Supersense tagging for Danish

Héctor Martínez Alonso‡ Anders Johannsen Sussi Olsen Sanni Nimb†
Nicolai Hartvig Sørensen† Anna Braasch Anders Søgaard Bolette Sandford Pedersen‡
Njalsgade 140, Copenhagen (Denmark), University of Copenhagen
†Christians Brygge 1,Copenhagen (Denmark), Danish Society of Language and Literature
‡alonso@hum.ku.dk, bspedersen@hum.ku.dk

Abstract

We describe the creation of a new Danish resource for automated coarse-grained word sense disambiguation of running text (supersense tagging, SST). Based on corpus evidence we expand the sense inventory to incorporate new lexical classes. We add tags for verbal satellites like collocates, particles and reflexive pronouns, to give account for the satellite-framing properties of Danish. Finally, we evaluate the quality of our expanded sense inventory in terms of variation in F_1 on a state-of-the-art SST system. The SST systems uses type constraints and achieves performance just under the upper bound of inter-annotator agreement. The initial release is a 1,500-sentence corpus covering six genres, made available under an open-source license.[1]

1 Introduction

Supersense tagging is a coarse-grained word sense disambiguation task, which bases its sense inventory on the top level of Princeton Wordnet (Fellbaum, 1998), taken from *lexicographer files*. A supersense is more general than a synset, grouping many related sense distinctions together, while keeping important semantic distinctions. The smaller number of supersenses (comparable to the size of a typical POS tag set) makes it possible for state-of-the-art taggers to be trained on datasets of moderate size.

Supersense tagging is similar to Named Entity Recognition (NER) in that the labels are comprised within spans of one or more tokens. NER, however, only recognizes a handful of entity types and does not extend beyond nouns, while supersenses may be defined for all part of speech and permit more granular semantic distinctions.

While coarse-grained semantic types find use in a range of applications, such as information retrieval, question answering (QA), and relation extraction, one of the main intended uses of the annotated corpus is building a semantic concordancer in the style of SemCor (Miller et al., 1994).

We base our annotation effort on the set of supersenses derived from Princeton Wordnet, which makes our annotations interoperable across many languages through the already existing linkings to Princeton Wordnet. However, we found several cases where the Princeton supersenses made overly broad distinctions that caused large groups of lexemes to be grouped together (e.g. buildings and vehicles falling under the ARTIFACT class).

The original sense inventory comprises a total of 41 senses, spread over 26 noun senses, and 15 verb senses, plus a single "catch-all" sense for adjectives, which is grammatically rather than semantically motivated. Based on lexical data from the corpus-based Danish wordnet (Pedersen et al., 2009), we introduce seven new noun senses, two verb senses, and four adjective senses. A complete listing is shown in Table 3. Importantly, these additions do not break compatibility with supersenses, because the extended senses add more granularity to existing senses. An additional sense can thus always be unambiguously mapped to an original sense. For instance, a DISEASE is a STATE. Details about the newly introduced senses are given in Section 2.

After an annotation task, we experiment with SST in order to gauge the quality of automatic supersenses annotations for the aforementioned semantic concordancer.

[1] The data is available at clarin.dk under *Danish Supersense Corpus*

2 Extended sense inventory

The current standard supersense inventory is the list of WordNet lexicographer files.[2] However, the Danish wordnet (DanNet) is not organized in the WordNet lexicographer files. Instead, each synset in DanNet is described by an an *ontological type*, namely an array of ontological properties that we have mapped to the standard and new supersenses. Table 2 provides three examples of such mapping.

Ontological type	Supersense
Property+Physical+Colour	ADJ.PHYSICAL
Liquid+Natural	NOUN.SUBSTANCE
Dynamic+Agentive+Mental	VERB.COGNITION

Table 1: Ontological type to supersense projection.

The standard set of noun supersenses expresses very general lexical semantic properties such as *state*, *event*, *animal*, *person*, and *cognition*. We extend the standard set with a few more fine-grained types, translating DanNet ontological types to supersenses. The new noun supersenses are BUILDING, CONTAINER, VEHICLE, DISEASE, ABSTRACT, and DOMAIN (for fields of expertise like *philosophy*). The noun senses COGNITION and COMMUNICATION cover processes as well as contents, and might result in low-agreement annotations. We have added the ABSTRACT and DOMAIN specified senses for COGNITION, and disregarded extending COMMUNICATION—although it could potentially be extended into a supersense for linguistic units like *word* or *speech*, and another one for semiotic artifacts like *book* or titles like *Crime and Punishment*

For verbs, we choose to extend the set with the supersense PHENOMENON in order to cover general verb event senses like *happen*, in line with the corresponding ones for noun senses in the standard set (covering noun events and noun natural phenomena). Verbs of natural events are, in our annotation experience, only covered partly by the standard supersense WEATHER.

For adjectives, we introduce four supersenses: one PHYSICAL (*green*, *tall*, *hard*), one MENTAL (*jealous*, *sensible*, *clever*), one SOCIAL (*democratic*, *Arabic*, *economical*), and finally one for TIME (*early*, *contemporary*), which includes intensional adjectives like *former*.

[2] http://wordnet.princeton.edu/man/lexnames.5WN.html

New category	Subsumed by
Noun	
VEHICLE	ARTIFACT
BUILDING	ARTIFACT
CONTAINER	ARTIFACT
DOMAIN	COGNITION
ABSTRACT	COGNITION
INSTITUTION	GROUP
DISEASE	STATE
Verb	
ASPECTUAL	STATIVE
PHENOMENON	CHANGE
Adj	
MENTAL	ALL
PHYSICAL	ALL
SOCIAL	ALL
TIME	ALL
Sat	
COLL	-none-
PARTICLE	-none-
REFLPRON	-none-

Table 2: Extensions to the sense inventory.

Danish is a typical satellite-framing language in Talmy's (1985) terms, because the verb in combination with a satellite (such as a particle) typically expresses a composite and often non-transparent meaning. To give account for verb-headed collocations, phrasal verbs, and reflexive verbs, which often occur as discontinuous constituents in running text, we have introduced three verb-satellite tags: COLL, PARTICLE, and REFLPRON. These are rather to be understood as morphosyntactic tags indicating that the given satellite contribute to the composite meaning of the verb in question.

In other words, these three tags are interpreted in combination with the verb introducing them, as in *han slog ordet op i ordbogen* (lit. 'he hit the word up in the dictionary') meaning 'he looked up the word in the dictionary', and as in *han satte ham på plads* (lit: 'he put him in place') meaning 'he corrected him harshly'.

Tagging of satellites allows for a composite semantic intrepretation of the verb-headed multiword expressions, interpreting thus the collocation *sætte på plads* as communication and not as motion, which the verb *sætte* ('put') would indicate in isolation. This interpreta-

Domain	$\overline{SL}$	$\frac{\text{tokens}}{\text{types}}$	Sentences
Blog	16.44	2.95	100
Chat	14.61	3.70	200
Forum	20.51	3.85	200
Magazine	19.45	2.95	200
Newswire	17.43	3.28	600
Parliament	31.21	5.00	200

Table 4: Supersense tagging data sets.

tion is annotated in the corpus in the following way: *han satte*(VERB.COMMUNICATION) *ham på*(COLL) *plads*(COLL).

3 Annotation process

This section the describes the annotation task for supersenses, including detailes on corpus, guidelines and resulting agreement scores. For further information, cf. Olsen et al. (2015).

3.1 Corpus

We have chosen to annotate from the Danish CLARIN Reference Corpus (Asmussen and Halskov, 2012), which consists of newspapers, magazines, oral debates, blogs, and social media.[3]

Table 4 lists the amount of training data (1,500 sentences in total) currently annotated for each domain. We describe each domain in terms of its average sentence length ($\overline{SL}$) and proportion of tokens per type, namely the average amount of repetitions for a certain type.

The final release will be made up of 600 sentences from all of the domains in Table 4, plus the test section of the Danish Dependency Treebank (Buch-Kromann et al., 2003).

All the data has been POS-tagged using the Stanford POS-tagger (Toutanova et al., 2003) trained on the Danish PAROLE corpus.[4] Note that we strictly use predicted POS instead of gold-standard to provide a more realistic setup for the evaluation of our system in Section 5.

3.2 Annotation guidelines

Sense inventory The guidelines for the supersense annotation comprise the list of supersenses provided with an explanation and examples for each supersense.

Application rules The second part of the guidelines consists of a set of more specific rules for each part of speech. The rules for nouns concern the delimitation of units to be annotated, how to treat multiword units (e.g. names of people, places,or book titles), compounds, figurative senses and metaphors, but also clarifications of how to interpret some of the supersenses that are closely related.

The rules for adjectives treat the language-specific issues of determining when a word is a participle, an adverb or an adjective, and how to annotate it in the later case. The rules for verbs concern the identification of grammatical phenomena like auxiliary verbs, and modal verbs—which are not annotated, because we only assign a supersense to the main lexical verb, e.g. in constructions like *"would have found"* only *found* would be annotated—, and the identification of words that participate in verb-satellite constructions.

Decision trees The sense inventory and the application rules are vertebrated into three (one per main part of speech) decision trees, that illustrate the ontological structure of the supersenses to use in case of sense subsumption like ARTIFACT vs. VEHICLE or CHANGE vs. PHENOMENON.

3.3 Sense distribution

Figure 1 shows the distribution of tags across all the parts of speech in absolute frequency. The plot is divided in high and low-frequency bands. All the new adjective supersenses appear in the high-frequency band. The senses NOUN.BUILDING and NOUN.VEHICLE fall respectively in the high and low band. As regards the verbal satellites, SAT.COLL is ranked 12.

Sense distributions vary across domains. Figure 2 shows the variation of frequency for four supersenses in all domains. While NOUN.PERSON is the overall most frequent sense for nouns, it is not in Forum (where the most frequent noun sense is NOUN.COMMUNICATION), while Magazine—being made up of tabloid text, where the life of celebrities is discussed—is made of 10% of person-type nouns.

3.4 Agreement

Each sentence in our dataset has been annotated by two of the four native annotators with a background in linguistics, and reviewed by one of the

[3] http://cst.ku.dk/Workshop311012/sprogtekno2012.pdf
[4] http://korpus.dsl.dk/e-resurser/paroledoc_en.pdf

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: Sense inventory with new senses introduced in this article marked in bold.

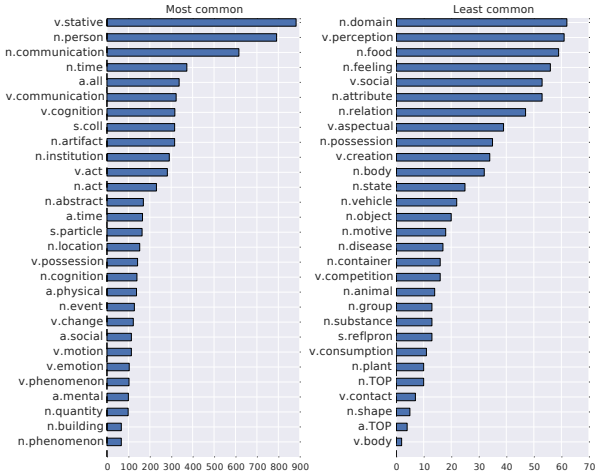

Figure 1: Distribution of senses in the high and low-frequency bands.

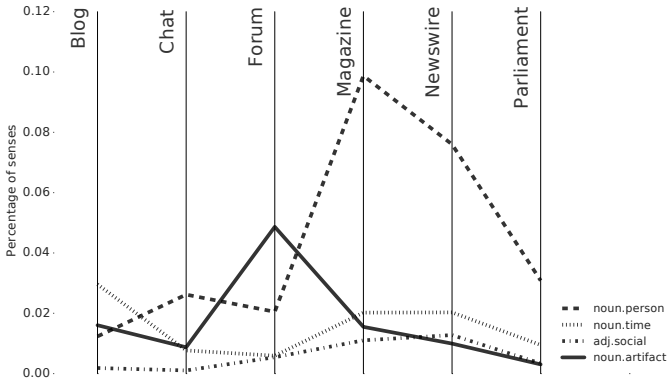

Figure 2: Variation of sense frequency across domains.

two adjudicators. We use WebAnno (Yimam et al., 2013) as annotation environment.

We calculate inter-annotator agreement using Cohen's κ. A first batch of documents were annotated by two of the annotators and later reviewed by one of the adjudicators. The agreement in the first documents was between 0.52 and 0.57. The causes of disagreement were principally verbal collocates, particle verbs and multiword units.

After discussion and refinement of the annotation guidelines, the agreement increased to 0.63. We also tested the agreement between adjudicators using the revised guidelines. The two adjudicators reached a κ of 0.7 on a 200-sentence sample. The remaining disagreement is mostly due to varying interpretations of the sentences (taken out of context) and to the delimitation of some of the abstract supersenses that overlap in some ways.

Figure 3 provides plots that illustrate the disagreement patterns between the annotators. Each row stands for the overall probability of any annotator assigning the sense listed. The size of the boxes indicate the probability that another annotator might have chosen another sense for the same word. We have calculated these probabilities on a 200-sentence sample from the Newswire domain.

Rows are sorted after the size of the diagonal value, and values in the diagonal indicate the proportion of agreement between two annotators for any given sense. Rows with many large, spread boxes indicate low-agreement senses. The sense NOUN.GROUP, for instance, has a smaller value in the diagonal than in the column for NOUN.QUANTITY. This difference indicates that these two senses are very often disagreeed upon,

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.

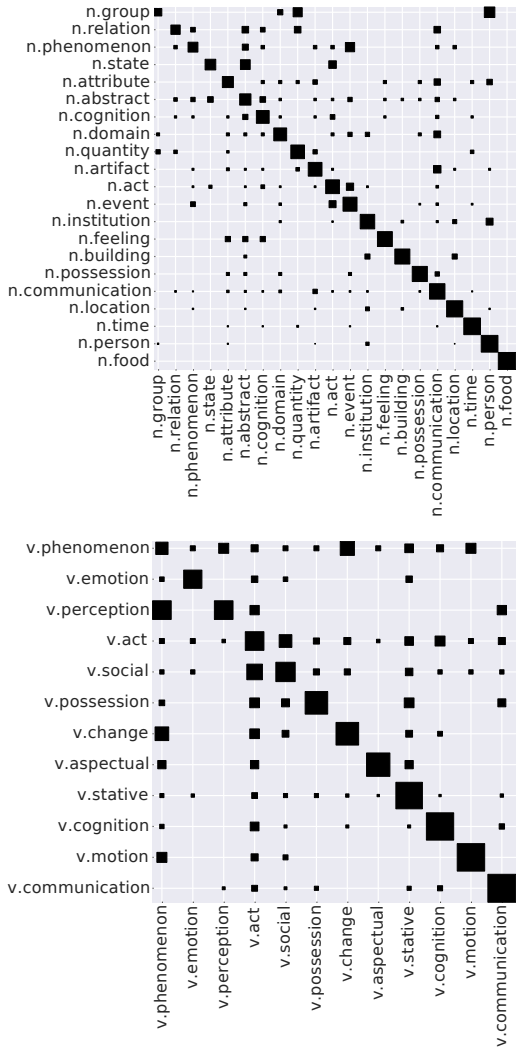

Figure 3: Disagreement plots for nouns and verbs.

We observe that there are some nouns with very good agreement, while there is much less general agreement on verb senses. The sense VERB.ACT has indeed a chance for being annotated with any other verb sense by the another annotator.

To compensate for the disagreement during the annotation step, there is an adjudication step. The following examples show cases of disagreement when annotating with our sense inventory. The word where annotators incur in disagreement is in italics, the two conflicting senses appear in brackets, with the adjudicated sense underlined.

(1) Findes der ikke et eneste stykke *papir* [ARTIFACT/COMMUNICATION] i Vatikanets kældre om alt det?
"Isn't there a single piece of *paper* in the Vatican's basements about all of this?"

(2) Så giver I bare fremmede frø mulighed for at *spire* [CHANGE/PHENOMENON].
"Then you are giving foreign seeds a chance to *sprout*."

(3) Togtrafikken mellem Vejle og Århus var i går *lammet* [PHYS/SOCIAL] i flere timer efter en personpåkørsel ved Horsens.
"The train traffic between Vejle and Århus was *paralyzed* yesterday for several hours after a human collision arounds Horsens."

(4) Thi først brød han *igennem* [COLL/PARTICLE] til det store publikum med den på mange måder uhyre vellykkede fimatisering af Umberto Ecos "Rosens Navn".
"Because of this, he broke *through* to the major audience with the, in many ways monstrously accomplished, filmatization of Umberto Eco's "Name of the Rose"."

4 SST model

The labels in supersense tagging are spans (defined using BIO notation) like *Hans/B-noun.person Hansen/noun.person*. Supersense tagging is typically cast as a sequential problem like POS tagging. However, the class distribution is more skewed than for POS tagging, given that in SST all the words that do not receive a supersense receive the outside-of-entity tag O. We use the feature model of Johannsen et al. (2014). For each word *w*, we calculate the following:

1. **The 2-token window** of forms, lemmas and POS tags around *w*, including *w*.
2. **2-token window** of forms, lemmas and POS tags around *w*, including *w*.
3. **The 2-token window** of forms, lemmas and POS tags around *w*, including *w*.
4. **Bag of words** of forms and lemmas at the *left* and *right* of *w*, marked for directionality so words at the left are different from words at the right.
5. **Morphology** of *w*, whether it is all alphanumeric, capitalized, contains hyphens, and its 3-letter prefix and suffix.
6. **Brown cluster** that *w* belongs to. We generate the 2-,4,6,8,10 and 12-bits long prefix

of the cluster bitstring from clusters from the ClarinDK corpus.[5]

7. **Embeddings** of w and its 2-word window context [6], using 100-dimension vectors, 5-word window sampling and 10-word negative sampling from the ClarinDK corpus. We calculate the weighted average of w and its four surrounding words, where w is weighted twice. For the five different embedding vectors, we also calculate the dimension-wise maximum and minimum. These three operations yield a total of 300 real-valued features. Moreover, we calculate the cosine similarity between w and its four context words.

The sequence-prediction algorithm for the system is on SEARN, search-based classification, with two passes over the data (Daumé et al., 2009).[7]

4.1 Type constraints

We implement distant supervision by only allowing a system to predict a certain supersense s for a given lemmatized word w with part of speech p with the following criteria:

1. If (w,s) has been observed in the training data, s is an allowed sense.

2. If (w) is not in the training data, but (w,p) appears in DanNet, we allow the most frequent sense for (w,p).

3. If w does neither appear in the training data or in DanNet, we make no assumptions and allow any sense to be assigned by the classifier.

We refer to this distant-supervision strategy as *type constraints*. Since SEARN decomposes sequential labelling into a series of binary classifications, we constrain the labels by simply picking the top-scoring sense for each token from the allowed set of senses.

5 Evaluation

In this section we evaluate the performance of the supersense tagging system (SST) against the MFS (most-frequent sense baseline). All our systems have been evaluated on 5-fold cross-validation on randomly shuffled sentences. All results are expressed in terms of micro-averaged F_1-score.[8]

We have trained and test the data using two variants of the training data: one where the verbal satellites where removed from the annotation replacing them with the O tag, and another where the annotations were kept intact. We evaluate only on the set of lexical supersenses (adjectives, nouns and verbs). The goal of this comparison is to establish whether adding the verb-satellite tags penalizes the performance of the system.

5.1 MFS baseline

For most word sense disambiguation studies, predicting the most frequent sense (MFS) of a word has been proven to be a strong baseline. Following this, our MFS baseline simply predicts the supersense for (w,p) in a manner similar to the one used to implement type constraints (Section 4.1), namely by calculating MFS from the training data and backing off to the value in DanNet if the word is not present in the training data. If a word is not present in either, it receives the most frequent sense for its part of speech.

5.2 System performace

Table 5 provides the micro-averaged F_1 for the SST system. The SEARN column reflects the classifier output before type contraints are applied, and +Constraints is the resulting F_1 after applying the type contraints described in 4.1.

	MFS	SEARN	+Constraints
SST	32.96	52.01	60.51

Table 5: F_1 scores for SST system.

The F_1 score between the two variants of the training data does not change, regardless of the presence of the verb-satellite tags. Thus, we consider that is viable to mantain the annotation of the verb satellites. Table 5 shows the micro-averaged F_1 score for the SST system with and without type contraints, and compared against the MFS baseline, using all the sense inventory (all the lexical senses and the verb satellites).

We have experimented with feature ablation, but the best final system contains the full feature set. In particular, embedding features provide an improvement of around 4.0 in F_1.

[5]We use Liang's implementation `https://github.com/percyliang/brown-cluster`

[6]We use Word2Vec `https://code.google.com/p/word2vec`

[7]SEARN in Vowpal Wabbit `https://github.com/JohnLangford/vowpal_wabbit`

[8]We have used the conlleval.pl script from the NER shared tasks

5.3 Constraint contribution

Applying type constraints contributes greatly to the performance of the system. Indeed, the +Constraints system has an F_1 score just below the expected maximum performance, namely the κ agreement coefficient of the data (0.63).

	SEARN	+Constraints
$\rho(\frac{\text{tokens}}{\text{types}}, F_1)$	0.74	0.27
$\rho(\text{tokens}, F_1)$	0.81	0.40

Table 6: Correlation scores for SST before and after applying type constraints.

Table 6 shows the Spearman's ρ between the F_1 of each individual supersense and its token-type ratio and number of tokens respectively, for both the SEARN and the +Constraints system. We observe that, before any constraint is applied, performance is highly correlated with token-type ratio, but even more so with the number of tokens.

Train	DanNet	Train+DanNet
0.49	0.34	0.16

Table 7: OOV rates for training data and DanNet.

Applying type contraints effective decorrelates the performance of the individual supersenses from the bias of the SST classifier. However, the correlation with the number of tokens remains higher, as it is also correlated with the coverage in DanNet for a certain supersense. That is, a high-frequency sense like NOUN.PERSON will contain more high-frequency words that will be covered in a wordnet (e.g. *person, child, sailor*).

5.4 POS-wise evaluation

This section provides tagwise evaluation in terms of precision (P), recall (R), and F_1. In addition, we provide the number of tokens (absolute frequency), the number of types, the token-type ratio for each supersense tag in tables 8, 10, 9 and 11.

Supersense	P	R	F_1	types	tokens	$\frac{\text{tokens}}{\text{types}}$
ADJ.ALL	59.8	62.1	60.9	246	341	1.39
ADJ.MENTAL	58.3	44.1	50.2	79	100	1.27
ADJ.PHYSICAL	56.5	46.3	50.9	98	138	1.41
ADJ.SOCIAL	68.8	69.4	69.1	92	114	1.24
ADJ.TIME	80.2	83.5	81.8	56	166	2.96

Table 8: Performance for adjectives.

Overall, the prediction of adjective supersenses fares fairly well, however ADJ.ALL makes up a 30% of the annotated adjectives senses, which is too large for a back-off sense. Also, ADJ.ALL is a low-agreement supersense tag. A further refinement of the annotation guidelines or an inclusion on an additional supersense—provided that we identify some internal semantic consistency—can reduce the amount of words labeled as ADJ.ALL.

Supersense	P	R	F_1	types	tokens	$\frac{\text{tokens}}{\text{types}}$
VERB.ACT	42.6	52.7	47.1	197	283	1.44
VERB.CHANGE	46.4	34.2	39.4	84	123	1.46
VERB.COGNITION	67.7	59.0	63.1	156	317	2.03
VERB.COMMUNICATION	75.5	72.7	74.1	158	323	2.04
VERB.CONSUMPTION	100.0	7.1	13.3	7	11	1.57
VERB.EMOTION	51.8	40.0	45.1	55	104	1.89
VERB.MOTION	39.8	48.5	43.7	76	114	1.5
VERB.PERCEPTION	47.4	51.4	49.3	25	61	2.44
VERB.PHENOMENON	39.3	34.2	36.5	75	103	1.37
VERB.POSSESSION	54.8	42.3	47.7	62	143	2.31
VERB.STATIVE	79.2	84.3	81.7	122	884	7.25

Table 9: Performance for the 10 most frequent verbs senses.

Overall performance for verbs is worse than for nouns. Even though there are fewer verbal senses, verbs are more difficult to annotate, as shown by the verb disagreement plot in Figure 3.

Supersense	P	R	F_1	types	tokens	$\frac{\text{tokens}}{\text{types}}$
NOUN.ABSTRACT	37.23	34.31	35.71	141	170	1.21
NOUN.ACT	56.9	61.34	59.03	189	233	1.23
NOUN.ARTIFACT	45.56	39.81	42.49	259	316	1.22
NOUN.COGNITION	49.44	53.61	51.45	112	141	1.26
NOUN.COMMUNICATION	41.24	52.49	46.19	399	618	1.55
NOUN.EVENT	43.21	29.41	35.0	107	128	1.2
NOUN.INSTITUTION	51.69	46.15	48.76	235	292	1.24
NOUN.LOCATION	67.37	70.09	68.7	130	155	1.19
NOUN.PERSON	66.72	75.04	70.64	579	795	1.37
NOUN.TIME	83.92	84.73	84.32	163	373	2.29

Table 10: Performance for the 10 most frequent noun senses.

The sense COMMUNICATION is the second most frequent noun sense, yet it fares much worse than that first sense, namely PERSON. Even though COMMUNICATION has lower support, its token-type ratio is higher than the one for PERSON, which should increase F_1. However, PERSON has a subset of well-defined proper names that are easy to identify automatically given features like capitalization.

For NOUN.COMMUNICATION, out of its 323 examples, 10% of them are hapaxes. The VERB.STATIVE class, however, with a 884 examples, is constituted by forms of the verb *være* (to be) in 76%. The low variety of lexical elements makes it an easy-to-predict sense, and yields an F_1 of 78.39, which is very high for word-sense disambiguation tasks. The three verbal satellites fare

very differently from each other. The most common tag, COLL, has a very low F_1 (14.35). Besides the already commented factors of number of tokens and token-type ratio, the predictability of these senses is also determined by how many different POS tags they can be applied to: REFLPRON is only for pronouns, PARTICLE encompasses prepositions and adverbs, whereas COLL can also contain nouns, verbs, and adjectives.

Supersense	P	R	F_1	types	tokens	$\frac{\text{tokens}}{\text{types}}$
NOUN.ABSTRACT	37.2	34.3	35.7	141	170	1.21
NOUN.ARTIFACT	45.6	39.8	42.5	259	316	1.20
NOUN.BUILDING	47.9	37.0	41.7	58	67	1.15
NOUN.CONTAINER	91.7	64.7	75.9	12	16	1.33
NOUN.DISEASE	73.3	55.0	62.9	14	17	1.22
NOUN.DOMAIN	63.3	28.8	39.6	49	62	1.27
NOUN.INSTITUTION	51.7	46.2	48.8	235	292	1.25
NOUN.VEHICLE	53.9	33.3	41.2	20	22	1.10
VERB.ASPECTUAL	77.8	32.6	45.9	27	39	1.45
VERB.PHENOMENON	39.3	34.2	36.5	75	103	1.37
SAT.COLL	37.9	7.7	12.8	120	316	2.63
SAT.PARTICLE	59.4	47.9	53.0	34	165	4.76
SAT.REFLPRON	69.6	76.4	72.9	4	13	3.22

Table 11: Performance for extended noun and verb supersenses, and satellites.

6 Related work

There has been relatively little previous work on supersense tagging, and it has mostly been limited to English newswire and literature (namely running on SemCor and SensEval data).[9] Nevertheless, the interest in applying word-sense disambiguation techniques to reduced, coarser sense inventories has been a topic since the development of the first wordnets (Peters et al., 1998). Kohomban and Lee (2005) and Kohomban and Lee (2007) also propose to use lexicographer file identifers from Princeton WordNet senses (supersenses) and, in addition, discuss how to retrieve fine-grained senses from those predictions.

The task of supersense tagging was first introduced by Ciaramita and Altun (2006), who used a structured perceptron trained and evaluated on SEMCOR via 5-fold cross validation. Johannsen et al. (2014) extend the SST approach to the Twitter domain, and include the usage of word embeddings in their feature representation.

Supersenses have been used as features in various tasks, such as preposition sense disambiguation, noun compound interpretation, metaphor detection and relation extraction (Ye and Baldwin, 2007; Tratz and Hovy, 2010; Tsvetkov et al., 2013; Søgaard et al., 2015). Schneider et al. (2012) annotated supersenses on Arabic Wikipedia articles . Princeton WordNet only provides a fully developed taxonomy of supersenses for verbs and nouns. Tsvetkov et al. (2014) propose an extension for adjectives, along the lines of the adjective sense of the German wordnet(Hamp and Feldweg, 1997).

To the best of our knowledge, the current work is the first SST approach to Danish, which also extends to less canonical, characteristically web-based text types like chats or fora.

7 Conclusions

We have presented a resource for SST that includes an extension of the English supersense inventory that can be used for any language, plus three additional tags that give account for characteristics of the syntax-semantics interface of a satellite-framing language like Danish.

We have conducted an annotation task on 1,500 sentences, reaching 0.63 κ score after refining the annotation guidelines. After annotation, the supersenses in our data has been adjudicated to resolve systematic disagreements. Later, we have trained an SST model that we have evaluated before and after applying type constraints. Our best system reaches a micro-averaged F_1 of 60.51, which is very close to the theoretical maximum of prediction performance set by the agreement score. This leads us to conclude that the system is mature enough to be used productively when the annotation process has finished.

Nevertheless, the performance is not even across all supersenses. Some of the high-frequency, low-variation supersenses show very high scores (above 81%), while other infrequent senses with a lot of variation or low agreement show lower scores. Some frequent senses like NOUN.COMMUNICATION might benefit from extension.

To the best of our knowledge , this article represents the first attempt to incorporate verb-satellite annotation in sense annotation to give account for verb-headed multiword expressions, which present more practical and theoretical difficulties than the span annotation for nominal multiwords typical of NER.

[9] http://web.eecs.umich.edu/ mihalcea/senseval/senseval3/

References

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

Matthias Buch-Kromann, Line Mikkelsen, and Stine Kern Lynge. 2003. Danish dependency treebank. In *TLT*.

Massimiliano Ciaramita and Yasemin Altun. 2006. Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In *Proc. of EMNLP*, pages 594–602, Sydney, Australia, July.

Hal Daumé, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Machine learning*, 75(3):297–325.

Christiane Fellbaum. 1998. *WordNet: an electronic lexical database*. MIT Press USA.

Birgit Hamp and Helmut Feldweg. 1997. Germanet-a lexical-semantic net for german. In *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15. Citeseer.

Anders Johannsen, Dirk Hovy, Héctor Martínez, Barbara Plank, and Anders Søgaard. 2014. More or less supervised supersense tagging of Twitter. In *Lexical and Computational Semantics (*SEM 2014)*.

Upali Kohomban and Wee Lee. 2005. Learning semantic classes for word sense disambiguation. In *ACL*.

Upali Kohomban and Wee Lee. 2007. Optimizing classifier performance in word sense disambiguation by redefining word sense classes. In *IJCAI*.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Proceedings of the workshop on Human Language Technology*, pages 240–243. Association for Computational Linguistics.

Sussi Olsen, Bolette Sandford Pedersen, Héctor Martínez Alonso, and Anders Johannsen. 2015. Coarse-grained sense annotation of danish across textual domains. In *COLING*.

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.

Wim Peters, Ivonne Peters, and Piek Vossen. 1998. Automatic sense clustering in eurowordnet. In *LREC*. Paris: ELRA.

Nathan Schneider, Behrang Mohit, Kemal Oflazer, and Noah A Smith. 2012. Coarse lexical semantic annotation with supersenses: an arabic case study. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 253–258. Association for Computational Linguistics.

Anders Søgaard, Barbara Plank, and Hector Martinez Alonso. 2015. Using frame semantics for knowledge extraction from twitter. In *AAAI*.

Leonard Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description*, 3:57–149.

Kristina Toutanova, Dan Klein, Chris Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *NAACL*.

Stephen Tratz and Eduard Hovy. 2010. Isi: automatic classification of relations between nominals using a maximum entropy classifier. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 222–225. Association for Computational Linguistics.

Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. *Meta4NLP 2013*, page 45.

Yulia Tsvetkov, Nathan Schneider, Dirk Hovy, Archna Bhatia, Manaal Faruqui, and Chris Dyer. 2014. Augmenting english adjective senses with supersenses. In *Proc. of LREC*.

Patrick Ye and Timothy Baldwin. 2007. Melb-yb: Preposition sense disambiguation using rich semantic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 241–244. Association for Computational Linguistics.

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.