Sanni Nimb, Ida Flörke, Sussi Olsen, Bolette S. Pedersen, and Nathalie C. H. Sørensen

COR.SEM, A NEW FORMAL SEMANTIC LEXICON FOR DANISH

Abstract We present the COR.SEM lexicon, an open-source semantic lexicon for general AI purposes funded by the Danish Agency for Digitisation as part of an AI initiative embarked upon by the Danish Government in 2020. COR.SEM describes the core senses of 34,000 Danish lemmas with formal semantic information, e.g., ontological type, hypernym, semantic frame, regular polysemy pattern, and polarity value; features which are in essence drawn and simplified from other existing resources. Lexical information from The Danish Dictionary DDO and the Danish Thesaurus DDB is also integrated, e.g., user examples, domain label, synonyms, and near synonyms. It provides direct links to synsets in the Danish WordNet DanNet, as well as to the morphological lemma information in COR, the Central WordRegister which is based on the Danish Orthographical Dictionary and DDO. The register's common numerical index at both lemma and sense level makes it is more straightforward to merge mono- as well as bilingual dictionaries with COR.SEM and thereby inherit the formal semantic information. At the website corsem.dsl.dk it is possible to browse the lexical entries and to download tailored extracts of data of your choice. We give examples of the use of COR.SEM in linguistic studies, in NLP tasks and in lexicographic projects.

Keywords semantic lexicon; lexical semantics; linked data; computational lexicography

1. Introduction

COR.SEM is a new semantic lexicon for Danish published in 2024 at ordregister.dk in the form of a dataset for free download (CC0 license). The dataset is linked to the indexed morphological lexicon COR which is based on RO (the Danish Orthographical Dictionary) and available at the same site. The Central WordRegister has been developed as a public state initiative with the overall aim of facilitating and easing the transition of Danish industry and public institutions towards an increased use of language technology and language-centric AI. The vision has been to support the development of the technology so that it can function seamlessly in a Danish societal context as well as in different professional settings. To this end, a central premise of the COR project is that the resource be based on existing high-quality dictionaries and thus should take advantage of the rich information sources already provided in and about Danish culture and society. In section 2, we describe the existing dictionaries and resources and how the COR.SEM was compiled based on linked data. The initiative has made it more straightforward to share and merge existing Danish lexical resources via the register's common numerical index, which ensures a homogeneous approach to all incoming resources, be they terminologically, historically or in other ways specific. This feature has in fact been requested from Danish companies and public institutions for a long time.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

COR.SEM gives open access to several semantic information types in a standardized and manageable way. This includes information such as a coarse-grained sense inventory of Danish, sentiment, and ontological information on Danish concepts as well as frame semantic information on Danish verbs and deverbal nouns. In section 2 we describe the lexicon. In section 3 the webpage CORSEM.DSL.DK is presented, and in section 4 we discuss different types of use of the resource before we conclude in section 5.

2. The COR.SEM Lexicon: Content and Method

The lexicon contains 34,020 lemmas (28,725 nouns, 2770 verbs, 2343 adjectives, and 177 adverbs) from the Danish Dictionary DDO (Hjorth & Kristensen, 2003–2005; ordnet.dk/ddo) and the Danish WordNet DanNet (wordnet.dk/dannet; Pedersen et al., 2009). Via links to DanNet in COR.SEM, more than 5,000 English equivalents can be identified. 11,300 COR.SEM lemmas are central lemmas in Danish being defined as either linked to one of the 5,000 base concepts in Princeton WordNet or being a central word (a highlighted keyword) in the Danish thesaurus DDB (Nimb et al., 2014; Nimb et al., 2014/2015) – or both. In this way we ensured that both the most central lemmas as well as a large variety of themes and semantic fields are well represented. The information on centrality is part of the resource. The degree of centrality is represented as a number (3, 2 or 1), in contrast to the 24,000 non-central lemmas (0).

6,000 of the lemmas are polysemous having a total of 15,000 COR.SEM senses. The sense inventories of these lemmas have been established by carefully studying the senses in the fine-grained DDO dictionary, discarding rare and uncommon senses and lump cases of semantically close senses, typically in the cases of narrower subsenses of a main sense. In this way, we have reduced the number of senses by 25% (see Section 2.1. for more details).

The sense information is expressed in the form of formalised information types assigned as a value from a closed list: ontological type, hypernym, frame, and polarity value. Furthermore, synonyms from the DDO and DanNet (in the form of synset_members of the same synset), as well as a small group of semantically related words extracted automatically from the thesaurus DDB contribute to the description of the sense.

All senses in COR.SEM have at least one, sometimes two ontological type values (8% of the cases), and at least one, sometimes two hypernyms (8% of the cases). The ontological type is chosen among 146 values from an established COR.SEM ontology, a simplified version of the DanNet and EuroWordNet ontologies (Vossen, 1999).

The frame values correspond to the frames in Berkeley FrameNet, where they are further specified, and the polarity values follow a scale from minus to plus 3 (0-values are not marked).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

The polarity values were transferred from working files from the sentiment lexicon project (*Det Danske Sentiment-leksikon*; Nimb et al., 2022). They were validated and equilibrated in the case of lumped DDO senses in COR.SEM. Also new senses were assigned a polarity value. In a supplementary resource at ordregister.dk, COR.SEM. EXT, definitions and examples from the DDO are available under a more restricted license (CC BY-NC-ND).

In Figure 1 the verb *håbe* ('to hope') is shown, illustrating the formal semantic information types. Its ontological type is 'Act+Mental', the semantic frames are 'Desiring' and 'Reliance_on_expectation'. The sentiment polarity value is +2 (on a scale from -3 to +3). Links to the equivalent synset and the hypernym in DanNet are also provided.

In Table 1 the numbers of lemmas, senses and values are shown.

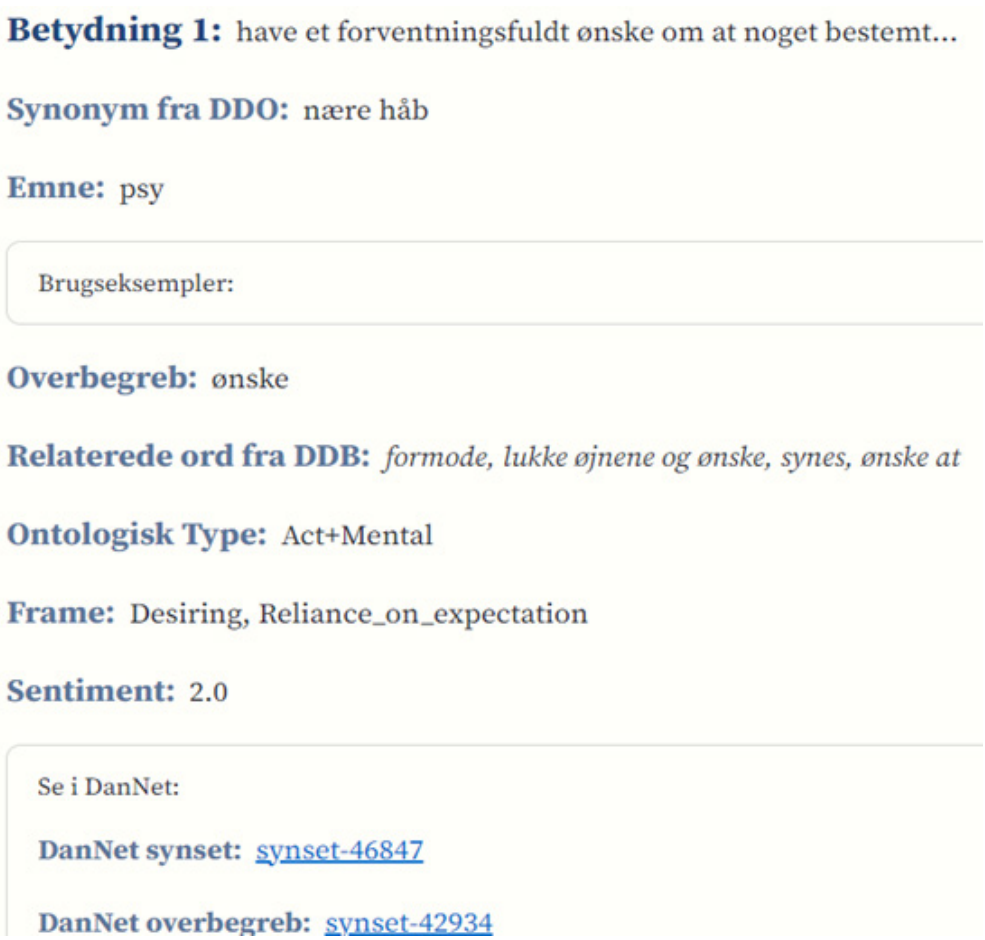

Fig. 1: The entry of the monosemous verb *håbe* ('to hope') with the (automatically shortened) DDO definition 'have an expectant desire for something specific' , the DDO synonym *nære håb*, the DDO domain 'psy' (psychology), the hypernym *ønske* ('to wish'), related words from DDB ('suppose', 'close the eyes and wish', 'think', 'wish to'), the ontological type 'Act+Mental', the frames 'Desiring' and Reliance_on_expectation', and the positive sentiment value 2. Links to the corresponding synset In DanNet as well as the hypernym synset in DanNet are also given

Table 1: The numbers of information types for lemmas and senses in COR.SEM

Information type	Lemmas	Senses	Values, total number
Synonyms from DDO	11,557	13,236	19,959
Related words from thesaurus	32,832	38,830	114,834
Usage examples from DDO	25,275	32,979	46,993
Frame values (Berkeley FrameNet)	6720	8806	12,891
Systematic polysemy	1100	1609	1609
Sentiment value	5730	6823	6823

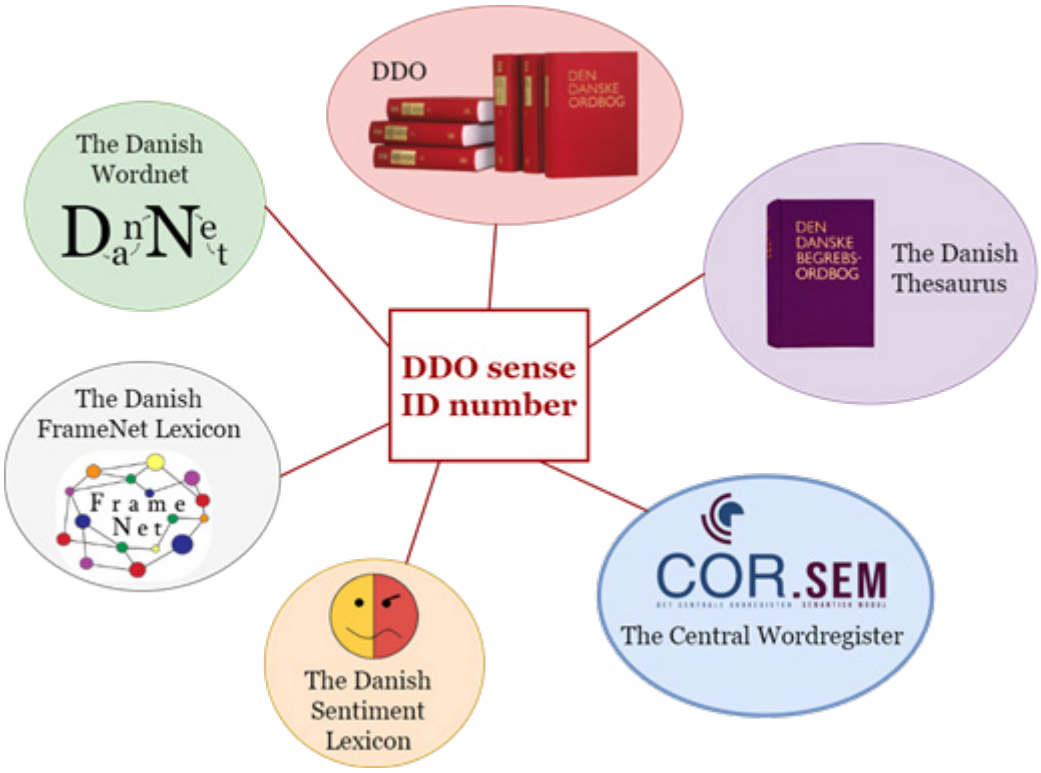

Fig. 2: The linked dictionaries and lexicons. Information from DDO, the thesaurus DDB, DanNet, The Danish FrameNet Lexicon and the Danish Sentiment Lexicon constitute the skeleton of COR.SEM

2.1 Linked Data

The COR.SEM lexicon was compiled based on already existing information in five dictionaries and formal semantic lexicons. In addition to the three mentioned above - DDO, DanNet, and the Danish thesaurus DDB - also The Danish FrameNet Lexicon (Nimb et al., 2017), and The Danish Sentiment Lexicon (Nimb et al., 2022) were used.

The five resources constitute a lexical network since they are all linked at sense level to the DDO sense inventory, see Figure 2. By keeping track of the shared sense id numbers, COR.SEM is today integrated in the network representing both semasiological and onomasiological information for approx. ⅓ of DDO's lemmas and more than half of the lemmas in DanNet and the RO orthographic dictionary.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Opposite to DanNet, which in a 'classical' wordnet manner does not necessarily contain all senses for each lemma, COR.SEM covers all common senses of a lemma, except fixed expressions.

The fine-grained sense distinctions that we find in dictionaries for humans such as DDO, constitute a problem for language technology and AI. Experience from former projects, e.g., SemDaX, see Pedersen, B.S. et al. (2016), shows that coarse-grained sense distinctions work much better. For that reason, in COR.SEM we collapsed the fine-grained DDO senses of many polysemous words into fewer senses. We describe the principles followed in the next section.

Fig. 3: The DDO and COR.SEM entries of the adjective *elektrisk* ('electric'). In DDO to the left, the adjective has five senses, in COR.SEM to the right it has only two since three of the DDO subsenses are lumped with the main sense

2.2 Compiling a Core-Grained Sense Inventory Based on a Fine-Grained Monolingual Dictionary

The main principle for lumping senses was to keep the DDO main senses of a lemma and collapse a subsense with its main sense since subsenses in DDO are closely semantically related to their main sense. If, however, a subsense is of a different ontological type, which will typically be the case for figurative senses of a lemma, it will not be collapsed with its main sense. For instance, for the word *elektrisk* ('electric') in Figure 3, COR.SEM has two senses while DDO has five: one main and four subsenses. The DDO main sense and three of its four subsenses are merged into one sense (Betydning 1 ('Sense 1')). These senses all describe electricity in its physical sense: concerning/functioning with electricity, and they share the ontological type: Property_Physical. The last subsense is a figurative sense: 'characterised by tension, intensity' with the ontological type: Property+Mental. This subsense has its own sense in COR.SEM (Betydning 2 ('Sense 2') in Figure 3).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

For cases that are not straightforward, other factors influence the decision of whether to collapse a sense or not. We use calculated weight scores for each sense, based on the amount of extra lexicographic information connected to the sense in DDO, e.g., frequency information, collocations, and number of examples (no example in DDO is often due to low corpus frequency), to decide whether the sense is essential and consequentially should not be collapsed with other senses.

In Figure 4 the COR.SEM entry of the noun *afsløring* ('disclosure', 'revelation', 'reveal') illustrates how five senses in DDO ends up having only two senses in COR.SEM. Betydning 1 (Sense 1) in COR.SEM covers both the DDO main sense 1 ('discovery of a (dishonorable) relationship that has been tried to be hidden') and the DDO main sense 2 ('an investigation and discussion of conditions that have hitherto been hidden or kept secret') since they both reflect the act of discovering something secret or hidden and of disclosing the information to someone. A strong argument for the lumping is that the two senses share ontological types (Act+Mental and Act+Communication). The subsenses of the main DDO sense 2 ('information about something previously kept secret'; 'detection by means of a survey etc.') can also be described by these two ontological types and are therefore included in the broad COR.SEM sense 1.

Fig. 4: In COR.SEM the noun *afsløring* ('disclosure', 'revelation', 'reveal') has only two senses because of the lumping of two main DDO senses as well as two DDO subsenses which all concern the revealing of information. The third main sense in DDO describes the unveiling of something concrete to an audience and is therefore kept separately. It constitutes sense 2 in COR.SEM.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Opposite, we consider the third DDO main sense 'ceremony where a memorial or the like is unveiled' (in DDO illustrated by the example 'the small circle of invitees was to witness the unveiling of a commemorative plaque') to be a different sense. It is therefore not lumped with sense 1, but inserted in COR.SEM as sense 2 (Betydning 2) with the ontological type Event+Social.

The FrameNet frames assigned to the two COR.SEM senses (as this is a verbal noun) consolidate the decision of merging the four DDO senses and keeping the last one apart. 'Reveal_secret' for the first COR.SEM sense and 'Cause_to_perceive' and 'Social_event' for the second sense correlate closely with the ontological types.

A specific problem regarding the lumping of senses regards the approach to systematic polysemy in DDO where polysemy sometimes results in two DDO senses, and sometimes in one, regardless the type. A set of 30 systematic polysemy patterns with specified guidelines on whether to lump or split the two senses ensures a homogenous treatment in COR.SEM (see Sørensen et al., 2023). The patterns are part of the information in the lexicon. Figure 5 has been generated from the visualisation function ('datavisning') at the website corsem.dsl.dk (see section 3), and it shows the distribution of systematic polysemy patterns for nouns and verbs. The two PROCESS / RESULT patterns, one which represents concrete results, where COR.SEM splits in two senses, and one which represents abstract results, where COR.SEM lumps the two senses into one, constitute more than half of the noun cases. Examples are *forventning* ('expectation') (two senses in DDO, one sense in COR.SEM), and *produktion* ('production') (two senses in DDO, also two senses in COR.SEM). Looking at the distribution for verbs, ACT / EVENT is by far the most common pattern. The (typically) two senses in DDO are in this case not lumped but both represented in COR.SEM.

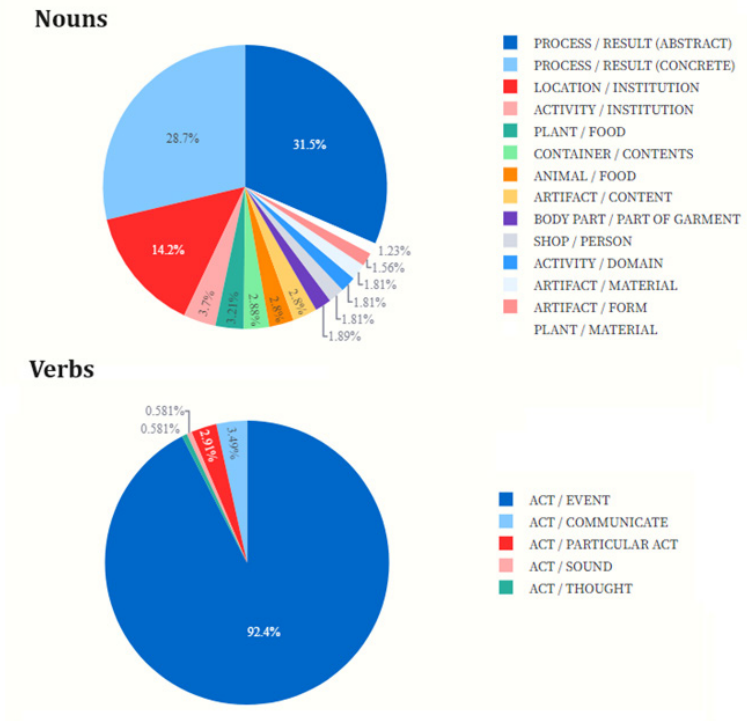

Fig. 5: The distribution of systematic polysemy patterns of nouns (top) and verbs (bottom) in COR.SEM

3. The Website CORSEM.DSL.DK

The cor.sem resource is published at ordregister.dk where it can be downloaded as a tsv-file[1]. We have also created a Danish webpage **corsem.dsl.dk**[2] aimed at linguists and lexicographers, allowing them to view and interact with the Danish data, as well as to download either the entire dataset or selected parts of their own choice.

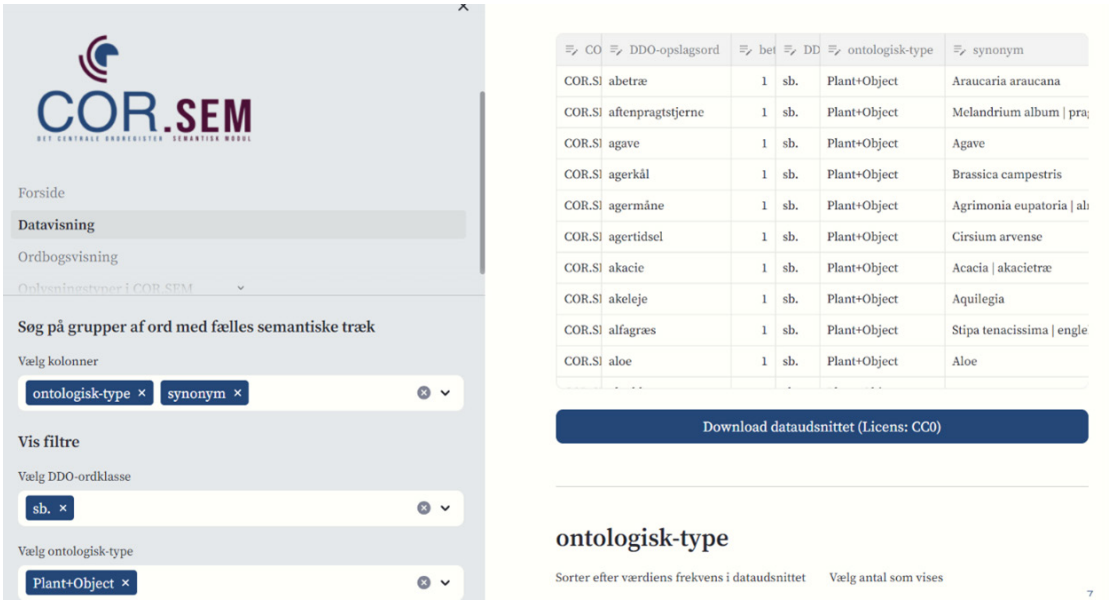

Fig. 6: The website corsem.dsl.dk and the function 'Datavisning' ('demonstration of the data'). The user has chosen ontologisk type = 'Plant+Object', the wordclass 'sb.' ('noun') and 'synonym', the synonym from DDO, which is in this case often a Latin term

The webpage contains two types of viewing functions. The first one shows the content in line with a traditional dictionary presentation, presenting all semantic information for a selected lemma (see Figure 1, 3 and 4 above). This also includes definitions and (if activated) usage examples from the DDO. The second viewing function (see Figure 7) is aimed at giving the user an idea of the different types of semantic information across the entire COR.SEM vocabulary. It focuses on data and combinations of data of your own choice and allows the user to search for specific words and select information types to be shown on the screen. The data is presented as a table where each column represents one of the selected information types. As default, the data table contains a) an ID-column with an ID-string that is unique for a specific sense and is linkable to the lemma across the COR-resource, b) the lemma from the Danish Dictionary (DDO), c) the sense number in the resource, and the word class from the Danish Dictionary (DDO). The user can select additional columns to be shown in the data table. Additionally, we have created a filter functionality allowing the user to filter each column for specific values. With this functionality, it is for example possible to extract all negative words for animals (the filters ontologisk type = 'Animal+Object' and

[1] COR.SEM is open source (license CC0), however the DDO-definitions and examples are kept in a separate file COR.SEM.EXT under the license CC BY-NC-ND

[2] It should be noted that no formal usability evaluation of the website has been carried out

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

sentiment < 0). Another example is shown in Figure 6, where the user has selected all nouns with the ontological type 'Plant+Object' combined with the synonym from DDO, which is in this case is typically a Latin term[3]. It is also possible to generate diagrams showing the frequency of every unique information type in the selected data, as well as to download the data as tsv-files for further elaboration and application.

4. Making Use of COR.SEM Data

COR.SEM can be integrated as an independent semantic component in Danish language technology applications to improve aspects of Natural Language Understanding (NLU). Depending on the requested functionality this could for example be for sentiment analysis, for refined entity recognition, or for relation extraction. It can however also be used for training and/or fine-tuning of large language models (LLMs). By training and fine-tuning, the resource can help remedy some of the language transfer bias that are seen with current Danish LLMs where a substantial part of the content is in fact translated from English and thus not representing the Danish language, culture, and society in an appropriate way.

COR.SEM can also be used for benchmarking large language models (LLMs) for Danish. In Pedersen et al. (2024a) we demonstrate how it can be used as a benchmark for sense disambiguation. The coarse-grained sense inventory of COR. SEM combined with the usage examples (originally from DDO) were used to test whether the LLMs can disambiguate words in contexts. The usage examples of two senses of a lemma were mixed and put together in groups of two, some of which were indeed from the same sense, others which were not. The models are asked to identify whether the two examples represented two different senses of the lemma or only one and the same sense.

In another case of benchmarking, the sentiment value of a sense combined with a user example (manually confirmed to have the same polarity as the sense) is used to test whether the models can identify the polarity of the example (see Pedersen et al., 2024b).

Since the focus is on creating the evaluation datasets, only two models are tested, namely ChatGPT3.5 Turbo and ChatGPT4. In the word-in-context case, ChatGPT 4.0 performs best with 66% correct responses on the polysemous data. In the case of the sentiment datasets, ChatGPT 3.5 Turbo achieves a Pearson coefficient of 69.6%, while ChatGPT 4.0 achieves a coefficient of 81.9% (where the perfect result is a coefficient of 100%).

[3] The Latin terms are not included as entries in DDO, only presented as synonyms to the Danish lemma sense.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Table 2: German verbs with Danish monosemous equivalents in a bilingual dictionary can inherit the formal semantic information in COR.SEM

German verb	equivalent COR.SEM monolingual verb	Frame (Berkeley FrameNet)	Ontological Type	Sentiment
dementieren	*dementere*	Respond_to_proposal | Statement	Act+Communication	-1.0
tagträumen	*dagdrømme*	Cogitation | Desiring	Act+Mental	
blanchieren	*blanchere*	Cooking_creation	Act+Physical+Purpose	
bevormunden	*beskænke*	Offering	Act+Possession	
dezentralisieren	*decentralisere*	Leadership	Act+Social	
desillusionieren	*desillusionere*	Stimulate_emotion	Cause+Mental	-2.0
diskutieren	*diskutere*	Discussion| Judgment_communication	Act+Communication	

Within lexicographic work, COR.SEM data can be applied for many different purposes. It can be used to identify the words and meanings that are most relevant to include in a bilingual dictionary where Danish constitutes the source language. In the case where Danish is the target language, the information in COR.SEM on monosemy (constituting 28,000 lemmas) can be used to identify whether the Danish equivalent in the dictionary is monosemous. If this is the case, the formalised information in COR.SEM, e.g., ontological type, frame value and sentiment value can be automatically transferred to the source language. See Table 2 for examples from German/Danish. The plants in Figure 6 which contain a Latin synonym, is another example.

In monolingual lexicographic work, COR.SEM data can also be useful. At the Society for Danish Language and Literature, we are currently working on developing a DDOLite where we aim at using COR.SEM information as a baseline. The large amount of information in the fine-grained distinctions that we find in DDO can be an obstacle for students, learners of Danish and people with reading deficiencies who want to look up the meaning of a word. In these cases, the coarse-grained sense inventory of COR.SEM can be applied. In connection with the future DDOLite, which targets the above demographics, COR.SEM provides both a valuable data source and a starting point, especially regarding the cases where we condensed the fine-grained DDO senses of a lemmas into fewer senses. An example of this is the adjective *elektrisk* ('electric') in Figure 3. DDO has five senses while COR.SEM reduces this to two: 'regarding electricity' and 'characterised by tension, intensity'. A DDOLite could predictably be built on top of this information baseline, thus avoiding the complexity learners and people with reading deficiencies encounter when visiting DDO today. This new platform would only unfold a lemma in its central senses.

Finally, it should also be mentioned that COR.SEM can be used to study the Danish vocabulary in new ways. For instance, it is possible to gain new insights into the distribution of regular polysemy-patterns (see Figure 5), as well as to look for patterns that are not yet described in the literature. Likewise, combinations of information

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

types in COR.SEM often reveal interesting facts, e.g., that monosemous verbs are most likely to have the ontological type 'Act + Communication' and at the same time negative polarity. Some examples are *bagvaske* ('to slander'), *bebrejde* ('to blame'), *anklage* ('to accuse'), *efterplapre* ('to repeat what somebody says', 'to parrot'), and *formane* ('to admonish'). The site CORSEM.DSL.dk facilitates this type of linguistic and lexical study for researchers, as well as for students and others with an interest in the Danish language.

5. Conclusion

COR.SEM constitutes a highly informative lexical semantic resource for Danish to be used in NLP as well as for lexicographic purposes. It also allows for new types of studies of the Danish language, facilitated by the strictly formalised semantic information types. The webpage corsem.dsl.dk presents the large amount of data and facilitates the studies and use of the content. We aim to expand the lexicon with more lemmas based on the vocabulary in the Danish orthographical dictionary RO which is updated once a year as well as bi-annual DDO updates. We also want to add frequent lemmas which are not yet covered, focusing on simplex words that are often part of composita, some of which are already included in the lexicon.

References

Baker, C. F., Fillmore, Ch. J., & Cronin, B. (2003). The structure of the FrameNet database. *International Journal of Lexicography*, *16*(3), 281–296.

Berkeley FrameNet. Retrieved May 27, 2024 from https://framenet.icsi.berkeley.edu/

DanNet. Retrieved May 27, 2024 from https://wordnet.dk/DanNet

Den Danske Ordbog (DDO) (2009-). Online *ordnet.dk/ddo*. Det Danske Sprog- og Litteraturselskab.

Det Danske FrameNet-leksikon. Retrieved May 27, 2024 from https://korpus.dsl.dk/resources/details/framenet.html

Det Danske Sentiment-leksikon. Retrieved May 27, 2024 from https://github.com/dsldk/danish-sentiment-lexicon

Fellbaum, Ch. (Ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Hjorth, E., & Kristensen, K. (Eds.). (2003–2005). *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab og Gyldendal.

Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S., & Søgaard, A. (2017). From Thesaurus to FrameNet. In I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček, & V. Baisa (Eds.), *Electronic Lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 1–22). Lexical Computing Ltd.

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. Institute for the Croatian Language.

Nimb, S., Lorentzen, H., Troelsgård T., & Theilgaard, L. (2015). *Den Danske Begrebsordbog*. Det Danske Sprog- og Litteraturselskab

Nimb, S., Olsen, S., Pedersen, B. S., & Troelsgaard, T. (2022). A Thesaurus-based Sentiment Lexicon for Danish: The Danish Sentiment Lexicon. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Language Resources and Evaluation Conference: LREC2022,* Vol. 2022 (pp. 2826–2832). Marseille. European Language Resources Association.

Nimb, S., Sørensen, N. H., & Troelsgård, T. (2018). From standalone thesaurus to integrated related words in the Danish Dictionary. In J. Čibej, V. Gorjanc, I. Kosem, & S. Krek (Eds.), *Proceedings from Euralex 2018* (pp. 915–923). Ljubliana, Slovenia. Znanstvena založba Filozofske fakultete Univerze v Ljubljani / Ljubljana University Press, Faculty of Arts

Nimb, S., Trap-Jensen, L., & Lorentzen H. (2014). The Danish Thesaurus: Problems and Perspectives. In A. Abel, C. Vettori & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 191–199). EURAC research

Pedersen, B. S., Braasch, A., Johannsen, A. T., Martinez, H. A., Nimb, S., Olsen, S., Søgaard, A., & Sørensen, S. (2016). The SemDax Corpus. Sense Annotations with Scalable Sense Inventories. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Moreno | J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (pp. 842–847). Paris: European Language Resources Association. https://aclanthology.org/L16-1136.

Pedersen, B. S., Sørensen, N. C. H., Nimb, S., Flörke, I., Olsen, S., & Troelsgård, T. (2022). Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR-Lexicon. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Language Resources and Evaluation Conference: LREC2022,* Vol. 2022 (pp. 51–60). Marseille: European Language Resources Association. https://aclanthology.org/2022.lrec-1.6.

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. *Language Resources and Evaluation, Computational Linguistics Series.* 10.1007/s10579-009-9092-1. 31 s.

Pedersen, B. S., Sørensen, N. C. H., Olsen, S., & Nimb, S. (2024b). Evaluering af sprogforståelsen i danske sprogmodeller – med udgangspunkt i semantiske ordbøger. In *Nydanske Sprogstudier, NyS 65*, 8–40. https://doi.org/10.7146/nys.v1i65.143072

Pedersen, B. S., Sørensen, N. C. S., Olsen, S., Nimb, S., & Gray, S. (2024a). Towards a Danish Semantic Reasoning Benchmark - Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models. In N. Calzolari, M. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of LREC-COLING 2024*, Torino, Italy.

Sørensen, N. C. H., Nimb, S., & Pedersen, B. S. (2023). How do We Treat Systematic Polysemy in Wordnets and Similar Resources? – Using Human Intuition and Contextualized Embeddings as Guidance. In G. Rigau, F. Bond, & A. Rademaker (Eds.), *Proceedings of Global WordNet Conference 2023*, San Sebastián, Spain: Global WordNet Association (pp. 117–126).

This paper is part of the publication: Despot, K. Š., Ostroški Anić, A., & Brač, I. (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress.* Institute for the Croatian Language.

Vossen, P. (Ed.). 1999. *EuroWordNet, A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers. https://doi.org/10.1017/S1351324903223299

Acknowledgements

The database structure of COR.SEM has been developed by Thomas Troelsgård, Det Danske Sprog- og Litteraturselskab.

Contact information

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

Ida Flörke
Society for Danish Language and Literature
if@dsl.dk
Sussi Olsen
Centre for Language Technology, NorS, University of Copenhagen
sol@hum.ku.dk

Bolette S. Pedersen
Centre for Language Technology, NorS, University of Copenhagen
bspedersen@hum.ku.dk

Nathalie C. H. Sørensen
Society for Danish Language and Literature
nats@dsl.dk