THE SOCIETY FOR DANISH LANGUAGE AND LITERATURE

CENTRE FOR LANGUAGE TECHNOLOGY
DEPARTMENT OF NORDIC RESEARCH
UNIVERSITY OF COPENHAGEN

A Thesaurus-Based Sentiment Lexicon for Danish – The Danish Sentiment Lexicon

Sanni Nimb[1], Sussi Olsen[2] , Bolette S. Pedersen[2], Thomas Troelsgård[1]

[1]The Society for Danish Language and Literature & [2]University of Copenhagen

sn@dsl.dk, saolsen@hum.ku.dk, bspedersen@hum.ku.dk, tt@dsl.dk

The Danish Sentiment Lexicon characteristics

- 13,859 Danish polarity lemmas
- -3 to +3
- Includes morphological information
- Freely available at https://github.com/dsldk/danish-sentiment-lexicon (licence CC-BY-SA 4.0 International)

Main idea

- Easy acces to comprehensive vocabulary in thesaurus → much higher lexical coverage than existing Danish sentiment lexicons
- Identify positive and negative thesaurus section titles as the starting point
- Calibrate the degree of polarity within sets of words denoting very similar concepts

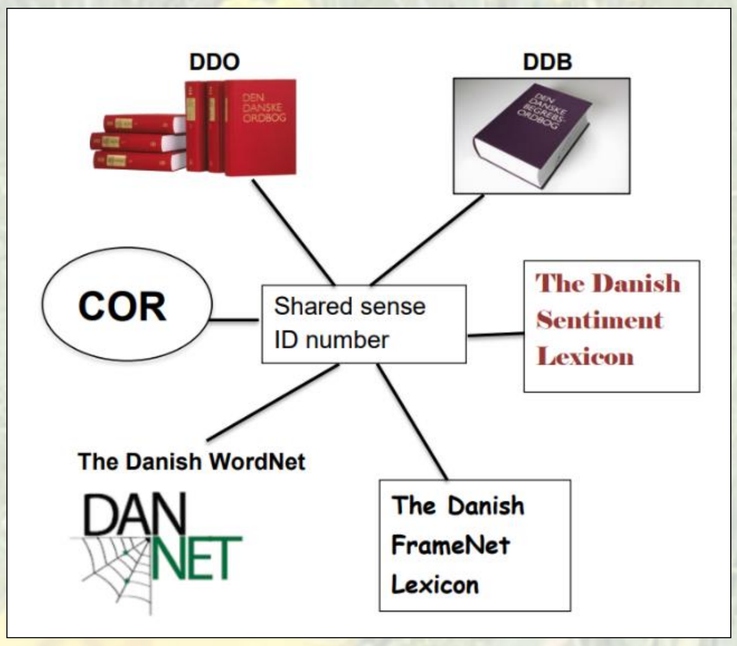
Linked data

- Based on 17,883 polarity annotated senses linked to other Danish lexical resources
- differentiates it from other sentiment lexicons for Danish
- Allows for future experiments where sentiment is combined with other types of information from lexicons and corpora

Linked Data

- Four computational lexicons linked to a monolingual dictionary DDO and to the thesaurus DDB via the sense ID numbers of DDO
- The WordNet linked to Princeton WordNet

‘



Creation of the dataset

•24% of the 888 sections judged by two lexicographers to contain polarity words based on the section name. Clear predominance of negative sections:

- 57% negative (e.g. sections *Unimportant*; *Sadness*)
- 37% positive (e.g sections *Important; Admire; Friendship*)
- 6% both negative/positive (e.g. sections Reputation; *Protest*, *Rebellion*).
- → in total 25,000 senses (not including MWU).

•Senses from identified sections extracted from DDO and as default assigned the section polarity value

•Combined with information from other resources

- DDO (definition, frequency, usage (e.g. *derogatory*; *offensive*)
- Values from AFINN (Danish, ~3000 words, -5 to +5, see Nielsen, 2018)
- Keyword and synonyms/near synonyms in thesaurus sections

7	9	19	Opgive, give afkald på	verb	3	abandonnere	28114378	abandonnere	opgive; give slip påopgive n		-2		0	give afkald på, give k
7	9	7	Tvivlrådig, forvirret	adj	3	åben	21098037	åben	(endnu) ikke afklaret, afgjo	0	-2	2	0	uvis, ukendt, uklar, u
7	9	20	Handle, agere	subst	3	åbning	21098095	åbning	begyndelse; indledningbegy	0	2		0	handling, forehavend
7	9	3	Være nødt til	adj	3	absolut	21000113	absolut	som er ubetinget og uafhæn p		-2	2	0	nødvendig, uundgåel
7	9	6	Beslutsom	subst	3	absoluthed	21000116	absoluthed	viljen til at gennemføre nog p		2		1	beslutsomhed, deterr
7	9	6	Beslutsom	subst	3	absolutist	21000119	absolutist	person som går ind for absc p		2		0	en handlingens manc
7	9	41	Vigtig	subst	3	accent	21000162	accent	(ofte)hovedvægt; eftertryk;	0	2		0	prioritering, vægtnin
7	9	41	Vigtig	verb	3	accentuere	21000167	accentuere	fremhæve; understregefren p		2		0	fremhæve, pointere,

Inter-annotator agreement on sample

•Two lexicographers annotated 400 words also present in AFINN.

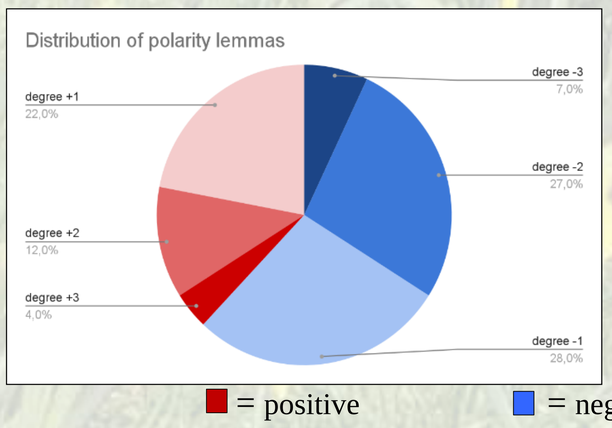
•Negative and positive values, but not degree → Cohen’s Kappa: 0.83

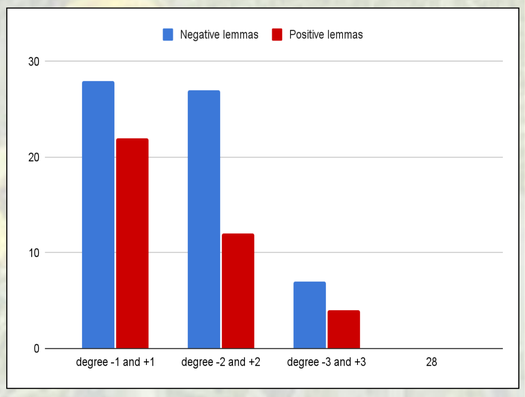
Annotation of the dataset

	Task	The judgement is based on
Step 1	sense polarity: 0, positive or negative?	• AFINN • Comparison of synonyms and near synonyms in the DDB thesaurus • Information in DDO
Step 2	degree of sense polarity: –3, -2, -1, +1, +2, +3?	• AFINN • Comparison of synonyms and near synonyms in the DDB thesaurus • Information in DDO
Step 3	harmonizing values of identical lemma senses in dataset	• Lemma / sense represented more than once in the dataset? (due to multiple representations in thesaurus)
Step 4	deciding upon polarity at lemma level	• Conflicting polarities of lemma senses? • Rare sense to be ignored? • Or lemma to be left out?
Step 5	validation of data	• 1/3 of annotated data • Comparison of all lemmas with same high degree (e.g. all +3 lemmas)

Results and conclusions

- Clear predominance of negative lemmas
  - 62% negative polarity
  - 38% positive polarity
- ~ same distribution as in Swedish SenSALDO (6,386 lemmas, also thesaurus-based, see Rouces et al., 2018a and b
- Devitt & Ahmad (2013): seems general for sentiment lexicons; in texts the relationship is thought to be reversed.





= positive = negative

→ When sentiment analysis methods are based on large lexicons (like ours) the variation of lemmas that are recognized is higher – and these are more likely to be negative than positive → overweight of negative polarity

→ When methods are based on smaller lexicons, it is more likely that the neglected lemmas are negative than positive → underrepresentation of negative polarity

- We manually annotated a literary piece of text containing many polarity conveying lemmas (Pedersen et al. 2021)
  → only ad hoc composita not covered → we should include morphemes with high polarity+ module for automatic splitting of unknown composita
- There might be undiscovered polarity lemmas in thesaurus/DDO
  - Hidden in neutral thesaurus section (e.g. ‘Man’, ‘Woman’).Should be added.
- We might consider including neutral lemmas (like SenSALDO)
- Plan: transfer of polarity information at sense level (not part of the open release)
  - to DDO → new presentations, e.g. of synonyms
  - to DanNet
  - to new formal semantic lexicon COR-S (Pedersen et al. LREC 2022)

The Danish Sentiment Lexicon has been created in the project “DanNet2” 2019-2022, granted by the Carlsberg Foundation