A Multi-domain Corpus of Swedish Word Sense Annotation

Richard Johansson*†, Yvonne Adesam†, Gerlof Bouma†, Karin Hedberg†

*Department of Computer Science and Engineering, University of Gothenburg, Sweden
†Språkbanken, University of Gothenburg, Sweden
richard.johansson@gu.se, yvonne.adesam@gu.se, gerlof.bouma@gu.se, kkmhedberg@gmail.com

Abstract

We describe the word sense annotation layer in *Eukalyptus*, a freely available five-domain corpus of contemporary Swedish with several annotation layers. The annotation uses the SALDO lexicon to define the sense inventory, and allows word sense annotation of compound segments and multiword units. We give an overview of the new annotation tool developed for this project, and finally present an analysis of the inter-annotator agreement between two annotators.

1. Introduction

In an ongoing project on natural language processing for Swedish, *Koala*, we are investigating the possibilities of combining different levels of annotation so as to improve overall quality of automatic processing. As part of this, we are constructing a manually annotated reference corpus, *Eukalyptus*, with annotation of morphological structure, lexical sense and syntactic structure. In previous papers, we have presented the morphological and syntactic annotation in this corpus (Adesam et al., 2015a), and discussed the methodological choices underlying our annotation of multiword units (Adesam et al., 2015b).

In this paper, we describe the word sense annotation in the *Eukalyptus* corpus, which will eventually be used to evaluate tools for automatic word sense disambiguation (WSD). We first introduce the corpus itself and its composition, and then describe SALDO, the Swedish semantic lexicon that provides the sense inventory we use for this annotation. We then describe the annotation process, including the new annotation tool that we have developed, and discuss the reliability of the annotation by making an inter-annotator agreement study.

2. The *Eukalyptus* Corpus

The *Eukalyptus* corpus consists of around 100,000 tokens of contemporary Swedish text. Since a goal of the project is to measure the domain sensitivity of Swedish NLP tools, we have collected text from five different genres, each of which contains about 20,000 tokens:

- Fiction: the first chapters from four novels,
- Encyclopedic: full articles from the Swedish Wikipedia, 100 to 3,000 tokens per article,
- Social media: blog entries from the SIC corpus (Östling, 2013),
- Political debates: proceedings from the European parliament (Koehn, 2005),
- Professional prose: a mix of different types of information from the goverment, and articles from *Arbetaren*, a Swedish weekly.

For all genres, we have made sure that the text can be distributed freeely, and the corpus and all its annotation will eventually be released under a Creative Commons license.

3. The SALDO Lexicon

The word sense annotation described in this paper uses SALDO (Borin et al., 2013) to define its sense inventory. While there are alternative Swedish sense inventories such as the Swedish WordNet (Viberg et al., 2003), SALDO has the advantage of being licensed under a Creative Commons license, and it is also the largest resource of this kind for Swedish: as of October 2015, it contains 131,020 entries organized into a single semantic network.

Compared to WordNet (Fellbaum, 1998), there are similarities as well as differences. Both resources are large, manually constructed networks intended to describe the language in general rather than any specific domain. However, while both resources are hierarchical, the main lexical-semantic relation of SALDO is *association* based on centrality, while in WordNet the hierarchy is taxonomical (based on hyponymy).

Every SALDO entry corresponds to a specific sense of a word, and the lexicon consists of word senses only. There is no correspondence to the notion of synonym set as in WordNet, except for a few cases such as spelling variants of the same word. In general, the sense distinctions in SALDO tend to be more coarse-grained than in WordNet, which reflects a difference between the Swedish and the Anglo-Saxon traditions of lexicographical methodologies.

As in other semantic networks such as WordNet, an entry in SALDO gets its meaning by means of its relations to other entries. Each entry except a special root is connected to other entries, its *semantic descriptors*. One of the semantic descriptors is called the *primary* descriptor, and this is the entry which better than any other entry fulfills two requirements: (1) it is a semantic neighbor of the entry to be

described and (2) it is more central than it. That two words are semantic neighbors means that there is a direct semantic relationship between them, for instance synonymy, hyponymy, antonymy, or meronymy. Most of the primary descriptors tend to be synonyms or hypernyms. Centrality is determined by means of several criteria. The most important criterion is frequency: a frequent word is more central than an infrequent word.

To exemplify, Figure 1 shows a fragment of the SALDO network (only primary descriptor relations are shown). In the example, there are some cases where the relation corresponds to hypernymy, such as *hard rock* having a primary descriptor *rock music*; but there are also other types of relations, such as the predicate–argument relation between *to play* and *music*.

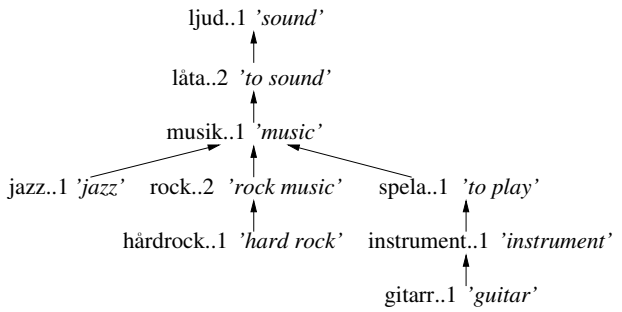

Figure 1: A part of the SALDO network.

SALDO's senses are not necessarily sorted by frequency, unlike in WordNet where they are sorted with respect to the frequency in a reference corpus. The first sense in SALDO is the one that the lexicographers regarded as the most important or basic, which in many cases, but not always, is the same as the most frequent one. Empirically, the first sense tends to dominate for most lemmas, so a first-sense baseline is nontrivial to beat for word sense disambiguation systems. For instance, Johansson and Nieto Piña (2015) found that a first-sense baseline achieved an accuracy of around 50% on two collections of examples hand-picked by lexicographers, and in preliminary investigations we have carried out on the material described in this paper, this tendency seems even stronger: the first sense seems to be the correct alternative for about 70% of the ambiguous words.

4. Word Sense Annotation Process

We processed the texts using the *Korp* infrastructure of Swedish NLP tools (Borin et al., 2012b). The texts were split into sentences and tokens, and we applied a lemmatizer to all tokens. Even though an automatic part-of-speech tagger was available, we preferred not to use it for lemmatization, so that tagging errors would not cause the correct lemma to be unavailable. Finally, lemmas were mapped to a set of possible SALDO senses. The annotator then had to select from this list of senses suggested by the automatic tool, or specify that none of the suggested senses was applicable.

The annotation takes into account that senses need to be annotated not only for single atomic tokens. Because compounding is very productive in Swedish, we need to annotate the SALDO senses of the individual *segments* of compounds, at least when the compounds are compositional and lack their own SALDO entries. Conversely, a token can be a part of a *multiword unit*: a group of words that is listed as a whole in SALDO because its meaning or syntactic behavior is not predictable from the individual words. So for each token, we list a number of possible senses of the token as a whole, of each segment of each possible compund segmentation, and of each multiword unit the token could possibly be a part of.

4.1. Annotation Tool

We developed a new annotation tool for this annotation project because we found no existing word sense annotation tool that could easily be adapted for our requirements of being able to annotate senses of compound segments and multiword units. Furthermore, developing our own tool made it easier to communicate with our lexicon infrastructure (Borin et al., 2012a) for explaining the senses to the annotators, and to design the tool for an efficient and ergonomic annotation process with a minimal use of the mouse.

Since the SALDO lexicon lacks definitions and glosses, the annotation tool instead explains each available sense for a word by displaying a set of neighbors in the sense network. For instance, the noun *fotboll* ('football') has two senses: one referring to the sport, and another to the ball used in that sport. When annotating an occurrence of this word in a text, the tool shows a list of types of sport and other sport-related terms when considering the first sense, while for the other sense we instead get a list of types of balls.

Figure 2 shows an example of the annotation user interface. The annotator has selected the token *fotbollsklubbar* ('football clubs'). Since this is a compound, the annotator needs to decide on a segmentation of the compound, and then select senses for each of the segments. In this case, the annotator selects the segmentation *fotboll+klubb*, and then the sport sense for the first segment of the compound (the second segment is monosemous). To help the annotator understand which sense the tool refers to, the tooltip shows a few sports-related terms.

4.2. Annotation Methodology

As of October, 2015, two annotators have annotated a total amount of 17,580 tokens, with an overlap of 2,919 tokens. We are aiming to produce as much double annotation as possible, so that we can more easily spot lemmas where the senses are hard to distinguish; at a later stage, we will adjudicate the conflicting annotations. So far, we have mainly focused on annotating the senses of nouns and adjectives, since verbs are more complex to annotate due to their frequent participation in multiword units, light verb constructions, etc.

To reduce the cognitive load and to improve the consistency of annotations, the annotators select one lemma at a time, try to understand the distinctions between the senses of that lemma, and then annotate all occurrences of the lemma in the corpus. We sorted the noun and adjective lemmas by the number of occurrences multiplied by the average number of

Figure 2: Example of how the compound *fotbollsklubbar* ('football clubs') is handled in the word sense annotation tool.

senses, so that the annotators start by annotating senses of lemmas that are both frequent and ambiguous. After annotators have moved long enough down the sorted list of lemmas, a lemma-based annotation process is no longer efficient, and the annotators will switch to a new process where the remaining tokens are considered sequentially.

When selecting the sense for a token, the annotator may select the option 'none of the above' if the SALDO lexicon lacks a suitable entry. In some cases, the new sense is very clearly distinct from the existing entries. In other cases, new meanings have been formed from existing senses through productive processes such as metaphor and metonymy, and it is a difficult question to decide exactly when a new sense is important or frequent enough for a new lexicon entry to be created (Cruse, 1986). Ideally, sense-annotated corpora should be produced by trained lexicographers (Kilgarriff, 1999; Artstein and Poesio, 2008); this is probably hard to achieve in practice, but it is important that the lexicographers behind the lexicon are available for consultation by the annotators.

4.3. Inter-annotator Agreement Analysis

The overlap between the two annotators allows us to measure their inter-annotator agreement, which we measured using the well-known κ coefficient (Cohen, 1960):

$$\kappa = \frac{P(a) - P(e)}{1 - P(e)}$$

where $P(a)$ is the estimated probability of agreement – that is, the proportion of tokens where the two annotators agree – and $P(e)$ the estimated agreement probability if the two annotators are assumed to be statistically independent. For practical reasons, we excluded tokens that had been annotated as a part of a multiword expression or non-lexicalized compound by at least one annotator; however, we have shown in previous work (Adesam et al., 2015b) that there is a fairly high agreement between the multiwords in the sense layer and those in the syntactic layer, even though these two layers are annotated using completely different tools and methodologies.

We first considered a number of lemmas separately. Table 1 shows the κ for the 10 lemmas where we had the largest number of double annotations. We excluded lemmas with highly skewed sense distributions ($P(e) \geq 0.95$) from this table. The table also shows the number of annotations (n_t), the average number of available senses (n_s), and the estimated agreement probability ($P(a)$) and chance agreement probability ($P(e)$).

Lemma	n_t	n_s	$P(a)$	$P(e)$	κ
man 'man; one; husband; …'	193	5.0	0.98	0.84	0.90
dag 'day; daylight'	98	3.0	0.98	0.82	0.89
fråga 'question; matter'	90	3.5	0.98	0.34	0.97
vatten 'water'	58	3.0	0.90	0.81	0.46
väder 'weather; scent'	52	4.0	0.96	0.93	0.49
land 'country; ground; …'	50	4.0	0.98	0.74	0.92
gång 'time; walk; pace'	41	3.5	0.93	0.84	0.55
mål 'goal; target; meal; …'	36	8.0	0.94	0.84	0.65
jord 'earth; ground; soil'	34	4.0	0.88	0.55	0.74
folk 'people'	33	4.7	0.15	0.12	0.03

Table 1: κ for the 10 lemmas with the largest number of double annotations, and with chance agreement probability $P(e) < 0.95$.

As we can see, κ ranges from moderate to high values for most lemmas. The exception here is *folk* ('people'), where the two annotators had difficulty distinguishing a generic pronoun-like sense from the sense denoting people as a mass, and the sense of people as inhabitants of a state from that of people as a tribe. This is one of the cases where the lack of precise definitions and examples in SALDO makes it difficult for annotators to draw the line between the senses.

Other lemmas with a low κ include *jobb* ('work'), where the annotators disagreed on whether a new lexicon entry was needed for 'work' as a location, and *stad* ('town; city'), where one annotator would have liked to see a new entry with a meaning equivalent to 'city center.' This exemplifies the methodological difficulty mentioned above: defining when a new sense formed by productive processes of polysemy is frequent and notable enough.

Over all the doubly annotated tokens, the estimated agreement probability was 0.90. Computing the κ globally is

nontrivial since the notion of chance agreement is harder to define and quantify. We estimated a chance agreement probability in this case by defining a global sense distribution: a single probability for selecting the first sense, the second sense, etc. This is of course a simplification of the true situation since the sense distribution is different for each lemma, but it gives a high chance agreement probability because the sense distributions for most lemmas are dominated by the lower-numbered senses. With this chance agreement probability, we get a κ of 0.70, a fairly high value.

5. Conclusion

We have presented the word sense annotation layer in the *Eukalyptus* corpus of contemporary Swedish, which consists of five subcorpora, each corresponding to a separate genre. The annotation uses the SALDO lexicon to define the sense inventory; using this lexicon makes the annotation easier than e.g. WordNet since the sense distinctions are more coarse-grained, but is complicated by the lack of definitions and examples in the lexicon.

We described the new annotation tool developed specifically for this project, and gave a brief overview of the annotation process. Finally, we investigated the inter-annotator agreement between the two annotators. The results showed that the agreement was high in general, except for some lemmas where the SALDO lexicon was not clear about the sense distinctions, or where there was disagreement about whether new senses should be added to the lexicon.

In future work, the *Koala* project will investigate how well off-the-shelf and newly developed graph-based and corpus-based WSD tools perform when evaluated on the new corpora described in this paper. We have already seen in preliminary studies that a first-sense baseline achieves a high accuracy (around 70%), which will likely be hard to outperform for unsupervised WSD tools. We should also add that the annotation in the *Eukalyptus* corpus enables us to consider other tasks that are traditionally neglected in WSD, such as the disambiguation between a compositional and noncompositional reading of a multiword expression or compound word. Moreover, the multi-layer annotation in this corpus opens up opportunities to explore the interaction between linguistic representation levels.

6. Acknowledgements

We are grateful to Martin Stenberg for developing the annotation tool. This work was carried out in the *Koala* project, funded 2014–2016 by Riksbankens Jubilemsfond, grant number In13-0320:1. The first author was also funded by the Swedish Research Council under grant 2013–4944, *Distributional methods to represent the meaning of frames and constructions*, and grant 2012–5738, *Towards a knowledge-based culturomics*. We also acknowledge the University of Gothenburg for its support of Språkbanken.

7. References

Yvonne Adesam, Gerlof Bouma, and Richard Johansson. 2015a. Defining the Eukalyptus forest – the Koala treebank of Swedish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 1–10, Vilnius, Lithuania.

Yvonne Adesam, Gerlof Bouma, and Richard Johansson. 2015b. Multiwords, word senses and multiword senses in the Eukalyptus treebank of written Swedish. In *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories*, pages 3–12, Warsaw, Poland.

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

Lars Borin, Markus Forsberg, Leif-Jöran Olsson, and Jonatan Uppström. 2012a. The open lexical infrastructure of Språkbanken. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 3598–3602.

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012b. Korp – the corpus infrastructure of Språkbanken. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 474–478.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

D. Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press.

Christiane Fellbaum, editor. 1998. *WordNet: An electronic lexical database*. MIT Press.

Richard Johansson and Luis Nieto Piña. 2015. Combining relational and distributional knowledge for word sense disambiguation. In *Proceedings of the 20th Nordic Conference of Computational Linguistics*, pages 69–78, Vilnius, Lithuania.

Adam Kilgarriff. 1999. 95% replicability for manual word sense tagging. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 277–278, Bergen, Norway.

Phillip Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86.

Åke Viberg, Kerstin Lindmark, Ann Lindvall, and Ingmarie Mellenius. 2003. The Swedish WordNet project. In *Proceedings of the Tenth EURALEX International Congress*, pages 407–412, Copenhagen, Denmark.

Robert Östling. 2013. Stagger: an open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.