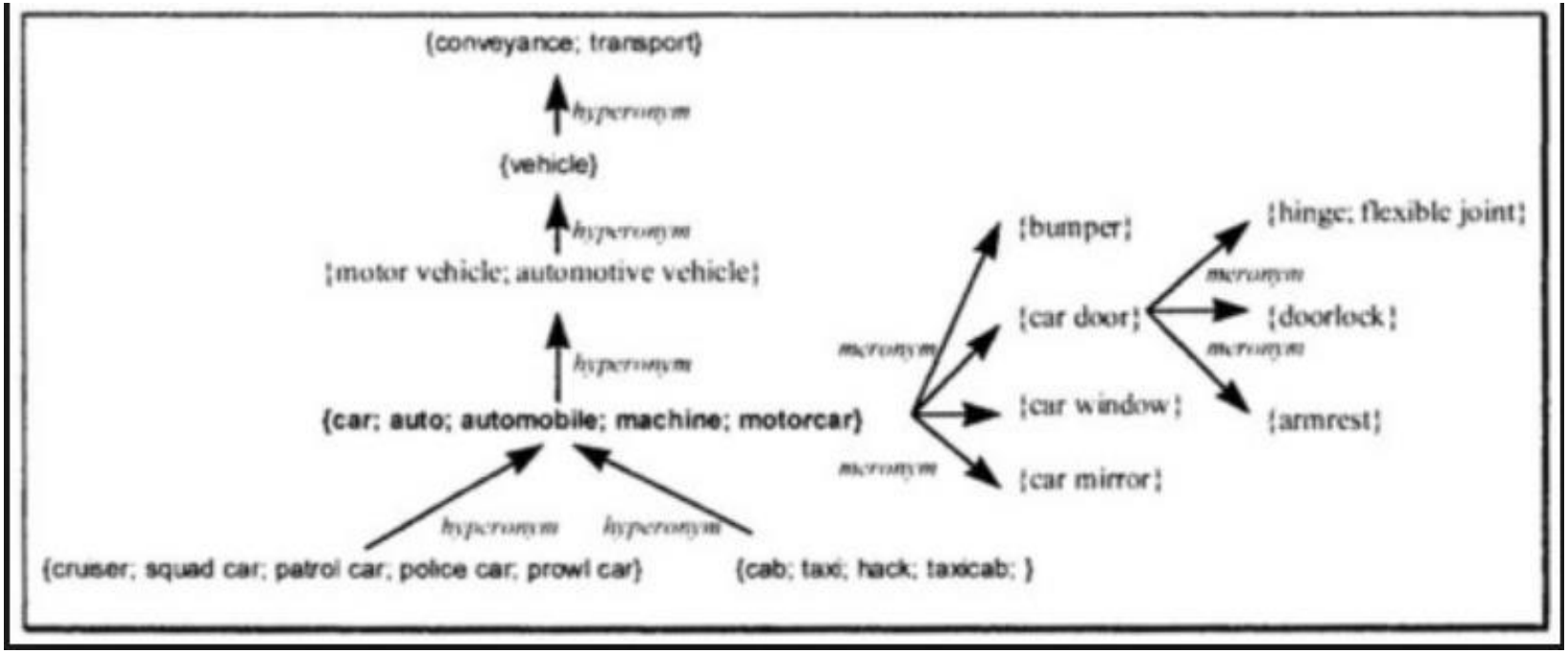
Semantic corpus of Danish

Danish FrameNet

DANNET

Danish wordnet

The wordnet: a formal concept hierarchy organised in terms synonymy clusters ("synsets") and of relations between synsets



From DDO to DanNet: Via genus proximum

Focus on table: **Selection for encoding (locked to user)**

	Expanded lemm	Pos	GenProx	Definition
1	bæltekøretøj_1	sb.	køretøj	køretøj der bevæger sig på bælter
2	cykel_1	sb.	køretøj	tohjulet køretøj som man driver frem ved at dreje to pedaler rundt
3	dræsine_2	sb.	køretøj	tohjulet, cykellignende køretøj som man drev frem ved at løbe me
4	ellert_1	sb.	køretøj	lille eldrevet køretøj
5	gaffeltruck_1	sb.	køretøj	kraftigt, som regel el- el. dieseldrevet køretøj som bruges til at løf
6	grusspreder_1	sb.	køretøj	særligt køretøj el. maskine der bruges til at sprede grus på stier og
7	kampvogn_1	sb.	køretøj	pansret militært køretøj på larvefødder, med en kanon monteret i
8	køretøj_1	sb.	transportmidddel	transportmiddel der bevæger sig på hjul på landjorden, evt. vha. m
9	løbehjul_1	sb.	køretøj	køretøj som består af et bræt med et lille hjul i hver ende og med
10	mandskabsvogn_	sb.	køretøj	køretøj til transport af mandskab
11	motorkøretøj_1	sb.	køretøj	køretøj med en motor som drivkraft, fx bil el. motorcykel
12	panserkøretøj_1	sb.	køretøj	køretøj med panser af hærdet stål
13	radiobil_1	sb.	køretøj	elkøretøj til indendørs bane som får strøm fra et trådnet i loftet via
14	redningsvogn_1	sb.	køretøj	køretøj med kran, spil e.l. til bjergning af forulykkede personer el.
15	salatfad_1	sb.	køretøj	stort, lukket køretøj som politiet bruger til transport af fanger el. m
16	sammenstød_1	sb.	køretøj	det at et køretøj, en genstand, en person el. andet støder voldsomt
17	skinnecykel_1	sb.	køretøj	cykellignende køretøj til kørsel på jernbaneskinner
18	slamsuger_1	sb.	køretøj	køretøj el. maskine med en tank, pumpe el. slange til opsugning af
19	slæde,1_1	sb.	køretøj	køretøj med meder som bruges til kørsel på sne el. is, som regel m
20	slæde,1_1_1	sb.	køretøj	mindre, fladt køretøj af plastic som bruges til at kælke med
21	sporvogn_1	sb.	køretøj	(eldrevet) køretøj til kørsel på skinner i en by, bestående af en mo
22	standsning_1	sb.	køretøj	det at et køretøjs bevægelse fremad standses
23	stridsvogn_1	sb.	køretøj	åbent, tohjulet, hestetrukket køretøj omgivet af et frontskjold som
24	sæbekassebil_1	sb.	køretøj	køretøj som et barn kan sidde i og lege med, fremstillet af forhånd
25	tag,1_1_1	sb.	køretøj	det der udgør den øverste flade på en bil, et tog el. andet køretøj
26	tender_3	sb.	køretøj	køretøj brugt af brandvæsenet til at transportere pumpe og slange
27	traktor_1	sb.	køretøj	køretøj med stor motorkraft, store baghjul og kraftige dæk, især b
28	trojka_1	sb.	køretøj	russisk køretøj trukket af et forspand på tre heste
29	vogn_1	sb.	køretøj	køretøj som består af en plade (lad), kasse e.l. forsynet med hjul,

Readjustment of inconsistent or underspecified hyponymies

Example: *fruits* and *vegetables*

Different definitions from The Danish Dictionary:

tomato is a fruit and a vegetable
aubergine is a vegetable
beetroot is a root vegetable
spinach is a plant
rhubarb is a stalk
artichoke is a flower bud

Readjustments..

Food taxonomy:

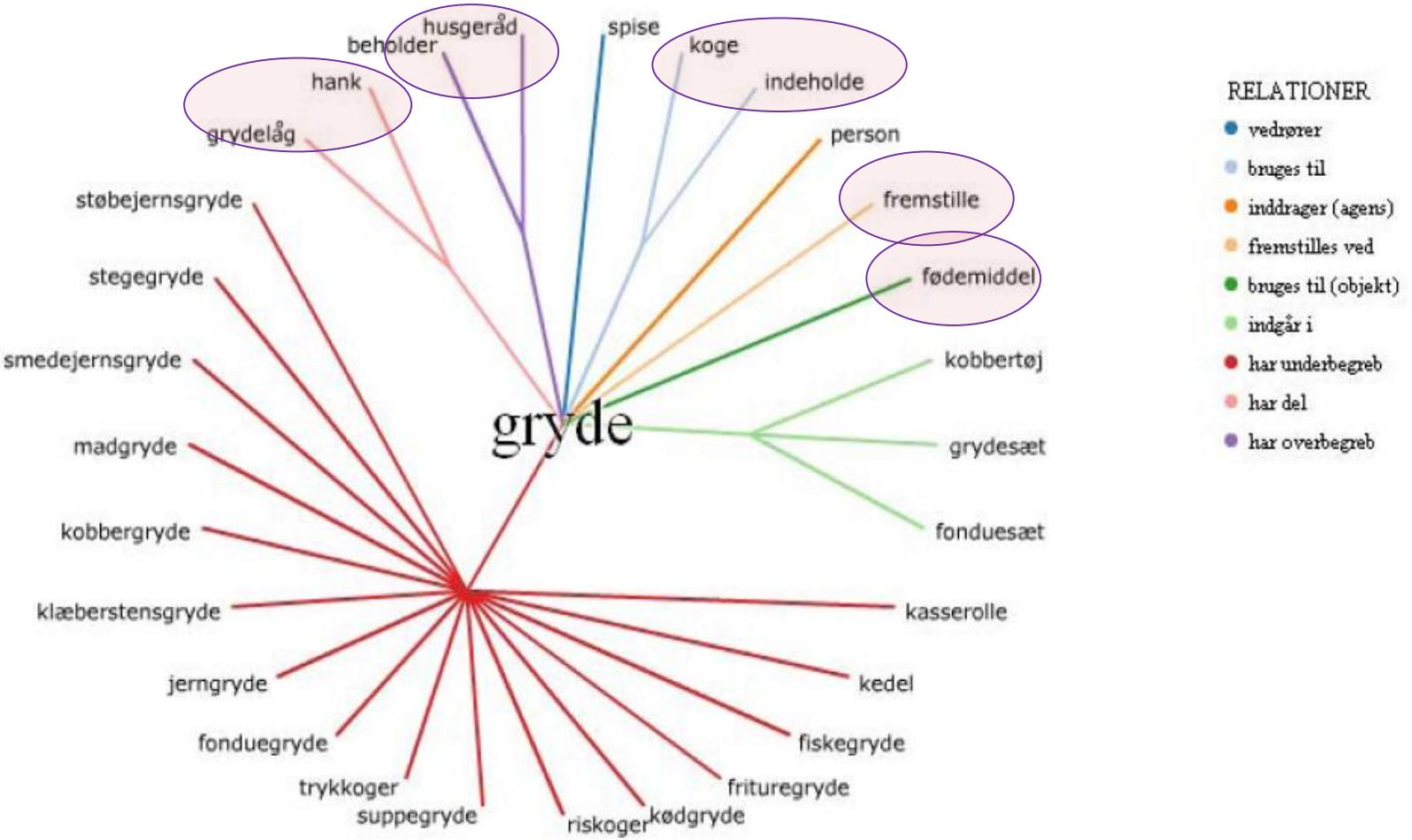
grøntsag (vegetable)
rodfrugt (root vegetable)
krydderurt (spice herb)
suppeurt (potherb)
..
fjerkræ (poultry)
flæsk (pork)
indmad (offals)

Natural taxonomy:

plante (plant)
skærmplante (umbelliferous plant)
rod (tuber)
stilk (stalk)
..
indvolde (entrails)

Wordnet relations from definitions

Definition of "pot": Container, usually with two handles and a lid used for cooking food



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

From thesaurus to FrameNet

The Danish Dictionary

The Danish Thesausrus

Common sense id number

SemDax

Semantic corpus of Danish

DANNET

Danish wordnet

Danish FrameNet

From thesaurus to FrameNet

{08_Vb_SbAfledning/has_hyperonym: ·vise sin vrede has_hyperonym: ·vredesudbrud involved_agent: ·person involved_patient: ·person}
▻·**skælde ud**, ·skrue bissen på, skænde, skælde, skælde (ud) for, ·tordne, ·tale dunder, tale med store bogstaver, skælde og smælde, ·udskælde, ·gennemhegle, ·give (med) grovfilen, give en gang lak, hegle igennem, ·sige et par borgerlige ord (til), give tørt på, skælde hæder og ære fra, skælde bælgen fuld, skælde huden fuld, rive hovedet af nogen, ·tage nogen i skole, ·slå i bordet, ·bruge mund, ·herse, overfuse, rise, dænge til◅; ▻·**få luft for sin vrede**, ·få afløb for sin vrede, ·bande nogen langt væk, ·rase ud◅; ▻·give luft for sin vrede, ·rase, ·fråde, ·se rødt, ·gå amok, ·springe/ryge i luften, ·koge over, ·eksplodere◅; ▻·**snerre**, ·bide ad, ·hvæse, ·spytte sætningen ud, ·sige vredt/bittert, ·råbe vredt◅; ·komme efter nogen, ·småskænde på, ·vreden løb af med ham, gl.·udøse sin vrede, ·skamme ud; ▻·komme med tilråb, ·fare i blækhuset, hvæsse/spidse pennen, hvæsse pennen◅; ·hvæsse kløerne, ·forløbe sig; ▻·**vredesudbrud**, raserianfald, udfald, ·vredesskrig, ·vris, ·snøft, ·gnaveri◅; ▻**udskældning**, ·skældud, udskæld, ·skænd, opsang, ·formaning, ·pegefinger, en sang fra de varme lande uform.◅; ▻·irettesættelse, røffel, gardinprædiken, moralprædiken neds.◅; ▻·overhaling, skideballe uform., sviner uform., ·et ordentligt pulver, møgfald, bredside, det glatte lag, balle, overfald, hak i tuden◅; ▻·hårde ord, knubbede ord, ·salut, ·svada neds., ·salve, dundertale, ·tordentale, ·afskedssalut◅; ·forløbelse; ▻·**skænderi**, ·større skænderi, ·ophidset diskussion, ·hidsig diskussion, ·skændsmål gl.◅; ▻ {syn} ·heftigt skænderi, ·kæmpeskænderi◅; ·familieskænderi

5316	skælde	Judgment_direct_address	ngn skælder (på ngn)
5317	skælde (ud) for	Judgment_direct_address	ngn skælder ngn (ud) for sb
5318	skælde nogen bælgen fuld	Judgment_direct_address	
5319	skælde nogen huden fuld	Judgment_direct_address	
5320	skælde nogen hæder og ære fra	Judgment_direct_address	ngn skælder ngn hæder og ære fra
5321	skælde og smælde	Judgment_communication	ngn skælder og smælder (over ngt/at../..)
5322	skælde ud	Judgment_direct_address	ngn skælder ud på ngn; ngn skælder ngn ud (for at../ngt); ngn s
5323	skældud	Judgment_communication	
5324	skæmte	Communication_manner	ngn skæmter (med ngn); ngn skæmter replik; ngn skæmter me
5325	skænd	Judgment_communication	
5326	skænde	Judgment_direct_address	ngn skænder (på ngn); ngn skænder replik

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

Danish FrameNet:

- 5300 verbs
- 6490 deverbal nouns

From FrameNet to a semantic frame classifier

A pilot communication corpus of 440 sentences annotated with frames from the FrameNet used for training a semantic classifier

Danish communication events

	Ours	*Random*
Statement	0.66	0.25
Opinion	0.69	0.15
Telling	0.52	0.11
Text_creation	0.86	0.09
Becoming_aware	0.43	0.07
Certainty	0.54	0.09

Supervised and unsupervised F1-scores on various frames.

English-Danish

	Ours	*Random*
Statement	0.31	0.20
Opinion	0.16	0.13
Telling	0.13	0.12
Text_creation	0.08	0.05
Becoming_aware	0.06	0.05
Certainty	0.08	0.05

SemDax – a corpus for semantic processing

The Danish Dictionary

The Danish Thesausrus

Common sense id number

SemDax

Semantic corpus of Danish

Danish FrameNet

DAN NET

Danish wordnet

Sense inventories in SemDax – searching for the right balance

Challenge: senses in traditional dictionaries are generally too finegrained for NLP

Aim:

- to develop principled methods for sense clustering which can make existing lexical resources practically useful in NLP
- not too fine-grained to be operational
- yet fine-grained enough to be worth the trouble.

Scalable sense inventories

Informativeness ↓ **Coarse-grained** → **Fine-grained**

Cross-linguality ↑ **Language independent** → **Language specific**

Supersense tagging

Reduced clusters of DDO/DanNet

Clusters of DDO/DanNet

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

Scalable sense inventories

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

SemDax - a corpus for semantic processing

A Danish human-annotated corpus annotated with sense inventories of *different granularity* based on our sense inventory (ids are kept!)

- 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.

Aim:

To assess the **reliability** of the different sense annotation schemes for Danish based on existing resources

SemDax - a corpus for semantic processing

New approach to semantic corpus annotation

- Not all disagreement is noise: contains valuable linguistic information that can improve annotation schemes and learning algorithms
- Double annotation of a larger part of corpus than usually seen
- The available corpus includes not only adjudicated files but also diverging annotations

SemDax - a corpus for semantic processing

SemDaX-Coarse

- All-words annotation (nouns, verbs, adjectives)
- Annotated with so-called supersenses derived from the list of WordNet's *lexicographical* files.
- Size: 90,000 words
- 60 % doubly annotated and adjucated

The annotation process

- Mapping of DanNet synsets to the 44 supersense classes (based on top level of Princeton Wordnet)

SemDax - a corpus for semantic processing

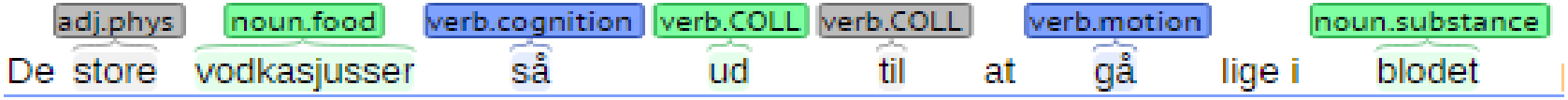
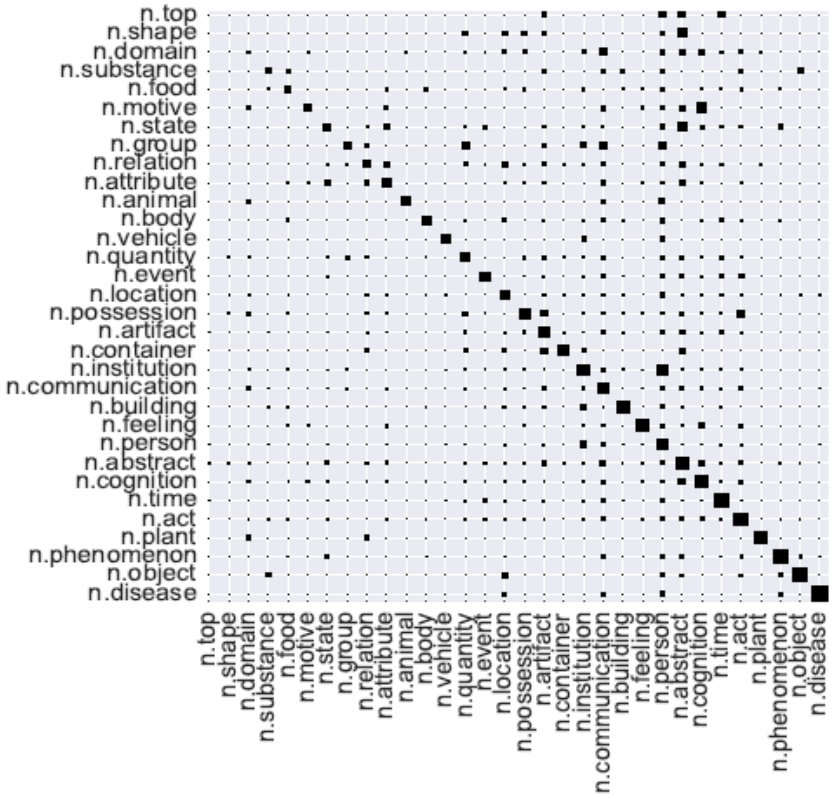


Fig. 1. Phrasal verbs with more than one particle (*se ud til* ('seem')) are annotated as collocations

with the sense label (here: verb.cognition) on the lexical kernel (*se*).

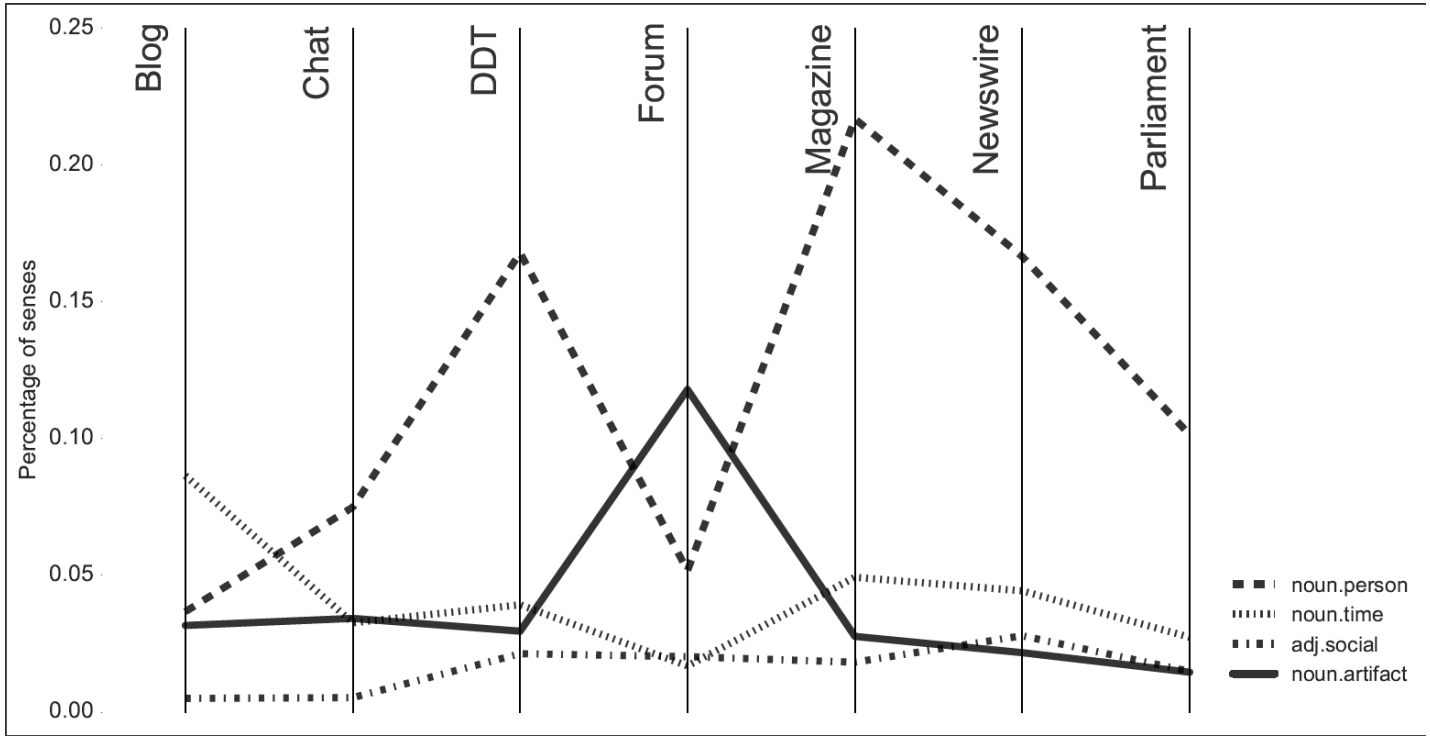
SemDax - a corpus for semantic processing

Evaluation: Where do annotators disagree?



SemDax - a corpus for semantic processing

Evaluation: How do text types differ?



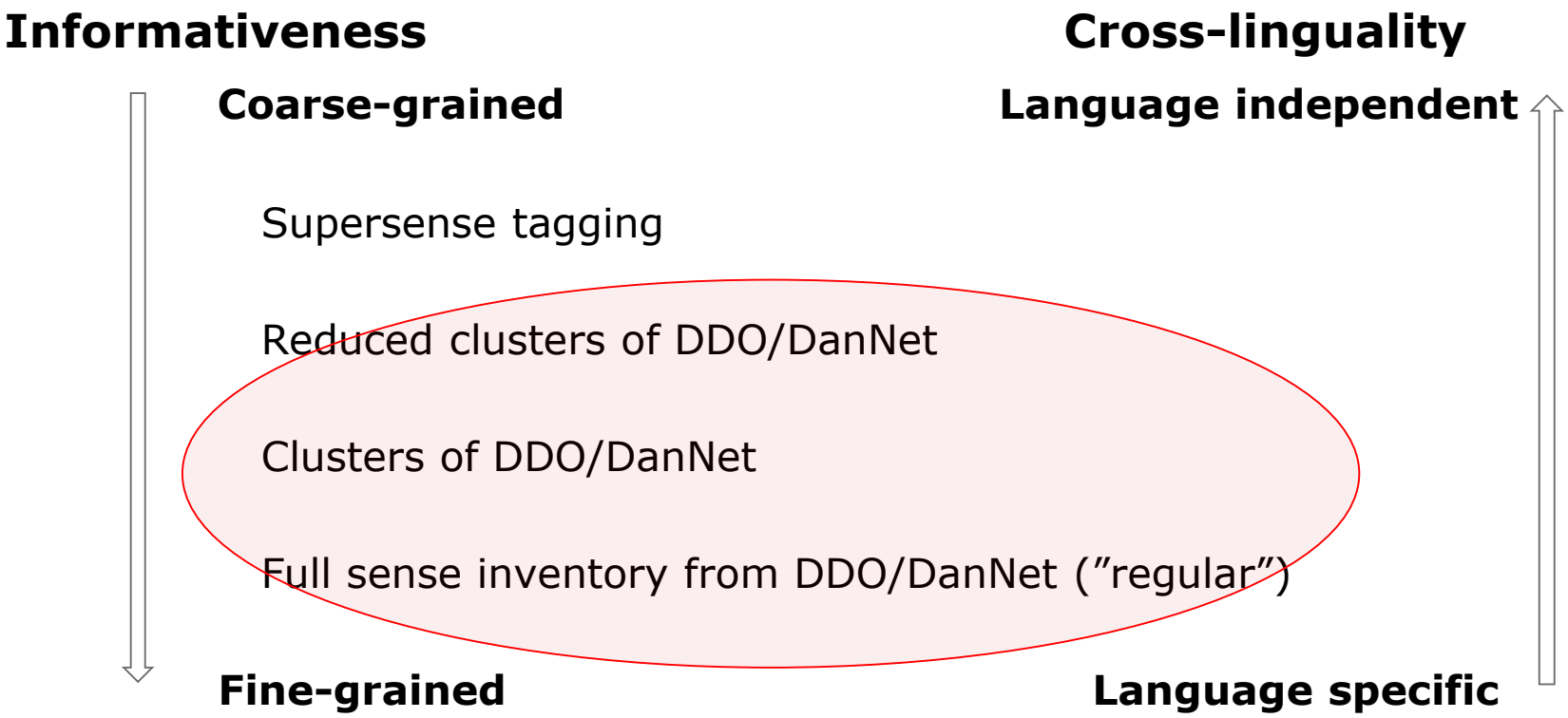
SemDax - a corpus for **semantic processing**

Development of a **sense tagger**:

- The corpus has been used for training and testing of a sense tagger that achieves an overall F1 score of 0.82 on heldout data, considering only the F1 of supersense labeling, micro-averaged score is ~0.65

Available at:
https://github.com/coastalcph/dsl_semtagger

Scalable sense inventories



SemDax - a corpus for semantic processing

SemDaX-LexicalSample

- Sense annotation of 20 highly ambiguous nouns (11 senses on average)
- Sense inventory derived from 1) The Danish Dictionary (DDO) and 2) DanNet
  combining main and subsenses from DDO and the top-ontological types from DanNet
- Clustering method: a reduction of senses of 23.5 % on average

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

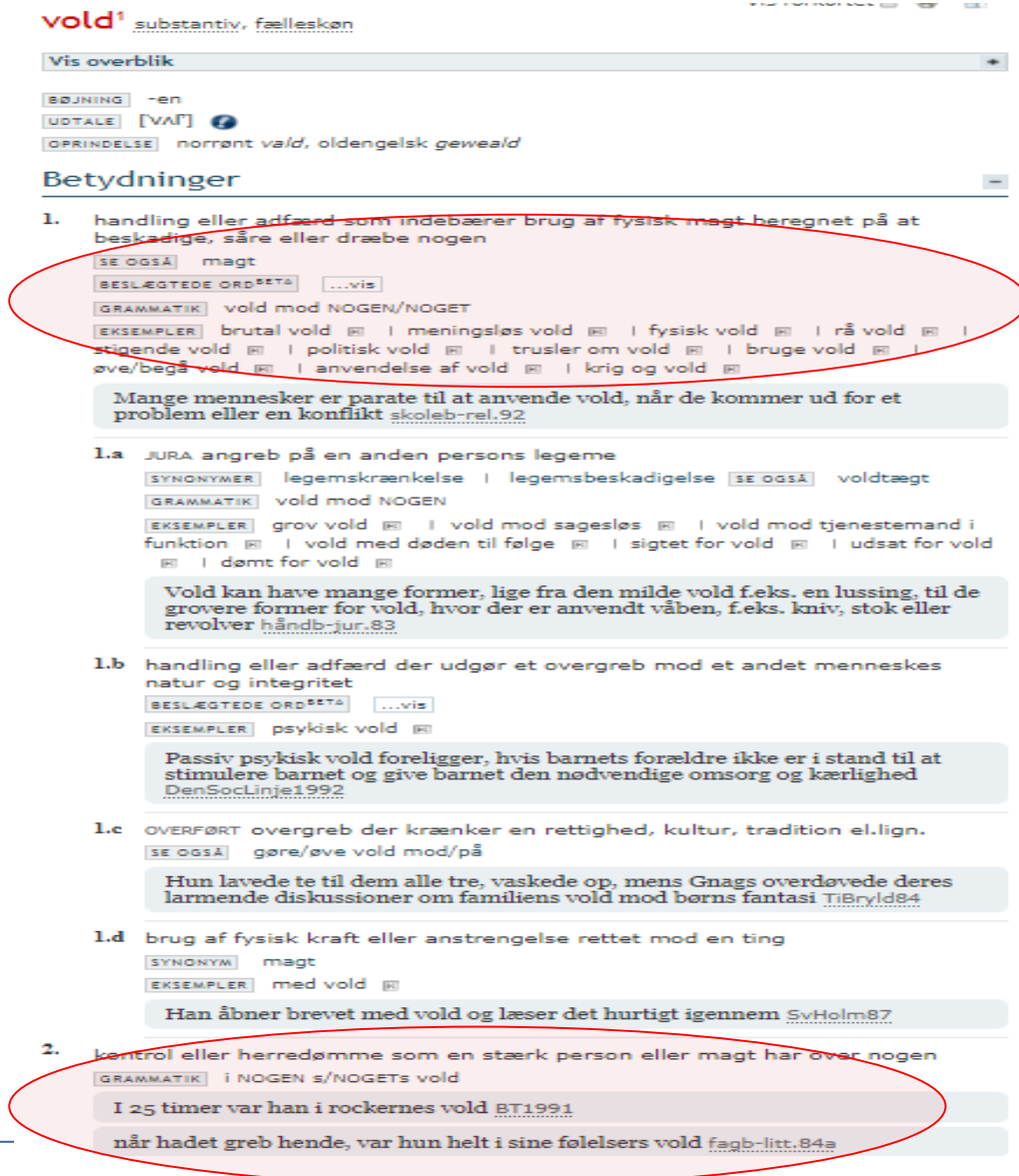
> Han åbner brevet med vold og læser det hurtigt igennem SvHolm87

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

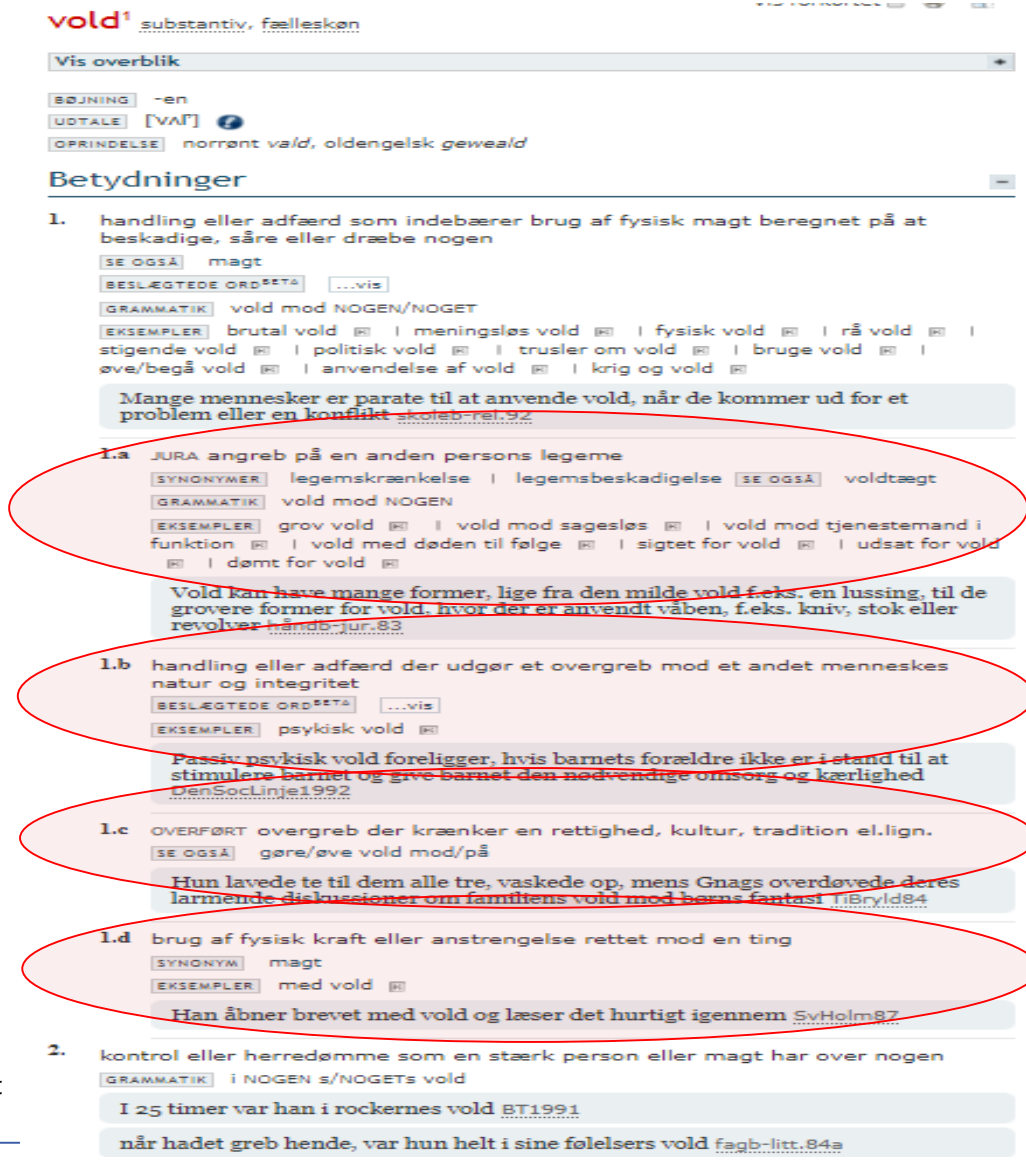
> Han åbner brevet med vold og læser det hurtigt igennem SvHolm87

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

Establishment of clusters

Exploiting semantic info from both sources

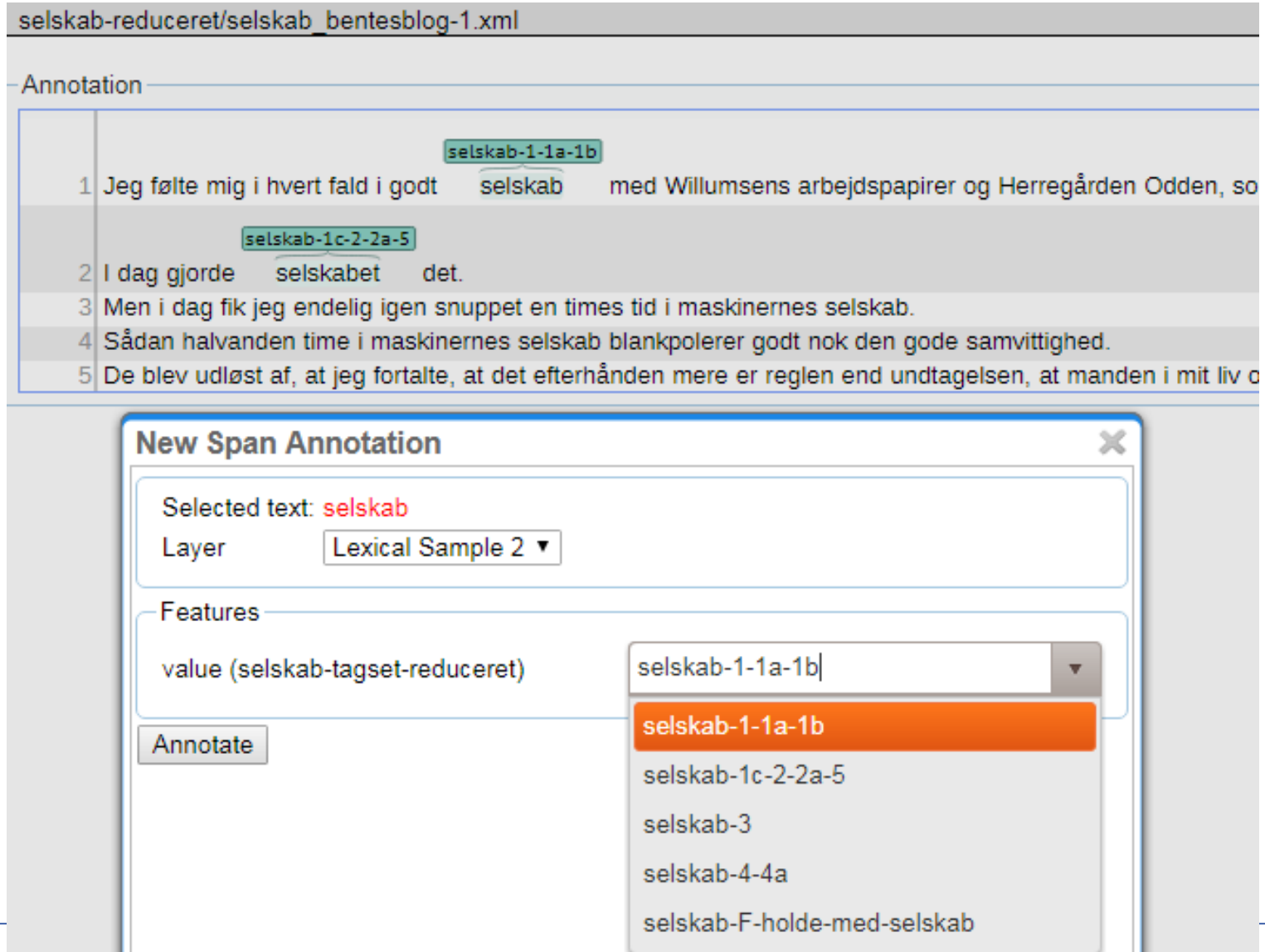
- **Experiment 1** ('regular') where all main and subsenses are maintained
- **Experiment 2** ('clustered') where subsenses are clustered if they are of the same ontological type
- **Experiment 3** ('clustered reduced') where also main senses are clustered if they are of the same ontological type.

Annotation of lexical samples

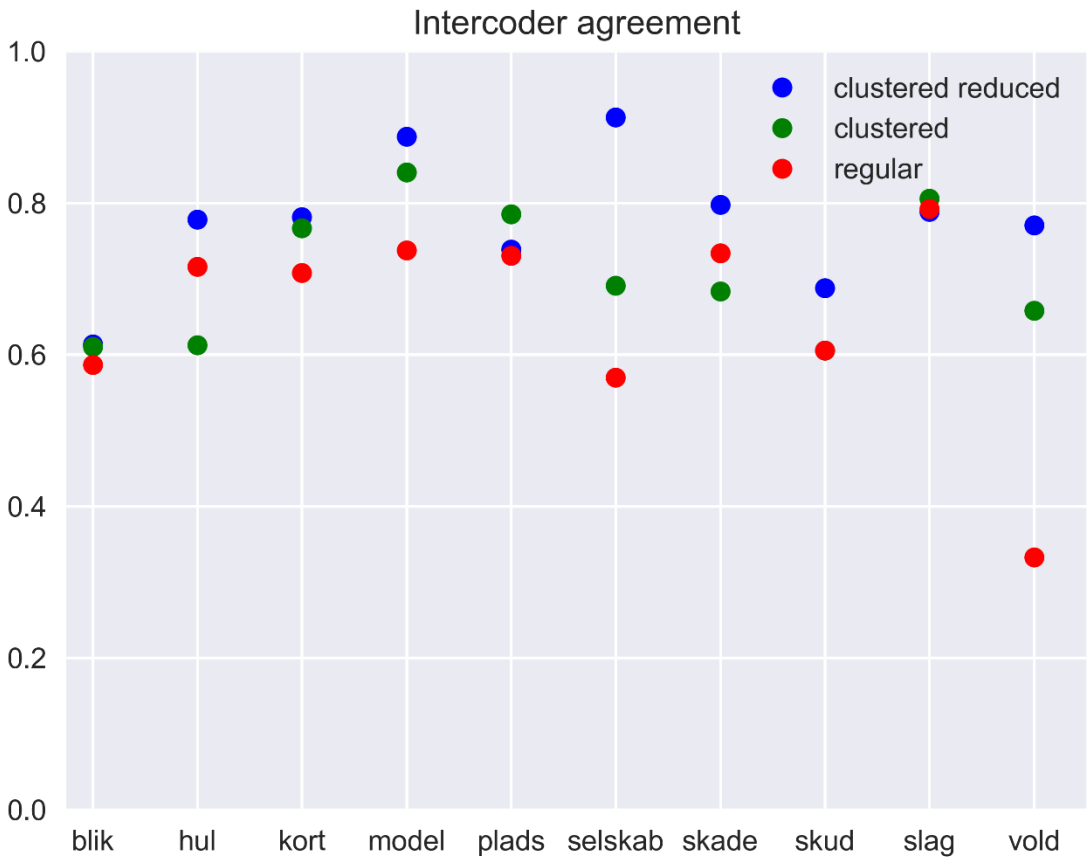
- The number of annotated sentences for each noun varies according to the number of DDO senses of the noun (100 + 15*no. of senses), resulting in from 175 to 600 sentences per noun.

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.

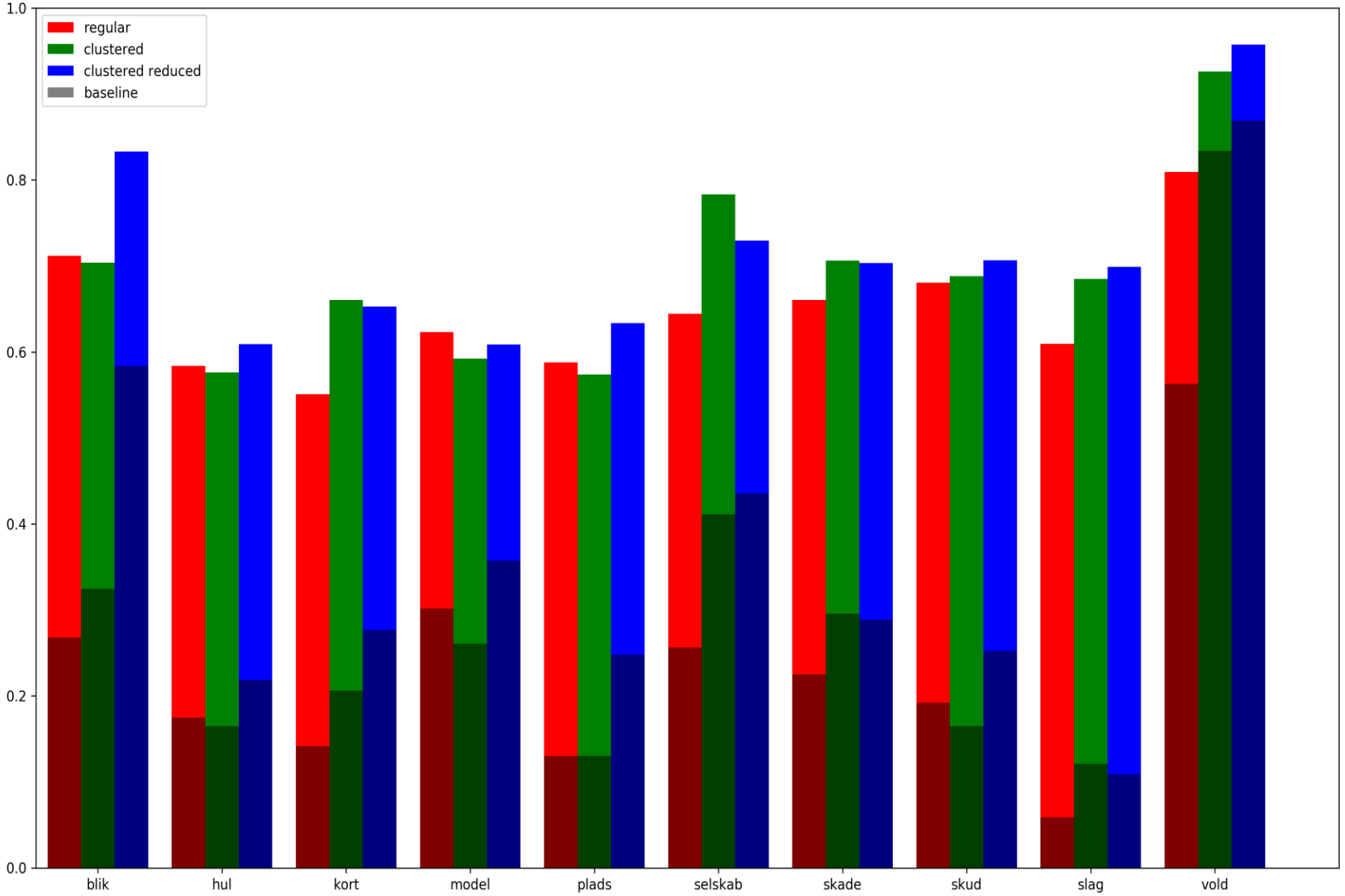
WSD using the LibLINEAR package

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

- Disambiguate the polysemous 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).


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

Word embeddings and word senses

- Ongoing studies to see how well word senses are resembled in Danish word embeddings
- Only recently reliable word embeddings are available for Danish
- Only recently evaluation data sets are ready for evaluating Danish word embeddings
- Interesting to examine to which extent reduced sense clusters are reflected in the distributional data

The ELEXIS project

H2020 Infraria project, started 2018

Institut "Jožef Stefan" Ljubljana, Slovenija

SAPIENZA Università di Roma

LEXICAL COMPUTING

ÖAW Österreichische Akademie der Wissenschaften

Institute of the Estonian Language

NUI Galway OÉ Gaillimh

Det Danske Sprog- og Litteraturselskab

Københavns Universitet

/instituut voor de Nederlandse taal/

Real Academia Española

KDictionaries

FCSH Faculdade de Ciências Sociais e Humanas Universidade Nova de Lisboa

Belgrade Center for Digital Humanities

Research Institute for Linguistics Hungarian Academy of Sciences

Kompetenzzentrum Trier Center For Digital Humanities (TCDH)

Institute for Computational Linguistics «Antonio Zampolli» National Research Council of Italy

Standards, data and tools in the ELEXIS project

Aim: Connecting the disconnected

- The lexicographic landscape is heterogeneous.
- There are stand-alone lexicographic resources, which are typically encoded in incompatible data formats due to the isolation of efforts
- This prohibits reuse of the data in natural language processing, linked open data and the Semantic Web, or in digital humanities.

Standards, data and tools in the ELEXIS project

Aim: Connecting the disconnected

- ELEXIS will introduce common standards,
- develop conversion tools and, most importantly,
- will interconnect the existing resources so that they can be used to develop new modern data which can be used in ways that new digital technologies need.

Summing up

- Employ existing high-quality lexical data in natural language processing and intelligent systems (AI)
- Comply with international standards to ensure linking and reuse potential
- Incorporate elements of language transfer from better resourced languages where relevant

Thank you!

Selected links:

DanNet: http://wordnet.dk/lang.html
SemDax:https://github.com/coastalcph/semdax
WordTies: http://wordties.cst.dk/
STO: http://cst.ku.dk/sto_ordbase/
Sense tagger:
https://github.com/coastalcph/dsl_semtagger

Selected references on our work

Pedersen, B. S., Aguirrezabal Zabaleta, M., Nimb, S., Olsen, S., & Rørmann, I. (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In *Proceedings of Global WordNet Conference 2018* Singapore: Global WordNet Association.

Krek, S, Kosem, I, McCrae, J, Navigli, R, Pedersen, BS, Tiberius, C & Wissik, T(2018), 'European Lexicographic Infrastructure (ELEXIS): Proceedings of the XVIII EURALEX International Congress' Paper presented at, Ljubljana, Slovenia, 16/07/2018 - 21/07/2018, pp. 881-892.

Pedersen, BS, Nimb, S, Søgaard, A, Hartmann, M & Olsen, S (2018), A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier. in *Proceedings of the 11th edition of the Language Resources and Evaluation Conference, Miyazaki, Japan.*

Nimb, S, Braasch, A, Olsen, S, Pedersen, BS & Søgaard, A 2017, From Thesaurus to Framenet. in I Kosem, C Tiberius, M Jabobicek, J Kallas, S Krej & V Baisa (eds), *Electronic Lexicography in the 21st Century : Proceedings of eLex 2017 conference.*Lexical Computing CZ, pp. 1-22, eLex 2017 - electronic Lexicography in the 21st Century , Leiden, Netherlands, 18/09/2017.

Pedersen, B.S., A.Braasch, A. Johannsen, H. Martínez Alonso, S. Nimb, S. Olsen, A. Søgaard, N. H. Sørensen (2016) The SemDaX corpus – sense annotations with scalable sense inventories. In *2016 LREC Proceedings*, Portorož, Slovenia.

Pedersen, B.S., S.Nimb, S.Olsen, A.Søgaard, N.Sørensen (2014) Semantic Annotation of the Danish CLARIN Reference Corpus. *Proceedings from isa-10, 10th Joint ACL - ISO Workshop on Interoperable Semantic Annotation* p. 25-29, Reykjavik, Iceland.

Pedersen, B.S. (2013). Coding semantic properties of words in computational dictionaries. In: Gouws, Heid, Schweickard, Wiegand (Eds.): Dictionaries: *An International Encyclopedia of Lexicography Supplementary volume: Recent Developments with Focus on Electronic and Computational Lexicography*. Berlin: Walter de Gruyter

Nimb, S. B.S. Pedersen, A.Braasch, N. H. Sørensen and T.Troelsgård (2013). Enriching a wordnet from a thesaurus. *Workshop Proceedings on Lexical Semantic Resources for NLP from the 19th Nordic Conference on Computational Linguistics*.(NODALIDA). Linköping Electronic Conference Proceedings; Volume 85 (ISSN 1650-3740)

Pedersen, B.S., L. Borin, M. Forsberg, K. Lindén, H. Orav, E. Rögnvalssson (2012) Linking and Validating Nordic and Baltic Wordnets- A Multilingual Action in META-NORD. In: *Proceedings of 6th International Global Wordnet Conference* pp.254-260. Matsue, Japan.

Pedersen, B.S, J. Wedekind, S. Kirchmeier-Andersen, S. Nimb, J.E. Rasmussen, L.B. Larsen, S. Bøhm-Andersen, H.Erdman Thomsen, P. J. Henrichsen,J. O. Kjærum, P. Revsbech, S.Hoffensetz-Andresen, B. Maegaard (2012). *The Danish Language in the Digital Age - Det danske sprog i den digitale tidsalder*. META-NET White Paper Series, Springer Verlag.

Pedersen, B.S, S. Nimb, J. Asmussen, N. Sørensen, L. Trap-Jensen, H. Lorentzen (2009). DanNet – the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation, Computational Linguistics Series, pp.269-299.* http://link.springer.com/article/10.1007%2Fs10579-009-9092-1