How do We Treat Systematic Polysemy in Wordnets and Similar Resources? – Using Human Intuition and Contextualized Embeddings as Guidance

Nathalie Sørensen[1], Sanni Nimb[2] & Bolette S. Pedersen[1]

Centre for Language Technology, NorS, University of Copenhagen[1], Society for Danish Language and Literature [2]
Emil Holms Kanal 2, 2300 Copenhagen S[1], Christian Brygge 1, 1219 Copenhagen K [2]
nmp828@hum.ku.dk, sn@dsl.dk, bspedersen@hum.ku.dk

Abstract

Systematic polysemy is a well-known linguistic phenomenon where a group of lemmas follow the same polysemy pattern. However, when compiling a lexical resource like a wordnet, a problem arises regarding when to underspecify the two (or more) meanings by one (complex) sense and when to systematically split into separate senses. In this work, we present an extensive analysis of the systematic polysemy patterns in Danish, and in our preliminary study, we examine a subset of these with experiments on human intuition and contextual embeddings. The aim of this preparatory work is to enable future guidelines for each polysemy type. In the future, we hope to expand this approach and thereby hopefully obtain a sense inventory which is distributionally verified and thereby more suitable for NLP.

1 Introduction

Systematic polysemy, also called regular polysemy, is a well-known linguistic phenomenon where a group of lemmas follow the same polysemy pattern (Apresjan 1974, Malmgren, 1988, Pustejovsky 1995, Nimb 2016 and several others). For instance, the lemmas *chicken* and *school* belong to the patterns ANIMAL/FOOD and LOCATION/INSTITUTION due to their inherently dual meanings with different ontological types.

The phenomenon is challenging to describe in theoretical linguistics as well as in practical lexicography where decisions need to be made regarding whether to split regular polysemous lemmas into several senses, or whether to see the meaning of these lemmas as inherently complex, with the individual context simply highlighting one or the other meaning. At times, a context does not specify any of the meanings and may highlight both equally. This kind of *underspecification* (Cruse, 1986) thus invokes two ontological types simultaneously, as seen in sentence a), where *taste* highlights a FOOD reading of *salmon*, while *lived a good life* draw attention to the ANIMAL reading:

a) You can taste if the **salmon** has lived a good life.

In lexicons, systematic polysemy can be dealt with in two ways (Vicente and Falkum, 2017, Ruhl, 1989). First, a *sense enumeration lexicon* can be established where different readings of a lexical item are listed under a single dictionary entry. In the case of *salmon*, such an approach would list both the ANIMAL and FOOD sense. This method is typically used in traditional dictionaries. Alternatively, it can be treated with a *one-representation approach* motivated by the fact that it is impossible in praxis to list all existing meanings of a lexical item. Instead, the lexicon describes regular patterns of sense alternations which also predict senses in a systematic way. A well-known example of the one-representation approach is provided in *The Generative Lexicon* (Pustejovsky, 1995). According to this approach, the *salmon* would be considered a *complex type* that denotes both the living animal and its corresponding meat.

This paper describes the challenges of achieving a homogenous approach to represent systematic polysemy in lexical resources and discusses when to rely on *a sense enumeration* approach and when to underspecify. We perform our studies within the framework of the COR[1] lexicon, which is based on previous lexical

[1] The Danish abbreviation of ‘the central word register'.

resources that were not consistent in their treatment of systematic polysemy. Overall, COR aims towards a restricted sense inventory where only distributionally 'verified' senses are maintained.

The new lexicon is primarily based on the corpus-based monolingual Danish dictionary: *Den Danske Ordbog* (DDO). Even though the dictionary mostly follows a sense enumeration approach, it occasionally uses a joint sense description for instances of systematic polysemy, typically in the case of less frequent lemmas in the corpus. In the COR lexicon, we rely heavily on our experience from compiling two other resources based on the DDO dictionary. First, in the Danish WordNet project *DanNet* (Pedersen et al., 2009), in which we took steps towards expanding the representations for specific systematic polysemy patterns, see Pedersen et al. (2010). Later, we compiled a Danish thesaurus based on senses in DDO and DanNet (Nimb et al., 2014, 2016). We also take inspiration from Alonso (2013), who examines expert and laymen annotations of the underspecified sense, however only on selected number of patterns.

In the COR project we aim at a homogenous treatment of similar polysemy patterns throughout the whole vocabulary, and with specific information on the type of pattern as part of the lexical semantic information. We adopt a similar idea to Nimb (2016) who suggests a method for systematic polysemy detection through lexical resources. The strategy is based on the initial hand annotations of a set of polysemous lemmas in DDO, which are again informed with information from DanNet. Thereby, we examine the vocabulary both bottom-up and top-down to establish a typology of Danish systematic polysemy patterns. The registered patterns lead to a set of rules stating whether the senses of a certain pattern must be reflected as either one or two COR lexicon senses. A subset of these rules is supplemented by two additional investigations, namely i) surveys on the human intuition, and ii) distributional investigations using a large, contextualised embedding model (BERT).

The idea of evaluating systematic polysemy by use of multiple information sources originates from the work of McCrae et al., (2022), who investigate an integrative method for distinguishing senses. They treat the sense distinction problem by including four perspectives: formal, cognitive, distributional, and multilingual. In our case, the combination of a formal semantic resource (DanNet), a study of the human intuition, and a distributional analysis, allows us to analyse systematic polysemy from different angles, including how the patterns are perceived by humans and used in texts. For instance, we investigate whether cases of systematic polysemy are conceptualised by humans as one or multiple senses by asking informants whether context pairs invoke the same or different senses. By using a distributional approach, we examine whether the ontological types in a pattern are represented in texts. This is particularly relevant in the application of NLP, as texts do not necessarily reveal the metonymic relationship between the senses in systematic polysemy, and distributional models may not be able to distinguish such senses.

The representation of systematic polysemy in lexical resources has been explored and discussed before (Peters & Kilgariff, 2000, Barque & Chaumartin, 2009). Although, to our knowledge, this is the first study to use both language models and informants to analyse systematic polysemy to compile valid encoding guidelines for a practical resource, in our case the COR lexicon. The study also gives valuable feedback to the treatment of systematic polysemy in DanNet and the DDO dictionary.

The structure of the paper is as follows. Section 2 introduces a typology of Danish systematic polysemy patterns. In Section 3 and 4, we present a preliminary study that analyses a selection of patterns in two ways, first using a survey of human intuition, then using a distributional model (BERT). In section 5, we discuss the interaction of the different approaches, and discuss how the treatment of systematic polysemy in lexical resources can benefit from the results.

2 A Typology of Danish Systematic Polysemy Patterns

In our annotation work, we have identified 28 Danish systematic polysemy patterns based on the compilation of the central vocabulary in the COR lexicon (Pedersen et al., 2022). The project is initiated by the annotation of ~3,300 polysemous lemmas in the DDO dictionary. We consider this a core vocabulary of Danish since they all have at least one sense which is linked to a core concept in the Princeton WordNet (PWN) (Fellbaum, 1998). In total, more than 15,000 senses are annotated.

In the dataset, all patterns of systematic polysemy are identified based on information in the sense definitions in DDO and the taxonomies in DanNet. The different patterns are analysed and discussed, resulting in a list of the most prominent systematic polysemy patterns in Danish.

As briefly mentioned above, an overall goal of the COR-project is not to reflect the fine-grained DDO sense inventory 1:1, but to compile a more coarse-grained sense inventory for Danish which is suitable for AI purposes and computational applications. By identifying the patterns, we can apply a homogenous analysis across multiple lemmas with the same patterns.

The starting point is the patterns registered in the projects DanNet (Pedersen et al., 2010) and the Danish thesaurus (Nimb, 2016), e.g., PROCESS/RESULT, PLANT/ FOOD, and ANIMAL/ FOOD. However in contrast to these projects, we consider the entire lemma information including all senses, and not just concepts represented as standalone DDO senses. This allows us to detect patterns of polysemy in a systematic way, lemma by lemma. For instance, it is typical for the lemmas that hold the pattern LOCATION/INSTITUTION to have ‘building/ location’ senses with similar definitions, which are typically listed under the same main sense as the ‘institution’ senses. By looking into DanNet, we can also compare the ontological types and thereby detect the patterns top-down.

During the discussion of the initially identified patterns, we questioned whether some patterns were actually cases of systematic polysemy or rather a case of the annotators being too eager to register patterns. Therefore, we include a pattern in the typology only if it fulfils the following three criteria:

a) At least five instances of the pattern can be found in the COR-dataset of ~3300 polysemous core lemmas.
b) The Danish Dictionary (DDO) or DanNet must distinguish between both senses of the pattern for most of the identified lemmas.
c) Each sense in a pattern must have distinct ontological types.

The criteria a) and b) ensure that a pattern is prominent in Danish by taking frequency and previous sense descriptions into account. If the pattern is systematic in Danish, we assume that it would be reflected in the core polysemous part of the Danish vocabulary. In the case of b), we must consider that the DDO in some cases prefers a single sense description. This is partly due to space limitations in the originally printed dictionary. The DDO typically uses sense enumeration when the lemma is frequent and a central simplex lemma, e.g., *bog* (‘book’), while for compound nouns (e.g., *kogebog* (‘cooking book’)) as well as more rare

Pattern	Examples
Group 1	1stOrder
ANIMAL / FOOD	*laks* ‘salmon’
PLANT / FOOD	*tomat* ‘tomato’
PLANT / MATERIAL	*eg* ‘oak’
ARTIFACT / MATERIAL	*sølv* ‘silver’
SHOP / PERSON	*bager* ‘bakery, baker’
ANIMAL (body part) / FOOD	*vinge* ‘wing’
BODY PART / GARMENT (part)	*ærme* ‘sleeve’
Group 2	2ndOrder (/1stOrder)
PROCESS / RESULT (concrete)	*bygning* ‘building’
ARTIFACT / ACTIVITY	*fodbold* ‘football’
ARTIFACT / PROPERTY	*sølv* ‘silver’
ACT / EVENT	*bøje* ‘bend’
Group 3	1stOrder / 3rdOrder
CONTAINER / CONTENTS	*glas* ‘glass’
LOCATION / INSTITUTION	*skole* ‘school’
ARTIFACT / FORM	*klokke* ‘bell’
ARTIFACT / CONTENT	*bog* ‘book’
ARTIFACT(s) / INSTITUTION	*arkiv* ‘archive’
OBJECT / SYMBOL	*hjerte* ‘heart’
COUNTABLE / UNCOUNTABLE	*øl* ‘(a bottle of) beer, (the liquid) beer’
Group 4	2ndorder / 3rdorder
PROCESS / RESULT (abstract)	*forandring* ‘change’
ACT / THOUGHT	*metode* ‘method’
ACTIVITY / INSTITUTION	*cykelløb* ‘bicycle race’
ACT / INSTITUTION (acting)	*administration*
ACT / COMMUNICATE	*pive* ‘whine’
EVENT / POINT IN TIME	slutning ‘ending’
ACT / SOUND	*klask* ‘smack’
Group 5	3rdOrder / 3rdOrder
DANCE / MUSIC STYLE	*disko* ‘disco’
TASK / INSTITUTION	*autoritet* ‘authority’
AREA OF KNOWLEDGE / SCHOOL SUBJECT	*matematik* ‘mathematics’

Table 1: Overview of the systematic polysemy typology. We group the 28 patterns based on Lyons’ semantic divisions (Lyons, 1977).

lemmas it includes both senses in only one definition (often indirectly, for example by referring to the genus proximum), e.g., *bog* ('book') which has two senses for *kogebog*.

Criterion c) excludes patterns found for adjectives describing people vs. objects or acts as in 'an ambitious student' vs. 'an ambitious jump'. In these cases, one could argue that the contrast lies within the described ('student' and 'jump') rather than the descriptor ('ambitious'). Another excluded pattern regards acts with or without a realized cognate object, e.g. at *svømme* ('to swim') and at *svømme crawl* ('to swim crawl').

For each pattern, we decide whether the sense descriptions should be enumerated or combined. A combined sense gets the ontological type of the most prominent sense, unless both senses are evaluated as being equally important. In that case, the merged sense will be assigned two ontological types. In all cases, the pattern is labelled explicitly in the lexicon The decisions are based on the available information from DanNet and DDO and supplemented by introspection and searches in corpora.

We further partition the 28 patterns into five groups based on Lyons' semantic divisions (Lyons, 1977). Thus, patterns that only include semantically concrete types fall into one group, while patterns that include a mix of concrete and abstract types fall into another. The groups are shown in Table 1[2].

3 Humans' intuition on systematic polysemy – an experiment

To support our set of polysemy rules, we first, examine the phenomenon by including investigations on the *human intuition*.

3.1 A systematic polysemy dataset

We limit this preliminary study to four patterns. First, we analyse ANIMAL/FOOD, and PLANT/FOOD as they have been considered during the compilation of DanNet (Pedersen et al., 2010). In addition, the ontological types in the patterns are all concrete (group 1) characterized by the contrast between the botanical/zoological world and the function as food. We examine two patterns that have an abstract INSTITUTION sense in common, i.e., the patterns ACTIVITY/ INSTITUTION (group 4) and LOCATION/INSTITUTION (group 3). These patterns are challenging since the meaning is quite often underspecified.

We compile a small dataset with contexts for eight target lemmas: *laks* ('salmon'), *jomfruhummer* ('langoustine'), *kål* ('cabbage'), *forårsløg* ('spring onion'), *badminton*, *ishockey* ('ice hockey'), *parlament* ('parliament'), and *hospital*. Each context is hand labelled with a broad ontological type (e.g., PLANT, FOOD, LOCATION). To facilitate this task, we restrict the target lemmas to those who have no more than two senses in the DDO dictionary, as well as no homonyms.

As previously mentioned, DDO is inconsistent in the treatment of systematic polysemy as it varies between a sense enumeration approach and a one-representation approach. Generally, a one-representation approach is used for low frequent lemmas, although it is not always the case. For instance, the high frequency lemma *hospital* is described as only having a LOCATION sense in DDO, even though it can be understood as both a LOCATION and/or an INSTITUTION. This might be an illustration of the duality of the systematic polysemy patterns – it is difficult to separate the senses as they co-exist. For this reason, we select two lemmas for each pattern: one with exactly two senses in DDO that corresponds to the senses in the pattern, and a DDO monosemous example.

We retrieve the contexts from *KorpusDK* – a Danish text corpus of 110 million words collection from the period 1985-2010. We randomly select 100-200 contexts for each target lemma. We hand-label approx. 60 with ontological types. The reduced number of annotated contexts is caused by three factors. First, we aim at having the same number of contexts for each target lemma. Secondly, we balance the labels to ensure a fair representation of each sense of a pattern, although this might not reflect the actual frequency distribution of the senses in use. For instance, it was challenging to find LOCATION examples of *parlament* 'parliament' in the 200 contexts. Lastly, we exclude contexts containing named entities with the target. In particular, the INSTITUTION patterns included several named entities (e.g., *Dansk Ishockey Union* and *Herlev Hospital*).

[2] The typology with additional examples and our strategy is available at https://github.com/kuhumcst/pycor/

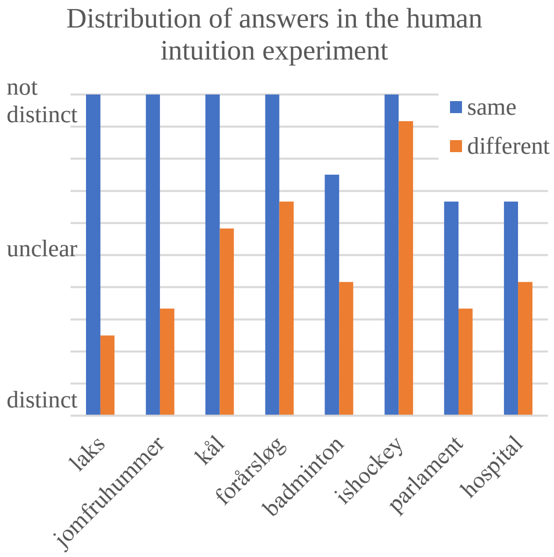

Figure 1: Distribution of answers across instances with either the same (blue) or two different (orange) ontological labels.

3.2 Experimental setup

The purpose of the experiment is to test the human intuition isolated from the task of creating a semantic lexicon. The question is whether the informants can distinguish senses of systematic polysemy given only minimal information. The experimental setup is inspired by the Word-in-Context task (Pilehvar & Camacho-Collados, 2018). The idea is that the participants are shown a target lemma and two contexts. The task is to answer whether the target lemma has the same sense in the two contexts.

The experiment is done through an online survey that consists of 24 context pairs and a few additional questions to ensure that the informants understand the task. Even if this is a low number of pairs, it resembles how intuitive the sense distinctions in patterns are. We frame the task as input to an automatic method for dictionary quote selection. Therefore, we ask the informants whether the two contexts would fit as quotes for the same sense entry.

Figure 1 shows the distribution of survey answers. The answers are divided by context pairs with the same ontological label (blue) or different labels (orange). A low column indicates intuitively distinct senses, while a tall column suggests no distinction of senses. Mid-range columns show cases without consensus among the informants.

We calculate a moderate agreement score of 0.49 using fleiss kappa (Fleiss, 1971). We see a large difference in the agreement depending on whether the contexts pairs have the same ontological label or not. In the pairs with the same label, we find a high agreement (0.72), while the agreement is drastically lower for pairs with different types (0.11). This means that the informants are close to guessing when the pairs differ in ontological type, and that it is indeed difficult to intuitively separate the senses of the patterns. This falls in line with the comments from some of the informants who comment that they are not consistent in their answers.

Some informants notice that the survey is related to systematic polysemy, and they report that the distinction in the concrete patterns (related to FOOD) is clearer than the more abstract patterns (related to INSTITUTION). This adds up with the actual results, where the most distinct pattern is ANIMAL/FOOD. The PLANT/FOOD is overall perceived as the same sense, although this is less clear as some participants still make the distinction. Generally, the INSTITUTION patterns are the least clear; they show lower agreement scores on instances where the ontological type is INSTITUTION for both contexts. We hypothesise that this can be caused by the patterns being even more complex due the relation between INSTITUTION and another ontological type, HUMAN_GROUP. We discuss this further in Section 6.

4 A distributional analysis with BERT

According to the distributional hypothesis, we can estimate the senses of a lemma from its distribution in language (Harris, 1954, Firth, 1957).. We investigate the distribution by performing a clustering experiment using the dataset described in section 0. The idea is to cluster the representations of a contextualised embedding model that has been pretrained on a large amount of textual data. If a systematic polysemy pattern is distinguishable in text, then the result will show separate clusters for each sense.

4.1 Model

We represent each occurrence of the target lemma with the base Danish BERT model which is

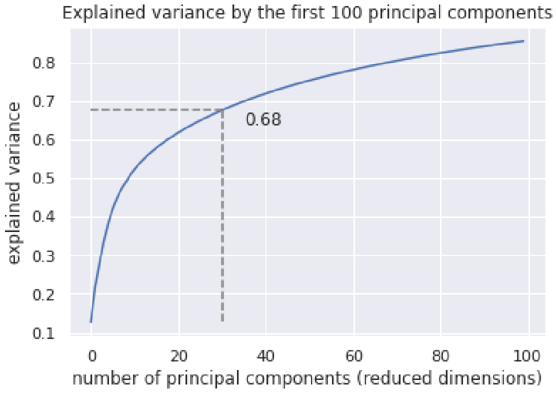

Figure 2: Explained Variance on the first 100 principal components from PCA experiment. The orange lines show the explained variance at 30 dimensions.

pretrained by Certainly[3]. The pretraining material included 1.6 billion words from different text sources (Common Crawl, Danish Wikipedia, web scraped forums, OpenSubtitles (Lison & Tiedemann, 2016)). To compute the contextualised target embedding, we first embed each context and then retrieve the token embedding corresponding to the target lemma. The token embedding is an average of the output of the last four layers.

4.2 Dimensionality reduction

Since our dataset contains a low number of samples (492) compared to the high dimensionality of the embeddings (768), it may be beneficial to reduce the dimensions in the embeddings[4]. The goal is to arrive at a level that retrains the most relevant information, but still reduces the complexity of the embedding space. We choose to reduce to 30 dimensions. We analyse this choice by an experiment with Principal Component Analysis (PCA). The purpose of PCA is to transform correlated dimensions into uncorrelated principal components that explain the most variance in the initial dimensions. Figure 2 shows how much variance each principal component can account for. We see that the first 30 principal components explain 68% of the variance in the 492 embeddings. Although, 32% of the variance is yet to be captured, any increase in the dimensionality does not give us drastic improvements. Instead, we attempt to retain more of the information by using a non-linear reduction technique: Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018). This technique has two advantages over other non-linear techniques: a) it takes more of the global structure of the data into account, and b) it can reduce to a higher number of dimensions (30 vs. 2-3). For the UMAP parameters, we use *cosine as* the distance metric and set min_dist to *0.0*.

4.3 Sense Clustering

We use a density-based clustering method: HDBSCAN[5] (Campello et al., 2013). The method is useful when we do not have any assumption about the shape, size, and number of clusters. We use the following parameter settings: *min_samples* =10 and *min_cluster_size*=15.

We apply the clustering method on the entire dataset and arrive at total of 11 clusters. The clusters are visualised in Figure 3 (FOOD related) and Figure 4 (INSTITUTION related) after further dimensionality reduction with UMAP. Of the eight lemmas, three have contexts distributed to multiple clusters: *laks* ('salmon'), *parlament* ('parliament'), and *hospital*. The remainder have a single cluster representation. To evaluate the clusters, we calculate an average silhouette score of 0.82 across all clusters. From this, we conclude that the clusters are distinct with a minimal to no overlap.

5 Discussion

In this section, we discuss how the formal, the intuition-based, and the distributional approaches, respectively, contribute to our understanding of the different cases of systematic polysemy. We start by analysing the how each pattern is formally represented in the lexical resource, DanNet.

ANIMAL/FOOD: All three approaches support that we separate our sense descriptions into an ANIMAL and FOOD sense. In DanNet, the pattern is consistently distinguished when both senses occur in DDO. Each sense has its own synset with non-overlapping taxonomic structures. In the survey, the participants are also able to recognize the contrast between the living animal and its meat.

[3] More information about the model is available here: https://github.com/certainlyio/nordic_bert

[4] This is to avoid the curse of dimensionality, where the high number of dimensions hinder the optimisation of algorithms.

[5] The python implementation is available here: https://hdbscan.readthedocs.io/en/latest/index.html

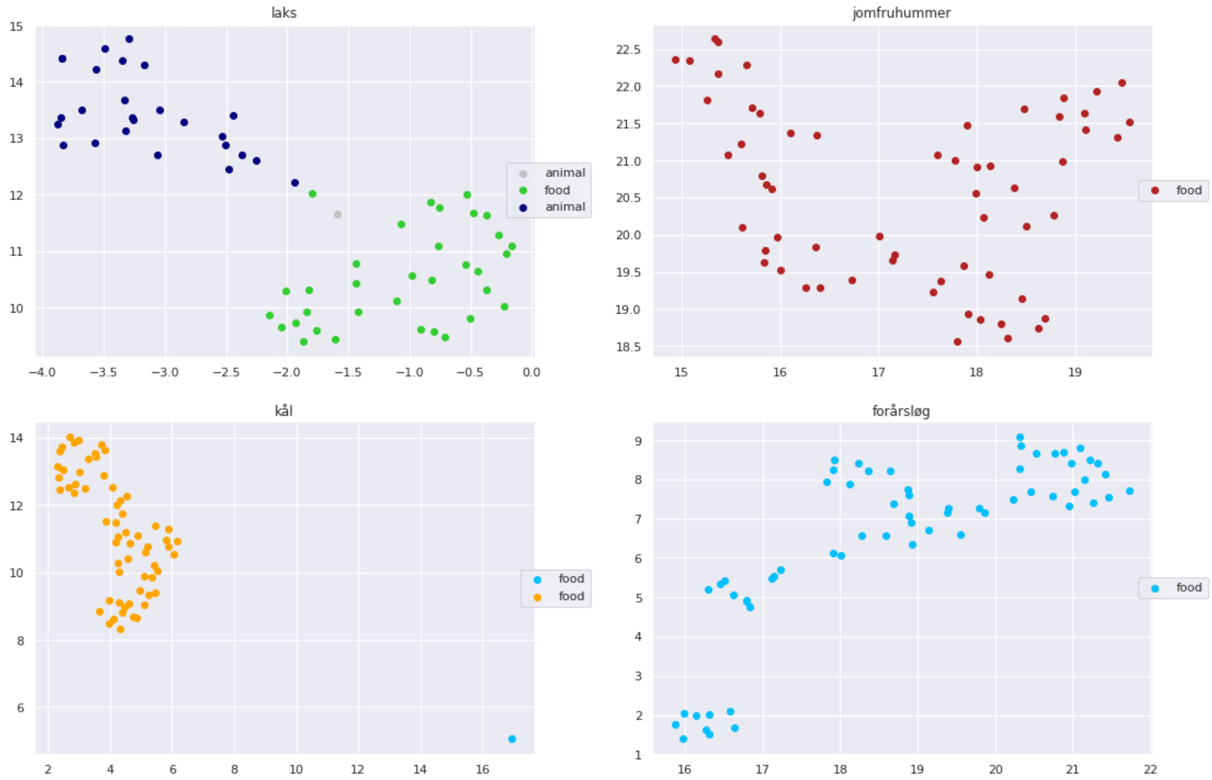

Figure 3: Sense clusters for lemmas with a FOOD sense and either one (right column) or two senses (left column) in DDO. The labels indicate the most common label in that cluster.

The distributional approach separates the senses of the pattern for the frequent target *laks* (‘salmon’) but does not separate the less frequent target *jomfruhummer* (‘langoustine’). We can explain the difference with the frequency of the FOOD sense of *jomfruhummer*. Since the BERT model is trained on mostly web crawled texts, we expect a high number of recipes and food reviews in the text collection. Therefore, the model might not have seen enough clear ANIMAL examples to create distinct representations. Unfortunately, we do not have access to the exact training data and cannot confirm this hypothesis. However, we do know that our lexical resources contain this missing real-world knowledge, although for the infrequent lemmas, we see a mismatch between the sense descriptions and the language use. In DanNet, the DDO genus proximum *dyr* 'animal' has led to only one sense, the ANIMAL sense, and the FOOD sense has not been included even though the example is food.

PLANT / FOOD: The approaches mostly support combined representation of the pattern. In DanNet, the specialist and folk taxonomies of plants are treated differently from animals. Here, the specialist and folk perspectives are merged in a single synset by using two hypernyms, related to PLANT and FOOD respectively. The dual taxonomies indicate that we can merge the pattern into a single representation that incorporates both ontological types depending on the situation.

The distributional analysis also supports a one-representation approach, although we see an error in the clustering. A single instance of *kål* ‘cabbage’ has been wrongly added to the same cluster of *forårsløg* ‘spring onion’. The confusion arises from the morphological similarity of the use of *kål* in that specific context and a typical use of *forårsløg*: the definite plural form (e.g., *kål -ene* and *forårsløg -ene*). This is one of the flaws of using a “black box” distributional model – we cannot guarantee that the BERT embeddings only include the semantic information and are not sensitive to other variation in the input. Still, a promising observation is the small sub-cluster on the bottom left of the blue cluster (‘spring onion’) on Figure 3. Here, we find an extra sense that we did not consider during the creation of the dataset. The PLANT sense can arguably be split into two: ‘the edible plant’ and ‘flower bulbs’ that are planted during the spring. With the current clustering parameters, this sub-cluster is too small to be represented as a separate

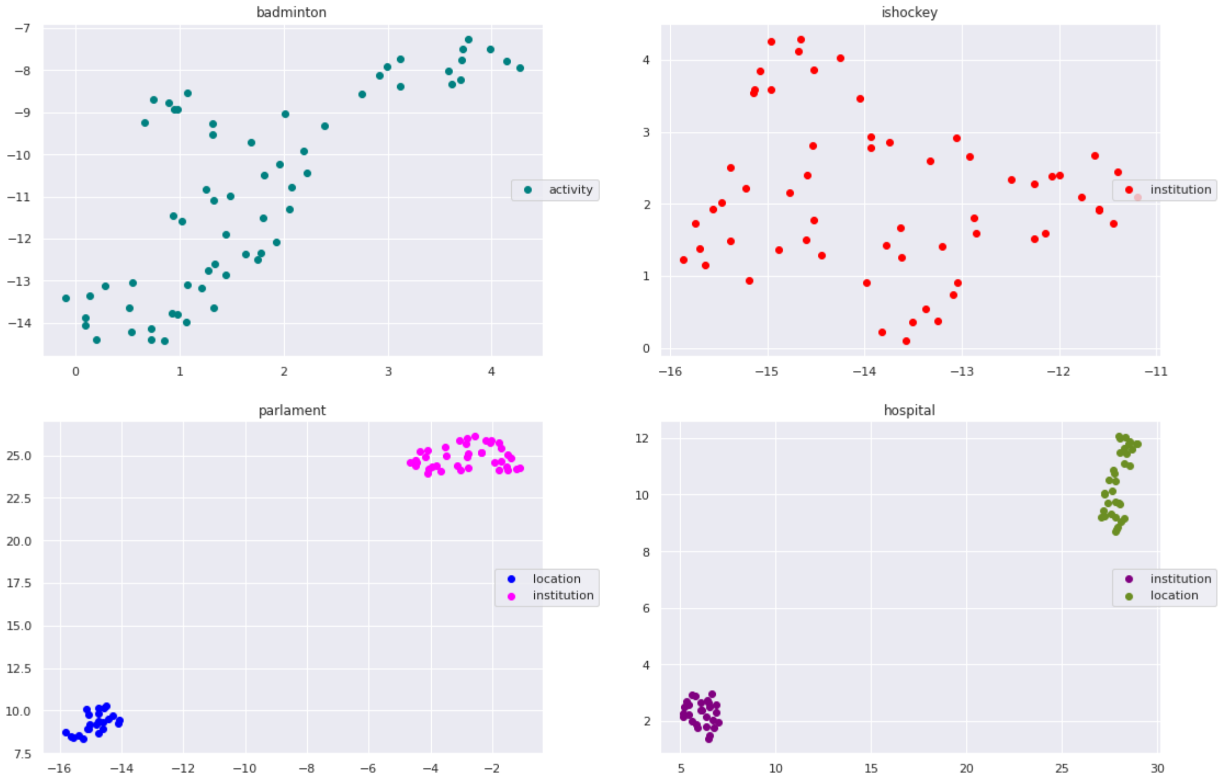

Figure 4: Sense clusters for lemmas with an INSTITUTION sense and either one (right column) or two senses (left column) in DDO. The labels indicate the most common label in that cluster.

cluster. The fact that they are still grouped together is a sign that BERT embeddings can detect this sense to some degree.

The human intuition survey gives a mixed result on this pattern. On the one hand, most of the informants do not distinguish between PLANT and FOOD. Yet, a few informants still detect the difference, and some even mention the pattern in their comments. The survey includes a low number of examples, and it is possible that the distinction is not expressed clearly enough in those examples. A further study with more contexts and target lemmas is needed for us to determine the human intuition on this pattern.

ACTIVITY / INSTITUTION: This pattern is not clearly distinguishable in neither DanNet, nor in the distributional analyses. Although DanNet includes both the ACTIVITY and INSTITUTION senses from DDO, we cannot find a contrast between these synsets as they are close to being structurally identical. Additionally, the hypernyms and ontological types only express the ACTIVITY sense. This questions why both synsets are maintained as they do not reflect the systematic polysemy patterns. In the survey, we see that *ishockey* ‘ice hockey’ is the only lemma where the informants almost all agree that there is no difference between the senses. In the case of *badminton*, about half of the informants distinguish between ACTIVITY and INSTITUTION. This can be related to *badminton* being a more widely known and played sport in Denmark and therefore more likely to be institutionalised. Along with the previously mentioned case of *jomfruhummer*, this shows the difficulty of making a top-down approach to polysemy. We must consider the story of each lemma and its presence in the language.

LOCATION / INSTITUTION: The possible third HUMAN_GROUP interpretation complicates the analysis of this pattern as is evident from the survey results. The complexity is also visible in DanNet, where three senses are sometimes included [6]. However, most often only the LOCATION/INSTITUTION contrast is maintained by a ‘concrete building’ synset and an ‘abstract institution’ synset, respectively.

[6] In some cases the dictionary that DanNet is based did not include all three senses, which means a manual effort has been put into DanNet to express this three-way pattern.

Surprisingly, *hospital* only has a LOCATION sense in DanNet. For cases like this, the distributional analysis tells us where we can improve our lexical resources, as the contrast between LOCATION and INSTITUTION is clearly reflected in the clusters. However, we note that there is no guarantee that the clusters can be directly mapped to distinct LOCATION and INSTITUTION senses. Being at the hospital is expressed by the preposition *på* 'on/at'. However, a strictly LOCATION reading could mean that one is physically on top of the building, whereas we usually mean that we are in a building. Thus, the LOCATION cluster may be a clustering of underspecified senses that superficially appears to highlight a concrete LOCATION. Likewise, if a politician is *in the parliament*, the context may highlight HUMAN_GROUP and/or INSTITUTION more than a LOCATION. To understand how we should interpret the clusters, we need to investigate which semantic information they contain and whether this corresponds to the sense descriptions in the lexical resources.

6 Future work

The approach described in this paper provides new insights into how to treat four frequent systematic polysemy patterns in the COR lexicon. A noticeable finding is that, as in the case of many other lexical phenomena, the patterns, and to some degree also lemmas within a pattern, tend to dispose quite individual properties. We would like to carry out similar investigations on the remaining part of the patterns in the typology we have presented, both in order to examine the prototypical behaviour for each pattern, and how this should correspondingly be represented in COR, but also to reveal the deviant cases. We think that by including information from both a survey among informants and statistical methods, we will be able to treat the many cases of systematic polysemy across the Danish vocabulary in a more appropriate manner.

References

Apresjan, J. D. (1974). Regular Polysemy. *Linguistics*, *142*(142), 5–32.

Barque, L., & Chaumartin, F. R. (2009). Regular polysemy in WordNet. *Journal for language technology and computational linguistics*, *24*(2), 5-18.

Campello, R. J., Moulavi, D., & Sander, J. (2013, April). Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining* (pp. 160-172). Springer, Berlin, Heidelberg.

Cruse, D. A., Cruse, D. A., & Cruse, D. A. (1986). *Lexical semantics*. Cambridge university press.

Devlin, J., Chang, M-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fellbaum, C. (ed.). (1998). *WordNet: An Electronic Lexical Database*. MIT press.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. *Studies in Linguistic Analysis.*

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146-162.

Lyons, J. (1977). *Semantics. Volumes 1-2*. Cambridge University Press.

Malmgren, S. G. (1988). On regular polysemy in Swedish. *Studies in computer-aided lexicology*, 179-200.

Martínez A., H. (2013). *Annotation of regular polysemy: an empirical assessment of the underspecified sense* (Doctoral dissertation, Universitat Pompeu Fabra).

McCrae, J. P., Fransen, T., Ahmadi, S., Buitelaar, P., & Goswami, K. (2022). Towards an Integrative Approach for Making Sense Distinctions. *Frontiers in Artificial Intelligence*, 3.

McInnes, Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.*

Nimb, S., Trap-Jensen, L., & Lorentzen, H. (2014). The Danish thesaurus: Problems and perspectives. In *Proceedings of the XVI EURALEX International Congress: The User in Focus* (pp. 15-19).

Nimb, Sanni (2016) Der er ikke langt fra tanke til handling. Simon Skovgaard Boeck & Henrik Blicher (eds.): *Danske Studier 2016*, København, Universitets-Jubilæets danske Samfund, pp. 25-59Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., & Lorentzen, H.

(2009). DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language resources and evaluation*, *43*(3), 269-299.

Pedersen, B. S., Nimb, S., & Braasch, A. (2010). Merging specialist taxonomies and folk taxonomies in wordnets-a case study of plants, animals and foods in the Danish wordnet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).*

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 Thirteenth Language Resources and Evaluation Conference* (pp. 51-60).

Peters, W., & Kilgarriff, A. (2000). Discovering semantic regularity in lexical resources. *International Journal of Lexicography*, *13*(4), 287-312.

Pilehvar, M. T., & Camacho-Collados, J. (2018). WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *arXiv preprint arXiv:1808.09121*.

Pustejovsky, J. (1995). *The Generative Lexicon.* MIT Press. Cambridge, Massachusetts.

Ruhl, C. (1989). *On monosemy: A study in linguistic semantics.* Albany: State University of New York Press.

Vicente, A., & Falkum, I. L. (2017). Polysemy. In *Oxford research encyclopedia of linguistics*.

Language Resource References

Danish Systematic Polysemy Typology and dataset: https://github.com/kuhumcst/pycor/tree/master/data/systematic_polysemy

DanNet: wordnet.dk; https://andreord.nors.ku.dk.

Den Danske Ordbog (DDO): Hjorth, E. & K. Kristensen red. (2003-2005). *Den Danske Ordbog*, volume 1-6, Det danske Sprog- og Litteraturselskab/Gyldendal, Copenhagen. Online: https://ordnet.dk/ddo

Den Danske Begrebsordbog (DDB): Nimb, Sanni, Henrik Lorentzen, Thomas Troelsgård, Liisa Theilgaard, Lars Trap-Jensen (2014). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab. Copenhagen.

KorpusDK. Det Danske Sprog- og Litteraturselskab. Copenhagen. korpus.dsl.dk/resources/details/korpusdk.html