Reusing the Danish WordNet for a New Central Word Register for Danish a Project Report

Bolette S. Pedersen[1], Sanni Nimb[2], Nathalie Carmen Hau Sørensen[1], Sussi Olsen[1], Ida Flörke[2], Thomas Troelsgård[2]

Centre for Language Technology, NorS, University of Copenhagen[1], Society for Danish Language and Literature [2]

Emil Holms Kanal 2, 2300 Copenhagen S[1], Christian Brygge 1, 1219 Copenhagen K [2]

{bspedersen, nmp828, saolsen}@hum.ku.dk, {sn,if,tt}@dsl.dk

Abstract

In this paper we report on a new Danish lexical initiative, the Central Word Register for Danish, (COR), which aims at providing an open-source, well curated and large-coverage lexicon for AI purposes. The semantic part of the lexicon (COR-S) relies to a large extent on the lexical-semantic information provided in the Danish wordnet, DanNet. However, we have taken the opportunity to evaluate and curate the wordnet information while compiling the new resource. Some information types have been simplified and more systematically curated. This is the case for the hyponymy relations, the ontological typing, and the sense inventory, i.e. the treatment of polysemy, including systematic polysemy.

1 Introducing COR and DanNet

The Central Word Register of Danish – with the acronym COR – is a lexicon project running from 2021 to 2023 as part of a Danish governmental language technology and AI initiative. The aim of the project is to coordinate, curate, combine and extend already existing lexical NLP resources – including the Danish wordnet – in a joint initiative in order to ease the use of NLP resources and thereby help boost NLP and language-centric AI for Danish.

The COR project is funded by The Danish Agency for Digitisation and led in collaboration by three of the main dictionary and LT institutions in Denmark: i) the Danish Language Council (DSN), ii) Society for Danish Language and Literature (DSL), and iii) Centre for Language Technology (CST) at the University of Copenhagen.

One of the main ideas in COR is to assign *a unique identifier*[1] to all lemmas[2]. The main resource consists of a lexicon of the general language vocabulary with basic morphology and semantics. Syntactic and phonological information is foreseen in subsequent phases of the project.

The lemma selection as well as the morphological information, the glosses and the usage examples are based on three 'classical' dictionaries, the orthographic dictionary *Retskrivningsordbogen* from DSN, the monolingual dictionary *Den Danske Ordbog* (The Danish Dictionary, DDO) and the thesaurus *Den Danske Begrebsordbog* (The Danish Thesaurus, DDB) from DSL.

The formal semantic information in COR (labelled COR-S), in contrast, relies to a large extent on the Danish wordnet, DanNet (Pedersen et al. 2009), but also includes data from the Danish

[1] Which can be seen as a parallel to The Danish Person Register (CPR) where all Danish citizens are assigned a unique id.

[2] See also the COR description on the website of The Danish Language Council (in Danish): https://dsn.dk/nyheder-og-arrangementer/dansk-sprognaevn-med-i-stor-sprogteknologisk-satsning/

FrameNet Lexicon (Nimb et al. 2017) and the Danish Sentiment Lexicon (Nimb et al. 2022).

DanNet was originally built on DDO, meaning that, instead of compiling the wordnet as a transfer and adjustment of Princeton WordNet, it is based on monolingual grounds and subsequently linked to Princeton WordNet (cf. Pedersen et al. 2019 for a description of the linking procedure). The sense definitions from the DDO were semi-automatically transformed into wordnet relations via the genus and differentia. The rather fine-grained sense inventory of DDO was more or less taken over in DanNet with some minor adjustments, however in a 'classical' wordnet manner (Fellbaum 1998), that is, with all senses equally described at synset level and thus not capturing the structure of main and sub-senses from the DDO – and not necessarily all its senses, either. In cases of synonymy, a wordnet approach was adopted of typically including synonyms as part of the same synset.

In the following sections we describe the role of DanNet in the compilation of COR and discuss which adjustments and simplifications have been performed to make the wordnet information applicable in a resource like COR. In Section 2 we describe the overall picture of COR in relation to other existing resources. Section 3-6 goes into depth wrt. which information types have been taken over in COR-S and how. In Section 7 we discuss the consequences that our revisions may have for a future DanNet, and in Section 8 we conclude.

2 COR-S as Related to Other Danish Lexical Resources

As has been described in previous accounts (Pedersen et al. 2022 and others), all NLP resources including DanNet are linked at sense level with the sense inventory of the DDO. This means that the semantic NLP resources are all conferring to the same sense and lemma inventory and that an integration of information types is therefore more or less straight-forward.

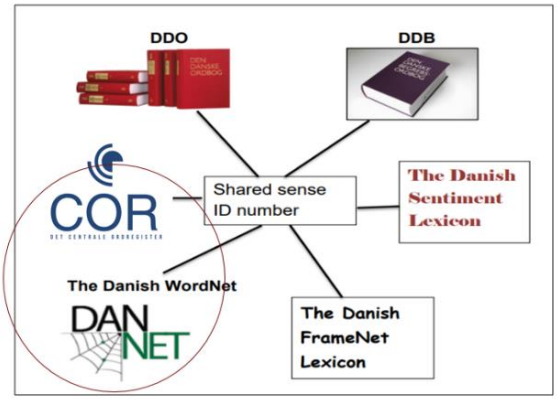

Figure 1: Danish lexical-semantic resources sharing the same sense ID number

As depicted in Figure 1, COR-S is mainly compiled on the basis of DanNet, but as mentioned above, including the integration of further information from primarily DDO, The Danish Thesaurus (surrounding words), The Danish Sentiment Lexicon (connotation polarity), and the Danish FrameNet (semantic frames on verbs and deverbal nouns).

3 Hyponymy revisited

The skeleton of the wordnet in the sense of its hyponymy structure is essentially taken over in COR, meaning that all senses in COR include a link to its most suited hypernym. Some adjustment has however taken place. For instance, very specialist taxonomies are simplified to a certain extent, reflecting now to a larger degree a layman's perspective to i.e., natural entities (e.g., plants and animals). The hypernyms of abstract and verbal entities in DanNet (denoted as 2nd and 3rd Order Entities, respectively, according to Lyons' semantic divisions (Lyons 1977) often relate to synsets that are based on highly polysemous DDO lemmas. These were therefore in some cases incorrectly assigned and have now been adjusted. For instance, the inventory of verbal hypernyms has been reduced to ensure consistency among verbs. Our goal is to cover all DanNet hypernyms in COR-S, and in the final phase to convert the synsets to the corresponding COR-S senses.

The task is somewhat complicated by the simultaneous overall reduction of senses in COR, meaning that two DanNet synsets might result in only

one COR-S sense according to a set of principled reductions rules (see Section 5).

4 A Slightly Simplified Ontological Typing

DanNet contains ontological typing on all synsets conferring to the EuroWordNet Ontology (Vossen 1999) with a few extensions, such as an additional type denoting body parts, which seems to a very frequent ontological type with specific characteristics.

For COR, however, we generally aim at a much simpler and more intuitive ontology that can easily be managed and understood also by non-experts and where a high degree of consensus can be achieved in a first encoding round.

To this end, the ontology has been radically simplified, reducing the number of types by 36% from 204 to 130. For example, approx. 1/8 of the EuroWordNet ontological values were only applied 10 times or less in DanNet signaling thereby their somewhat unconsolidated status. Therefore, we decided to remove these in COR-S[3]. Since the aspectual distinction between bounded and unbounded events is rarely lexicalized in Danish (but rather determined by the surrounding adverbs, adverbial particles, or prepositional phrases), we decided to neglect this meaning component in COR-S, a fact that also reduces the number of types significantly.

DanNet	COR-S
UnboundedEvent	Event
BoundedEvent	
UnboundedEvent+Agentive	Act
BoundedEvent+Agentive	
Dynamic+Agentive	
3rdOrderEntity+Mental+ Purpose	Abstract+Purpose
3rdOrderEntity+Mental+ Purpose+Manner	
BoundedEvent+Agentive+ Purpose+Possession	Act+Possession
BoundedEvent+Agentive+ Purpose+Possession+Social	

Table 1: Ontological types in DanNet converted into simpler types in COR-S

[3] Examples of removed types are Artifact+ Substance+Part, Container+Artifact+Object+Group; and 3rdOrderEntity+Relation.

Some of the most complex 2nd Order types describing several meaning components at a time in different combinations (purpose, social, as well as possession, for example) were also omitted. Instead, the lexicographer must decide on the most prominent meaning aspect when assigning a type. Finally, the names of the types were in some cases changed into more intuitive ones (3rd OrderEntity is changed to Abstract, 2ndOrderEntity+Agentive to Act and so forth). See Table 1 for examples of simplifications[4].

Where most transfer from the EuroWordNet Ontology to the COR Ontology is done fully automatically (many -> one), a few are left for manual inspection to select the most prominent meaning component among several. This is the case for instance where both the meaning components Purpose and Social are encoded in the source, and where we in COR select what we consider to be the most prominent, as in *drille* (to tease): Social.

5 A Reduced Sense Inventory Suitable for NLP

Another characteristic feature of COR-S compared to most other available lexical resources for Danish, is its *reduced sense inventory*. This feature has been suggested by NLP developers to ease word sense disambiguation and overall make the resource more directly applicable in practical NLP tasks, an approach that corresponds well to positions put forward for instance by Kilgarriff (1997), and Pedersen et al. (2018).

In Pedersen et al. (2022) we report on the lexicographical principles behind this sense reduction in COR to what we label core senses, and which can be summarized as follows:

Delete a DDO main or sub-sense if it

- is marked as rare, historic, colloquial, or slang in DDO[5]
- is marked as domain specific in DDO

[4] The entire COR-S Ontology will be released in late 2023 with the full resource.

[5] It could be argued that slang and colloquial senses would be relevant for COR, for instance for processing social media. However, it proves to be indeed very hard to keep up to date with slang meanings, and in several cases, suggested slang senses in DDO have proven to be by far outdated and thereby more confusing than helpful for NLP.

- has a low sense weight score, amounting to how much info is given about the sense in terms of examples etc.

Merge/cluster a DDO sub-sense with its main sense

- unless it diverges from the main sense in ontological typing (from DanNet) (typically concrete ontological types versus abstract types, as is the case of most figurative senses.)

The reduction is done manually for the most complex (i.e. most polysemous) part of the vocabulary[6], whereas automatic methods are used for treating the least polysemous part of the vocabulary (2-4 senses per lemma), using however, the hand-coded examples as a gold standard. We apply a rule-based method, a word2vec model (Mikolov et al. 2013) and a BERT model (Devlin et al., 2019) for our automatic merges (cf. Pedersen et al. 2022: Section 4). Since accuracy does not exceed 0.82 for any of our automatic methods, however, all merged vocabulary is carefully manually curated before admitted into COR-S.

6 Systematic Polysemy in a Reduced Sense Inventory

Systematic polysemy constitutes a particular case of ambiguity where multiple lemmas show the same, regular pattern of polysemy. The phenomenon is well described in literature (Apresjan 1973, Malmgren 1988, Pustejovsky 1995 and others) and has been dealt with in both lexicography and in linguistics more broadly, relating to whether you tend to represent the phenomenon by splitting or merging the senses – or by something in between [7]. A general aim in all approaches is to try and treat the phenomenon *consistently*, which, however, is not as easy as it sounds at least not in a fully-fledged lexicon.

For instance, the merge principles defined in Section 5 cannot really serve as guidance here since the DDO as source applies mostly extra-linguistic principles for describing lemmas that are systematically polysemous, such as space principles in the original printed dictionary in combination with the frequency of the lemma[8].

Therefore, to get an overview of the phenomenon in Danish, and to subsequently outline consistent merge or split principles for COR-S (as well as to encode the pattern value as part of the semantic information for each sense), we have been through a large set of lexicographical material based on the aforementioned core vocabulary and have identified more than 20 patterns of systematic polysemy. For an in-depth account of this work, see Sørensen et al., (2023).

For each pattern we have decided whether to keep the distinction of the senses or merge them into a single sense. Here, we reused the work already done in DanNet with respect to clarifying systematic polysemy (see Pedersen et al. 2010) since the patterns become obvious both from the ontological types and from the hypernym structures. For instance, the distinction between living and non-living entities in the DanNet taxonomy reveals the ANIMAL/FOOD pattern, and these senses are maintained in COR-S since they are quite clearly distinguished in use. In addition, the frequency of a particular sense type plays a role. To this end, in the related pattern ANIMAL BODY PART/FOOD the principle says to merge due to the proportionally much higher frequency of the FOOD sense here (we only very rarely talk about e.g. chicken breasts or chicken wings outside the cooking scenario). For the PROCESS/RESULT pattern, to give another example, we only split senses when the result is a concrete artifact and thus distinguishes itself clearly from the process (as is the case for *konstruktion* 'construction').

7 COR as Feedback to DanNe

In the COR project, a lemma that is only represented in one of its senses in DanNet is considered from a *semasiological* perspective, meaning that a con-

[6] The reduction is done manually for the 3,300 lemmas in DDO of which at least one sense is linked to the so-called core concepts in PWN (https://wordnetcode.princeton.edu/standoff-files/core-wordnet.txt) via DanNet, and which constitute a highly polysemous part of the vocabulary.

[7] Pustejovsky (1995) suggests so-called 'dot types' as a means to represent under-specification in systematic polysemy.

[8] In other words: Frequent lemmas tend to be 'unfolded' in the DDO with both meanings explicitly represented, whereas rare lemmas are only provided with one sense.

siderable amount of supplementary information is encoded to it.

When the COR project ends in 2023, we will therefore consider which of this curated information should be transferred back to DanNet with the aim of improving the wordnet. The id numbers ensure that this should not be too difficult a task.

First of all, DanNet does not contain all senses of a lemma in the way that COR-S does (even if for COR-S, we merge senses). This is a flaw of DanNet, which was produced under hard time constraints and which had hypernyms with many hyponyms as a driving principle leaving sometimes quite central senses untreated.

Secondly, the DanNet senses that have being deemed rare or too domain specific via the examinations in COR-S (approx. 10%) should be labeled as such in DanNet since the information is relevant for several purposes.

Some senses in DanNet are lumped together in COR-S, and it should be considered whether also to reflect this in DanNet in some way. The validation of the hypernyms in COR-S also provides useful feedback to DanNet and calls for a similar curation in the original resource.

Last but not least, it might be fruitful to adopt the more coarse-grained version of the EuroWordNet Ontology developed in COR, and in this case also transfer the validated ontological types from COR back to DanNet to ensure a higher lexical quality of the wordnet, especially in the case of 2nd and 3rd Order Entities where the EuroWordNet Ontology has proven somewhat complex to use in practice.

Sentiment values will already be directly included in DanNet based on the underlying data of the sentiment lexicon (Nimb et al. 2022) (describing values at sense level). Finally, the integration of semantic frames in DanNet is still under consideration as a way to improve the verbal descriptions in the resource.

8 Concluding Remarks

Building a new lexical resource like COR is an expensive and extremely time-consuming task. The COR project is primarily meant to serve the NLP-related AI industry by providing an easy-to-use, open-source resource with unique identifiers. In such a case, it seems indispensable to look around in the language community for resources that can be easily reused for that particular purpose. As well as to consider lexicographical standards that can ease transfer and alignment, as underlined in the lexicographic ELEXIS infrastructure (Krek et al. 2018).

As has been shown in this paper, we have had the great advantage in the Danish language community of having several substantial semantic resources interlinked via a unified sense id structure and relying on international standards. This has enabled us to easily transfer information into the new resource. In particular, the Danish wordnet, DanNet (in combination with DDO) has proven useful for this task.

While going through DanNet for the purpose of compiling COR, we have further taken the opportunity to also consider which revisions we would like to transfer back to the wordnet at a later stage in order to improve this stand-alone resource. In this way, the COR project has given us the chance to actually curate a resource that was compiled more than 10 years ago as part of a research project with only limited resources

References

Apresjan, J. (1973). Regular polysemy. In: *Linguistics 142*, pp 5-32.

Devlin, J., Chang, M-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1. pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics

Fellbaum, C. (ed) (1998). *WordNet – An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, London.

Kilgarriff, A. (1997). I Don't Believe in Word Senses. In: *Computers and the Humanities.*Vol. 31, No. 2 (1997), pp. 91-113.

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen B. S., Tiberius, C. & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In : *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, Ljubljana.

Lyons, J. (1977). *Semantics.* Cambridge University Press.

Malmgren, S. (1988). On Regular Polysemy in Swedish. In: *Studies in Computer-Aided Lexicography*, Almquist & Wiksell, Stockholm.

Mikolov, T., Yih, W. & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In : *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 746-751).

Nimb, S. (2016). Der er ikke langt fra tanke til handling. In: S. Skovgaard Boeck & H. Blicher (eds) *Danske Studier 2016*, København, Universitets-Jubilæets danske Samfund, pp. 25-59. Copenhagen.

Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S., & Søgaard, A. (2017). From Thesaurus to Framenet. In: I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (red.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 conference* (s. 1-22). Lexical Computing CZ.

Nimb, S., Olsen, S., Pedersen, B. S., & Troelsgaard, T. (2022). A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. In: *Proceedings of the Language Resources and Evaluation Conference: LREC2022* (Bind 2022, s. 2826--2832). European Language Resources Association.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N., Trap-Jensen, L., & Lorentzen, H. (2009). DanNet - the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation*, *43*, 269-299.

Pedersen, B., M. Agirrezabal, S. Nimb, I. Olsen, S. Olsen (2018). Towards a principled approach to sense clustering – a case study of wordnet and dictionary senses in Danish. In: *Proceedings of the 9th Global Wordnet Conference*. Singapore.

Pedersen, B. S., Nimb, S., Olsen, I. R., & Olsen, S. (2019). Linking DanNet with Princeton WordNet. In: *Global WordNet 2019 Proceedings, Wroclaw, Poland* Oficyna Wydawnicza Politechniki Wroclawskiej.

Pustejovsky, J. (1995). *The Generative Lexicon.* The MIT Press.

Sørensen, N., S. Nimb & B. Pedersen (2023). Validating Systematic Polysemy in WordNets by Means of Contextualized Embeddings. In: *Global WordNet Conference 2023*, Donostia, Spain.

Vossen, P. (ed.) (1997). *EuroWordNet: A multilingual database with lexical semantic networks*. Springer Verlag.