A Thesaurus-Based Sentiment Lexicon for Danish – The Danish Sentiment Lexicon
Sanni Nimb¹, Susi Olsen², Bolette S. Pedersen², Thomas Troelsgård¹
¹The Society for Danish Language and Literature & ²University 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/dslk/danish-sentiment-lexicon (licence CC-BY-SA 4.0 International)

Main idea
• Easy access 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

Annotation of the dataset

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

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

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 Röcker et al., 2018a and b)
• Devitt & Ahmad (2013): seems general for sentiment lexicons; in texts the relationship is thought to be reversed.

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

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

Distribution of polarity lemmas

→ 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 sections (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)