Coarse-Grained Sense Annotation of Danish across Textual Domains

Sussi Olsen Bolette S. Pedersen Héctor Martínez Alonso Anders Johannsen
University of Copenhagen, Njalsgade 140, Copenhagen (Denmark)
{saolsen, bspedersen, alonso, ajohannsen}@hum.ku.dk

Abstract

We present the results of a coarse-grained sense annotation task on verbs, nouns and adjectives across six textual domains in Danish. We present the domain-wise differences in intercoder agreement and discuss how the applicability and validity of the sense inventory vary depending on domain. We find that domain-wise agreement is not higher in very canonical or edited text. In fact, newswire text and parliament speeches have lower agreement than blogs and chats, probably because the language of these text types is more complex and uses more abstract concepts. We further observe that domains differ in their sense distribution. For instance, newswire and magazines stand out as having a high focus on persons, and discussion fora typically include a restricted number of senses dependent on specialized topics. We anticipate that these findings can be exploited in automatic sense tagging when dealing with domain shift.

1 Introduction

It is commonly observed that word meanings vary substantially across textual domains, so that an appropriate sense inventory for one domain may be inappropriate or insufficient for another (Gale et al., 1992). This essential quality of the lexicon poses a huge challenge to natural language processing and underlines the need for developing systems that are generally less sensitive to domain shifts. The present work is framed within a project that deals with sense inventories of different granularity and across textual domains.

The overall goal is to discover what sense inventories and algorithms are manageable for annotation purposes and useful for automatic sense tagging. In this paper we experiment with coarse-grained annotations, and we analyze how reliable the annotations are and how much they vary over textual domains.

In Section 2 we present the backbone of our scalable sense inventory based on a monolingual dictionary of Danish. In Sections 3 and 4 we present the data, describing the different corpora, as well as the coarse-grained sense inventory. In Section 5 we present the differences in inter-coder agreement across the textual domains and discuss how the applicability and validity of the sense inventory vary depending on the kind of textual domain. Section 6 is devoted to comparisons of the relative frequency of selected supersenses across the six domains, and Section 7 describes the relation between specific senses via pointwise mutual information. Section 8 provides the conclusion for the article.

2 Scalable sense inventory

We operate with a sense inventory derived from the Danish wordnet (DanNet), which bases its sense inventory on a medium-sized Danish dictionary, Den Danske Ordbog (DDO). This is a pragmatic decision that leaves the more theoretical discussion aside of whether it is at all possible to define where one word sense starts and another begins (Kilgarriff, 2006). The ontological labels encoded in DanNet, based on the EuroWordNet top ontology as described in Pedersen et al. (2009) and Vossen (1998), have enabled us to automatically the word senses defined for the Danish vocabulary onto the cross-lingual supersenses. These are based on the Princeton Wordnet lexicographical classes[1] and have become a popular choice for coarse-grained sense tagging with the advantage of being applicable across languages.

[1] https://wordnet.princeton.edu/man/lexnames.5WN.html

Sussi Olsen, Bolette S. Pedersen, Héctor Martínez Alonso and Anders Johannsen 2015. Coarse-grained sense annotation of Danish across textual domains. *Proceedings of the workshop on Semantic resources and semantic annotation for Natural Language Processing and the Digital Humanities at NODALIDA 2015*. NEALT Proceedings Series 27 / Linköping Electronic Conference Proceedings 112: 36–43.

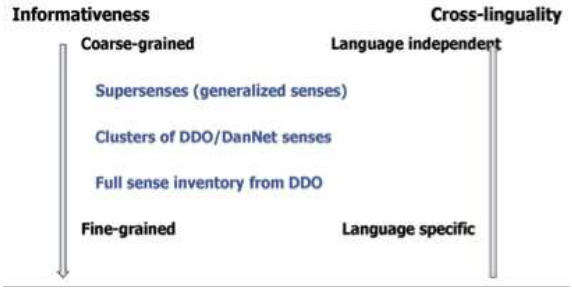

Figure 1: Scales of sense granularity.

All corpora have been automatically pre-annotated on the basis of this mapping, allowing the annotator to choose the appropriate supersense in context.

Figure 1 shows three points on the continuum of word sense granularity applied in the project, spanning from the supersense annotation experiment presented in this paper over clusters of DDO senses, to the highly fine-grained full sense inventory of DDO applied to lexical samples experiments (Pedersen et al., 2015).

3 Corpora across domains

In this paper we use the term *domain* (or textual domain) for text type or genre, and not for subject domain; i.e. our domains are categories like BLOG, CHAT, and MAGAZINE, instead of Politics, Geography or Literature. The texts for annotation have been selected from the Danish CLARIN Reference Corpus (Asmussen and Halskov, 2012), which is a general-language corpus of 45M words spanning several text types or domains, although with a predominance of newswire texts (48%) . We have taken care to include a broad range of domains in our annotation data set.

Table 1 lists the domains and text sources that have been selected for manual annotation from each domain. The rightmost column shows the names of the domains in this paper.

3.1 Corpus characteristics

We have characterized aspects of language use in the different textual domains with regard to average sentence length and the token/type ratio. The results of the analysis can be seen in Table 2.

Average sentence length is considerably larger for PARLIAMENT. These texts are originally speeches, written down by professional secretary staff, and long sentences are common in this genre. Apart from this, differences in sentence length between the textual domains are small. We initially expected the texts produced by professionals (NEWSWIRE and MAGAZINE) to have longer sentences than user-generated texts (BLOG, CHAT and FORUM), but found that for the user-generated content domains the language was similar to spoken language, and punctuation was less used, which may account for the longer sentences.

Domain	Av. sent.length	$\frac{\text{token}}{\text{type}}$	# sentences
BLOG	19.83	3.88	600
FORUM	22.22	3.22	300
CHAT	18.66	3.83	600
MAGAZINE	20.58	2.90	600
PARLIAMENT	32.49	5.07	600
NEWSWIRE	19.47	2.66	600

Table 2: Language characteristics of the textual domains.

The token/type ratio measures the variety of the vocabulary, or more precisely the average number of repetitions of each type. A higher token/type ratio thus means a less varied choice of vocabulary. PARLIAMENT is the domain with the highest token/type ratio. The domains BLOG and CHAT also have a rather high token/type ratio, which fits well with the annotators' impression that the language in these textual domains was homogenous with lots of repetitions. We find the highest lexical variation in the newswire domain.

3.2 Annotation process

The texts in our analyses were manually annotated by trained students. Our students annotate using WebAnno, a web-based annotation tool developed by Technische Universität Darmstadt for the CLARIN community (Yimam et al., 2013). Using WebAnno allows monitoring and measuring the progress and the quality of the annotation projects in terms of inter-annotator agreement.

More than half of the sentences have been annotated by two or more annotators in order to measure inter-annotator agreement, and most of these sentences have been adjudicated by a trained linguist. The remaining sentences have only been annotated by one annotator. Three annotators worked on the newswire texts, and two of them did the annotations on the remaining texts. Although these two annotators are skilled and, as demonstrated by the adjudication process, adhered closely to the instruction guidelines, the low num-

Source	Description	Domain
Bentes blog	A blog written by a woman in her forties	BLOG
Selvhenter	A chat forum mostly used by young people	CHAT
Se og Hør	A celebrity gossip magazine	MAGAZINE
Folketingstaler	Speeches from the Danish Parliament written down by professionals	PARLIAMENT
Mangamania	A chat forum for persons who love manga	FORUM
Politiken	A large Danish newspaper	NEWSWIRE

Table 1: The domains and texts included in the annotation data set.

ber of annotators may have adversely affected the results, leading to slightly biased data (see Section 5).

4 The extended supersense inventory

Basing the supersense inventory on the Princeton Wordnet lexicographical classes has the advantage of being inter-lingually comparable and interoperable, because wordnets for a wide range of languages are linked to Princeton Wordnet.

However, the supersense classes were not originally designed for sense annotation. During the annotation process, we discovered that some senses are needlessly coarse and in fact confound important distinctions. Therefore, we refine the Princeton supersense inventory with additional senses in cases where these cover large groups of easily detectable word senses in DanNet, such as diseases, body parts, institutions and vehicles. Because this is a process of refinement, we maintain compatibility with Princeton Wordnet. A new sense is introduced by subdividing an original sense and can thus always be unambiguously mapped back to the original sense.The full set of supersenses with the extensions can be seen in Table 3.

In total, the standard supersense set has been extended with seven noun categories and two verb categories. For adjectives, which only have a *catch-all* sense in Princeton Wordnet, we have added four high-level categories covering mental, social, physical and time-related property senses. The inspiration for the new adjective senses came from the four major sense groupings from the Danish wordnet. Finally, three tags for verbal satellites have been introduced to account for collocations, particles, and reflexive pronouns. While these satellite tags seemingly do not carry semantic meaning but are more grammatical in nature, they obtain a semantic interpretation in conjunction with a verb. In particular they ensure that a certain particle, pronoun or element of a collocation is understood as a lexical unit in conjunction with its preceding verb. To exemplify, Figure 2 shows how the phrasal verb *sige fra* (lit. say from, 'cancel') receives the supersense verb.communication, while the particle fra received the particle tag.

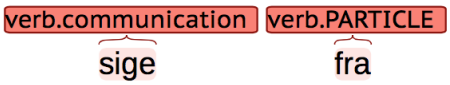

Figure 2: Annotation of phrasal verbs.

Ide and Wilks (2006), Brown et al. (2010) and more recently Melo et al. (2012) discuss coarse-grained sense distinctions for natural language processing, and Ciaramita and Johnson (2003) provide one of the first to use lexicographical classes as sense inventory for an automatic prediction task.

5 Inter-annotator agreement across domains

Over 50% of our data, 1900 sentences in total, has been doubly annotated with the aim of measuring and controlling annotator consistency. The disagreements inform on the validity of the sense inventory in general as well as for the different domains. They also provide hints about problematic, document-specific issues. Such issues were found for BLOG, for instance, which includes a frequent number of meta remarks where a certain feed can be found, as in:

> *Dette indlæg blev udgivet den tirsdag, 21. september 2010 kl. 10:14 og er gemt i Min have. Du kan følge alle svar til dette indlæg via RSS 2.0-feedet. (Bentes Blog)*

ADJ.ALL	NOUN.FOOD	**SAT.PARTICLE**
ADJ.MENTAL	NOUN.GROUP	**SAT.RELFPRON**
ADJ.PHYS	**NOUN.INSTITUTION**	VERB.ACT
ADJ.SOCIAL	NOUN.LOCATION	**VERB.ASPECTUAL**
ADJ.TIME	NOUN.MOTIVE	VERB.BODY
NOUN.TOPS	NOUN.OBJECT	VERB.CHANGE
NOUN.ABSTRACT	NOUN.PERSON	VERB.COGNITION
NOUN.ACT	NOUN.PHENOMENON	VERB.COMMUNICATION
NOUN.ANIMAL	NOUN.PLANT	VERB.COMPETITION
NOUN.ARTIFACT	NOUN.POSSESSION	VERB.CONSUMPTION
NOUN.ATTRIBUTE	NOUN.PROCESS	VERB.CONTACT
NOUN.BODY	NOUN.QUANTITY	VERB.CREATION
NOUN.BUILDING	NOUN.RELATION	VERB.EMOTION
NOUN.COGNITION	NOUN.SHAPE	VERB.MOTION
NOUN.COMMUNICATION	NOUN.STATE	VERB.PERCEPTION
NOUN.CONTAINER	NOUN.SUBSTANCE	**VERB.PHENOMENON**
NOUN.DISEASE	NOUN.TIME	VERB.POSSESSION
NOUN.DOMAIN	**NOUN.VEHICLE**	VERB.SOCIAL
NOUN.FEELING	**SAT.COLL**	VERB.STATIVE

Table 3: The standard supersense inventory with the added senses/satellite types in bold.

Domain	κ-agreement	% double annotated
BLOG	0.66	50 %
FORUM	0.54	66 %
CHAT	0.68	66 %
MAGAZINE	0.61	33 %
PARLIAMENT	0.59	33 %
NEWSWIRE	0.59	100 %
All domains	0.63	

Table 4: Inter-annotator agreement κ across domains together with the percentage of double annotated files.

> This feed was published Tuesday September 21 at 10:00 and is saved under My garden. You can follow all comments to this feed via the RSS 2.0 feed.

In such cases the annotators reached a consensus on how to tag the blog-specific metadata.

Table 4 shows that even if agreement results are generally good for the task (Artstein and Poesio, 2008), not all textual domains are equally easy to annotate. NEWSWIRE and PARLIAMENT show the lowest agreement, which is a somewhat surprising finding, because these texts are the most canonical and elaborate and thus arguably easier to understand and annotate. FORUM has 300 sentences, unlike the other domains, which have double the amount. This difference has an impact in the chance-correction measure of the κ coefficient, making the chance-adjustment more severe. However, NEWSWIRE has more semantic types than

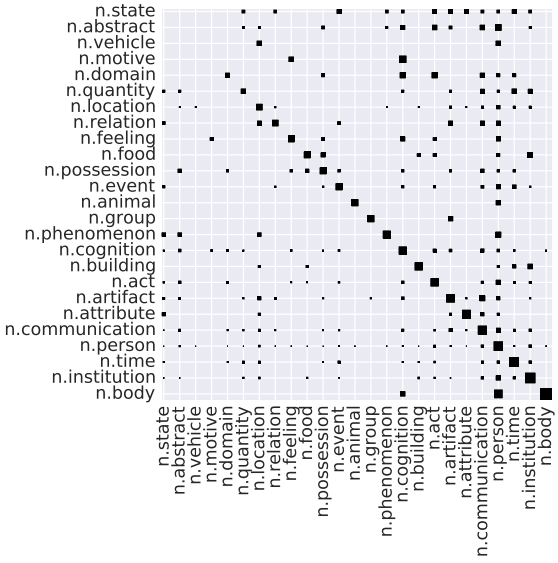

Figure 3: Disagreement for noun senses in NEWSWIRE.

e.g. BLOG (see Figure 3, 4 and 5), and the more varied the text, the more difficult it will be to annotate and achieve high agreement. Furthermore, PARLIAMENT texts are in a higher register than texts from BLOG or CHAT and include more abstract words (verb.cognition, noun.abstract).

Figures 3 to 5 illustrate the patterns of disagreement between annotators. The matrix is constructed by first gathering all of the words tagged by at least one annotator as, say, noun.abstract, observing what the other annotators tagged the same words as. Each cell in the plotted matrix measures the number of times two annotators tagged

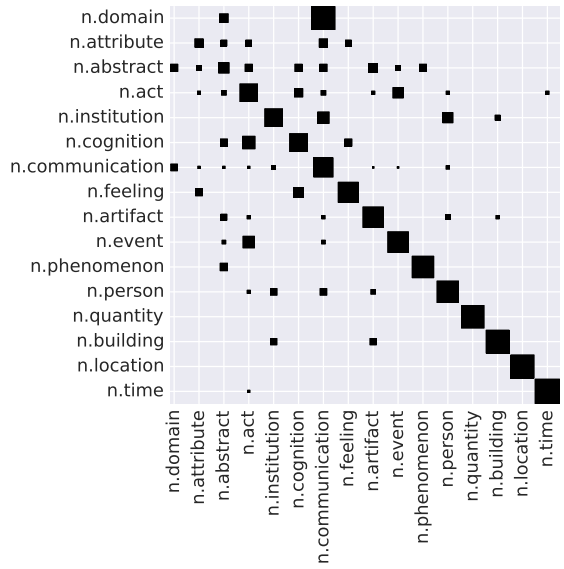

Figure 4: Disagreement for noun senses in BLOG.

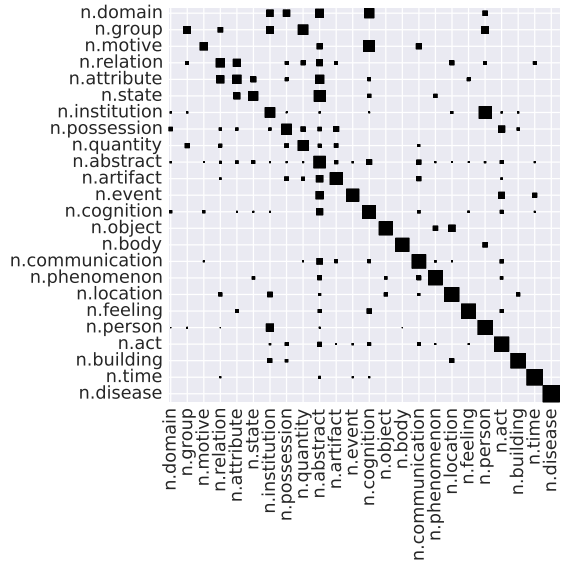

Figure 5: Disagreement for noun senses in PARLIAMENT.

Word	Conflicting annotation
musik (music)	noun.communication
dans (dancing)	noun.act
natradio (night radio)	noun.communication
design (design)	noun.attribute
kultur (culture)	noun.cognition

Table 5: Examples of disagreement between noun.domain and another supersense.

a word with a given combination of tags (e.g. one annotator chose noun.abstract and another chose noun.body). Large entries on the diagonal indicate agreement, while off-diagonal entries mean that two senses are confused. Furthermore, the matrix is normalized by row, and rows are sorted after the size of the diagonal value. Thus the senses with the worst disagreement appear first while the best senses are located near the bottom of the matrix.

For instance, the sense noun.group has a smaller value in the diagonal than in the column for noun.quantity. This difference indicates that annotators often disagree about these senses, and that there is little agreement on when to assign the sense noun.group. Other senses like noun.food have perfect or near-perfect agreement. In all three disagreement plots, covering the NEWSWIRE, BLOG and PARLIAMENT, we find that the supersense noun.domain is problematic to the annotators. This supersense has a smaller value in the diagonal than in the column for communication and cognition.

Table 5 shows some examples of this disagreement, where nouns have been annotated with noun.domain and some other sense respectively. As a consequence this supersense should either be better explained and exemplified in the annotator guidelines, or it should be discarded from the extended list altogether.

We also observe that some of the very frequently used types are easier to annotate in NEWSWIRE than in BLOG and PARLIAMENT debates. This is true for supersenses such as noun.institution and noun.communication (for supersense frequency see Section 6) where the number of off-diagonal boxes are lower for NEWSWIRE than for the other textual domains. More metaphorical language in political speeches, which is generally harder to annotate, could explain this difference, as well as frequent reference

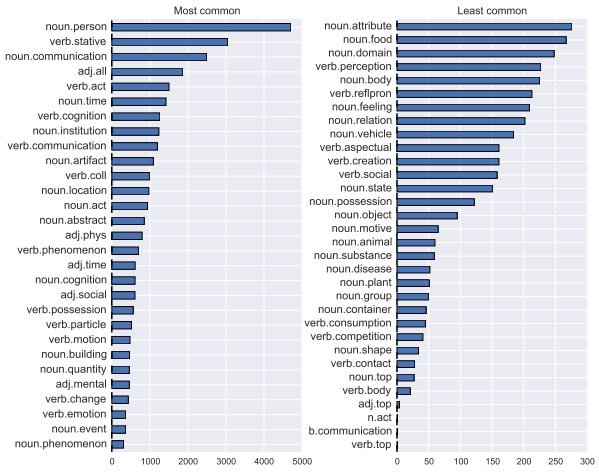

Figure 6: Most and least frequent supersenses in the complete annotated corpus.

Sense 1	Sense 2	PMI
verb.consumption	noun.food	2.71
verb.contact	noun.body	2.26
noun.food	noun.container	2.04
verb.body	noun.body	1.39
noun.disease	noun.body	1.29
verb.competition	noun.event	1.13
verb.motion	verb.contact	1.10
verb.contact	noun.artifact	1.08
noun.substance	noun.object	1.07
noun.shape	noun.body	1.06
noun.vehicle	noun.substance	0.79
verb.competition	noun.relation	0.75

Table 6: Mutual information for supersenses.

to institutions of a very different status.

6 Sense distributions

We now analyze the variation across domains for the top 15 supersenses. Figure 7 provides a picture of which senses are the dominating in each selected domain compared to the sense distribution in the complete annotated corpus in Figure 6.

We observe that noun.person is by far the most frequent tag in MAGAZINE and NEWSWIRE where references to people make up a large portion of the text. The MAGAZINE domain is mostly tabloid content in which the life of famous people is discussed. In contrast, the annotated blogs refer only sparingly to people but focus on personal reflections on life. The tag noun.communication is frequent in BLOG, partly influenced by the meta comments exemplified in Section 5.

The CHAT domain is the only one where the first most frequent sense is not a nominal sense, but instead the verb.stative supersense (mainly forms of the verb *være*, to be). In this domain pronominal subjects are about three times as common as in the NEWSWIRE domain, and many of the syntactic slots (e.g. subject) that would otherwise be satisfied by noun.person in other types of text are satisfied by pronouns in this domain. This explains why noun.person is only the fourth most frequent sense in this domain.

In FORUM, noun.artifact is the second most frequent sense, because the members of the forum discuss *things*: publications, computer parts, and collectible card games. More abstract concepts like movies or games are often referred to in their physical incarnation. The high frequency of noun.artifact is a result of the specialized topic of the forum.

The PARLIAMENT texts are special in several ways, which we see reflected in the annotations. Abstract concepts and verbal states are frequent for this text type, which is not the case for the other text types. Moreover, this text type has more words per sentence and the highest token/type ratio (as seen in Table 2) and thus the least varied language.

7 Relation between senses

This section offers an overview on how supersenses co-occur. To give account for relevant associations between senses, we use PMI (pointwise mutual information), which is an information-theoretical measure of association between variables. Higher PMI values indicate stronger association, i.e. variable *A* is more predictable from variable *B*.

Table 6 shows the twelve pairs of supersense with the highest pointwise mutual information calculated across sentences. We observe that some of the associations are prototypical selectional restrictions like verb.comsumption + noun.food as in:

> *Hvad drikker I af sodavand , hvis I gør?*
> What kind of soda (noun.food) do you drink(verb.consumption), if you do?

Other associations are topical, regardless of parts of speech, like verb.competition and noun.event:

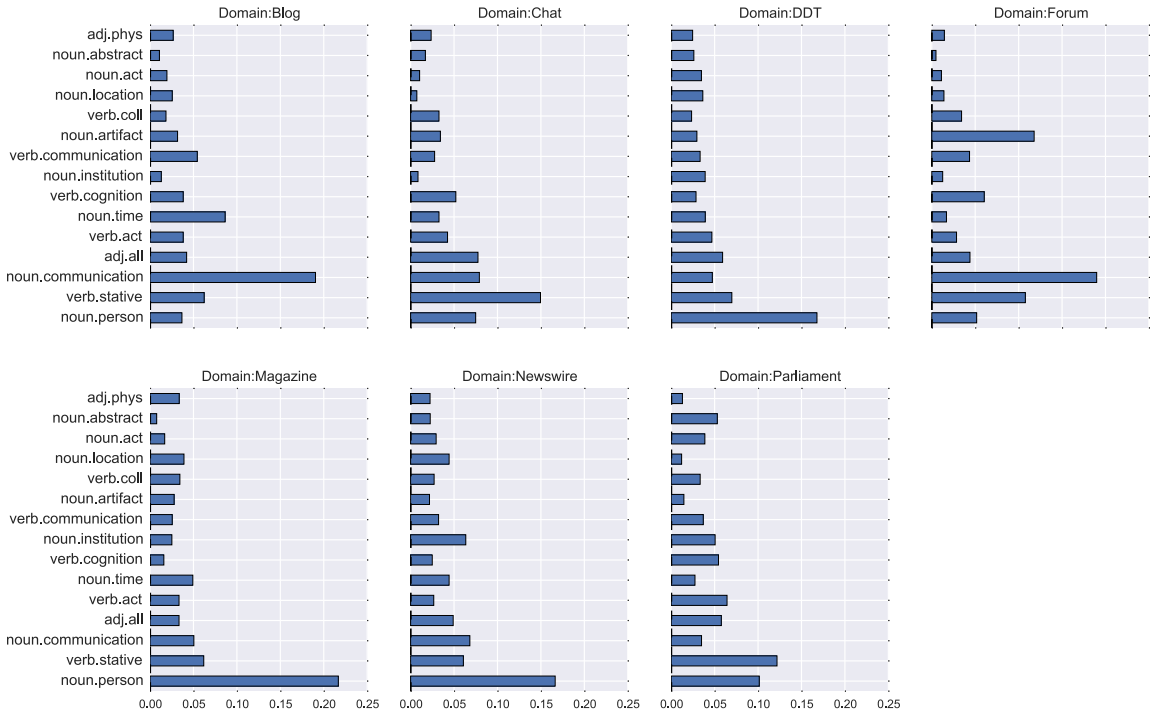

Figure 7: Variation across domains in the top 15 supersenses.

FCK har vundet pokalfinalen.
FCK has won(verb.competition) the cup final(noun.event).

Finally, some of the associations appear for the same part of speech, like noun.disease and noun.body, or noun.food and noun.container. In these associations, one sense is a strong indicator for the other at the topic level (diseases are bodily; food is kept somewhere, etc).

8 Conclusion

We observe that domain-wise agreement is not linked to factors such as how canonical the text is, and whether the text is professionally edited or not. The NEWSWIRE and PARLIAMENT domains, which contain the most thoroughly edited text in the corpus, have the lowest agreement, which is somewhat unexpected. Here we suggest that certain words and sense variations are intrinsically more difficult, e.g. abstract senses. In comparison, FORUM has a clear topic, constraining the discourse elements and their semantic type and thus making annotation easier.

The annotation task yields good agreement for supersense annotations across a number of domains, matching or exceeding the level of agreement found in previous, comparable studies. However, a few supersenses are hard to apply uniformly across all domains, calling for further analysis and perhaps an adjustment of the sense inventory. Abstract noun supersenses as well as verb supersenses related to cognition were generally harder to annotate consistently than more concrete supersenses.

By examining the top 15 supersenses of each domain, we have also shown how textual domains differ in their sense distribution. These observations can later be exploited in automatic sense tagging when dealing with domain shift. One way to do this is pre-estimating the most-frequent sense of the target domain using a lexical knowledge base like DanNet.

For experiments with automatic tagging of Danish data based on the annotations, we refer to Martínez Alonso et al. (2015a) and Martínez Alonso et al. (2015b).

Acknowledgements

We wish to thank the anonymous reviewers for their valuable comments. Likewise, we thank all the project staff, as well as our team of annotators.

The research resulting in this publication has been funded by the Danish Research Council under the *Semantic Processing across Domains* project: `http://cst.ku.dk/english/projekter/semantikprojekt/`.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jørg Asmussen and Jakob Halskov. 2012. The CLARIN DK Reference Corpus. In *Sprogteknologisk Workshop*.

Susan Windisch Brown, Travis Rood, and Martha Palmer. 2010. Number or nuance: Which factors restrict reliable word sense annotation? In *LREC*.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 168–175. Association for Computational Linguistics.

Gerard De Melo, Collin F Baker, Nancy Ide, Rebecca J Passonneau, and Christiane Fellbaum. 2012. Empirical comparisons of masc word sense annotations. In *LREC*, pages 3036–3043.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237. Association for Computational Linguistics.

Nancy Ide and Yorick Wilks. 2006. Making sense about sense. In *Word sense disambiguation*, pages 47–73. Springer.

Adam Kilgarriff. 2006. Word senses. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation*, pages 29–46. Springer.

Héctor Martínez Alonso, Anders Johannsen, Anders Søgaard, Sussi Olsen, Anna Braasch, Sanni Nimb, Nicolai Hartvig Sørensen, and Bolette Sandford Pedersen. 2015a. Supersense tagging for danish. In *Nodalida*.

Héctor Martínez Alonso, Barbara Plank, Anders Johannsen, and Søgaard. 2015b. Active learning for sense annotation. In *Nodalida*.

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

Bolette Pedersen, Anna Braasch, Sanni Nimb, and Sussi Olsen. 2015. Betydningsinventar - i ordbøger og i løbende tekst, forthcoming. In *Presentation at the 13th Conference on Lexicography in the Nordic Countries*.

Piek Vossen. 1998. *EuroWordNet: A multilingual database with lexical semantic networks*. Springer.

Seid Muhie Yimam, Iryna Gurevych, Richard Eckart de Castilho, and Chris Biemann. 2013. Webanno: A flexible, web-based and visually supported system for distributed annotations. In *ACL (Conference System Demonstrations)*, pages 1–6.