A Danish FrameNet Lexicon and an Annotated Corpus Used for Training and Evaluating a Semantic Frame Classifier

Bolette S. Pedersen[1], Sanni Nimb[2], Anders Søgaard[3], Mareike Hartmann[4], Sussi Olsen[5]
University of Copenhagen [1,3,4,5] & The Danish Society for Language and Literature[2]
Njalsgade 136, DK-2300 Copenhagen S[1,5,] Christians Brygge 1, DK-1219 Copenhagen K[2], Universitetsparken 1, DK-2200 Copenhagen,[3,4]
bspedersen@hum.ku.dk, sn@dsl.dk, soegaard@di.ku.dk, hartmann@di.ku.dk, saolsen@hum.ku.dk

Abstract

In this paper, we present an approach to efficiently compile a Danish FrameNet based on the Danish Thesaurus, focusing in particular on cognition and communication frames. The Danish FrameNet uses the frame and role inventory of the English FrameNet. We present the corresponding corpus annotations of frames and roles and show how our corpus can be used for training and evaluating a semantic frame classifier for cognition and communication frames. We also present results of cross-language transfer of a model trained on the English FrameNet. Our approach is significantly faster than building a lexicon from scratch, and we show that it is feasible to annotate Danish with frames developed for English, and finally, that frame annotations – even if limited in size at the current stage – are useful for automatic frame classification.

Keywords: Danish FrameNet, concept dictionary, frame-annotated corpus, low-resourced languages, semantic frame classifier

1. Danish as an under-resourced language

The META-NET white papers, which discussed the most urgent risks and chances of the European languages in the digital age, illustrated that several of our languages are severely under-resourced for the ongoing and coming digital revolution; Danish being no exception (cf. Pedersen et al. 2012).

To this end, several players in the Danish language and language technology community have in recent years focused on methods for building language technology resources and tools that employ *both* existing Danish lexical data *and* language transfer from better resourced

Figure 1: A wordnet (DanNet), a framenet and a semantically annotated corpus (SemDax) expanded from two dictionaries via common sense ids

languages (cf. Pedersen et al. 2009, Nimb et al. 2017, Johannsen et al. 2015, Levy et al. 2017). In order to enable this combination of methods, quite a lot of effort has been put into relating the resources to international standards (see for instance Martinez et al. 2016).

Most recently, effort has been put into compiling a Danish Berkeley style (Ruppenhofer et al. 2016) frame lexicon (BFN) by extracting semantic data from The Danish Thesaurus (DT) and The Danish Dictionary (DDO), (cf. Nimb et al. 2017). The FrameNet is one of several LT resources being built from a common sense id inventory first established with The Danish Dictionary and further employed in The Danish Thesaurus. Figure 1 illustrates the complex of interrelated resources, including also a Danish WordNet, DanNet, and a semantically annotated corpus, SemDaX.

In this paper we focus on the evaluation of our method of using linked data to compile new lexical resources to be used for semantic annotation and processing. At the current state, the lexicon contains 5,300 verbs (80 % of the verb lemmas in DDO) and 6,490 verbal nouns represented in 33,930 different expressions[1]. These are given either in the form of just the lemma itself, or in terms of a collocation from DDO, or an infinitive phrase with grammatical elements expressed as pronouns (based on the DDO valency patterns), or a multiword expression from DDO, all assigned a frame value from Berkeley FrameNet. The words and expressions represent 1/7 of the senses in DDO. Each verb lemma has an average of 3.3 frames. The noun lemmas have half the number of frames per lemma, namely 1.7. In total, 671 different frame values from Berkeley FrameNet have been applied, and among the most used ones are the ones describing acts from the semantic areas of motion, emotion, communication and cognition.

The paper is organized as follows. Below we sketch out how we, in order to test the strength of the lexical working method, started by compiling a pilot frame lexicon based on only two selected semantic domains in existing lexica (Section 2). Section 3 describes how we used the resulting set of frames to annotate selected

[1] The Danish frame lexicon is now freely available at https://github.com/dsldk/dansk-frame-net. It will be presented in more detail at The International FrameNet Workshop 2018, Multilingual FrameNets and Constructions at LREC 2018 (Nimb, submitted for review).

corpus examples within the same two domains, and finally we present in Section 4 how we used the frame and role annotated corpus data for training and evaluating a semantic frame classifier.

2. From a Thesaurus to a FrameNet Lexicon

The Danish Frame Lexicon is being built by exploiting the thematic divisions of DT and the fact that each subdivision includes groups of semantically closely related words (see Figure 2). In the pilot project we focus on groups of verbs, including idiomatic multiword units (typically phrasal verbs) and in the case of act groups, also deverbal nouns which in a semantic content correspond quite well to the ontological groupings of acts and events in FrameNet.

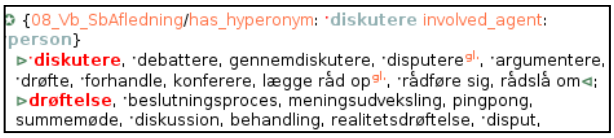

Figure 2: Synonyms and near synonyms of the verb *diskutere* ('to discuss') in DT, including verbal nouns and annotated with type 08 (acts) and semantic relations.

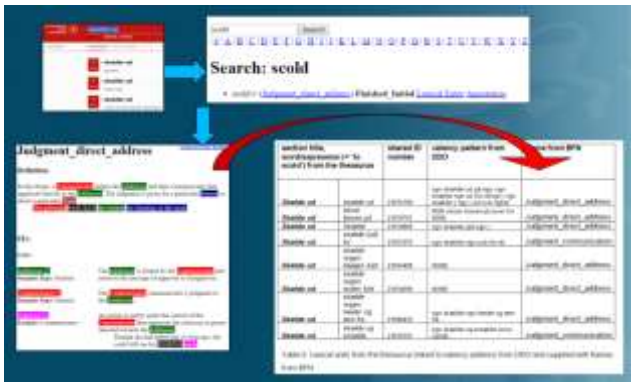

Figure 3: Mapping Berkeley frames onto thematically ordered verb groups from DT, in this case synonyms and near synonyms of *skælde ud* 'to scold'.

Having detected a group of closely related verbs and verbal nouns and furthermore supplied the verbs with valency patterns from DDO via the common id numbers, the editor would search for an appropriate frame in BFN and assign this to the particular group (shown in Figure 3). In this way, a relative large framenet lexicon is being compiled with relatively little effort[2].

Not surprisingly, however, not all groups were equally easy to assign frames to since the relation between DT and BFN was obviously not one-to-one in all semantic areas. A classic example is the discrepancy related to antonymous word senses such as remembering and forgetting which are covered by one frame (seen as the same scenario) in BFN but situated in two different groups in the thesaurus which has antonymy as an important criteria for section division. All such examples obviously required careful adjustment by the editors.

Some situations proved to have a higher degree of lexicalization in Danish than in English, for instance we found no frame which covered the Danish word *hemmeligholde* with the sense 'to refrain from telling'/'to keep as a secret'. The fact that the compilation of Berkeley FrameNet is still in progress and that all areas are not covered yet, probably also caused a lack of frames in some cases. Approx. 10 % of Berkeley FrameNet's ~1000 frames came into play when covering the two semantic areas in Danish, and the number of frames was more or less evenly distributed between the two.

3. Annotating Frames and Roles for Communication and Cognition

Behind the idea of a framenet lies not only the identification of a particular semantic frame for a particular verb or deverbal noun sense, but also the identification of the semantic roles/frame elements that are activated with a particular frame.

We rely on two assumptions:

- that our frame lexicon will ease annotation considerably since a very limited set of possible frames for a given word is presented to the annotator via the annotation tool, and
- that BFN frames for English can be more or less directly transferred to Danish; in other words, that the same frame elements or semantic roles can be identified in a Danish textual context with a particular frame. (A similar approach is taken for most other framenets being built for a number of languages, cf. Heppin & Gronostaj 2012, 2014 for Swedish, Candito et al. 2014 for French, Ohara 2014 for Japanese).

In order to test this approach, we annotated 440 sentences from the corpus with their corresponding frames and frame elements. The sentences from SemDax cover a variety of text types such as blog, chat, forum, magazine, Parliament debates (written down by professionals), and newswire, of which the latter constitutes almost half of the corpus.

In order to easily access examples which would evoke frames relating to communication and cognition, we took advantage of the coarse sense annotations available in SemDax corpus (Pedersen et al 2016) and extracted all sentences annotated with either cognition and communication events (or both)[3]. This extraction also enabled us to prove whether a frame lexicon based on thesaurus vocabulary was actually extensive enough.

We used an open source, browser-based framenet annotation tool (https://github.com/andersjo/framenet-annotation). For each verb or verbal noun relating to

[2] The project is supported by the Danish Research Council and by The Carlsberg Foundation.

[3] In SemDax all verbs, nouns and adjectives are annotated with so-called supersenses, cf. Martinez et al. (2016).

cognition or communication, the annotator has access to a small set of relevant frames, depending on the previous semantic annotation of the word. Once a frame is chosen, the annotator can assign the frame elements pertaining to this frame to the other words in the sentence by writing the word's position in the role boxes, see Figure 4.

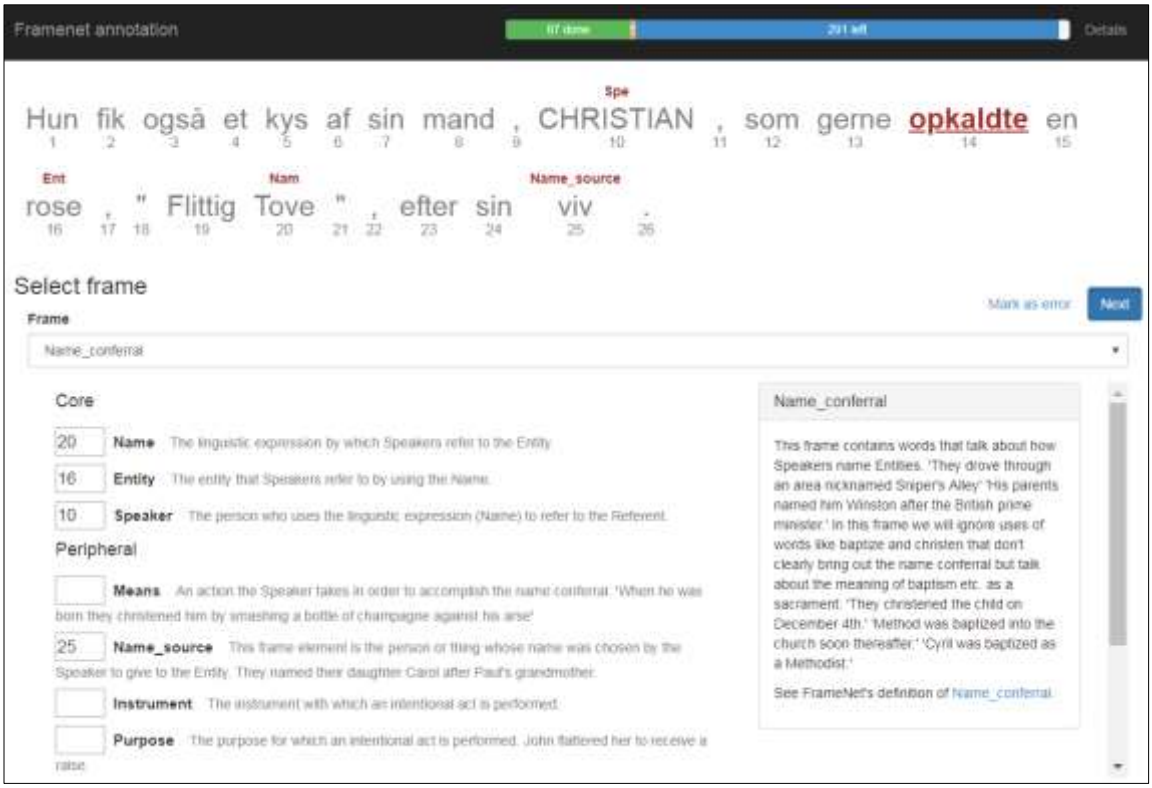

Figure 4: The FrameNet annotation tool shows the sentence with the verb to be annotated in red, and the assigned frame elements in red above the words. Below is the box for choosing the frame and boxes for assigning words to the frame elements.

Several observations were made during annotation:

- Semantic roles were generally straight-forward to detect and corresponded well to the ones suggested by the frame, however, in hardly any of the examples would we find all roles present; leaving for many of the roles to be implicit in the text.
- The annotators would sometimes disagree upon which precise frame to choose, often depending on whether to interpret the act in a concrete way or not (would the purpose of the act be more important to annotate than the concrete way of carrying it out, for example).
- The lexicon facilitated to a high degree the annotation task since the annotator only had to consider a very restricted set of frames in each case.

However, for some verbs, frames were missing because the specific sense had not been foreseen in the frame lexicon based on the DT vocabulary. The largest part of these were ad hoc (figurative) senses not to be included in the frame lexicon (nor in the dictionaries), but there were also cases which led us to expand the lexicon, e.g. cognition verbs with communication senses in corpus but not included in the thesaurus chapters on communication (for good reasons since the sense depends completely on a very specific discourse context).

The 440 frame annotated sentences were validated and in all cases supplied with the most appropriate frame before we used the corpus for training and evaluation.

4. Training and Evaluation of a Multilingual Semantic Frame Classifier

In order to demonstrate the value of our resource in terms of training and evaluating NLP models, we used the 440 frame-annotated sentences of the SemDax corpus to train and evaluate a semantic frame classifier. Since the corpus is relatively small and only contains a subset of the frames contained in the Danish FrameNet Lexicon, we experiment with cross-lingual frame label prediction in order to exploit more data. We reduce the frame-semantic parsing problem to a sentence classification problem, i.e. we train a model that predicts one frame label per sentence. The top frames subsume a total of 6 distinct frames, making this a 6-way multinomial classification problem.

Inspired by recent semantic parsing models, e.g., Zhou and Xu (2015), we use a set of binary deep, bi-directional Long Short Term Memory (LSTM) networks to predict frame labels. Each network predicts a single label, and we evaluate each network individually by computing sentence-level F1-scores.

In the following, we report implementation details and results for our experiments in the frame prediction task. We ran experiments in three setups.

- First, we trained the model on the Danish data using 10-fold cross validation and predicted Danish test data.
- Second, we did a cross-lingual experiment, where we trained on English FrameNet example sentences annotated with the equivalents of the frame labels in the Danish corpus, and predicted the Danish test data. This setting is unsupervised in that we did not use any Danish training data.
- The third setting was meant to provide a reference point for comparing the performance of the Danish model. Here, we trained on English frames and predicted English test data. As a baseline, we used a model that randomly assigns labels to the test data.

In all experiments, we represented sentences by pre-trained cross-lingual word embeddings. The embeddings are 40-dimensional, computed on 59 languages using MultiCCA.[4]

Our results for the six most frequent frames[5] are presented in Table 1. For all experiments, the LSTM hyper-parameters are tuned on English development data in a supervised set-up. The best hyper-parameter values that were used in the final experiments is a single hidden layer with a dimensionality of 20, the optimization algorithm was Adam with initial learning rate 0.001, and we used tanh activation functions and softmax for the output layer. The maximum sequence length during training is set to 20.

English-English

	Ours	*Random*
Statement	0.81	0.37
Opinion	0.39	0.06
Telling	0.49	0.05
Text_creation	0.60	0.02
Becoming_aware	0.57	0.05
Certainty	0.47	0.07

Danish-Danish

	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

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

Table 1: Supervised and unsupervised F1-scores for the 6 most frequent frames.

As mentioned, we rely only on the Danish Framenet annotated corpus, not on the frame lexicon as such. Das et al. (2010), for example, use heuristics to detect frame triggers and pick the appropriate frame using a classifier that only considers the frames licensed by the (English) Framenet. Li et al. (2012), working on POS tagging, use a tag dictionary to constrain the output space; the Danish Framenet could be used in a similar way for frame-semantic parsing. The cross-lingual parsing performance would probably improve a bit from using such constraints. On the other hand, learning to associate trigger words and frames is the easiest part of the frame detection problem; disambiguation is the hardest. In other words, if we assume that we can solve the disambiguation problem with more data, we should be able to trivially learn which frames are adequate for each verb. Also, we are already limiting the search space by considering only a subset of the total set of frames.

As can be seen from Table 1, our F1-scores are much lower when relying exclusively on cross-lingual signals (English-Danish). This shows that only part of the signal from the English data transfers. The fact that we perform better than the random baseline across all frames, shows that transfer is possible, as also suggested by Johannsen et al. (2015). However, the gap between Danish-Danish and English-Danish shows the value of annotating data in low-resource languages – and, in particular, the need for scaling up the annotated resource. Since performance correlates with support in the data – i.e., our model performs better on frequent frames – it may be beneficial to consider active learning as a strategy to efficiently annotate more data (Martínez et al., 2015).

5. Concluding Remarks

Building language resources for LT is cumbersome and expensive, and relying on existing lexicographical resources is often a challenging and not always straight forward business, in particular if – at the same time – you want to conform to international standards.

Our experiments with the compilation of a Danish framenet show that – with the given high-quality background resources, DDO and DT – it is actually feasible to build a framenet on top of an existing resource and to start from the lexicon part and move onwards to the corpus annotation. Further, adapting the role inventory from BFN to Danish proves unproblematic, with a few exceptions.
In our training experiments we have shown that language transfer of semantic frame information is possible, but that improvements are considerable when

[4] http://128.2.220.95/multilingual/data/

[5] Due to data scarcity, only the six most frequent frames were used for training and evaluation in this experiment.

combining language transferred data with annotated data of the specific target language, in this case Danish.

6. Bibliographical References

Candito, M., Amsili, P., Barque L., Benamara, F., De Chalendar, G., Djemaa, M., Haas, P., Huyghe, R., Yannick Mathieu, Y., Muller, P., Sagot, B. & Vieu, L. (2014). Developing a French FrameNet: Methodology and First Results. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland.

Das, D., N. Scneider, D. Chen & N. Schmith. (2010). Probabilistic Frame-Semantic Parsing. In *Proceedings opf ACL 2010*.

Heppin, K. & M. Gronostaj. (2012). The Rocky Road towards a Swedish FrameNet - Creating SweFN. In: *Proceedings of the Eight International Conference on Language Resources and* Evaluation. Istanbul, Turkey

Johannsen, A., Alonso, H. M., Søgaard, A. (2015). Any-langøøuage frame-semantic parsing. Empirical Methods in *Natural Language Processing* 2015. Lisbon, Portugal.

Li, J. V. Graca, B. Taskar. (2012). Wiki-ly Supervised Part-of-Speech Tagging. In: *Proceedings of ACL 2012.*

Martínez Alonso, H., Johannsen A, Olsen S, Nimb S and Pedersen BS (2016). An empirically grounded expansion of the supersense inventory. *Proceedings of the 8th Global Wordnet Conference*, Bucharest, Romania.

Martínez Alonso, H. M., Plank, B., Johannsen, A., Søgaard, A. (2015). Active learning for sense annotations. *The 20th Nordic Conference on Computational Linguistics*. Vilnius, Lithuania.

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

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.

Nimb, S. (Submitted for review). The Danish Frame Lexicon : Method and Lexical Coverage. In *Proceedings from The International FrameNet Workshop 2018 at LREC/Japan: Multilingual FrameNets and Constructicons*.

Ohara, K. H. (2014). Relating Frames and Constructions in Japanese FrameNet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* Reykjavik, Iceland pp. 2474-2477.

Levy, Omer; Søgaard, Anders; Goldberg, Yoav. 2017. A strong baseline for learning cross-lingual word embeddings from sentence alignments. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL).* Valencia, Spain.

Pedersen, B. A. Braasch, A. Johanssen, H. Martínes Alonso, S.Nimb, S. Olsen, A. Søgaard, N. Sørensen (2016). The SemDaX Corpus - sense annotations with scalable sense inventories. In *Proceedings of the 10th edition of the Language Resources and Evaluation* Conference, Portorož, Slovenia.

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.*

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.

Zhou, J., Xu, W. (2015). End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China.*

7. Language Resource References

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

The Danish Frame Lexicon (2018): available at https://github.com/dsldk/dansk-frame-net.

The Danish Thesaurus (DT): Nimb, S., Lorentzen, H., Troelsgård T., Theilgaard, L., Trap-Jensen, L. (2014a): *Den Danske Begrebsordbog*, Society for Danish Language and Literature, Copenhagen, Denmark.