Sanni Nimb, Nathalie C. H. Sørensen, and Jonas Jensen

MAKING DANISH THESAURUS DATA AVAILABLE TO RESEARCHERS

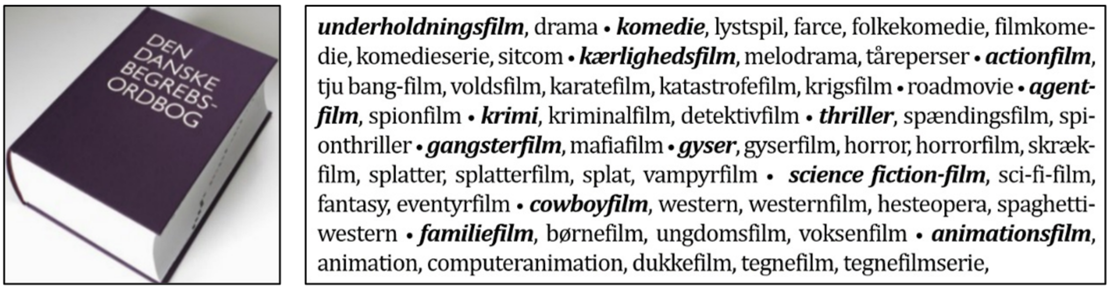
The WebDDB project

Abstract This study presents a project aiming to make thesaurus data available under an academic licence. The project is based on the printed thesaurus Den Danske Begrebsordbog (DDB) which covers approx. 80% of the Danish dictionary DDO (ordnet.dk/ddo). It presents more than 100,000 different words and expressions categorised and ordered semantically in 22 thematic chapters, and 888 named sections. The data is now downloadable at a webpage where it can be supplemented with different types of lexical information from other resources of choice, e.g., information on valency, etymology, or ontological type. The supplementation is possible due to shared sense id-numbers between the lemmas in the digital thesaurus manuscript, the Danish online dictionary DDO, the semantic lexicon COR.SEM, and a WordNet (DanNet). The webpage allows for new types of studies of the Danish vocabulary with semantic similarity as the starting point. As part of the project, more lemmas from the DDO were added to the digital manuscript which today covers 95% of the dictionary. The vocabulary as well as certain sections and lemmas denoting nationality, sexual orientation, gender identity etc. are thoroughly revised due to the change of attitudes towards this vocabulary in the last decade.

Keywords thesaurus; linked data; semantics; webpage

1. Introduction

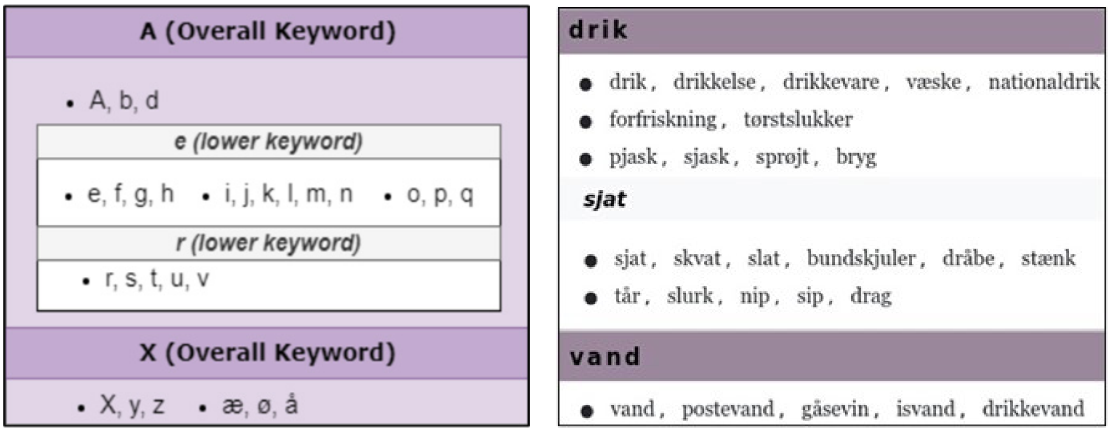
The print version of the Danish thesaurus *Den Danske Begrebsordbog* (Nimb et al., 2014, Nimb et al., 2014/2015) covers of approx. 80% of the Danish dictionary DDO (ordnet.dk/ddo). It contains more than 100,000 different words and expressions categorised in 22 thematic chapters (e.g., 'Arts and Culture' and 'Sports and Leisure'), and 888 named sections (e.g., 'Films and Cinema' and 'Football'). In Figure 1., a selection of nouns from 'Films and Cinema' is shown.


DEN DANSKE BEGREBS-ORDBOG

underholdningsfilm, drama • ***komedie***, lystspil, farce, folkekomedie, filmkomedie, komedieserie, sitcom • ***kærlighedsfilm***, melodrama, tåreperser • ***actionfilm***, tju bang-film, voldsfilm, karatefilm, katastrofefilm, krigsfilm • roadmovie • ***agentfilm***, spionfilm • ***krimi***, kriminalfilm, detektivfilm • ***thriller***, spændingsfilm, spionthriller • ***gangsterfilm***, mafiafilm • ***gyser***, gyserfilm, horror, horrorfilm, skrækfilm, splatter, splatterfilm, splat, vampyrfilm • ***science fiction-film***, sci-fi-film, fantasy, eventyrfilm • ***cowboyfilm***, western, westernfilm, hesteopera, spaghettiwestern • ***familiefilm***, børnefilm, ungdomsfilm, voksenfilm • ***animationsfilm***, animation, computeranimation, dukkefilm, tegnefilm, tegnefilmserie,

Fig. 1: Nouns from the section Films and cinema: "**entertainment movie**, drama • **comedy**, farce, folk comedy, sitcom • **romance**, melodrama, tearjerker • **action movie**, high-octane movie, violent film, karate film, disaster film, war film • road movie • **agent thriller**, spy film • **crime film**, detective film • **thriller**, suspense drama, espionage thriller • **gangster film**, mafia film • **horror**, horror film, slasher, slasher film, gore, vampire film • **science fiction film**, sci-fi film, fantasy, adventure film • **western**, western movie, western film, oat opera, spaghetti western • **family film**, children's film, teen movie, adult film • **animated film**, animation, computer animation, puppet film, cartoon, cartoon series

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.



drik

- drik, drikkelse, drikkevare, væske, nationaldrik
- forfriskning, tørstslukker
- pjask, sjask, sprøjt, bryg

sjat

- sjat, skvat, slat, bundskjuler, dråbe, stænk
- tår, slurk, nip, sip, drag

vand

- vand, postevand, gåsevin, isvand, drikkevand

Fig. 2: The structure of the DDB thesaurus is shown to the left. Each letter represents a word. To the right, the nouns *drik* ('beverage', overall keyword), *sjat* ('small amount of a beverage', lower keyword), and *vand* ('water', overall keyword), all three followed by groups of near-synonymous nouns, illustrate the structure

The words are listed in semantic order in groups of persons, artefacts, acts, properties etc., some of which are initiated by a highlighted word which thereby functions as a keyword and headline to the following vocabulary, see Figure 2. The hierarchical structure provides valuable formalised knowledge of the varying degrees of semantic similarity between senses.

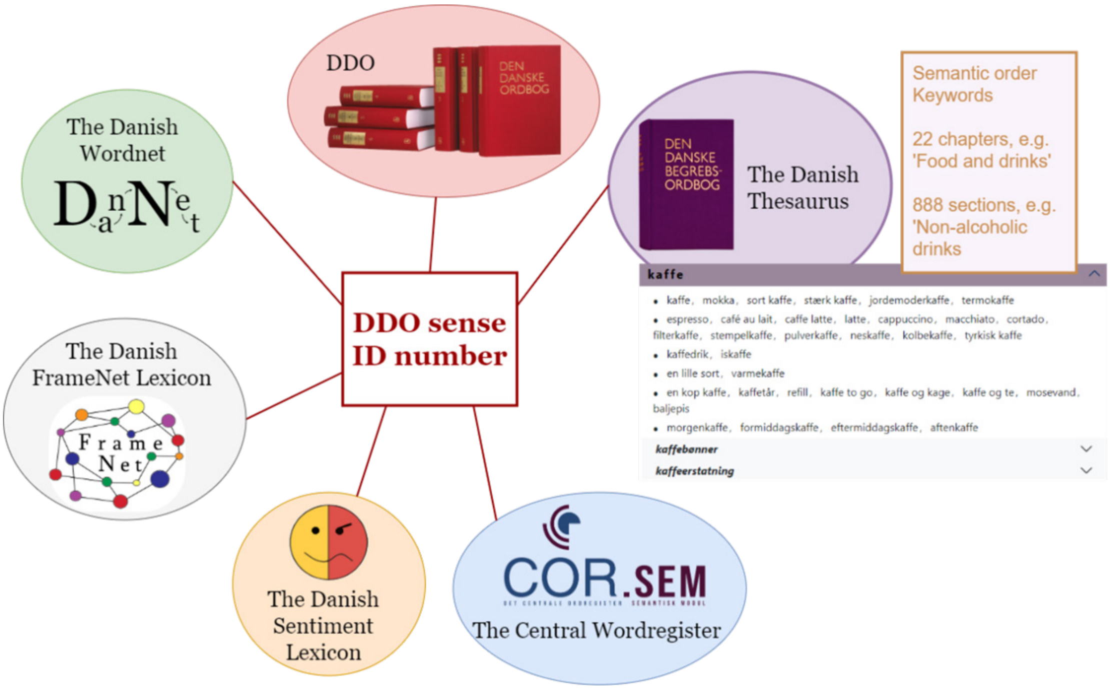
The structured and semantically ordered DDB data has proved useful in several internal projects due to links at sense level from the thesaurus to the DDO. In section 3 we describe the linked data and use of the thesaurus in different tasks and, in section 4, the new webpage. The next section focuses on the extension and review of the thesaurus manuscript before the online publication.

2. Review of the DDB Thesaurus 10 Years After: Additions and Outdated Section Vocabulary

Following the 2014/15 publication of the print version of the DDB, more than 20,000 words and expressions have been added to the digital manuscript. Within the same time span, language attitudes have changed considerably due to an increased focus on equality in society, having the consequence that in some cases the vocabulary, usage labels, keywords and section names have become outdated. To secure the quality and integrity of the data on the webpage we therefore took the opportunity to review the manuscript. This review brought attention to a number of larger issues, e.g., the discussion of whether to include obsolete words such as *dragesæd* (a rare term of ancient Greek origin meaning 'hostility' or 'conflict') and whether *plantemælk* (~plant milk) would be an appropriate keyword following a 2017 ruling by the Court of Justice of the EU which prohibits the use of the word 'milk' in the advertising and marketing of purely plant-based products. In the first case, *dragesæd* has been removed. We aim at covering only modern Danish corresponding to the DDO vocabulary (1950 until today). In the latter case, the synonym *plantedrik* ('plant-based drink') has been upgraded to keyword instead. In the thesaurus, we prefer neutral keywords, i.e., politically correct and non-derogatory keywords, whenever possible. The reviewers were instructed to scour existing sections for potential incongruities, e.g., instances in which derogatory words or racial slurs featured next to neutral or unmarked terms denoting nationality, sexual orientation, gender identity, etc. In one such case, *trans*, a derogatory word denoting ‘crossdresser’, has been removed from the section ‘Dislike, antipathy’) and is now only part of the section ‘Clothes’. Likewise, the word *mobbeoffer* (’victim of bullying’) was removed from the section ‘Ridiculous’ where it was grouped with derogatory words such as *fjog* (‘fool’), *nar* (‘jester’), *jubelidiot* (‘jubilant idiot’), and *klovn* (‘clown’). On the webpage, it is only presented in the section ‘Sarcasm, mockery’. Furthermore, a few section titles have been changed in the new version of the thesaurus data, e.g., the section ‘Ligestilling’ (‘Gender Equality’) is renamed ‘Ligestilling, diskrimination’ (‘Gender Equality, Discrimination’).

3. Linked Data

All words in the DDB thesaurus are linked to the corresponding DDO sense. The linked thesaurus and dictionary data have proven useful in several respects.


DDO
DEN DANSKE ORDBOG
The Danish Wordnet
DaNet
The Danish FrameNet Lexicon
FrameNet
DDO sense ID number
DEN DANSKE BEGREBS-ORDBOG
The Danish Thesaurus
Semantic order
Keywords
22 chapters, e.g. 'Food and drinks'
888 sections, e.g. 'Non-alcoholic drinks
The Danish Sentiment Lexicon
COR.SEM
The Central Wordregister

kaffe

- kaffe, mokka, sort kaffe, stærk kaffe, jordemoderkaffe, termokaffe
- espresso, café au lait, caffe latte, latte, cappuccino, macchiato, cortado, filterkaffe, stempelkaffe, pulverkaffe, neskaffe, kolbekaffe, tyrkisk kaffe
- kaffedrik, iskaffe
- en lille sort, varmekaffe
- en kop kaffe, kaffetår, refill, kaffe to go, kaffe og kage, kaffe og te, mosevand, baljepis
- morgenkaffe, formiddagskaffe, eftermiddagskaffe, aftenkaffe

kaffebønner

kaffeerstatning

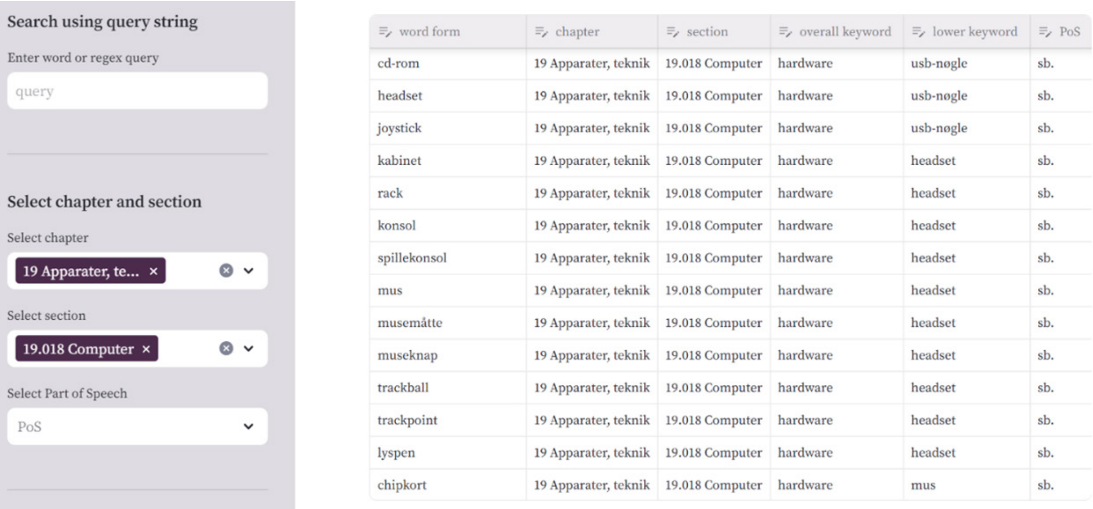
Fig. 3: The linked dictionaries and lexicons. The DDO dictionary and the DDB thesaurus are compiled by DSL and DanNet, the FrameNet lexicon, the sentiment lexicon and COR.SEM by DSL and CST, the University of Copenhagen

It allowed for the compilation of a FrameNet lexicon, based on thematic sections and semantic groups in the thesaurus combined with the valency patterns in the DDO which were assigned frames from Berkeley FrameNet (Nimb et al., 2017; Baker et al., 2003). A sentiment lexicon was developed from the vocabulary in negative and positive DDB sections combined with DDO information on style labels (Nimb et al., 2022). Recently the COR.SEM lexicon has been compiled based on the links (Pedersen et al., 2021; Nimb et al., 2024). Altogether, the resources constitute a lexical network at sense level which also includes the Danish WordNet DanNet (Pedersen et al., 2009), see Figure 3. The links between the thesaurus and DDO are also used in the online DDO to present extracts of thesaurus section data based on the structure and keyword scope (Nimb et al., 2018). Recently, datasets based on the thesaurus have been used as LLM evaluation data as part of a Danish semantic reasoning benchmark (Pedersen et al., 2024). Based on the thesaurus where the closest word is the synonym (neighbour word), the next words are from the same section or chapter, and the word furthest away is from a different chapter, the task is to select the correct synonym or to identify the outlier out of 4 words (see also Nielsen & Hansen, 2017, and Camacho-Collados & Navigli, 2016).

4. The Webpage

The thesaurus data is available on a purpose-built webpage webddb.dsl.dk to be published in December 2024. We have used the semantic links between the DDB manuscript and the lexical resources to precompile a selection of datasets which can be combined with the selected vocabulary and downloaded under a research licence. Prior to this, a group of researchers specified the combined data which would be relevant in their respective research projects. For instance, linguists were interested in semantic groups combined with the DDO information on derivation patterns of the lemmas (verbs and the corresponding verbal nouns) to be able to study whether semantics influence patterns. Literary researchers showed interest in thesaurus section names for the purpose of automatic topic detection. In addition, formalised data (e.g., the thesaurus structure linked to additional formal information such as ontological type) was of interest to data scientists aiming to carry out automatic sense disambiguation. Format and other technical issues were also discussed.

On the webpage, researchers can access the data either by searching for words with a query string or by selecting a chapter and then a section to retrieve all words within that section. The data is shown as a table with the default columns: word form, chapter, section, overall keyword, lower keyword, and Part of Speech (see Figure 4).



Search using query string

Enter word or regex query

query

Select chapter and section

Select chapter

19 Apparater, te... ×

Select section

19.018 Computer ×

Select Part of Speech

PoS

word form	chapter	section	overall keyword	lower keyword	PoS
cd-rom	19 Apparater, teknik	19.018 Computer	hardware	usb-nøgle	sb.
headset	19 Apparater, teknik	19.018 Computer	hardware	usb-nøgle	sb.
joystick	19 Apparater, teknik	19.018 Computer	hardware	usb-nøgle	sb.
kabinet	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
rack	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
konsol	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
spillekonsol	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
mus	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
musemåtte	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
museknap	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
trackball	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
trackpoint	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
lyspen	19 Apparater, teknik	19.018 Computer	hardware	headset	sb.
chipkort	19 Apparater, teknik	19.018 Computer	hardware	mus	sb.

Fig. 4: The search interface, webddb.dsl.dk

Researchers can also select additional columns with information retrieved from either the DDO or COR.SEM. Furthermore, each column can be filtered for specific values so that researchers can easily select a subset of the data - for instance by filtering for verbs in PoS and then a specific valency pattern. The available information is shown in Table 1.

Examples of filtered subsets of data are near synonym verbs in DDB with the sense 'to think', combined with information in DDO on which language they are borrowed from: ræsonnere ('to reason'): French and Latin; deducere ('to deduce'): Latin; konkludere ('to conclude'): Latin; tænke ('to think'): Middle Low German; forestille sig ('to imagine'): Low German or German.

Table 1: Information types to be added to thesaurus groups of your own choice, webddb.dsl.dk

Information type	**Source**	**Example**
Grammar: valency/ construction pattern	DDO	verb *smage* ('to taste'): NGN *smager* NGT ('somebody tastes (something)'); NGN *smager på* NGT ('somebody tastes Ø something')
Etymology: borrowed from which language(s)?	DDO	verb *smage* ('to taste'): nedertysk ('low German') *smaken*
Etymology: dating (year)	DDO	noun *computer*: 1957
Morphology: Derivation patterns	DDO	verb *smage*: noun *smager* 'a taster', noun *smagning* ('tasting')
Sentiment value	COR.SEM	verb *lugte* 'to smell': '-1', negative polarity
Ontological type	COR.SEM	verb *smage* ('to taste'): Experience+Physical
Semantic frame (values based on Berkeley FrameNet)	COR.SEM	verb *smage* ('to taste'): frame=Tasting
Synset in the Danish WordNet	DanNet	https://wordnet.dk/dannet/data/synset-72897

Another example is near synonymous words for the noun computer ('computer') combined with information on when the words first appeared in Danish (provided that this information is available in DDO): datamaskine: 1953; elektronhjerne (lit. 'electron brain'): 1954; computer: 1959; datamat: 1966; terminal: 1968; lommeregner (lit. 'pocket calculator'): 1970; pc: 1984; laptop: 1990; client/server: 1991; desktop: 1993. Lastly, the table shown on the webpage can also be downloaded as a tsv-file to save specific searches for further studies and applications.

5. Conclusion

The thesaurus webpage will be accessible to researchers in December 2024. Access to the lexical data under an academic licence allows for a wide range of new applications in the research community, either based on standalone thesaurus data or combined with data from DDO and formal semantic lexicons. DSL is also going to publish a second version of the thesaurus in 2025/2026 based on the extended and revised manuscript. Unlike the print version from 2014, this will be published in the form of an open access online thesaurus linked to DDO and containing approx. 20% more lemmas and senses.

References

Baker, C. F., Fillmore, C. 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/

Camacho-Collados, J., & Navigli, R. (2016). Find the word that does not belong: A Framework for an Intrinsic Evaluation of Word Vector Representations. In *Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP* (pp. 43–50). Association for Computational Linguistics.

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

Hjorth, E., & Kristensen, K. (Eds.). (2003–2005). *Den Danske Ordbog*, Society for Danish Language and Literature & Gyldendal.

Nielsen, F. Å., & Hansen, L. K. (2017). Open semantic analysis: The case of word level semantics in Danish. In *Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 415–419).

Nimb, S., Braasch, A., Olsen, S., Pedersen, B. S., & Søgaard, A. (2017). From Thesaurus to FrameNet. In *Electronic Lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 1–22).

Nimb, S., Flörke, I., Olsen, S., Pedersen, B. S., & Sørensen, N. C. H. (2024). COR.SEM, a new formal semantic lexicon for Danish. In K. Š. Despot, A. Ostroški Anić, & I. Brač, *Proceedings of the 21st EURALEX International Congress, Lexicography and Semantics* (this issue).

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 Proceedings of the Language Resources and Evaluation Conference: LREC2022* (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 of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 915–923). Ljubljana: 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).

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, Computational Linguistics Series.* 10.1007/s10579-009-9092-1. 31 s.

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 *Proceedings of the Language Resources and Evaluation Conference: LREC2022* (pp. 51–60). Marseille: European Language Resources Association.

Pedersen, B. S., Sørensen, N. C. S., Olsen, S., Nimb, S., & Gray, S. (2024). Towards a Danish Semantic Reasoning Benchmark – Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models. In *Proceedings of LREC-COLING 2024*, Torino, Italy.

Contact information

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

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

Jonas Jensen
Society for Danish Language and Literature (DSL)
jj@dsl.dk