Lexical-Semantic Resources as a Culture-Aware Basis for Benchmarking and Evaluation of LLMs

Nathalie Norman [1], Sanni Nimb [2], Sussi Olsen [1],

Nina Schneidermann[1], Bolette S. Pedersen[1],

[1] University of Copenhagen, Njalsgade 80, DK-2300 S
[2] The Society for Language and Literature, Christians Brygge 1, DK-1219 K

E-mail: naha@hum.ku.dk, sn@dsl.dk, saolsen@hum.ku.dk, ninasc@hum.ku.dk,

bspedersen@hum.ku.dk

Abstract

Large Language Models (LLMs) tend to expose severe language and cultural biases when working in medium- and low-resourced languages. In this paper, we present our work on Danish benchmarking and evaluation of LLMs to more precisely diagnose and potentially remedy such bias. To this aim, we apply available lexical-semantic resources to compile a set of Natural Language Understanding (NLU) tasks in Danish that reflect the breadth and nuances of the Danish vocabulary, thereby capturing also implicit traits of Danish values and culture. Currently the benchmark comprises nine NLU tasks, including tasks such as disambiguating words in context, determining semantic outliers, inferencing and interpretation tasks based on semantic relations, as well as selecting the correct explanation of culture-related metaphorical idioms. The large-scale benchmark (currently approx. 8,000 data instances) is supplemented by a selection of a much smaller dataset prepared for human evaluation of LLM-generated explanations, thereby enabling a more careful study of the language generation and interpretation abilities of the models from a lexical-semantic perspective.

Keywords: benchmarks, Large Language Models, evaluation, lexical resources.

1. Introduction

It is well-known that Large Language Models (LLMs) as included in chatbots like ChatGPT or Gemini expose severe language and cultural biases when working in medium- and low-resourced languages (Zhang et al., 2022), as is also the case of Danish and the other Scandinavian languages (see Pedersen et al. 2025, Einarsson et al. 2025 among others). If we want to remedy such biases via cultural adjustment of the pretrained models, it is an indispensable first step to develop principled methods for compiling benchmark datasets that can detect cultural strengths and weaknesses of the models in a systematic and preferably large-scale fashion (Hershcovich et al., 2022) as

exemplified for the Scandinavian languages by the following pioneering benchmarks: Berdicevskis et al. (2023), Samuel et al. (2023), and Pedersen et al. (2024). Without access to language resources of a considerable quality and size in the target languages, this adjustment exercise becomes both cumbersome, methodologically hard, and in the worst case completely random and sporadic.

In the following, we present our work on Danish benchmarking of LLMs where we apply available lexicographical resources to compile nine Natural Language Understanding (NLU) tasks in Danish[1]. Currently, the benchmark comprises of tasks such as disambiguating words in context, determining semantic outliers, inferencing and reasoning tasks based on semantic relations (from our WordNet and FrameNet resources), as well as selecting the correct explanation of culture-relevant metaphors and metaphorical idioms.

Our large-scale benchmark is currently supplemented with a much smaller dataset meant for human evaluation of LLM-generated explanations. In this dataset we distinguish between culture-specific and cross-cultural metaphorical expressions to investigate how well Danish culture-specific features are captured relative to cross-cultural aspects. Overall, such human-evaluation datasets allow for more precise validation of the language generation and interpretation abilities of the models.

The language resources applied for the benchmark include the monolingual Danish dictionary, DDO (Det Danske Sprog- og Litteraturselskab, 2025), a thesaurus organising the DDO senses into thematic and semantic groups (Nimb et al. 2014, Nimb et al. 2024b), and several semantic lexicons where the dictionary senses from DDO are supplied with semantic relations, ontological types, and information on polarity values and semantic frames. These semantic lexicons include COR.SEM (Nimb et al. 2024a), the Danish wordnet, DanNet (Pedersen et al., 2009), the Danish Sentiment Lexicon (Nimb et al., 2022) and the Danish FrameNet Lexicon (Nimb et al., 2017). Links at sense level, i.e. shared sense id-numbers, between all the resources allow us to combine data in many ways. For example, to test the capabilities of LLMs judging semantic similarity and relatedness, we can automatically generate different groups of words representing close synonymy (based on DDO information), near synonymy (based on thesaurus sense groups), relatedness (based on chapters and sections in the thesaurus), or completely unrelated (again based on thesaurus information). We can also combine corpus examples from the DDO sense descriptions with the polarity information of the same senses in the sentiment lexicon to test a model's ability to identify the polarity of the senses in specific usage contexts.

Structured information on figurative language in the DDO dictionary also comes into play. Metaphors are explicitly labelled as such in the dictionary and furthermore organised as subsenses of the basic, concrete sense they relate to. Since the concrete

[1] Our benchmark dataset is available at: https://github.com/kuhumcst/danish-semantic-reasoning-benchmark under password protection.

sense is thematically labelled in the underlying manuscript, we can identify groups of somehow related metaphors, e.g. metaphors based on lemmas within the same domain such as 'anatomy', 'building', 'family', 'food', or 'education'. This also facilitates the identification of Danish culture specific metaphors, for example metaphors based on specific types of food in Denmark. We then test the models by asking them to explain the metaphorical use in the corpus examples which illustrate the figurative sense in DDO.

In Section 2, we describe the methods behind our approach and give examples from each of the benchmark datasets. Further, Section 3 describes a dataset meant for human evaluation of the interpretation of metaphors, which tests the language generation abilities and more subtle nuances of metaphor interpretation. Finally, in Section 4 we conclude and describe some ideas for further work.

2. A selection of NLU Tasks Derived from Lexical-Semantic Resources

2.1A Lexical Approach to Danish NLU Benchmark

The rationale behind the benchmark is to provide a collection of tests that can be used for assessing the NLU abilities of LLMs to reason[2] on lexical-semantic aspects of Danish based on the information already available through high-quality lexical resources. The aim is that all tasks should include approx. 1000 test instances, which is a highly time-consuming effort if compiled manually from scratch. However, much of the relevant data is already collected and interlinked in multiple resources, although originally with a different purpose. By repurposing and restructuring existing data, we transform the manual workload from primarily annotating to validating the new data structure.

The tests are primarily designed to enable direct and automatic comparisons of models by constraining the output format to a standardised structure, which both minimises ambiguity and ensures consistent evaluation that is runnable via an API/evaluation leaderboard. In our case, we chose to frame the tasks as a number on an interval scale (sentiment), a binary answer (True/False), or as multiple choice. In the multiple-choice setup, a test instance typically consists of a correct example (utterance or concept) and one or more distractors, that is utterances/concepts that are inappropriate to the context. The approach on how to generate the true interpretation and the corresponding distractors differs from task to task; some are compiled automatically from neighbouring concepts, where others require a human in the loop. Below, the particular compilation method for each task is described in more detail.

[2] In this paper, we use “reason” to refer to interpreting and drawing on lexical-semantic relations and knowledge, and not the specialised “reasoning” modules in LLMs.

2.2The Synonymy Task

In the synonymy task, the aim is for the LLM to select the synonym of a target word from a list of candidates. The candidates consist of a narrow synonym according to information in the DDO dictionary, and a number of distractors automatically selected from different word groups in the Danish Thesaurus. The thesaurus manuscript is continuously updated and currently contains 95% of the DDO lemmas and senses, i.e. also the less frequent part of the Danish vocabulary (Nimb et al., 2024b). This makes it very suitable for testing language models' understanding of Danish in all its nuances. The thesaurus is divided into named chapters, one of each being divided into named sections, where the word senses in DDO are listed semantically in core groups initiated by 'keywords' (words functioning as headlines) at two levels, a 'low' and a 'high' level, the last type having scope over the first (Nimb et al., 2018). In that way, the most closely related senses to a target word (synonyms and near-synonyms) are always represented by the nearest words in the same core group in a section, while the least related words are found in a different chapter.

The current version is an expansion of the pilot dataset presented in Pedersen et al. (2024) and includes 1094 test instances[3] with five distractors representing five different semantic distances from the target word as illustrated in table 1. The most similar distractor is a near synonym as it originates from the same core group but is not noted as a synonym to the target in DDO. The second most similar distractor is from the same 'high' keyword. Hereafter, we have two related distractors: one from the same section and another one from the same chapter. The final distractor is an unrelated random word from another chapter.

Target	**Synonym**	**Distractors**				
		Core group	**keyword** (wine)	**section** (alcohol)	**chapter** (food & drinks)	**Random**
bobler 'bubbly '	*boblevin* 'bubbly wine'	*cava* 'cava'	*altervin* 'altar wine'	*ølkrus* 'beer mug'	*tatar* 'tartare'	*læsekursus* 'reading course'

Table 1: The correct synonym and the five distractors for the target word *bobler* 'bubbly' which in The Danish Thesaurus is placed in a core group of words about sparkling wine.

2.3The Semantic Similarity Task

[3] By processing the entire thesaurus, we can compile up to 12,000 test instances.

Evaluating semantic similarity is framed through an outlier detection task (also called word intrusion) where we automatically add an outlier to a core group of five semantically similar words (Nielsen & Hansen, 2017; Camacho-Collados & Navigli, 2016; Pedersen et al., 2024). The dataset currently includes 1250 examples and three granularities depending on the distance of the outlier to the semantic core group.

Core group	**Outlier**	
it-udstyr, edb-udstyr, maskinel, hardware, dataudstyr 'IT equipment, computer equipment, machinery, hardware, data equipment'	**Fine**	*motherboard* 'motherboard'
	Medium	*gigabyte* 'gigabyte'
	Coarse	*programmør* 'programmer'

Table 2: Example of a core group and the three granularities of outliers.

The data is also in this case extracted from the Danish Thesaurus. We are thereby basing the dataset on hand annotated semantic similarity. The core group consists of words placed manually next to one another in a structurally marked group at the 'lowest' level in the thesaurus, i.e. a number of synonymous and near-synonymous words which are very close to each other in meaning, some of which might also belong to informal or historic language (see table 2). Different types of outliers were added based on the thesaurus structure to represent degrees of semantic distance to the core group (the five general words for IT equipment in Table 2). For instance, a fine-grained outlier ('motherboard' in Table 2) was selected from one of the other core groups under the same keyword, while a medium distance outlier ('gigabyte' in Table 2) was selected from a core group initiated by a different 'low' keyword ('data unit'), but still sharing the same 'high' keyword ('hardware'). Finally, the most distant outlier ('programmer' in Table 2) was selected among the words under another 'high' keyword than 'hardware' in the section, namely 'IT employee' which initiates the group of words describing persons working with IT. In that way, the dataset represents quite subtle nuances of meaning. All data was manually validated by two lexicographers to ensure that the automatically added outlier was in fact very different in meaning from the core group and to remove core groups consisting of co-hyponyms rather than near-synonyms.

Additionally, we performed a small human experiment on 450 random test instances (150 of each granularity) to verify that the different types of outlier granularities worked as intended. Two other lexicographers were presented with lists of words and asked to identify the outlier. They were allowed to consult a dictionary for unknown terms, record each lookup[4], and were instructed to give their best guess when in doubt. The results showed a clear correlation between granularity and task difficulty: 93% of coarse

[4] Approximately 1% of all words were checked in the dictionary.

outliers were correctly identified, compared to 86 % of medium outliers and 53% of fine outliers.

2.4The Word Sense Disambiguation (Word-in-Context) Task

We represent word sense disambiguation as a Word-in-Context task (Pilehvar & Camacho-Collados, 2019), which we automatically compile from the COR.SEM lexicon (Nimb et al., 2024a; Pedersen et al., 2022). COR.SEM is manually compiled based on the rather fine-grained DDO dictionary. Closely related lemma senses are lumped in COR.SEM in order to obtain a granularity which is better suited for language technology purposes (i.e. a narrow or more restricted subsense is lumped with its main sense). COR.SEM includes corpus examples from the DDO, and in the case of lumped senses the examples from all the covered DDO senses are grouped together.

Context 1	**Context 2**	**Label**
Man har mellem 100.000 og 150.000 ***hår*** *på hovedet* 'You have between 100,000 and 150,000 **hairs** on the head'	*Gulvtæppet er fuldt af hundens* ***hår*** 'The carpet is full of the dog's **hair**'	Same sense
han [havde] allerede fået sportsvogn og ridehest og ***båd*** 'He [had] already received a sports car and a riding horse and a **boat**'	*Skræl æblerne og skær dem i* ***både*** 'Peel the apples and cut them into **wedges**'	Different senses

Table 3: 'Same sense' and 'different senses' examples from the Word-in-Context dataset

In the Word-in-Context task, two context sentences with the same target word are presented to the LLM. The task is then to determine whether the two context sentences represent the same or two different senses of the target word as illustrated in table 3. Since the target words have a varying number of senses and context sentences (DDO corpus examples) for each sense, we limit the number of instances for each target word to a maximum of 3. Additionally, we compile two versions of the dataset: a monosemous dataset where the label is always 'same sense', and a polysemous dataset with an equal split between the two labels. The monosemous dataset may appear overly simplistic by only containing a single label. However, it serves as a useful tool for assessing whether an LLM can perform the task reliably. For instance, an accuracy of 50% or lower would indicate that the model is simply responding at random rather than demonstrating task competence. In fact, in earlier experiments with GPT-3.5-turbo, we observed that

the model consistently outputted “different senses” for nearly all monosemous instances (Pedersen et al., 2024). We call this dataset DanWiC and both versions include 1098 instances.

2.5 The Sentiment of Words in Context Task

The 'sentiment in context' dataset is based on the COR.SEM lexicon[5] and preliminary sense polarity annotations from the Danish Sentiment Lexicon (Nimb et al., 2022)[6]. The aim of the task is to estimate the sentiment of individual words within a given context. This is different from the more typical task of sentence-level sentiment classification, as the overall sentiment of the context may differ from the sentiment of the target word. Thus, the goal is to evaluate how well the sentiment of each word in the lexicon can be identified, using the surrounding context to disambiguate the word's specific sense. The majority of the sentiment values and contexts are automatically compiled from the COR.SEM lexicon. Additionally, we manually added a number of neutral (sentiment: 0) instances. In total, we include 1041 examples across 7 sentiment values, from -3 to 0 to +3 (see table 4 for examples). Each context has been manually validated to ensure that the sentiment of the word from the sentiment lexicon is reflected in the context.

Context	Sentiment
*Bandemedlemmer mødte op med køller for at **afstraffe** den 19-årige.* 'Gang members showed up with clubs to **punish** the 19-year-old.'	-3
*Flyt bollerne over på en **bagerist** og lad dem afkøle.* 'Transfer the buns to a **baking rack** and let them cool.'	0
Elskede, du må ikke dø fra mig. Jeg elsker dig af hele mit hjerte. '**Beloved**, you must not die on me. I love you with all my heart.'	+3

Table 4: Examples of sentiment in context with three polarity values (-3, 0, +3). The target word is in **bold**.

2.6 The Inference Task on Ontological Affiliation and Characteristics

[5] The resource is available at https://corsem.dsl.dk/

[6] The Danish Sentiment lexicon contains only lemma polarities. However, the lexicon is based on initial sense polarity annotations of the lemmas; these are transferred to the COR.SEM lexicon.

The inference task is semi-automatically compiled from the ontological affiliations and characteristics made explicit in the Danish wordnet, DanNet, via *semantic relations*, which are again semi-automatically derived from the sense definitions given in the DDO (Pedersen et al., 2009). The task tests the models' general understanding of taxonomical ordering of concepts, in particular for the natural and functional kinds (see Cruse, 1986). Following broadly the genus and differentia of the dictionary definitions, which have however been slightly adjusted in the wordnet according to a more systematic taxonomy, the task includes 970 instances of two true utterances referring to these ontological characteristics, like *en gaffel er et spiseredskab* ('a fork is a piece of cutlery'), and *en ske er et spiseredskab* ('a spoon is a piece of cutlery'), followed by a similar utterance for the language model to assess whether it is either true or false, as in: *en hat er et spiseredskab* ('a hat is a piece of cutlery') = FALSE (See table 7). The distractors are typically generated from neighbouring concepts, as in this case, another kind of artifact.

Semantic Relation	**Ontological Type**	**Example of a true Utterance**	**Example of a query Utterance**	**Truth Value of Query**
Has_hyphonym	ANIMAL	*En hund er et pattedyr* ('A dog is a mammal')	*En edderkop er ikke et pattedyr* ('A spider is not a mammal')	*TRUE*
Used_for/ Purpose_of	GARMENT	*Man tager en frakke på for at holde sig varm* ('You wear a coat to keep warm')	*Man tager en ring på for at holde sig varm* ('You wear a ring to keep warm')	*FALSE*
Has_meronym	BODY PART	*En hånd kan ikke have et øje som del* ('A hand cannot have an eye as a part')	*En hånd kan have en finger som del* ('A hand can have a finger as its part')	*TRUE*

Table 5: Examples of semi-automatically compiled utterances about the semantic relations of the ontological types: ANIMAL, GARMENT and BODY PART.

In addition, purpose relations (used_for, has_purpose) are applied for the inference task, such as stating what artifacts are typically used for, as in the case of cutlery: 'used to eat with' = TRUE.

Finally, we include relations of origin and part/whole in the inference benchmark: how

things have come about, and what they are typically a part of, as shown in the examples *man laver en hat ved at sy den,* ('you make a hat by sowing it') =TRUE, and *en hånd kan have et øje som del* ('a hand can have an eye as its part') = FALSE.

2.7 The Entailment Task

A standard task within Natural Language Inference (NLI) is the task of determining whether a statement logically follows from another (Bowman et al, 2015). In our case, we focus on semantic frames and the result of an event or an activity. The underlying data is the Danish FrameNet Lexicon (Nimb et al., 2017). In this lexicon, Danish Lexical Units (LUs) are assigned one or more frame values based on the frame inventory listed and described in the online resource Berkeley FrameNet for English (Berkeley FrameNet, 2025; Ruppenhofer et al., 2016). The detailed frame value descriptions in the 'Frame Index' in the online resource allow us to (manually) infer the kind of activities, transactions and situations connected to the Danish LUs.

FRAME	Statement	Label
CAUSATION	P: *Nedsat hørelse hos småbørn kan bevirke at sprogudviklingen går i stå.* H: *Nedsat hørelse har ingen betydning for små børns sprogudvikling* P: 'Hearing loss in young children can cause language development to stall.' H: 'Hearing loss has no effect on young children's language development.'	Contradiction
CAUSE_HARM	*P: Kaffen skoldede Peters højre arm.* *H: Peters arm blev forbrændt af kaffen.* P: 'The coffee scalded Peter's right arm.' H: 'Peter's arm was burned by the coffee'	Entailment
ACTIVITY_FINISH	*P: Han sluttede altid måltiderne med frisk frugt.* H: *Han spiste altid friske æbler.* P: 'He always ended his meals with fresh fruit.' H: 'He always ate fresh apples.'	Neutral

Table 6: Examples of statements and true or false queries with verbs from the frames CAUSATION, CAUSE_HARM and ACTIVITY_FINISH

For a selected group of frames, we create two standardised Danish sentences: a premise and a hypothesis (see table 6). The premise is an example of the activity or event that

a verb or verbal noun in a specific frame evokes. For instance, the verb *give* has the frame 'Giving' as in "I give the book to Ella", and the frame is described as: "the Donor first has possession of the Theme. Following the transfer the Donor no longer has the Theme and the Recipient does". The hypothesis then either entails or contradicts the result of that activity or event (entailment: "Ella has the book", contradiction: "I have the book). We also add some neutral statements as the "hypothesis" to check whether the task is completed as intended. The premises are based on usage examples found in the COR.SEM resource, which also includes frames from the Danish FrameNet, although some sentences have been shortened or simplified. The hypothesis statements were manually written with the constraint that the entailment and contradiction must directly target the frame. In total, we include 502 instances. We also note that this manual task is far more time-consuming than the more automatically compiled tasks.

2.8The Idiom Interpretation Task

To evaluate the capabilities of LLMs to correctly interpret Danish idioms, we use the Danish Idiom Dataset (Sørensen et al., 2025) which was created with support from the Danish Agency of Digital Government for this specific purpose. The dataset consists of a correct dictionary definition from the DDO supplemented with three different types of distractors: a literal incorrect definition (what is the "face value" of the induvial words), an abstract incorrect definition (what could this idiom mean if it was based on a different type of metaphor), and a random definition from another idiom. Thereby, the task is framed as a multiple-choice scenario as illustrated by table 5.

Idiom: *Køre med klatten* 'ride with the blob'			
Dictionary Definition	**Literal Distractor**	**Abstract Distractor**	**Random Distractor**
være suveræn til noget, brillere	*køre rundt med en klat'*	*Køre på en hensynsløs og uansvarlig måde*	*den negative side af noget positivt'*
'be excellent at something, shine'	'ride around with a blob'	'drive in a reckless and irresponsible manner'	'the negative side of something positive'

Table 5: Example from the Danish Idiom Dataset

The 1000 idioms were semi-automatically identified from the 11,500 fixed expressions in the DDO. First, a selection of expressions was extracted using COR.SEM information on centrality in the vocabulary as well as the number of fixed expression with the word in DDO (e.g. *hånd* 'hand' is a central noun (a core concept) and has a

high number of fixed expressions, some of which were selected for the final dataset). Additionally, explicit information on proverbs in the DDO was also used. From this initial extraction, a team of lexicographers then selected the final 1000 idioms.

The most time-consuming task was to manually compose the literal and abstract distractors. The abstract distractor was in particular difficult to create since this task depends to some extent more on the annotator's creativity than on specific lexicographical knowledge. The random distractor (one of the other 999 idiom definitions) had to be manually validated due to a rather high risk of overlapping meanings among the idioms in the dataset.

One of the qualities of the dataset is that it is useful in scenarios other than just the multiple-choice. For instance, the pairing between the idiom and the dictionary definition can be used in a generative task, where the dictionary definition corresponds to a gold standard. We can also use the distractors to test the ability to identify correct versus false information in a prompt by asking questions like "Why does [IDIOM] mean [DISTRACTOR]?" or "Is [DISTRACTOR] a plausible definition for the idiom [IDIOM]?". In other words, it pays to build in flexibility and extra metadata when designing a dataset. This way, it ensures that the dataset can be utilised in future work.

3. Evaluating a selection of models on the benchmark

The main purpose of any evaluation benchmark is to allow us to distinguish between models based on information that is important to be correctly represented in the LLMs. Thus, we should strive for compiling datasets in which the task is not so difficult or confusing that it is impossible to answer, nor is too easy for the majority of the models we want to evaluate. It is important to note that the success of an evaluation dataset is not diminished if the best models are shown to perform well on it, when we use the same dataset to reveal the quality gap in smaller, monolingual, and open-source models.

To test the practical applicability of the individual datasets for model evaluation and to investigate whether the effort of compiling them is justified, we apply each dataset to three LLMs. The aim is to carry out a robustness check of the benchmark. If the datasets consistently differentiate between models, it supports their usefulness as evaluation tools. Additionally, it also provides an insight into strengths and limitations of the current state of LLMs on Danish lexical semantic tasks, giving us inspiration to where to focus future work. We have selected three instruction-tuned models that represent complementary design choices: GPT-4.1 (the newest non-reasoning model), GPT-oss 120B (an open-weight reasoning model), and Llama 3.3-70B (a model from a different family). The results can be seen on Figure 1.

Overall, we see a pattern of GPT-4.1 achieving the highest scores, followed by GPT-

oss 120B and then Llama 3.3-70B. The stable ranking across the tasks indicates that the datasets differentiate between models in a reliable way, although some do so more effectively than others. The idiom interpretation task shows the largest performance gap with the open models (Llama 3.3-70B and GPT-oss 120B) getting noticeably lower scores. This is a positive outcome, as compiling the dataset required substantial manual effort compared to the other tasks. In contrast, the entailment dataset, which also relied on some manual effort, shows only minor differences between models, with a gap of just 0.04 between the best and worst model.

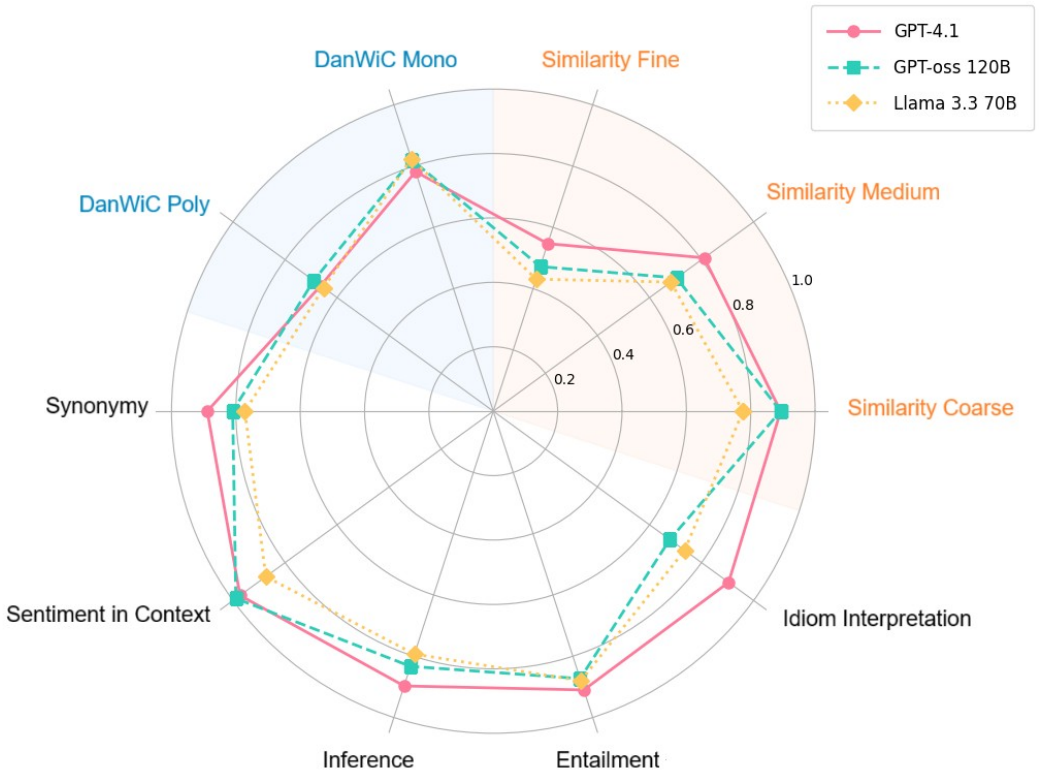

Figure 1: Evaluation results on the benchmark. Performance is given by accuracy, except for sentiment-in-context, which is reported by Mean Squared Error (MSE). Note, the three similarity granularities are marked with orange and the two DanWiC versions with blue.

The performance on DanWiC is low compared to the results on the other tasks. On the simpler monosemous dataset, the models can perform the task fairly reliably, with accuracies from 0.78 to 0.82. The performance drops on the polysemous data to 0.65-0.69. Thus, this task seems to be somewhat equally challenging to the models, which consequently makes it harder to use this task to differentiate between the models. This is surprising given that the senses in the dataset are derived from the COR.SEM resource, where the sense inventory has been manually simplified. Nevertheless, the models still display distinct output patterns: GPT-oss 120B shows a bias towards “same

sense" answers, GPT-4.1 leans towards "different sense" answers, while Llama 3.3 70B produces the most balanced distribution.

Looking at the synonymy and semantic similarity datasets, we see that the semantic closeness encoded in the dictionary and thesaurus is reflected in the datasets. In the synonymy dataset, the most frequently chosen (and thus most effective) distractor across all three models is the near synonym from the same sense group in the thesaurus. This distractor accounts for 70% of GPT-4.1's incorrect answers, 52% for GPT-oss 120B, and 59% in the case of Llama 3.3 70B. The distribution of the remaining errors follows on the scale of semantic similarity: keyword-level distractors are chosen second most often, followed by section-level distractors, and so forth. Likewise, results on the outlier detection datasets ("similarity" on Figure 1) show that outliers placed furthest away in the thesaurus structure are the easiest to detect. However, the granularity of sense distinctions can be too fine for reliable evaluation. As with the human participants, the models struggle with identifying the fine-grained outlier, with accuracies ranging between 0.43 and 0.55. This resonates with Kilgarrif's (1997) point that sense distinctions are not absolute, but dependent on the task at hand. In line with this, our findings suggest that LLMs have difficulty handling fine sense distinctions, highlighting the need for datasets that include semantically close yet still distinguishable examples.

4. A Metaphor Dataset Prepared for Human Evaluation

4.1Why Introduce Human Evaluation?

Not all lexical reasoning capabilities of LLMs can be appropriately assessed in a multiple choice-scenario. A more thorough evaluation of the models' abilities to reason on lexical-semantic issues through a model-generated coherent paragraph of text still calls for human evaluation, even if part of the evaluation process can automatised.

This additional dataset, which is relatively small with only 150 metaphor test instances, is designed to prompt models on two different prompts: "What is the metaphorical meaning of the Danish word/ expression X, and what has it got to do with the basic meaning of X"? And "What does the Danish word/words X mean in the following example: [CONTEXT]?" Prompts about Danish metaphors are given in both Danish and English, since previous studies have shown that models tend to perform better when asked in English about a lower-resourced language than when asked in the target language (Zhang et al. 2023).

Another important aspect of the dataset is that it is divided by topics of the source domain, basically into metaphors originating from the agriculture domain and those originating from the nautical domain; domains which are derived from the DDO domain labels, and which are particularly central for the Danish 'self-understanding' of our

culture and society. The context examples provided in the prompts are likewise extracted from the DDO sense descriptions in most cases.

Even more importantly, a further study by informants of English proficiency have categorised the dataset into metaphors that go across culture between English and Danish and those that are unique or culture-specific to the Danish society. This enables to examine if models' metaphor interpretation differs when exposed to a metaphor that can be learned from English compared to those that are specific to our own language and culture and that can therefore only be learned from Danish data.

In Pedersen et al. (2025), we let LLM-generated explanations of these metaphors undergo human evaluation when prompted on ChatGPT-4o-mini and an instruction-tuned Llama 3.1 405B. The study shows how culture-specific metaphors are indeed more problematic for the models to explain, and how features are often erroneously taken over from English to Danish resulting in wrong explanations. In particular the sarcastic tone of many Danish metaphors is often misinterpreted and understood positively by the models, learning from a parallel English metaphor. For instance, when we use *at sejle* ('to sail') in Danish referring to a chaotic, negative situation, where it refers to a positive situation in English, like in *smooth sailing*. The study also shows, not so surprisingly, that the models perform better when asked about Danish in English than when prompted in Danish about Danish, and that the more context they are provided with, the better they perform.

5. Concluding remarks and future perspectives

In this paper we have shown how lexicographical resources can be applied to semi-automatically compile large-coverage benchmarks for evaluating language models that work in Danish. Our claim is that such benchmarks supplement the more extrinsic and often translated benchmarks for evaluating language models, which typically focus more on the ability to solve downstream tasks than on more nuanced lexical interpretation. The possibilities of extending and updating the benchmarks are more or less endless, and by extending to also less frequent and figurative vocabulary, they can hopefully facilitate that models are evaluated and subsequently adjusted to better capture and reproduce the variety of each particular language instead of being strongly homogenised towards English. Moreover, we greatly benefitted from the data cleanliness of the repurposed lexical-semantic resources. In particular, the consistent maintenance of the entry and sense ID numbers across the different resources enabled a straightforward linking process, allowing us to selectively compile lexical information into new datasets. Retaining preliminary annotations also proved valuable. Although these annotations may have played a background role in the original project, the information they contain may be repurposed in new contexts. We emphasise the importance of high-quality metadata and flexible design principles when creating datasets, as the same data may be reused in different contexts in the future, as seen for example with the metadata on domain.

Future work will encompass also human tests of the compiled benchmarks in order to assess how humans perform in comparison with the models. Informants will have good language proficiency but not be experts as such (i.e. not be skilled lexicographers nor professional linguists). Another line of research to be pursued is to examine how model proficiency in lexical- semantic tasks align with proficiency in more extrinsic tasks like summarisation, knowledge extraction, machine translation and the like.

6. References

Berdicevskis, A., Bouma, G., Kurtz, R., Morger, F., Öhman, J., Adesam, Y., Borin, L., Dannélls, D., Forsberg, M., Isbister, T., Lindahl, A., Malmsten, M., Rekathati, F., Sahlgren, M., Volodina, E., Börjeson, L., Hengchen, S., & Tahmasebi, N. (2023). Superlim: A Swedish language understanding evaluation benchmark. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8137–8153. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.506

Berkeley FrameNet. Accessed at: https://framenet.icsi.berkeley.edu/ (22 September 2025)

Bowman, S. R, Angeli, G., Potts, C. & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, pp. 632-642. Association for Computational Linguistics.

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.

Comșa, I., Eisenschlos, J., & Narayanan, S. (2022). MiQA: A Benchmark for Inference on Metaphorical Questions. *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 373-381. https://doi.org/10.18653/v1/2022.aacl-short.46

Cruse, D.A. (1986). *Lexical Semantics*. Cambridge University Press.

Det Danske Sprog- og Litteraturselskab (2025). *Den Danske Ordbog.* Accessed at https://ordnet.dk/ddo (July 17 2025)

Einarsson, H., Simonsen, A., & Nielsen, D. S. (2025). *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*. Association for Computational Linguistics. https://aclanthology.org/2025.nbreal-1.0

Hershcovich, D., Frank, S., Lent, H., de Lhoneux, M., Abdou, M., Brandl, S.,

Bugliarello, E., Cabello Piqueras, L., Chalkidis, I., Cui, R., Fierro, C., Margatina, K., Rust, P., & Søgaard, A. (2022). Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6997-7013). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.482.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities, 31*(2), 91–113. https://doi.org/10.1023/A:1000583911091

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., Flörke, I., Olsen, S., Pedersen, B.S., Sørensen, N.C.H. (2024a). COR.SEM, a New Formal Semantic Lexicon for Danish. In *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress,* Cavtat, Crotia.

Nimb, S., Lorentzen, H., T. Troelsgård, L. Theilgaard (2014). *Den Danske Begrebsordbog.* Det Danske Sprog-og Litteraturselskab.

Nimb, S., Sørensen, N.C.H., Jensen, J. (2024b). Making Danish Thesaurus Data Available to Researchers – The WebDDB project. In *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress,* Cavtat, Crotia.

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. et al. 2017. From thesaurus to framenet. Electronic lexicography in the 21st century. *Proceedings of eLex 2017*, 1-22. Brno: Lexical Computing CZ s.r.o. https://elex.link/elex2017/wp-content/uploads/2017/09/paper01.pdf.

Pedersen, B., Sørensen, N., Nimb, S., Hansen, D., Olsen, S., & Al-Laith, A. (2025). Evaluating LLM-Generated Explanations of Metaphors – A Culture-Sensitive Study of Danish. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)* (pp. 470-479). University of Tartu Library. https://hdl.handle.net/10062/107190

Pedersen, B. S., Sørensen, N. C. H., Nimb, S., Flørke, 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 13th Language Resources and Evaluation Conference,* Marseille, France (s. 51-60). European Language Resources Association.

Pedersen, B., Sørensen, N., 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 the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16353-16363). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1421

Pilehvar, M. T. & Camacho-Collados, J. (2019). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Volume 1 (Long and Short Papers)*, pp. 1267-1273. Minneapolis, Minnesota: Association for Computational Linguistics.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. International Computer Science Institute. https://framenet.icsi.berkeley.edu/the_book

Samuel, D. et al. 2023. NorBench – a benchmark for Norwegian language models. Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa), 618–633. Tórshavn: University of Tartu Library. https://aclanthology.org/2023.nodalida-1.61.

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G (2023). Don't Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7915-7927). Singapore. Association for Computational Linguistics. https://aclanthology.org/2023.emnlp-main.491/