The Danish FrameNet Lexicon: Method and Lexical Coverage

Sanni Nimb
The Society for Danish Language and Literature
sn@dsl.dk

Abstract

This paper presents and discusses the results of compiling a comprehensive Danish frame lexicon compliant with the Berkeley FrameNet standard by making use of linked lexical data from two Danish resources, namely the semantic and thematic grouping of verbs and verbal nouns in a Danish thesaurus with the valency patterns of the same verbs in a monolingual Danish dictionary. The frame lexicon covers a large number of Danish lemmas, including phrasal verbs and multiword units, and furthermore gives information on one or more phrases illustrating the typical textual context in which the lemma evokes the frame in question. The overall aim is to supply annotators of semantic frames and roles in Danish texts in future research projects with a restricted and thereby manageable set of possible frames to choose from. We present the content of the lexicon in detail, including a comparison with the frame coverage of Berkeley FrameNet.

Keywords: thesaurus, frame lexicon, Danish

1. Lexical resources as input

In order to compile a Danish frame lexicon compliant with the international standard resource Berkeley FrameNet (Ruppenhofer et al., 2016, henceforth BFN) we combine the valency information on verbs in a comprehensive monolingual dictionary with the semantic and thematic grouping of the same verbs and related verbal nouns in a thesaurus. The dictionary we use (*Den Danske Ordbog*, henceforth the DDO dictionary[1]) contains approx. 100,000 lemmas and 136,000 senses. The thesaurus (*Den Danske Begrebsordbog* ('The Danish Concept Dictionary', Nimb et al., 2014 a & b, henceforth the thesaurus) is based on and linked to the lemma senses in the DDO dictionary and covers 80 % of the senses. The links between the two resources allow us to combine all sorts of lexical information and use it for different purposes, in this case semantic relatedness on the one hand and syntactic information on the other hand used as input to the manual assignment of frame information, see figure 1. To a high degree, the valency patterns in the DDO dictionary reflect the semantic role inventory as described in BFN, and thereby help us select and assign the most appropriate frame.

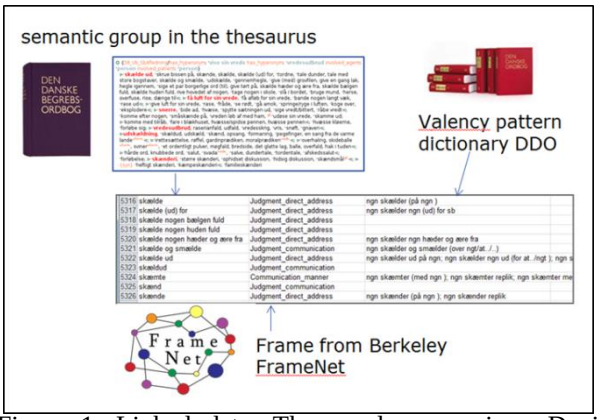

Figure 1: Linked data: The word groups in a Danish thesaurus combined with the valency information in a Danish dictionary constitute the background for the framenet

The semantically related verbs and verbal nouns in the thesaurus are typically assigned one of a rather restricted set of BFN frames, making it possible instantaneously to compile large amounts of lemmas and expressions within the same semantic area. Due to the close relation between English and Danish we assume that the frame descriptions and the role inventory from BFN can be transferred directly and used in future Danish annotation tasks. This was confirmed when testing part of the frame lexicon in a pilot project: we did not encounter any problematic cases of role assignment (Nimb et al., 2017; Pedersen et al., 2018). The frame lexicon project was carried out at DSL in collaboration with the University of Copenhagen and funded by the Carlsberg Foundation 2016-2017.

2. The method

We use the source xml-structured document of the thesaurus which arranges the Danish vocabulary in 22 chapters and 888 named sections[2] and furthermore in an average of 9-10 annotated semantic groups in each section where words and expressions are grouped according to semantics, not word classes (Nimb et al., 2014 b), making it very useful for our purpose since verbs and their corresponding verbal nouns are grouped together. Since all semantic groups are formally annotated with coarse-grained semantic information (i.e. 'property', 'act', 'event', 'person' etc.), we can easily identify and thereby focus on acts and events exclusively in order to create a core frame lexicon describing the part of the Danish vocabulary that occurs with semantic roles. If we take a closer look at the different semantic groups of a section, the section with the title 'Crying', for example, contains one group with words meaning 'person who cries', formally annotated with the type 'person'. Another group represents the verbs and verbal nouns with the meaning 'to cry', and is annotated with 'act/involved agent' information, while a third group lists adjectives describing persons who (easily) cry. In total there are approx. 8,300 semantic groups in the thesaurus. 1/5 of these groups (1487 semantic groups), containing more than 42,000 words and expressions, were identified as being of the type 'act' or 'event' via the coarse-grained formal semantic annotations of each group.

[1] DDO was compiled as a printed dictionary in the 1990s. Today the dictionary is online and regularly extended with new words and expressions.

[2] The section and chapter division is inspired by a German thesaurus (Dornseiff, 2004), but adjusted to the Danish language community of today.

The vocabulary of the groups includes not only single lemmas but also many of the collocations, for example support verb constructions, which are described in the DDO dictionary based on corpus statistics. In the case of the verbal noun *skrig* ('scream'), sense 1 in DDO (see figure 2), we find the collocational expressions *give et skrig fra sig* ('give/utter a cry') and *udstøde et skrig* ('utter a cry/cry out/shriek'), and they are both included in the thesaurus data and thereby also assigned their corresponding frames from BFN when compiling the frame lexicon.

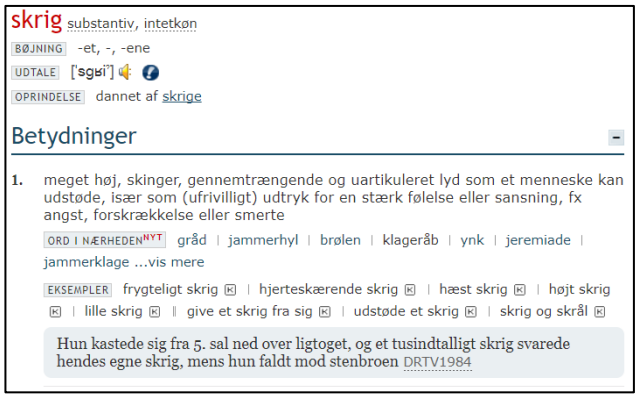

Figure 2: The entry of the verbal noun *skrig* ('scream') in the DDO dictionary with several collocations (listed after EKSEMPLER ('examples')). Apart from the lemma itself, two of these are included in the same semantic group in the thesaurus: *give et skrig fra sig* (lit. 'give a scream from oneself') and *udstøde et skrig* ('give a scream'), both meaning 'to scream'.

Many words and expressions in the DDO dictionary are part of more than one section in the thesaurus and therefore also listed more than once in the extracted data material used as input to the frame assignment. E.g. the verb *guide* ('to guide') is part of 4 different sections in 3 different chapters in the thesaurus and has therefore been assigned four frame values in the frame lexicon: Assistance, Cotheme, Leadership and Telling. Likewise the verb *cruise* ('to cruise/move easily') which occurs in seven different sections (in three different chapters) has been assigned five different frames, namely Motion, Self_motion, Operate_vehicle, Finish_competion (in the sense 'to win easily in sports') and Personal_relationship (in the sense 'to search for a partner'). The fact that the collocations in the DDO dictionary are statistically corpus-based and that representation in more than one thesaurus section often reflects different aspects of the same sense, similar to what a selection of corpus examples would do, allows us to consider the extracted data as a sort of 'condensed' corpus data in the form of small representative bits of phrases we would typically find in Danish texts if we were going to annotate a set of corpus examples with frames from BFN.

The thesaurus data and the corresponding valency patterns from the DDO dictionary were identified and extracted into a spreadsheet (carried out by Thomas Troelsgård, DSL). In figure 3 we present a small extract of the combined data, including the frames that we manually assigned after having translated them into English by use of a Danish English dictionary and afterwards having looked up the lexical_unit equivalents and their corresponding frames in BFN.

From the thesaurus : word/expression with the meaning 'to cry/to scream'	Shared sense id number	From the DDO dictionary: valency pattern 'somebody cries (+ manner) (because of something)'	Assigned frame from BFN
klage sig	21034458	ngn klager (sig) over ngt	Make_noise
jamre	21074699	ngn jamrer (sig) (over ngt)	Make_noise
jamre over	21074699	ngn jamrer (sig) (over ngt)	Judgment_communication
jamre sig over	21090433	ngn jamrer (sig) (over ngt)	Judgment_communication
græde	21074701	ngn græder (+ måde) (af noget)	Make_noise
skrige	21074700	ngn skriger (+ måde) (af noget)	Make_noise
skrige af smerte	21074700	ngn skriger (+ måde) (af noget)	Make_noise
give et skrig fra sig	21074701	NONE (*skrig* = verbal noun)	Make_noise
udstøde et skrig	21010806	NONE (*skrig* = verbal noun	Make_noise
sætte i et hyl	21033375	NONE (*hyl* = verbal noun)	Make_noise

Figure 3: The spreadsheet with the extracted data, in this case words and expressions with the meaning 'to cry' (some are verbs, others support verb constructions), linked to their corresponding valency patterns from the DDO dictionary via shared id numbers. The right colon presents the assigned frames from BFN.

In an initial pilot project (Nimb et al. 2017) we focused on the vocabulary from only two semantic areas, namely communication and cognition. The sections and semantic groups covering these two areas were easy to identify in the thesaurus due to the chapter names, and constitute approx. 16 % of all act and event groups in the thesaurus. We assigned a total of 104 different frames and tested the data in an annotation task where the supersenses of the verbs (verb.communication or verb.cognition) were already manually identified. The overall conclusion was that the compilation method was very efficient. By focusing on one semantic area at a time, which was made possible via the section and chapter grouping in the thesaurus, the lexical data considered was likely to be assigned the same frame, or at least a closely related frame, from BFN. When annotating with the frames, the decision-making was largely facilitated by the restricted number of possible frames for each verb in the text. But we also concluded that some of the most frequent verbs were lacking important frame values due to the fact that not all senses of highly polysemous verbs were represented in the thesaurus. Nimb et al. (2017) describes the pilot project in detail. In this paper, however, we focus on the lexical coverage of the entire Danish frame lexicon.

3. Lexical coverage and the distribution of frames

The lexicon consists of 23,260 unique words or expressions. Of these, 12,142 are single lemmas (e.g. the verb *chartre* ('to chart') and the noun *chartring* ('charting')), while 11,118 are expressions from the DDO dictionary consisting of two or more words. Most of these are fixed verbal expressions, including phrasal verbs, but we also find lexical collocations, mostly verbs with a typical object (*nippe til maden* ('to pick at the food'), *mene det modsatte* ('to mean the opposite')) or nouns with a typical support verb (*fatte en beslutning* ('to make a decision')). We also find verb phrases with one or more obligatory arguments represented by pronouns (e.g. *tale noget igennem* ('to talk something through'). The single lemmas and the fixed expressions correspond to Lexical Units in BFN (e.g. 'give up' and 'nip in the bud' are fixed expressions in BFN). We estimate that the frame lexicon covers approx. 20,000 Lexical Units, but it should be mentioned that the borderline between fixed expressions and collocations is elusive.

There are 33,930 unique combinations of word/multiword expression and frame value. A lemma or expression might be included two times or more in the lexicon, depending on how often it is represented in the different thesaurus sections. Due to this, there are 42,270 combinations of word/expression + frame value + thesaurus group number. The group numbers represent a different and often more fine-grained semantic relatedness of the data than the frame divisions do and are e.g. useful in the case of negative/positive words within the same frame group. I.e. they make it possible to divide words with the frame Remembering_experience into two groups, those meaning 'to forget' and those meaning 'to remember'.

Lemma in DDO dict.	Sense in DDO dict.	Sense also SynSet member in the Danish WordNet DanNet	Unique lemmas + expressions with frame value	Lemmas + expressions with unique combination of frame value and thesaurus group number
12,142	21,812	6,877	33,930	42,270

Table 1: Statistics on data in the frame lexicon

	Lemmas from DDO dict.	Senses from DDO dict.	Also in the Danish WordNet DanNet	Frame values
nouns	6,490	8,372	2,063	11,032
verbs	5,300	12,354	4,750	17,731

Table 2: Statistics on nouns and verbs in the frame lexicon

The words and expressions which are represented stem from 20,820 different senses from 12,124 lemmas. Some of the covered senses are also linked to synsets in the Danish WordNet DanNet (Pedersen et al., 2009), namely 38 % of the 12,354 verb senses and 25 % of the 8,372 noun senses. The 5,300 different verb lemmas have altogether been assigned 17,731 frame values. This means that 80 % of the verb lemmas in the DDO dictionary are represented in the frame lexicon with an average of 3.3 frames per verb. If we look at nouns, a total of 6,490 are represented in the lexicon and assigned a total of 11,032 frames (1.7 frames per noun). In tables 1 and 2 we present some statistics, and in figure 4 we list a small part of the frame lexicon entries, exemplified with a selection of Danish verbs originating in the English language.

	Lemma		Frame
v.	*chartre*	'to rent a plane or boat'	Renting
n.	*chartring*	'renting a plane or boat'	Renting
v.	*chatte*	'to chat via internet'	Communication_means
v.	*chippe*	'to move a ball'	Cause_motion Sports_jargon
v.	*coache*	'to guide wrt personal career'	Education_teaching
v.	*crashe*	'to have an accident by car'/ 'to hit' / 'to participate in a party without being invited'	Catastrophe Impact Drop_in_on

Figure 4: A small extract from the Danish frame lexicon sorted by lemma. The frames are assigned to both verbs and verbal nouns. Valency patterns are not included in the release of the lexicon, but example phrases and group numbers from the thesaurus are (not illustrated here).

Many, but not all the nouns are both single lemmas in the lexicon (with one or more frame assignments) as well as part of a verbal phrase, typically combined with a support verb (with one or more frame assignments). Also a number of adjectives and adverbs are represented in the lexicon but in this case always as part of a verbal phrase. In both cases the verb in the verbal phrase is identified in a specific data field.

3.1 The frame inventory used for Danish

671 different frame values from BFN (~2/3 of all frame values) have been assigned to the Danish vocabulary. We have not yet compared the two sets of frames but plan to do so in order to identify English frames which have not been applied. We expect to find cases where such frames might have been better choices. Due to our method the lexicographer became rather confident with the different frame possibilities in BFN, and this guarantees at least to a certain degree that the Danish frame assignment is homogeneous across the lexicon.

The most frequent frame in the Danish lexicon is Self_motion (2% of the data). Subsequently, we find Experiencer_focused_emotion, Statement, Stimulate_emotion, Judgment_communication, and Cogitation (all between 1 and 2% of the data). An additional 36 frames are assigned to between 0,5 and 1 % of the data, covering areas such as sports (Sports_jargon), acts in general (Removing, Filling, Processing_materials, Bungling, Intentionally_act), eating and drinking (Ingestion), and communication (Text_creation, Request, Respond_to_proposal). The remaining frames (~ 630) used to describe the Danish verbs and verbal nouns are only applied on less than 0.5 percent of the vocabulary in

the thesaurus, respectively. Of these, almost 200 are used only 10 or less times, and approx. 50 are used only once.

At a first glance into the statistics of the applied frames for Danish, the most frequently used ones describe the semantic areas of motion, emotion, act, communication and cognition. In supersense-annotated Danish texts (Martínez et al. 2015) act, communication and cognition are also among the most frequent, while motion and emotion are less frequent. While the far most frequent verb sense in texts is 'stative', we do not find a large variety of lemmas or frames with this sense in neither the thesaurus, nor the frame lexicon. Similarly, there are only a few lemmas and frames with the sense 'possession' in lexicons compared to the high frequency of the sense in corpora.

3.2 Semantic areas covered by the Danish thesaurus but not (yet) by Berkeley FrameNet

Due to the fact that the Danish thesaurus represents more or less the entire vocabulary of a comprehensive corpus-based Danish dictionary which covers all general semantic areas in Danish, it is interesting to compare its coverage with BFN and study the cases where it was difficult to find corresponding frames to assign to the Danish words. In figure 5 we list the cases where we found it hard to find appropriate frames for one or maybe more verbs with a given sense in Danish, either because English conceptualization seems to differ from Danish, or because BFN does not cover the sense yet. It should be mentioned that we still need to study the cases in more detail and validate the data in order to find out whether we have simply misinterpreted the coverage of already existing frames in BFN. We exclude cases in which BFN states that frames are planned to be created (e.g. acts in sports and many scientific domains).

Areas and concepts covered by the Danish thesaurus, but not (yet) by Berkeley FrameNet
Not a human act
a calm situation (note: opposite to the frame Chaos); to go well, to be solved (note: about situation/problem); a machine carrying out a function; biological reproduction (note: both animals and plants); plants growing; animals living and acting
'General' human acts
to have a habit/to carry out a habit; to delimit something; to exaggerate when carrying out an activity, to overdo something; to hurry when carrying out an activity; to repeat an activity
Cogitation
to change your opinion; to mentally accept/adapt to something
Cleaning/polluting/recycling
to make something clean (note: the frame Removing is too broad in its sense, we find); to ventilate/clean out the air; to make something dirty, to pollute; to throw out something (note: as garbage); to protect nature (note: we have used Protecting but find it too broad); to recycle something/reuse
Creation
knitting, sewing etc.; concrete repairing (note: in both cases we find Processing_ materials too broad)
Social acts
to force somebody to do something (note: without using violence), to defend somebody (note: by speaking, not physically); to mediate/act as a mediator; to celebrate something; to meet somebody by coincidence; to stay in a place without staying overnight; to feed/give food to other persons; to take care of children/to babysit; the act of flirting with somebody
Body activities
to do sports, run, ride, surf (note: without competing, focus instead on pleasure/health purposes); to play games, to play for fun; gambling; to bathe for fun, e.g. in the sea; the act of masturbating; to go to bed (note: to get up is covered); not to eat/to be on a diet; to do nothing, to relax
Domain-specific acts
to plant trees, flowers, foresting (note: the frame Agriculture is too narrow, we find); sterilization of animals; to dig, to make holes; to parcel out/subdivide a piece of land; economics : raise money on; mortgage; laws : defend in the court
Supernatural acts/events
to practice witchcraft, to conjure; to haunt a place; to tell fortunes

Figure 5: Cases where we found it hard to find appropriate frames in BFN

4. Conclusions and future work

The freely available lexicon which can be downloaded at https://github.com/dsldk/dansk-frame-net contains data on the lemma, its word class and its frame value, a typical phrase or collocation, and the group number from the thesaurus. In cases of noun lemmas with verbal phrase examples, the verb is furthermore identified. The valency patterns from the DDO dictionary are not part of the data.

While the number of different verb lemmas is very high in the thesaurus and thereby also in the frame lexicon, verb polysemy as it is represented in the DDO dictionary is less extensively covered. We therefore plan to supply especially the highly polysemous verbs - which are also the ones occurring very often in texts - with more frames, and in this case base the compilation on the DDO dictionary's sense descriptions. So far only the frames regarding cognition and communication have been used for annotation. Our hope is to use the frame lexicon in future annotation projects. We furthermore plan to integrate the frame data in the Danish WordNet (which is also linked to the senses of the DDO dictionary) and use the frame values to improve the hierarchies of verbs in the WordNet.

5. Bibliographical References

The Danish Dictionary (DDO dictionary) (2008-): online dictionary, ordnet.dk/ddo, Det Danske Sprog- og Litteraturselskab, Copenhagen, Denmark.

DanNet: andreord.dk; wordnet.dk.

Danish English Dictionary: ordbog.gyldendal.dk, Gyldendal, Copenhagen.

Dornseiff, Franz (2004) *Der deutsche Wortschatz nach Sachgruppen*, W. De Gruyter, Berlin; New York.

Martínez Alonso, H., Johannsen, A., Olsen S., Nimb S., Sørensen N., Braasch, A., Søgaard, A. & Pedersen, B. S. (2015). Supersense tagging for Danish. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015. Vol. 109*, Linköping University Electronic Press, NEALT Proceedings Series, Vol. 23.

Nimb, S., Lorentzen, H., Theilgaard, L., Troelsgård, T. (2014 a). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab, Copenhagen, Denmark.

Nimb, S., Lorentzen H., Trap-Jensen, L. (2014 b). The Danish Thesaurus: Problems and Perspectives. In: Abel A., Vettori C. & Ralli N. (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen 2014: EURAC Research, pp. 191-199.

Nimb, S., A .Braasch, S.Olsen, B. S. Pedersen, A. Søgaard (2017). From thesaurus to framenet. In: *Proceedings of eLex 2017*, Leiden.

Pedersen, Bolette Sandford; Nimb, Sanni; Asmussen, Jørg; Sørensen, Nicolai; Trap-Jensen, Lars; Lorentzen, Henrik (2009) DanNet - the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation, Vol. 43*, p. 269-299.

Pedersen, B.S., Nimb, S., Søøgaard, 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 LREC 2018*.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice* (Revised November 1, 2016.) https://framenet.icsi.berkeley.edu/fndrupal/the_book.