The DA-ELEXIS Corpus
- a Sense-Annotated Corpus for Danish with Parallel Annotations for Nine European Languages

Bolette S. Pedersen[1], **Sanni Nimb**[2], **Sussi Olsen**[1], **Thomas Troelsgård**[2],
Ida Flörke[2], **Jonas Jensen**[2], **Henrik Lorentzen**[2]
[1]University of Copenhagen, Centre for Language Technology, NorS,
[2]The Society for Danish Language and Literature
[1]Emil Holms Kanal 2, 2300 Copenhagen, [2]Chr. Brygge 1, 1219 Copenhagen
[1]{bspedersen, saolsen}@hum.ku.dk, [2]{sn,tt,if,jj,hl}@dsl.dk

Abstract

In this paper, we present the newly compiled DA-ELEXIS Corpus, which is one of the largest sense-annotated corpora available for Danish, and the first one to be annotated with the Danish wordnet, DanNet. The corpus is part of a European initiative, the ELEXIS project, and has corresponding parallel annotations in nine other European languages. As such it functions as a cross-lingual evaluative benchmark for a series of low and medium resourced European language. We focus here on the *Danish* annotation process, i.e. on the annotation scheme including annotation guidelines and a primary sense inventory constituted by DanNet as well as the fall-back sense inventory namely The Danish Dictionary (DDO). We analyse and discuss issues such as out of vocabulary (OOV) problems, problems with sense granularity and missing senses (in particular for verbs), and how to semantically tag multiword expressions (MWE), which prove to occur very frequently in the Danish corpus. Finally, we calculate the inter-annotator agreement (IAA) and show how IAA has improved during the annotation process. The openly available corpus contains 32,524 tokens of which sense annotations are given for all content words, amounting to 7,322 nouns, 3,099 verbs, 2,626 adjectives, and 1,677 adverbs.

1 Introduction

Even today in the era of neural language models, high-quality, sense-annotated corpora that are openly accessible prove to be highly requested for the training and evaluation of semantically related NLP tasks, in particular tasks such as word sense disambiguation (WSD) and natural language understanding (NLU).

In spite of numerous initiatives in the field during the last decades, such corpora are still in short supply for many lower-resourced languages, including to some extent the Nordic languages. Two main factors lie at the root of this scarcity:

- Freely available sense inventories (vis-à-vis dictionaries) with a suitable level of sense granularity are often not readily available for the task.
- Even with a suitable sense inventory available, the annotations are extremely costly to compile since they require substantial manpower, preferably from experienced linguists or lexicographers.

The former factor plays a particularly important role, since most curated dictionaries are not open for such use in practice, and since those that *are* available, may not be well-suited for several reasons.

We present here the DA-ELEXIS Corpus, which is one of the largest sense-annotated corpora available for Danish, and the first one to be annotated with the Danish wordnet, DanNet (Pedersen et al., 2009). The corpus is compiled as part of a larger European initiative, the ELEXIS project, (Krek et al., 2018) and corresponding parallel annotations have taken place in nine other European languages. As such it functions as a cross-lingual evaluative benchmark for a series of low and medium resourced European language. For a preliminary presentation of the design of the joint initiative, cf. Martelli et al. (2021). The initiative was led by The Artificial Intelligence Laboratory, Jozef Stefan Institute in Ljubljana and the Department of Computer Science, Sapienza University of Rome; each language group was, however, responsible for their own annotation procedures and sense inventories.

In this paper, we focus mainly on the Danish annotation process, including the Danish annotation scheme and the issues that arose during annotation with regards to calibration and agreement among annotators etc. The corpus is freely available and can be downloaded from CLARIN, www.clarin.si under a CC-BY-SA 4.0 license.

The paper is structured as follows: In Section 2 we give an account on related work, and in 3 we present the corpus and its annotation layers previous to the semantic annotation and provide examples where specific Danish adjustments were required. Section 4 discusses the Danish sense inventories applied for the task, and in Section 5 we describe the annotation process in more detail, whereas Section 6 discusses issues on inter-annotator agreement of the annotations. Finally, in Section 7, we conclude and discuss potential future investigations and development.

2 Related Work

Since the early days of SemCor (Landes et al., 1998), which is one of the first sense-annotated corpora for English based on the Princeton WordNet sense inventory (Fellbaum, 1998), there has been a continuous request in the NLP community for sense-annotated corpora for the world's languages.

Hence, semantic annotation projects have been carried out for a variety of languages; some are based on purely monolingual grounds, while others have adopted different kinds of multilingual approaches. Petrolito and Bond (2014) provides an overview of SemCor corpora and other corpora annotated with wordnets for different languages, and Bentivogli and Pianta (2005) provides more detail on the multilingual SemCor approaches. The Ontonotes corpus (Weischedel et al., 2011) which comprises English, Chinese and Arabic, also uses wordnet as a starting point for its sense annotations of the English part whereas the Chinese and Arab parts base the sense annotations on various lexical sources.

Newer initiatives experiment with semi-automatic approaches to sense annotation in order to overcome the scarcity of such data sets, among others the OneSec corpora created by Scarlini et al. (2020) in five languages, namely English, French, German, Italian, and Spanish. These corpora consist of Wikipedia texts with between 1.2 and 8.8 M sense annotations of nouns per language.

If we look into the Scandinavian languages, the Swedish Eukalyptus corpus (Johansson et al., 2016) is a sense annotated corpus of 100,000 tokens annotated with the senses from the SALDO lexicon (Borin et al., 2013), a Swedish lexical-semantic resource based on a concept of 'centrality' instead of being based on the hyponymy relation, which is a central organisational relation in wordnets. For Norwegian, sense tagging has mostly focused on named entity tags, see e.g. (Jørgensen et al., 2019) whereas a SemCor-like resource does not exist to the best of our knowledge.

In the case of Danish, there have been but a few previous initiatives concerned with sense annotation. Pedersen et al. (2016) presents the SemDax Corpus, which comprises 86,786 tokens, of which the 34,421 content words are sense annotated. It is important to notice, however, that the sense inventory applied in SemDaX refers to the so-called supersense inventory, which is a very coarse-grained, multilingual sense inventory derived from the list of WordNet's first beginners or lexicographical files – corresponding roughly to top-ontological types. In contrast, the DA-ELEXIS Corpus applies a fully-fledged sense inventory derived from monolingual sources, as we will describe in more detail below.

3 The Corpus and its Annotation Layers

DA-ELEXIS consist of 2024 sentences that were extracted from WikiMatrix3 (Schwenk et al., 2019). WikiMatrix3 is an immense open-access collection of parallel sentences derived from Wikipedia covering a diverse set of technical domains from this resource. The WikiMatrix Corpus overall covers Wikipedia articles in 96 languages, resulting in 1620 language pairs. The ELEXIS sense-annotated parallel corpora have been extracted from this collection for 10 European languages, namely Bulgarian, *Danish*, Dutch, English, Estonian, Hungarian, Italian, Portuguese, Slovene, and Spanish (Martelli et al., 2021). Only sentences in English with a counterpart in one of the other nine languages were extracted from WikiMatrix, and missing translations were either retrieved automatically and validated manually or translated manually.

To ensure that the sense annotation (also for multiword entities (MWEs)) can be performed in a flexible and consistent way, the ELEXIS corpus contains a total of five annotation layers, four of which are completed prior to the sense annotation:

- a tokenisation layer
- a sub-tokenisation layer
- a lemmatisation layer, and
- a POS tagging layer

These first four layers were annotated automatically following the Universal Dependency guidelines for each language and afterwards checked manually. See Martelli et al. (2021) for a more detailed account of this annotation process. In other words, before getting to the semantic annotation, several adjustments and decisions needed to take place for each language.

A challenge for Danish was how to deal with compounds, which, as for most Germanic languages, are quite common and relatively dynamically generated, and more importantly: they are written as a single word. We adopted the approach that conventionalized compounds found either in DanNet or in The Danish Dictionary (henceforth DDO) (Hjorth and Kristensen, 2003) should be kept as such, while compounds not found in any of these resources should be split into lemmas included in the resources, in order to enable them to be semantically tagged. When splitting compounds with a binding element, e.g. ‘s’ in *forsøgsperiode* (’trial period’), it was decided to keep the binding element during the sub-tokenisation and POS-tagging phase and to finally remove it in the lemmatization phase.

In several other cases, decisions were required at the POS-tagging level in order to facilitate the semantic tagging. For example in cases when participles are used as adjectives. Participles with adjective entries in the dictionary were lemmatised as such, e.g. *udstrakt* (’outstretched’, fig: ’extensive’), while those that had only verb entries in the dictionary were lemmatised as verbs, e.g. *samlede* (’assembled’, fig: ’total’).

4 Sense Inventories for the Danish Annotation

The sense inventory applied for the annotation, is mainly constituted by the DanNet resource[1] (Pedersen et al., 2009), which is an open-source wordnet compiled semi-automatically on the basis of DDO.

DanNet covers 70,000 Danish lemmas and includes approximately half of the DDO senses from the first, printed edition of the dictionary, mainly from nouns and verbs. It contains a slightly simplified sense inventory, where some DDO senses are collapsed into one when they have been considered very close in meaning. Senses are organised in synonym sets, each one called a synset, which constitute the basic building blocks in a standard wordnet, cf. Fellbaum et al. (eds) 1998. DanNet has taken over the sense definitions and usage examples from DDO, but due to copyright retrictions, definitions are given in an abbreviated form where only the first 50 characters are represented.

In contrast to DDO, DanNet is open-source (CC BY-SA 4.0) allowing the sense-annotated corpus to be freely used and integrated in all kinds of pipelines and applications. This was a prerequisite defined by the ELEXIS project for participating in the annotation task.

Since not all senses are covered in DanNet, a current online version of DDO was used as fallback and new senses from this resource were established via the annotation tool when required. This version of the dictionary covers approximately 100,000 lemmas and 150,000 senses and is continuously updated and published online since 2009 at ordnet.dk/ddo.

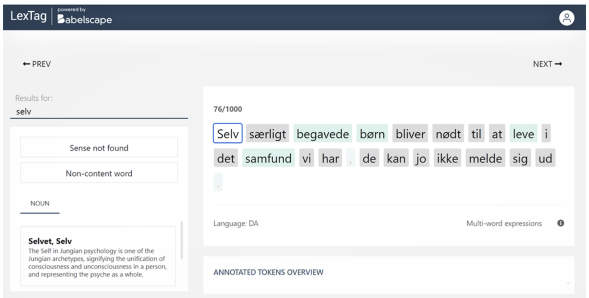
5 The Sense Annotations

5.1 The Annotation Tool

Due to the complex requirements of the many languages involved, a web-based annotation tool, LexTag, was developed for the sense annotation by the company Babelscape. As shown in Figure 1, the tool brings the annotator through each token in the sentence and presents all available senses for the token in question to the annotator.

In the case where a sense is not present in any of the resources, the annotator can add a new sense directly in the tool, including a definition. The new sense is thereafter part of the existing sense inventory – for other annotators to use.

[1] Link to the DanNet resource: `https://repository.clarin.dk/repository/xmlui/handle/20.500.12115/24`



Selv særligt begavede børn bliver nødt til at leve i det samfund vi har , de kan jo ikke melde sig ud .

Selvet, Selv
The Self in Jungian psychology is one of the Jungian archetypes, signifying the unification of consciousness and unconsciousness in a person, and representing the psyche as a whole.

Figure 1: LexTag annotation tool for sense annotation

The tool also facilitates the encoding of MWE with a single semantic label, as in the case of e.g. phrasal verbs (*spise op* – 'eat up'), which are very frequent in Danish. When MWEs occur discontinuously in the corpus, the tool also enables flexible annotation, as in *spise frokosten op* – lit: 'eat the lunch up'. In such cases, the entire MWE is subsequently looked up in the lexical resources.

5.2 The Annotators

The annotations were completed by seven different annotators, all experienced traditional lexicographers and/or computational lexicographers.

5.3 The Annotation Guidelines

Annotation guidelines were developed across language groups during the first annotation phase. Several zoom meetings among partners were required to achieve consensus on the most basic annotation principles to be used. In the Danish group, however, further language-specific guidelines were compiled in collaboration based on the first rounds of annotation of Danish. These included principles on defining word classes in unclear cases, on when to consider something a MWE, on when and how to enter new senses to the tool, etc.

5.4 Annotation Issues and Amendments to the Sense Inventory

Figure 2 and 3 illustrate how well the available sense inventories covered the corpus at sense level and token level, respectively. Overall, it can be seen that DanNet covers quite well, but that DDO has been consulted in more than 20 % of the cases due to missing senses in DanNet. In 2.5 % of the annotated examples, a completely new sense had to be established given that it was not found in any of the existing resources. In Figure 3 we observe 5% non-content words. These are words that were originally pos-tagged as content word, but which during the sense tagging process were found to be non-content words and thus were not tagged with a sense.

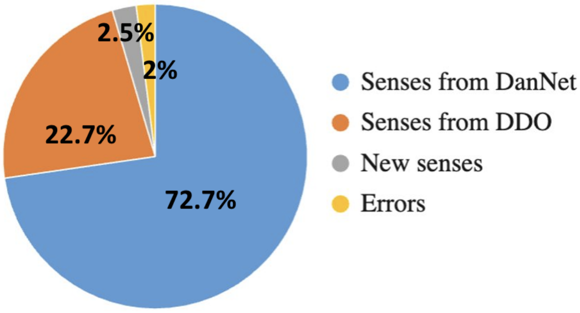


Figure 2: Distribution of lexical resources used for annotating calculated at sense level (excluding proper nouns)

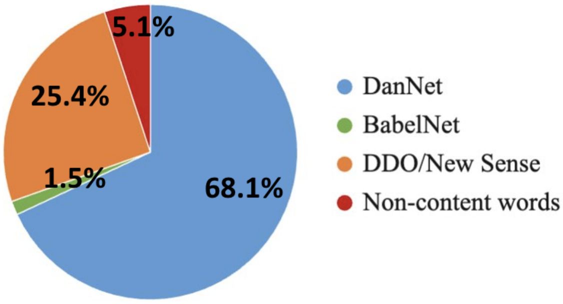


Figure 3: Distribution of lexical resources used for annotation calculated at token level (including proper nouns from BabelNet)

If we take a closer look at the annotations, around 4000 different DanNet senses came into use while 1500 other senses occurring in the corpus were not covered by DanNet. In these cases, the sense definitions were established on the fly by consulting the DDO online dictionary. However, in a number of cases (approx 250 cases) no suitable sense description was found in either resource. In some of these cases (110 senses), a sense could not be identified due to POS tagging errors in the corpus, i.e. the POS diverged from the one given in DanNet/DDO, in other cases a compound had not been split correctly at the syntactic level and could therefore not be identified in the dictionaries. Only in the remainder of the cases (140 examples), the available sense inventories proved to be insufficient.

Approximately half of the senses not covered by DanNet only appear once in the corpus, while the other half is represented two or more times. 125 of these occur at least five times, and 15 senses more than 20 times, e.g., two senses of the verb *være* (‘to be’) as well as one sense of the adjective *stor* (‘big’).

The annotation task gives very useful feedback regarding the vocabulary and sense inventory of both DanNet and the DDO. As expected, we find many adjective and adverb senses among the senses that are not covered by DanNet. This relates to the fact that adverbs are not part of DanNet, and that only a subset of adjectives were given priority when compiling the resource. Among senses occurring more than once, 70 are adverbs, and only six of these are covered in DanNet, typically in the form of an adjective. 71 are adjectives, and in half of the cases the lemma is not included in DanNet. Of the 125 most frequent senses (five or more) we find 41 adverbs and 33 adjectives, and four adjectives and 5 adverbs are among the top 15, the negation adverb *ikke* (‘not’) being the most frequent with 101 occurrences.

More surprisingly, we discovered that of the 76 verb senses missing in DanNet, only nine were down to a lack of the lemma itself, the rest being down to missing verbal senses of an already existing verb. In the case of the 33 noun senses occurring more than once, 13 are not lemmas in DanNet, the rest represent a sense which is not included in spite of the fact that the noun in question is already included in the WordNet. This gives us useful feedback on which senses and lemmas to add to DanNet in the future.

Also in the case of the DDO, useful feedback was provided. By making use of the corpus-based dictionary (covering more than 100,000 lemmas) as the default backup lexical resource, we expected to cover a very high percentage of the lemmas and their senses in the corpus, also given the fact that ad hoc compounds would be split beforehand at the syntactic level. This turned out to be correct. However, 3 % of the lemma senses occurring twice or more, were not described in the DDO. As expected, the number is higher for rarer senses (those that only occur once in the corpus): Approximately 20 % of these are not represented in our lexical resources. Sometimes for good reason since the DDO focuses on general language. Highly domain specific lemmas and senses such as *bro* ‘bridge’ in the sense ‘geometric figure that connects two things’, *aurora* in the sense ‘northern and southern lights’), as well as the lemmas *cefalexin* (a form of medicine) and *cleveit* (a mineral), are therefore not found in DDO. We also see cases where the lemma in DDO only contains morphological, not semantic information, e.g. in the cases of *rabbinsk* (‘rabbinical’, *tektonisk* (‘tectonic’), and *underudvalg* (‘subcommittee’).

Still, in spite of these explanations, a surprisingly large part of the missing lemmas are candidates to be included in the DDO, e.g. *affaldsindsamling* (‘waste collection’), *adfærdsmæssig* (‘behavioural’), *1980’erne* (‘the 1980s’), and *cloudbaseret* (‘cloud based’). All in all, a list of around 100 good lemma candidates for both DanNet and the DDO are identified through the annotation task.

A lesson learned was also the fact that POS annotations in the corpus should be calibrated well with the POS information of the lexical resources. It may seem surprising, but actually agreeing on part of speech is not as evident as one may expect. In fact, 10 % of the lemma senses which could be directly linked to a DanNet or DDO sense, had a diverging POS annotation in the corpus, e.g. *3D* is tagged as an adjective in the corpus, but as a prefix *tre-d-* in the DDO; and *beregnet* is tagged as an adjective in the corpus, but is explained in the DDO as a fixed expression *beregnet på* (‘intended for’) in the entry of the verb *beregne* (‘calculate’) in the DDO).

6 Agreement among Annotators

It is one thing whether a sense is actually described in the sense inventory at hand, another, however, is to what extent annotators agree on which sense tags to use for a given example. To study this aspect, we have, in accordance with consensus for semantic annotations tasks, triple annotated a little over 5% of the corpus, amounting to 108 sentences. This triple annotation has enabled us to calculate inter-annotator agreement and examine differences among annotators.

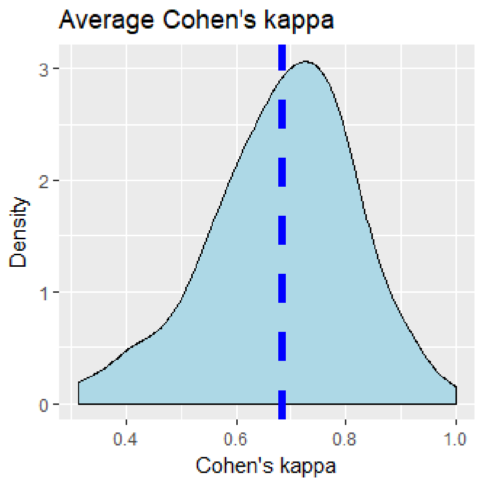


Figure 4: Average of agreement between all three annotators. The more the plot tends to slant towards the right, the more agreement there is

Inter-coder agreement reveals interesting things about several aspects of the annotation task. A number of issues come into play, such as:

- the pre-processing of the corpus, e.g. whether there is an overall agreement on the POS tag set and on how to employ it,
- the coverage of the sense inventory,
- the granularity of the sense inventory (fine-grained or coarse-grained),
- the depth of the annotation guidelines, and finally
- the overall proficiency of the annotators

We calculate an average Cohen's kappa agreement for the triple annotated data of 0.68 between all three annotators as seen in Figure 4. We would have liked also to employ the Krippendorph alpha measure, which takes into account the fact that it is generally easier to agree on few labels than on many, but this measure proved impossible to calculate in practice since it requires a full list of all possible senses (including MWEs) that every word can occur in, and such MWE lists are not provided in DanNet.

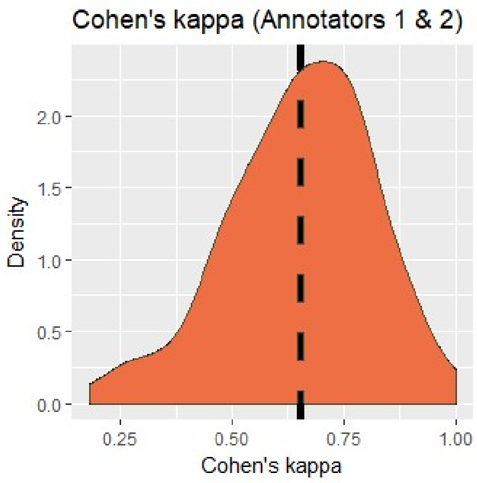


Figure 5: Average of agreement between an annotation made early in the project, and one done late in the project)

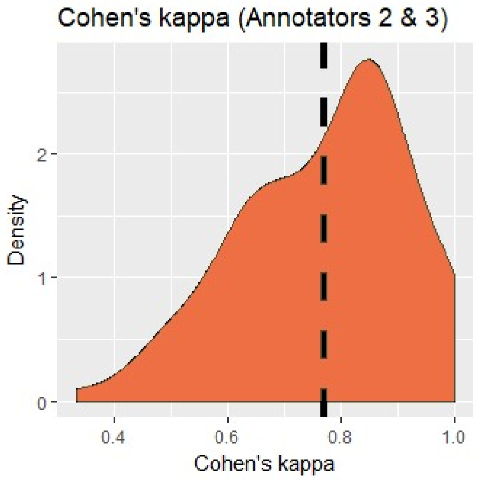


Figure 6: Average of agreement between two annotations made late in project

However, since the first annotation of the 108 sentences was performed in the first phase of the annotation project (corresponding to annotator 1 in Figure 5), several issues regarding delimitations of MWE, diverging POS tags, and problems with the encoding of compounds etc. had not yet been clarified. The two other annotations (corresponding to annotators 2 and 3 in Figure 5 and 6) were performed at the end of the project when more or less all pending issues had been clarified. Here we achieve an inter-coder agreement of 0.78. This can be considered a quite substantial level of agreement for this kind of task.

As revealed by the bullet points above, most issues come into play in the diverging annotations. As already touched upon, *preprocessing* proves to have caused some divergences, in particular in relation to MWE, and since these are relatively frequent, in particular for verbs (phrasal verbs), we positively know that this has caused divergences in a number of cases. With regards to *coverage*, it can be assessed to be relatively good if we see DanNet and DDO all together even if a small percentage of the examples could not be tagged with existing senses.

The *granularity* of our sense inventory is quite high as is the case for most dictionaries, in fact we have an average of 4.2 senses per lemma in the DDO overall. We can see that the fine granularity causes agreement problems in some cases, in particular when distinguishing between a high number of verbal senses.

In spite of problems with fine-grained senses, differences in agreement early and late in the project show that the *guidelines* improved substantially after the first round. Finally, the annotators involved in the task were professional lexicographers or computational linguists with overall *high proficiency* in annotating, and no annotators proved to protrude in the overall quality of their annotations compared to others.

7 Conclusions and Future Work

In this paper, we have presented the DA-ELEXIS Corpus, which is one of the largest sense-annotated corpora available for Danish and the first one to be annotated with the Danish wordnet, DanNet. We have described the careful preprocessing and preparation necessary to ensure a high quality of the resulting resource, and we have presented a series of difficulties in relation to sense coverage and in achieving a high inter-annotator agreement. In particular, the limitation and semantic description of MWE have proven to cause divergences among annotators even if these were reduced during the development of the annotation guidelines.

Even if we know that the *granularity* of our sense inventory is also crucial to the inter-annotator agreement, we have chosen not to go too deeply into the theoretical discussion of this issue here. This is in spite of the fact that sense coverage and granularity is highly discussed in lexical semantic literature (Cruse (1986), Fillmore and Atkins (1992) and many others) and that some even claim that word senses don't exist, or at least only relative to specific tasks (Kilgarriff, 1997). Along the same lines, fine-grained sense inventories (like the ones we have applied here) have been deemed somewhat unsuitable for NLP tasks such as WSD and NLU. In this context, we are currently examining ways to achieve a high-quality, *coarse-grained* sense inventory for Danish, since we are building a new lexical resource, the Central Word Register of Danish, COR, particularly for NLP, (Nimb et al., 2022). Still based on – and linked to – the same sources, namely DanNet and the DDO, we are developing principled and semi-automatic procedures for reducing the inventory with more than 40 % (Pedersen et al., 2022). Thus, we expect to have a more suitable inventory for sense annotation available in future projects, probably by the end of 2023.

We also plan to experiment with automatically added semantic and thematic information to the corpus, based on the available information in different semantic lexicons linked at sense level to the DDO and DanNet, including thesaurus information on topics and themes.

Finally, another aspect that deserves further attention in future work is the potential of the *multilingual* setup of the ELEXIS corpus. Having parallel sense annotations in nine aligned languages, several of them being generally low-resourced, provides valuable information not only for each individual language (Danish, in our case), but also for cross-lingual studies in NLP and lexicography. This provides us with a valuable cross-lingual evaluation benchmark to be applied for future WSD and NLU tasks.

Acknowledgements

Thanks to all of our ELEXIS colleagues for the fruitful collaboration on the ELEXIS corpora, in particular to the Artificial Intelligence Laboratory, Jozef Stefan Institute (Simon Krek and Jaka Čibej) and Department of Computer Science, Sapienza University of Rome (Federico Martelli and Roberto Navigli).

References

Luisa Bentivogli and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Natural Language Engineering*, 11:247–261.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. Saldo: a touch of yin to wordnet's yang. *Language Resources and Evaluation*, 47.

David Alan Cruse. 1986. *Lexical semantics*. Cambridge university press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. The MIT Press.

Charles J Fillmore and Beryl T Atkins. 1992. Toward a frame-based lexicon: The semantics of risk and its neighbors. *Frames, fields and contrasts: New essays in semantic and lexical organization*, 75:102.

Ebba Hjorth and Kjeld Kristensen, editors. 2003. *Den danske ordbog, bd. 1-6*, volume 6. Gyldendal. Isbn=samlede værk bd.1-6.

Richard Johansson, Yvonne Adesam, Gerlof Bouma, and Karin Hedberg. 2016. A multi-domain corpus of Swedish word sense annotation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3019–3022, Portorož, Slovenia. European Language Resources Association (ELRA).

Fredrik Jørgensen, Tobias Aasmoe, Anne-Stine Ruud Husevåg, Lilja Øvrelid, and Erik Velldal. 2019. NorNE: Annotating Named Entities for Norwegian.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Simon Krek, Iztok Kosem, John P. McCrae, Roberto Navigli, Bolette S. Pedersen, Carole Tiberius, and Tanja Wissik. 2018. European lexicographic infrastructure (elexis). In *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts*, pages 881–891, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Shari Landes, Claudia Leacock, and Christiane Fellbaum. 1998. *Building Semantic Concordances*, pages 199–216. The MIT Press.

Federico Martelli, Roberto Navigli, Simon Krek, Carole Tiberius, Jelena Kallas, Polona Gantar, Svetla Koeva, Sanni Nimb, Bolette Sandford Pedersen, Sussi Olsen, Margit Langements, Kristina Koppel, Tiiu Üksik, Kaja Dobrovolijc, Rafael-J. Ureña-Ruiz, José-Luis Sancho-Sánchez, Veronika Lipp, Tamas Varadi, András Györffy, Simon László, Valeria Quochi, Monica Monachini, Francesca Frontini, Rob Tempelaars, Rute Costa, Ana Salgado, Jaka Čibej, and Tina Munda. 2021. Designing the elexis parallel sense-annotated dataset in 10 European languages. In *eLex 2021 Proceedings*, eLex Conference. Proceedings. Lexical Computing CZ. Null ; Conference date: 05-07-2021 Through 07-07-2021.

Sanni Nimb, Bolette S. Pedersen, Nathalie Carmen Hau Sørensen, Ida Flörke, Sussi Olsen, and Thomas Troelsgård. 2022. COR-S – den semantiske del af det centrale ordregister (cor). *LexicoNordica*, 29.

Bolette Pedersen, Nathalie Carmen Hau Sørensen, Sanni Nimb, Ida Flørke, Sussi Olsen, and Thomas Troelsgård. 2022. Compiling a suitable level of sense granularity in a lexicon for AI purposes: The open source COR lexicon. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 51–60, Marseille, France. European Language Resources Association.

Bolette Sandford Pedersen, Anna Braasch, Anders Trærup Johannsen, Hector Martinez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard, and Nicolai Sørensen. 2016. The SemDaX Corpus - sense annotations with scalable sense inventories. In *Proceedings of the 10th conference of the Language Resources and Evaluation Conference*, pages 842–847. European Language Resources Association.

Bolette Sandford Pedersen, Sanni Nimb, Jørg Asmussen, Nicolai Hartvig Sørensen, Lars Trap-Jensen, and Henrik Lorentzen. 2009. Dannet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43(3):269–299.

Tommaso Petrolito and Francis Bond. 2014. A survey of WordNet annotated corpora. In *Proceedings of the Seventh Global Wordnet Conference*, pages 236–245, Tartu, Estonia. University of Tartu Press.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sense-annotated corpora for word sense disambiguation in multiple languages and domains. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5905–5911, Marseille, France. European Language Resources Association.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2019. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*, 3(3):3–4.