DDS: The Danish Sentiment Lexicon

DDS ("Det Danske Sentimentleksikon" - 'The Danish Sentiment Lexicon') is a comprehensive sentiment lexicon for Danish based on two dictionaries, "Den Danske Begrebsordbog" (DDB, The Danish Thesaurus) and "Den Danske Ordbog" (DDO, The Danish Dictionary), and compiled by use of lexicographic methods.

The dataset was compiled by
Sanni Nimb, Sussi Olsen, Thomas Troelsgård
version 0.1
2021-04-14

If you use this dataset or derivatives of it, please refer to the publishers: Det Danske Sprog- og Litteraturselskab (DSL, Society for Danish Language and Literature) and Center for Sprogteknologi, Københavns Universitet (CST, Centre for Language Technology, University of Copenhagen)

The DDS dataset presents a negative or positive polarity value of 13.859 Danish headwords from DDO. It is available in two formats, one format where the headword is the entry and where the word forms (full forms, text forms) of it are presented as a list (list 2), and one format where the word forms are entries (list 2a):

Format, filename **2_headword_headword_polarity**:
<headword>tab<homograph number>tab<part of speech>tab<DDO headword ID>tab<polarity label headword>tab<list of word forms>

Example,: the noun *konflikt* ('conflict'), the verb *konflikte* ('to make a conflict'), the noun *glæde* ('happiness') and the verb *glæde* ('make happy')

headword	homograph number	POS	DDO headword ID	polarity label headword	list of word forms
konflikt		sb.	11027274	-2	konflikt;konflikten;konfliktens;konflikter;konflikterne;konflikternes;konflikters;konflikts
konflikte		vb.	11027275	-2	konflikt;konflikte;konfliktede;konfliktende;konflikter;konfliktes;konfliktet
glæde	1	sb.	11018246	3	glæde;glæden;glædens;glæder;glæderne;glædernes;glæders;glædes
glæde	2	vb.	11018247	3	glæd;glæde;glædede;glædende;glæder;glædes;glædet

Format, filename **2a_fullform_headword_polarity**:
<word form>tab<headword>tab<homograph number>tab<part of speech>tab<DDO headword ID>tab<polarity label headword>

Example,: the noun *konflikt* (‘conflict’) and the verb *konflikte* (‘to make a conflict’)

word form	headword	homograph number	POS	DDO headword ID	polarity label headword
konflikt	konflikt		sb.	11027274	-2
konflikt	konflikte		vb.	11027275	-2
konflikte	konflikte		vb.	11027275	-2
konfliktede	konflikte		vb.	11027275	-2
konflikten	konflikt		sb.	11027274	-2
konfliktende	konflikte		vb.	11027275	-2
konfliktens	konflikt		sb.	11027274	-2
konflikter	konflikt		sb.	11027274	-2
konflikter	konflikte		vb.	11027275	-2
konflikterne	konflikt		sb.	11027274	-2
konflikternes	konflikt		sb.	11027274	-2
konflikters	konflikt		sb.	11027274	-2
konfliktes	konflikte		vb.	11027275	-2
konfliktet	konflikte		vb.	11027275	-2

<polarity label headword> = "-3" (highest degree negative) |"-2" (high degree negative) | "-1" (negative) | "1" (positive)| "2" (high degree positive)| "3" (highest degree positive)

<homograph number> = NONE (no homographs) | "1" | "2" | "3" | "4" | "5"

<part of speech>= | "adj." | "sb." | "sb. pl." | "vb." |“adv."|”udråbsord”|”sidsteled”|”egennavn”|”fork.”|”førsteled”|”konj.”|”lydord”|”pron.”|”præfiks”

<list of forms> = conjugated forms of the headword, separated by “;”, example:konflikt;konflikten;konfliktens;konflikter;konflikterne;konflikternes;konflikters;konflikts

Statistics

Part of speech: 7.803 sb., 3.884 adj., 1.906 vb., 96 udråbsord, 78 adv., 57 sb.pl., 1 sidsteled *(-narkoman*), 3 egennavn (*Pærekøbing, Udkantsdanmark, Waterloo*), 3 fork. (*dir., dr., m.v.h.*), 14 førsteled, 6 lydord, 2 pron. (*nada, nul*)), 1 konj. (*jamen*))

13.859 headwords assigned either the polarity value 1,2, 3,-1,-2, or -3.

22 % of the headwords have the value "1" (positive)
12 % of the headwords have the value "2" (high degree positive)
4 % of the headwords have the value "3" (highest degree positive)
28 % of the headwords have the value "-1" (negative)
27 % of the headwords have the value "-2" (high degree negative)
7 % of the headwords have the value "-3" (highest degree negative)

62 % of the headwords have a negative polarity (-1, -2, or -3)
38 % of the headwords have a positive polarity (1, 2, or 3)

50 % of the headwords, approx. 6,900 headwords, have a high or very high polarity degree (either 2, 3, -2 or -3).

Method

The DDS dataset assigns polarity value to 13.859 headwords. It was compiled in the following way:

Step 1: Extraction of lexical data from DDO and DDB: Approx. ¼ of the 888 sections in the thesaurus DDB were selected as either negative or positive sections based on the section title. E.g. the section with the title "Vrede" ('Anger') was estimated to contain negative words, and the section with the title "Medfølelse" ('compassion') was estimated to contain positive words, Also thesaurus sections which were estimated to contain either negative or positive words, or both, were included. Words in DDB are linked at sense level to the DDO dictionary, and via the links all the DDO-senses (from single headwords, not MWU's) represented in the selected DDB sections were extracted for annotation. Furthermore the lexical data to be annotated was supplemented with not already included DDO senses labeled 'derogatory'. The total inout data to be annotated consisted of 19,000 unique senses.

Step 2: Annotation: The polarity values -1 (negative), 1 (positive) or 0 (neutral polarity) were assigned manually to all the extracted DDO senses (listed in groups as they appear in the sections in the thesaurus DDB, some of them appearing in more than one sections. The senses were annotated by only one annotator (two persons half of the data each). In order to ensure interannotator agreement, the senses of 400 headwords which were represented in the sentiment lexicon AFINN (Nielsen, F. Å. (2018). Danish resources https://bit.ly/2NDHcbW (Section 4.8 Sentiment analysis)), were initially double annotated. The interannotator agreement (AFINN included) was 97 %.

Step 3: Upgrading of polarity values: The assigned values -1 and 1 were afterwards upgraded manually to -2 or -3, respectively 2 or 3 by one annotator based on the study of near-synonym groups in DDB where at least one of the words had a stronger value than 1 or -1 in AFINN. Validation: Approx. 25 % of the upgraded values were validated by the second annotator.

Step 4: Removal of irrelevant or disturbing senses: A) Headwords of which all annotated senses had the value 0, were removed from the dataset. B) Rare, historic, or technical/domain specific DDO senses with the value 0 were removed from the dataset, based on information in DDO. All other senses with the value 0 were kept on the list.

Step 5: Direct inclusion of a part of the annotated data: The result of the annotation process was a dataset consisting of 17.833 senses (of 14.444 headwords) with polarity values. The dataset did not include all DDO senses of the headwords, however the missing DDO-senses were likely to be neutral senses since they were not part of any of the polarity DDB sections, nor labelled 'derogatory' in DDO. Headwords of which all the senses annotated in the dataset had the same value (negative, neutral and/or positive as well a degree of negativity or positivity), were directly included in the final dataset.

Step 6: Extra annotation of diverging data: Opposite to this, headwords of which the annotated senses diverged w.r.t. polarity (negative, neutral and/or positive as well a degree of negativity or positivity) were studied and annotated a second time in order to assign a polarity value at headword level. The assignment of polarity value at headword level was based on studies of the sense descriptions in DDO, on corpus studies of the words, but also on subjective judgements taking into consideration the neutral DDO-senses which were not part of the annotated data. In this process, A) headwords with a common neutral sense and a rare polarity sense were removed from the dataset. B) Headwords with two opposite polarity senses which were both estimated to be frequent were removed from the dataset (e.g. *frelst, sej, skarp, overlegen, glat*). And 3) homographs having diverging polarity values were removed from the dataset if they were both frequent words in Danish. A total of 595 headwords were in this way removed from the sense annotated dataset due to diverging polarity values of their senses or homographs. The remaining headwords were included in the final dataset.

The result was 13.859 headwords assigned either the polarity value 1,2,3,-1,-2, or -3.

Remarks

The first version of the DDS dataset does not contain all the DDO headwords that have at least one polarity sense, but most of them. Negative words which are not part of the extracted sections in DDB (estimated to contain polarity words), neither labelled 'derogatory' in DDO, might still have to be added.

There is a many-to-many relationship between headwords and text word forms.This means that a text word form may have more than one sentiment label attached to it if a certain word form is connected to more than one headword. For example, the text word form /fred/ will correspond to the noun "fred..1" ('peace; quiteness etc.') as well as the verb "frede..1" ('to protect etc.'), with polarity labels "2" and "1", respectively.

The dataset includes fullforms of the headwords (corresponding to the text word forms) on the basis of the conjugational information on the headwords in DDO. It also includes rare (or non-existing) full forms such as the imperative form /a/of the verb /ae/, the superlative form /ufredeligst/ of the adjective /ufredelig/ and the form /mavesurts/ singular, genitive case, neuter of the adjective /mavesur/