Towards a Danish Semantic Reasoning Benchmark - Compiled from Lexical-Semantic Resources for Assessing Selected Language Understanding Capabilities of Large Language Models

Bolette S. Pedersen[1], **Nathalie Carmen Hau Sørensen**[2], **Sussi Olsen**[1], **Sanni Nimb**[2], **Simon Gray**[1]
Centre for Language Technology, NorS, University of Copenhagen[1],
The Society for Danish Language and Literature[2]
Emil Holms Kanal 2, 2300 Copenhagen S[1], Christians Brygge 1, 1219 Copenhagen K[2]
{bspedersen, saolsen, simongray}@hum.ku.dk, {nats, sn}@dsl.dk

Abstract

We present the first version of a semantic reasoning benchmark for Danish compiled semi-automatically from a number of human-curated lexical-semantic resources, which function as our gold standard or 'ground truth'. Taken together, the datasets constitute a benchmark for assessing selected language understanding capacities of large language models (LLMs) for Danish. This first version comprises 25 datasets across 6 different tasks and include 3,800 test instances. Although still somewhat limited in size, we go beyond comparative evaluation datasets for Danish by including both negative and contrastive examples as well as low-frequent vocabulary; aspects which tend to challenge current LLMs when based substantially on language transfer. The datasets focus on features such as semantic inference and entailment, similarity, relatedness, and ability to disambiguate words in context. We use ChatGPT to assess to which degree our datasets challenge the ceiling performance of state-of-the-art LLMs, average performance being relatively high with an average accuracy of 0.6 on ChatGPT 3.5 turbo and 0.8 on ChatGPT 4.0.

Keywords: reasoning benchmark, large language models, lexical semantic resources, Danish resources and datasets

1. Introduction

Striking performance improvements in most recent large language models (LLMs) have challenged existing benchmarks and encouraged the community to develop new and harder evaluation datasets that go deeper into semantic reasoning and language understanding, and further into the peculiarities of a given language's vocabulary. Where most well- and medium-resourced languages have a series of evaluation datasets that can be used to assess the language models' performance on downstream tasks such as Named Entity Recognition, sentiment analysis, question-answering, and summarization, there is a recurring requirement for more generalised benchmark datasets that go broadly into assessing the models' capacity of inferring and drawing conclusions from text. Along the same lines, there is a call for datasets that include negated and contrastive examples and go beyond the most prototypical and frequent part of the vocabulary and include also the more subtle and specialised concepts of a given language.

It is extremely time-consuming and cumbersome to handcraft such datasets from a monolingual perspective and at a large scale. Our claim is, however, that with existing human-curated lexical-semantic resources at hand, we have, in fact, much of the information needed to compile them. Taking a monolingual approach also enables us to avoid linguistic bias in our tests; bias that are often introduced when translating datasets from other languages.

Our approach is thus to take advantage of the rich semantic information already provided in our lexical-semantic resources, see them as our gold standard and make use of them for compiling a number of semantic reasoning datasets. How well do LLMs infer that when something is 'eaten up', there is no more food left on the plate? Or that if I have given you an object, you have it and not I? And that a pet, even if prototypically referring to a cat or a dog, can in fact be any kind of animal if somebody assigns this role to it? How well do they distinguish true synonyms from other similar concepts, and can they actually disambiguate subtle meanings from each other? These are examples of the more specific tasks that we address with our datasets.

We make use of four specific lexical-conceptual resources developed at the Centre for Language Technology, UCPH and the Society for Danish Language and Literature during the last decades, several of them in close collaboration. These include:

- The Danish wordnet, DanNet (Pedersen et al., 2009), which relates all concepts to an ontology and encodes an internal taxonomy of hyponymy together with an additional number of semantic relations describing the core mean-

LREC-COLING 2024, pages 16353–16363

ing components of a given concept.

- The Danish Thesaurus (Nimb et al., 2015, 2014) where the ordering of the words in chapters, sections and subgroups depending on their topic and semantic relatedness can be used to deduce semantic similarity and synonymy.
- The Danish FrameNet Lexicon (Nimb et al., 2017; Nimb, 2018) where all verbs and deverbal nouns are assigned a reference to the semantic frames inventory from Berkeley's FrameNet (Baker et al., 1998) thus, enabling the extraction of e.g. change-of-state verbs, communication and mental verbs.
- The Central Word Register for Danish (Nimb et al., 2022; Pedersen et al., 2022) a recently developed computational lexicon for Danish with a simplified (i.e. coarse-grained) sense inventory.

The rest of the paper is organised as follows. With a specific focus on Danish and Scandinavian datasets, we describe in Section 2 related work on benchmark data to assess LLMs on basic reasoning and language understanding. In Sections 3 to 6, we describe the methods used to compile our different benchmark datasets from the lexical-semantic resources. Section 7 describes how we test our datasets on ChatGPT 3.5 turbo and 4:0 via OpenAI's API and an interface constructed for the task. The testing is done in order to check whether the tasks are hard enough to challenge the performance of state-of-the-art models. In Section 8, we discuss how, as a side effect, the compilation of the datasets gives valuable feedback to the lexical resources, opening for adjustments. Finally, in Section 9 we conclude and present the lines for extending the dataset to Version 2 in near future.

All compiled datasets and API framework are made available through github[1].

2. Related Work

The GLUE and SUPERGLUE test suites for English ((Wang et al., 2018), (Wang et al., 2020) are probably the most well-known testbeds for assessing LLMs for their general language understanding abilities. The most recent SUPERGLUE benchmark comprises eight selected language understanding tasks including QAs on common sense reasoning and entailment as well as more complex tasks of coreference resolution and word sense disambiguation tasks (word-in-context). The SQuAD2.0 benchmark (Rajpurkar et al., 2018) is another well-known benchmark consisting of reading comprehension questions and answers based on Wikipedia articles.

(Conneau and Kiela, 2018) describe a toolkit for universal sentence representation, while several semantic tasks have been designed over the years for the SEMEVAL workshops, the latest from 2023 reported on in (Ojha et al., 2023) and covering a number of semantic tasks ranging from domain-specific tasks of clinical entailment identification to visual word sense disambiguation, and detection of persuasion techniques in news articles.

For the Scandinavian languages, a few initiatives have been embarked to provide evaluation benchmarks on semantic and related tasks. First of all, SUPERLIM deserves mentioning (Berdicevskis et al., 2023), being a recent Swedish language understanding evaluation benchmark following basically the idea of SUPERGLUE and providing a number of entailment and word meaning tasks for Swedish. For Norwegian, the benchmark NorBench was presented at the NODALIDA Conference in 2023 (Samuel et al., 2023) including mostly datasets for a number of downstream tasks like machine translation, sentiment analysis and QA.

A few semantic benchmarks for Danish, Swedish and Norwegian are provided in the Scandeval platform (Nielsen, 2023), including sentiment classification and linguistic acceptability tasks, and specifically for Danish, the Danish Foundation Models platform include a few semantically oriented benchmark data for Danish, like benchmarks for document embeddings[2].

Finally, previous benchmark work has been done on semantic relatedness and sentiment prediction for Danish (Nielsen and Hansen, 2017), and for Danish word sense disambiguation (Pedersen et al., 2016, 2023), and word similarity (Schneidermann et al., 2020).

It is worth mentioning that recent studies in NLP critically discuss which language understanding and reasoning capabilities are actually being evaluated with current benchmarks ((Tedeschi et al., 2023), since some models have been claimed to possess almost superhuman understanding capabilities against them. The studies underline the importance of being very careful when creating benchmark data and ensure that they i) do not suffer from instruction bias, ii) balance and evaluate easy and hard test set items, iii) increase annotation accountability, iv) complement automatic evaluations with human judgements, and v) do not introduce language and cultural bias via translations from other languages.

[1] https://github.com/kuhumcst/danish-semantic-reasoning-benchmark.

[2] kennethenevoldsen.github.io/scandinavian-embedding-benchmark/.

3. Inference Dataset Based on Semantic Relations from the Danish Wordnet, DanNet

With the datasets based on DanNet (Pedersen et al., 2009), we aim at providing a benchmark for assessing the models' ability to infer the ontological status and other core meaning components of a given concept. DanNet provides us with a gold standard for this information, since the ontological types and relations are encoded based on the word definitions provided by the Danish Dictionary (Hjorth, Ebba and Kristiansen, Kjeld, 2005) (70,000 lemmas are encoded in DanNet with approx 350,000 semantic relations)[3]. Ontological status is given implicitly via the genus proximum (i.e. the hypernym) of the definition as in 'a house is a building') and is standardised via reference to the EuroWordNet Ontology (Vossen, 1999).

Other meaning components are provided via the differentia of the definition, realised through a set of 15 predefined semantic relations (as in 'a pot is a container used for cooking food'. Following the theory of Pustejovsky's Generative Lexicon (Pustejovsky, 1998), the relations are organised into so-called qualia roles (or 'core' meaning components) relating to i) how a concept came about (the agentive role), ii) its function (telic role), and iii) its part-whole relation (constitutive role) or eventual other characteristic features. Table 1 presents examples of the compiled inference datasets, organised along the dimensions of the qualia roles and ontological types.

Test instances are generated from a generic template constructed for each ontological type under each qualia role. For instance, for the telic role (function) with the ontotype 'Instrument', we use the template *Man bruger en X til at Y med* (you use a X for Y-ing). We negate a selected number of utterances and try to contrast with examples from different parts of the ontology, keeping, however, always track of the truth value. A general observation when compiling the tests is that utterances generated from concrete concepts work far better than those generated from more abstract concepts. This is not so surprising, since the development of the wordnet posed the same challenges - abstract synsets were difficult to place in the taxonomical structure and semantic relations were harder to assign - even if in all cases they were adapted from the definitions given in the Danish Dictionary . As a consequence, utterances compiled from abstract concepts in the wordnet need to be curated more carefully than utterances relating to concrete concepts.

Another general factor that influences the datasets relates to word ambiguity. In particular where (rare) metaphorical or colloquial word meanings tend to mess up the intuitive truth value of the utterances. For instance, a test instance was generated stating as true that 'a chicken is a drinkable liquid' due to the (quite rare) meaning of *kylling* ('chicken') referring to a small bottle of brandy. Likewise with *flod* ('river') with the lexicalised metaphorical sense 'chaotic feeling', which generates the utterance: 'a river is a feeling' and labels it as 'true'.

4. Entailment Tests based on Semantic Frames from The Danish FrameNet Lexicon

We also aim to assess to which degree LLMs grasp the result of a particular event, in particular of causative events. If it is stated that somebody kills somebody, it entails somebody being dead. For this purpose, we make use of the Danish FrameNet Lexicon (Nimb et al., 2017; Nimb, 2018) as our gold standard. This Lexicon contains 671 different frames assigned to 5,300 Danish verbs and 6,490 deverbal nouns, and refers to the semantic event ontology of Berkeley FrameNet's (Baker et al., 1998) where it is spelled out what kind of event a frame refers to, what kind of result is achieved (if any), and which frame elements are typically evoked. For each of the tested frames, generic templates have been designed, like for BUYING *Peter X bogen af/fra Y* ('Peter X the book from Y'), see also Table 2.

Admittedly, entailment utterances are not easily compiled for all frames, and thus in this first version we have primarily dealt with the straight-forward ones including frames that refer to causative events (like e.g. CUTTING events) on more or less concrete items and where the final result of the event is relatively unambiguous. However, we believe that meaningful utterances can also be compiled for activities and mental frames even if they may be more subtle and ambiguous.

5. Datasets on Similarity and Relatedness based on the Danish Thesaurus

As mentioned previously, we aim to capture also the more fine-grained and culture-specific nuances of the Danish vocabulary in our semantic benchmark; nuances that materialise when looking into a large variety of synonyms and otherwise semantically related words as they are described in comprehensive monolingual Danish resources.

To this end, we develop three datasets with the purpose of testing the models' understanding of synonymy and semantic relatedness in Danish.

[3] The Danish Dictionary: ordnet.dk/ddo

Qualia and Ontotype	Generated utterances	Translation
FORMAL Feeling; Creature	P: *Sympati er en følelse; Tryghed er en følelse* Q: *Spøgelse er en følelse* (false)	P: Sympathy is a feeling; Safety is a feeling Q: Ghost is a feeling (false)
FORMAL Liquid; Quantity	P: *En bouillon er en væske; En drik er en væske* Q: *En slurk er ikke en væske* (true)	P: A broth is a liquid; A drink is a liquid Q: A sip is not a liquid (true)
AGENTIVE Semiotic; Artifact	P: *Man laver en roman ved at skrive den; Man laver et essay ved at skrive det* Q: *Man laver ikke en hat ved at skrive den* (true)	P: You make a novel by writing it; You make an essay by writing it Q: You don't make a hat by writing it (true)
AGENTIVE Food; Liquid	P. *Man laver et tog ved at fremstille det; Man laver en ret ved at tilberede den* Q: *Man laver te ved at pochere den* (false)	P. You make a train by manufacturing it; You make a dish by cooking it Q: You make tea by poaching it (false)
TELIC Garment; Artifact	P: *Man tager en frakke på for at holde sig varm; Man tager en hue på for at holde sig varm* Q: *Man tager en ring på for at holde sig varm* (false)	P: You put on a coat to keep warm; You wear a hat to keep warm Q: You wear a ring to keep warm (false)
TELIC Instrument	P: *Man bruger en kniv til at skære med; Man bruger en hammer til at hamre med* Q: *Man bruger ikke et rivejern til at rive med* (false)	P: You use a knife to cut with; You use a hammer to hammer with Q: You don't use a grater to grate with (false)
CONSTITUTIVE BodyPart; Part	P: *En hånd kan ikke have et øje; Et ansigt kan have en mund* Q: *Et fly kan have en propel* (true)	P: A hand cannot have an eye; A face can have a mouth Q: A plane can have a propeller (true)

Table 1: Examples of test utterances (precondition (P) and query (Q)) compiled from DanNet focusing on different qualia roles and ontological types.

Semantic frame	Generated utterance	Translation
BRINGING	P: *Peter bringer mad ud til Pia.* *Q: Pia har nu mad.* (true)	P: Peter brings out food to Pia. Q: Pia now has food (true)
CAUSE	P: *Eksplosionen medførte svære skader på bygningen.* Q: *Efter eksplosionen var bygningen ubeskadiget.* (false)	P: The explosion resulted in severe damages on the building Q: After the explosion the building was undamaged (false)
BUYING	P: *Peter købte bogen af Anne* Q: *Nu ejer Anne bogen* (false)	P: Peter bought the book from Anne, Q: Now Anne owns the book (false)
CUTTING	P: *Pia klipper rebet over.* Q: *Pia har nu to kortere reb.* (true)	P: Pia cuts the robe, Q: Pia now has two shorter robes (true)
TELLING	P: *Peter fortalte Pia om forlovelsen*, Q: *Pia kender nu til forlovelsen* (true)	P: Peter told Pia about the engagement, Q: Pia now knows about the engagement (true)

Table 2: Examples of test utterances (precondition (P), and query (Q)) compiled from The Danish FrameNet Lexicon focusing on different semantic frames.

The datasets are based on the Danish Thesaurus, which contains 22 chapters and 888 sections with more than 100,000 lemmas and 130,000 senses from the Danish Dictionary divided into groups with up to three levels of semantic similarity and relatedness marked in the structure. At the most fine-grained level, synonyms and near-synonyms, including co-hyponyms, are grouped in semantic word order. Furthermore, some of the words initiating a group are marked as either an overall keyword or a lower keyword [4], indicating two levels of semantic scope across the structure [5].

The thesaurus informs us on the degree of semantic similarity and relatedness among Danish words allowing us to consider it a gold standard from which we generate a number of datasets. From each of the 22 chapters, two sections with at least 20 noun subgroups are selected, each containing at least three nouns. We supplement

[4] Keywords in this context denote prototypical words. In the printed version of the thesaurus, keywords are highlighted graphically and function as references in the alphabetic index.

[5] For more details on the structure, see (Nimb et al., 2018) which describes how it is used to automatically present semantically related words in the online Danish Dictionary.

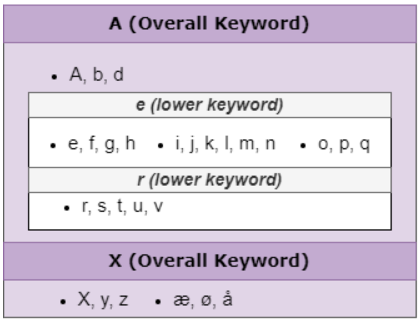

Figure 1: The structure of a subgroup and its narrow subgroups in the Danish Thesaurus. Each letter represents a word. The overall keyword A has scope over the lower keywords e and r, as well as their narrow subgroups.

the data with (automatic) information on synonymy from the Danish Dictionary, and manually remove cases where the core words consist of only co-hyponyms.

5.1. Synonymy Selection Dataset

The synonymy selection dataset is used to evaluate the models' knowledge of synonymy in Danish. The best synonym for a target word from a list of four candidates is to be identified, based on automatic extraction of sets of nouns from the thesaurus structure following a scale from the most similar word to the least similar. The closest synonym is a neighbour word in the same thesaurus subgroup and also a synonym in the Danish Dictionary. The second most similar word is from another subgroup in the section, but has the same keyword. The third most similar word is from another section, but still from the same chapter. Finally, the last and least similar noun is from any other chapter in the thesaurus.

For instance, the target word *havluft* ('ocean air') from the section *Vejr, luft* ('Weather, air') in the chapter *Natur og Miljø* ('Nature and Environment') has as most similar word *søluft* ('sea air') which occurs next to it in the thesaurus and at the same time is a synonym in the Danish Dictionary. From another subgroup with the same keyword (*luft* 'air'), we find the second most related noun *stratussky* ('stratus cloud'). The third most related noun is *øhav* ('archipelago') occurring in another thesaurus section (*kystområde, ø* ('coastal area, island')), however still in the same chapter. Finally, we randomly select the least similar noun (*kontamination*('contamination')) from a different chapter, *Apparater, teknik* ('Appliances, technology').

5.2. Similarity and Relatedness Word Intrusion

We frame the semantic similarity and relatedness evaluation as a word intrusion task. Four words are presented, one of which is an outlier compared to the other three. We take inspiration from the dasem word intrusion dataset where a list of three semantically related words and an outlier is given (Nielsen and Hansen, 2017). The task is to identify the outlier. By giving a list of words, they together create a context. Another advantage is that the task circumvents the need for a similarity score. Thereby, we can automatically generate the dataset from our lexical resources. Examples from the similarity and relatedness datasets can be seen in table 3

We address similarity (i.e. how synonymous are the words?) versus relatedness (i.e. how thematically related are the words?) by utilising the different semantic and thematic levels in the thesaurus described above. The three core words are from the same subgroup (synonyms and near-synonyms in the thesaurus), while the outlier is more or less similar in meaning, depending on where it occurs outside of this subgroup in the thesaurus structure. In the **medium** subset, the outlier has the same overall keyword as the core words, but is from a subgroup with a different lower keyword (meaning that there is a rather clear semantic shift between the three related nouns and the outlier). In the **fine** subset, the outlier has the same lower keyword as the three core nouns, and the semantic difference is therefore rather subtle.

In the relatedness dataset, the three core nouns in the **medium** subset are randomly selected among all nouns in the same thesaurus section, while the outlier is from another section within the same chapter. In the **fine** subset, the core nouns share keyword, while the outlier has another keyword in the same section. In this dataset, it turned out that many words were not clearly related to a specific theme, and often had a high degree of polysemy, making it all together difficult to identify the set of core words as well as the outlier. We therefore manually selected the best examples from each chapter.

6. Datasets on Word Sense Disambiguation Based on The Central WordRegister of Danish

The synonymy, similarity, and relatedness datasets focus on word level semantic nuances. However, it is also relevant to evaluate whether a model can actually disambiguate words in contexts. For this purpose, we include the task of word sense disambiguation in the form of a

Subset	Core group	Outlier
Similarity medium	*ildkugle, stjerneskud, meteor* 'fireball', 'shooting star', 'meteor'	*komethale* 'comet tail'
Similarity fine	*ildkugle, stjerneskud, bolide* 'fireball', 'shooting star', 'bolide'	*meteorit* 'meteorite'
Relatedness medium	*halvmåne, meteorit, gassky* 'crescent', 'meteorite', 'gas cloud'	*frontdannelse* 'front formation'
Relatedness fine	*måneår, selenologi, månefase* 'lunar year', 'selenology', 'moon phase'	*natside* 'nightside'

Table 3: Examples from the different subsets of the word intrusion task.

Word-in-Context (WiC) task.

To this end, we compile a Danish word-in-Context (Pilehvar and Camacho-Collados, 2019) dataset based on the sense inventory of the COR.SEM (Central Word Register, the lexical-semantic component) lexical resource (Nimb et al., 2022; Pedersen et al., 2022), available from `https://ordregister.dk/`. Although a Danish WiC dataset is already available through the XL-WiC benchmark (Raganato et al., 2020), we compile a new one based on the sense inventory of COR.SEM, which is slightly more coarse-grained and thus better suited for the task.

From the COR.SEM resource, we extract both the monosemous lemmas with at least two usage examples and the polysemous lemmas with at least one usage example. Then we create all possible combinations of usage example pairs within the same lemma. The label depends on whether the usage examples are taken from the same or different COR.SEM senses. Since a lemma can have a varying number of senses and a sense can have a varying number of linked usage examples, we restrict the number of instances per lemma to a maximum of three ”same sense” cases and three ”different sense” cases. We also split the final data into a monosemous and polysemous subset.

7. Performance Ceiling of Selected State-of-the-art LLMs

In order to evaluate whether our datasets are actually challenging enough for some of the most recent LLMs, we prompt ChatGPT 3.5 turbo and ChatGPT 4.0 with our test instances.

The prompts consist of a task descriptions and a template. The task descriptions follows a structure of (a) a few sentences that describes the input from the dataset (i.e., *I will give you a list of words* or *I will give you three sentences*), (b) a description of what we expect the model to do (i.e., *You need to identify the outlier* or *You need to answer whether the sentence is True or False*, and (c) instructions to limit the models answers (i.e., *You must answer ”True” or ”False”*. The latter is important to minimise the amount of post-processing necessary for evaluating the model output.

The template consists of a few lines with placeholders for the input from the datasets. An example of the template for the WiC-tasks looks like this: *Target word is [input_1] and the sentence pair is [input_2] and [input_3]*. How much information is inserted into the template depends on the task. For the inference and entailment tasks, we run a few-shot setup with one or two examples given as a kind of precondition. The remaining four tasks are all run as zero-shot. To prevent the models from learning from previous examples, we open a new chat for each data instance.

We prompt exclusively in Danish, since our focus in on the Danish language. We have performed some preliminary tests with both English and Danish prompts. Our overall observations are that the models are better at following instructions when prompted in English (e.g., returns a single word when we ask for it), however if we clean the responses, the Danish prompting results in similar if not better results.

7.1. Test Results

Figure 2 shows the average accuracy of the two ChatGPT models across the six tasks. Overall, the larger ChatGPT 4.0 model outperforms the 3.5 turbo model on all tasks, achieving accuracies ranging from 67% (Relatedness) to 97% (Synonymy). When running the tests, we also observe that ChatGPT 4.0 was more precise to follow the instructions and thereby produce less variable outputs. For instance, ChatGPT 3.5 turbo occasionally return a phrase like *the answer is* even when instructed to only answer with only a single word.

7.1.1. Results on Inference and Entailment from DanNet and FrameNet

Figure 3 shows the models' performance across the four Qualia Roles (function, origin, parts etc). We observe that ChatGPT performs perfectly on the AGENTIVE subset relating to how things have

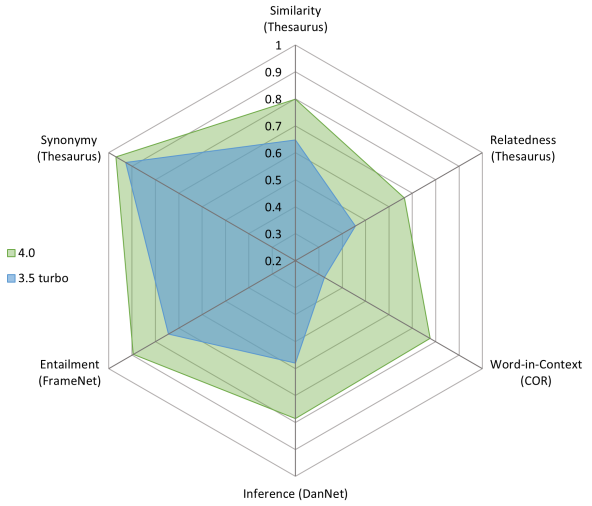

Figure 2: Model performance overview across tasks. The performance is calculated as the average accuracy for all datasets within the task.

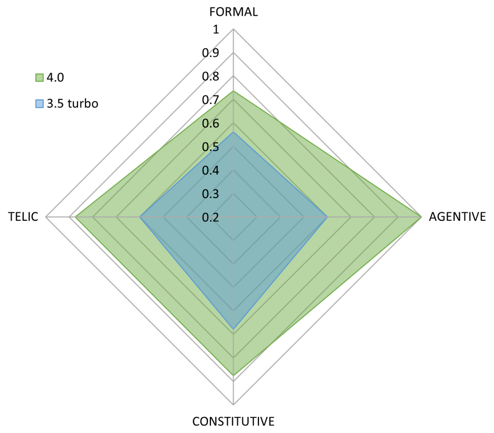

Figure 3: Model performance on Qualia Roles.

come about and gets decent results on the CONSTITUTIVE subset referring to the part-whole relations. Figure 4 shows how concrete ontological types achieve better results than abstract types like time and feelings. Animals and plants have remarkably high performance, probably due to international consensus on biological taxonomies.

Figure 5 shows the accuracy of the models on each semantic frame subset. ChatGPT 4.0 reaches a high accuracy on the TEMPORARY LEAVESTAY frame (100%) and the CAUSE frame (95%), which is quite impressing. TELLING frames have the lowest performance on both ChatGPT 3.5 turbo and 4.0 indicating that the result of something being told to somebody is not as obvious as the result of more concrete events like bringing and leaving events.

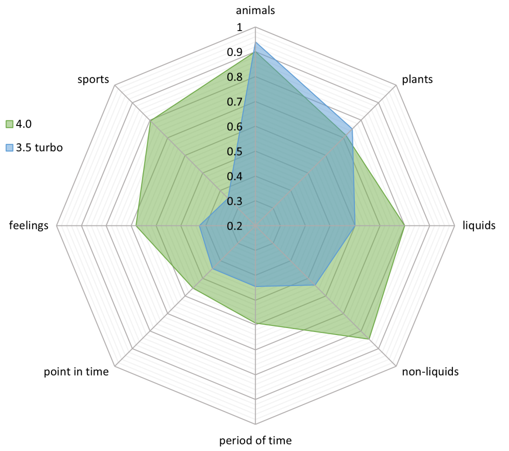

Figure 4: Model performance on a selection of ontological types (FORMAL Role).

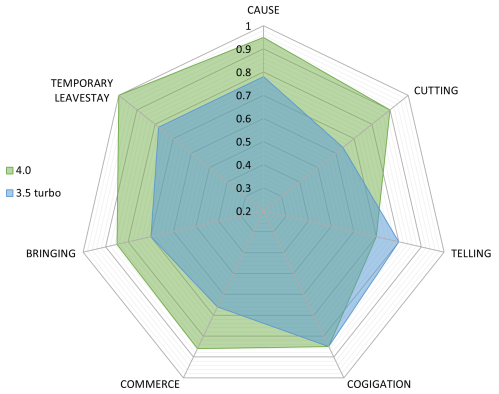

Figure 5: Model performance on a selection of semantic frames.

7.1.2. Results on Synonymy, similarity, and relatedness from the Danish Thesaurus

Figure 6 shows the performance of the models on the tasks related to sense granularity. First of all, it proves evident that both models reach a high performance on the synonymy task, which indicates that the current version of our task is not challenging enough for the latest LLMs. When we investigate the 10 errors made by ChatGPT 4.0, we find that it is possible to confuse the model with similar words (i.e., the next closest word in the list), in particular if the target or synonym is not a commonly used word.

In the similarity word intrusion task, we see that the fine-grained subset is more difficult for the models, as we expected it to be. When investigating the instances where both models are wrong, we are able to identify a number of possible rea-

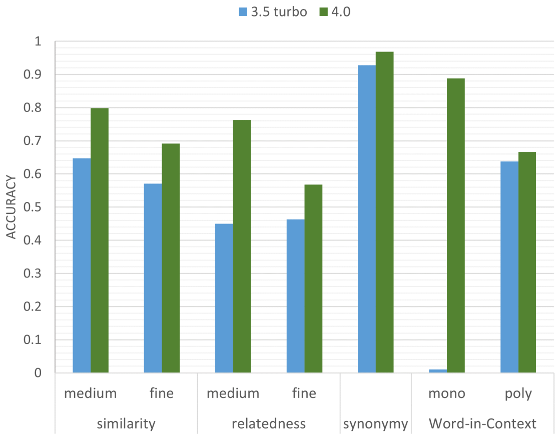

Figure 6: Model performance across the similarity, relatedness, synonymy, and DanWiC datasets

sons. 30% of the errors are probably due to the metaphoric sense of one or more of the words. In the set consisting of [*kvantespring* ('quantum leap' / 'quick development'), *lavine* ('avalanche' / 'quick development'), *cyklus* ('cycle'), *tigerspring* ('tiger jump' / 'quick development')] where all core words are metaphors with the same sense, the models wrongly estimate the outlier to be *lavine* instead of *cyklus* ('cycle'). When the outlier is thematically related to the core words, the risk of mistake seems to be higher, e.g. in the case of the set [*elefantvæddeløb* ('elephant race / overhaul among trucks'), *væddeløbskørsel* ('racing drive / overhaul'), *bykørsel* ('city driving'), *overhaling* ('overhaul')]. The outlier is *bykørsel*, however the models estimate it to be *elefantvæddeløb*, probably not considering the metaphorical sense 'overhaul among trucks' of the word and in the same estimating the outlier *bykørsel* to relate to traffic as do two of the core words. Highly polysemous words seem to cause problems to the model. So do rare words, whether they are old, domain specific, or not common in written language (e.g. slang). When it comes to domains, the models perform worse on the 10% of the data which is extracted from the section 'Sex, sexual desire'. Here we find 15% of the errors, maybe due to the many slang words.

The relatedness word intrusion task shows to be even more challenging than the similarity task. This mirrors an observation made during the validation of the data. The related core groups are less homogeneous than the similar core groups. Thus, it is more difficult to see the connection between the words.

7.1.3. Results on Word Sense Disambiguation Based on DanWiC from COR.SEM

The results on the Danish WiC task are also shown in figure 6. Here, we can see that ChapGPT 3.5 turbo almost exclusively answers "different sense" in the monosemous subset, resulting in an accuracy of only 1%. When we examine the answers on the polysemous subset, we see that the model only uses the "different sense" category. Thus, the model can not solve the task in the current setup. On the other hand, ChatGPT 4.0 does answer more evenly across the two categories and even get a 88% accuracy on the monosemous subset. However, the polysemous subset still appears to be challenging.

8. Feedback to the Lexical-Semantic Resources

A side-effect of the compilation of our datasets, is that it has in fact provided us with interesting feedback regarding the shape of our lexical resources.

For DanNet, we have become painfully aware that very peculiar word meanings (e.g. colloquial or old-fashioned use) cause confusion or at least lead to contra-intuitive utterances in several cases, and it seems more beneficial to exclude these meanings prior to compiling the dataset. It has also been verified that abstract entities do not intuitively follow a taxonomical structure to the same extent as do concrete entities, even if they are organised in a highly structured way in the wordnet.

Regarding the FrameNet Lexicon we have become more aware of the discrepancies wrt. telic and atelic verbs, i.e. whether a result is accomplished inherently by the verb itself or whether some adverbial particles in the surrounding context are decisive for telicity. This confusion has lead to the production of some odd utterances in several cases. In fact, assigning frames to the Danish verb vocabulary is a challenge. Many verbs are highly polysemous, also regarding frames, and they have a high tendency of mutating into phrasal verbs. It is therefore to a very high degree the textual context which defines the frame.

The use of data from the thesaurus confirms that it might be useful to improve the semantic structure in the subgroups by adding information on close synonymy (automatically extracted from the Danish Dictionary), as well as on co-hyponymy in order to distinguish such cases from other types of near-synonymy. The manually annotated data from our experiments constitute a first step.

With our experiments we have also tested the level of sense granularity of COR.SEM and have learned that some adjustments are needed, for instance it would be beneficial to clean up old spellings in the usage examples for each sense.

9. Conclusions and Future Work

Lexical-semantic resources contain valuable knowledge relating to basic reasoning, and in this paper we have pursued the hypothesis that this information can be used as a hand-crafted gold-standard or 'ground truth' for compiling evaluation data to assess LLMs' ability to cope with basic meaning aspects in text. We have presented the first version of such a semantic benchmark for Danish comprising six different semantic tasks, along 25 datasets and with 3,800 test instances. Although still relatively limited in size, our method has proven feasible and interesting in that it uncovers aspects of meaning and vocabulary not tested for Danish in any other available datasets, to our knowledge.

We will in future focus on scaling up the datasets to include more test instances and to thereby test more broadly the mastering of vocabulary and different semantic and ontological categories. We will work on making each task harder, for instance by testing the models' ability to capture the appropriate 'tone of voice' for synonyms and by testing narrower topics for the relatedness task. Last but not least, we would like to expand to other, harder reasoning tasks, including the comprehension of metaphorical expressions and other figurative patterns of speech, and in addition, make experiments as to test human attempts on selected tasks.

Experiments with ChatGPT have given us an indication of the fact that state-of-the-art LLMs have not yet achieved the performance ceiling of our datasets, although we are quite close with ChatGPT 4:0 for some semantic categories. Future work will include a more careful study of how also smaller language models trained on less data perform on our datasets; and, for instance, how well our datasets can identify model flaws or limitations in training data size and number of parameters. Further, it will be interesting and relevant to examine to which degree the models' performance on our benchmark correlate with their overall performance on regular downstream tasks.

10. Bibliographical References

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet Project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Aleksandrs Berdicevskis, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen, and Nina Tahmasebi. 2023. Superlim: A Swedish Language Understanding Evaluation Benchmark. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8137–8153, Singapore. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hjorth, Ebba and Kristiansen, Kjeld, editor. 2005. *Den Danske Ordbog*. Det Danske Sprog- og Litteraturselskab & Gyldendal. Online: ordnet.dk/ddo.

Dan Saattrup Nielsen. 2023. ScandEval: A Benchmark for Scandinavian Natural Language Processing. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 185–201, Tórshavn, Faroe Islands. University of Tartu Library.

Finn Årup Nielsen and Lars Kai Hansen. 2017. Open semantic analysis: The case of word level semantics in Danish. *Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 415–419.

Sanni Nimb. 2018. The Danish FrameNet Lexicon: method and lexical coverage. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Sanni Nimb, Anna Braasch, Sussi Olsen, Bolette Sandford Pedersen, and Anders Søgaard. 2017. From Thesaurus to Framenet. *Proceedings of eLex 2017*, pages 1–22.

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.

Sanni Nimb, Nicolai H. Sørensen, and Thomas Troelsgård. 2018. From Standalone Thesaurus to Integrated Related Words in The Danish Dictionary. In *Proceedings of the XVIII EURALEX*

International Congress: Lexicography in Global Contexts, pages 916–923, Ljubljana, Slovenia. Ljubljana University Press, Faculty of Arts.

Sanni Nimb, Lars Trap-Jensen, and Henrik Lorentzen. 2014. The Danish Thesaurus: Problems and Perspectives. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, pages 15–19.

Sanni Nimb, Lars Trap-Jensen, Henrik Lorentzen, Liisa Theilgaard, and Thomas Troelsgaard. 2015. *Den Danske Begrebsordbog*. Det Danske Sprog- og Litteraturselskab.

Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors. 2023. *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics, Toronto, Canada.

Bolette Pedersen, Anna Braasch, Anders Johannsen, Héctor Martínez Alonso, Sanni Nimb, Sussi Olsen, Anders Søgaard, and Nicolai Hartvig Sørensen. 2016. The Semdax Corpus Sense Annotations with Scalable Sense Inventories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Bolette Pedersen, Sanni Nimb, Sussi Olsen, Thomas Troelsgård, Ida Flörke, Jonas Jensen, and Henrik Lorentzen. 2023. The DA-ELEXIS corpus - a sense-annotated corpus for Danish with parallel annotations for nine European Languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 11–18, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

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, 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*, 43:269–299.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A Multilingual Benchmark for Evaluating Semantic Contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel, and Anna Palatkina. 2023. NorBench – a Benchmark for Norwegian Language Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 618–633, Tórshavn, Faroe Islands. University of Tartu Library.

Nina Schneidermann, Rasmus Hvingelby, and Bolette Sandford Pedersen. 2020. Towards a Gold Standard for Evaluating Danish Word Embeddings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4754–4763.

Simone Tedeschi, Johan Bos, Thierry Declerck, Jan Hajič, Daniel Hershcovich, Eduard Hovy, Alexander Koller, Simon Krek, Steven Schockaert, Rico Sennrich, Ekaterina Shutova, and Roberto Navigli. 2023. What's the meaning of superhuman performance in today's NLU? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12471–12491, Toronto, Canada. Association for Computational Linguistics.

Piek. Vossen, editor. 1999. *EuroWordNet. A Multilingual Database with Lexical Semantic Networks.* Kluwer Academic Publishers.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.