Koala – Korp's Linguistic Annotations
Developing an infrastructure for text-based research with high-quality annotations

Yvonne Adesam, Lars Borin, Gerlof Bouma, Markus Forsberg, Richard Johansson

Språkbanken, Department of Swedish
University of Gothenburg
firstname.lastname@gu.se

1. Introduction

The corpus infrastructure Korp at Språkbanken (http://spraakbanken.gu.se) offers sophisticated access to a large collection of Swedish texts, of many different types and ages. They range from 13th century law text, over 19th century novels and 20th century news paper texts, to 21st century blogs and discussion forum posts. The linguistic annotations of the texts allow for refined queries, and the quality of these annotations is crucial to get good search results.

The Koala infrastructure project, financed by Riksbankens Jubileumsfond 2014–2016, aims to enhance the automatically created annotations and to ensure the quality of them as the corpus collection grows. (The project is mainly concerned with contemporary Swedish, although the older texts will also benefit from the improvements.) We therefore need to address three issues. First, automatic annotation is traditionally handled in a pipeline, where one tool feeds the next, without interaction among the various annotation steps. Such interaction can increase quality. Secondly, most of these annotations are created by statistical tools, which we hope to improve by incorporating the explicit linguistic knowledge resources available at Språkbanken. Finally, the tools are not general enough to provide high annotation quality across text types and language varieties.

The high-quality annotations can immediately and directly improve the information retrieved by the Korp search interface, as well as the possible queries a user can enter and consequently the research questions which can be explored with the help of the texts in Korp. The whole infrastructure, including the corpus query tool, the annotated corpora, and the annotation tools, are and will remain freely available.

2. Annotation Tools and Quality Control

A key element in the project are the tools and data formats used. Some parts of the infrastructure need to be adapted to adhere to standard formats, and more metadata would help, e.g. for knowing if an annotation was made by a human or automatically, by which tool, model version etc. Additionally, we would like the system to handle ambiguity better. In the current implementation, the output of each processing component is a hard decision that will not change as new information becomes available in a later stage of analysis. In the project, we will develop a module which defers decisions until they can be made with optimal certainty, by flexibly weighting together the output of the various annotation tools, as well as allowing for adding new types of annotation and new annotation tools.

Evaluation and quality control is an important part of creating an infrastructure for text-based research, and essential for high-quality annotations. We will use methods for consistency checking of annotated data (Dickinson and Meurers, 2003; Dickinson and Meurers, 2005; Loftsson, 2009) to improve the gold standard annotation, as well as for detecting problematic areas when developing and improving the annotation tools. Most importantly, however, a number of annotation tools are currently used in Korp, which, although state-of-the-art when the annotation pipeline was set up, have not been evaluated for the wide variety of text types now available through Korp. To properly evaluate the current annotations, as well as the final result after improving the annotation tools, we are manually annotating a corpus of about 100,000 tokens. As part of this work, we have started to define the categories to be used for annotation. This includes redefining the SUC tag set (Ejerhed et al., 1992) to make it more in line with SAG (Teleman et al., 1999), and creating guidelines for a syntactic structure which defines both phrases and functions. The new corpus will be freely available for any use.

3. Lexical Analysis

The lexical analysis of Korp involves a number of subtasks: tokenization, identification of formal lexical units, lemmatization, compound analysis, and sense disambiguation. The large-scale semantic lexicon SALDO (Borin et al., 2013) is at the heart of all subtasks except tokenization. SALDO also serves as a pivot lexicon in Språkbanken, providing links to all other lexical resources in Språkbanken.

Tokenization has long been considered a trivial and solved task, but this view has been increasingly challenged in recent years (Chiarcos et al., 2009; Dridan and Oepen, 2012). We will replace tokenization with "lexing", identifying the fundamental units using a lexicon (SALDO), supplemented with an external rule set to deal with units outside the lexicon (e.g., phone numbers, dates, URLs, etc.). This also includes a focused effort on improving the overall quality of the lexical analysis.

The current annotation pipeline does not include word sense disambiguation. We will develop sense disambiguation tools using a combination of different methods: supervised methods for the most frequent and polysemous words, unsupervised methods for low-frequency words, and sense clustering methods to detect outliers and to discover previously undescribed senses. We have a version of the SUC corpus annotated with word senses (Järborg, 2003), which will serve as training and evaluation material in the development of a word sense disambiguation tool. In addition, we will use the annotations originating in the Swedish FrameNet project (Borin et al., 2010) for these purposes. The tools will be integrated in the Korp annotation pipeline.

4. PoS Tagging and Syntactic Analysis

Part-of-speech tagging in Korp is currently handled by the HunPoS-tagger (Halácsy et al., 2007), pre-trained on the SUC corpus (Megyesi, 2009). The tagger has achieved an accuracy of 97% on data similar to the training data, but it is still unclear how well it fares on, for instance, the multi-billion token blog and discussion forum corpora in Korp. These corpora contain non-standard spelling and a large amount of new words, which is difficult to handle for a statistical system.

A statistical system is always limited by the training data and research has shown that adding linguistic information (Loftsson et al., 2011) can improve results. We thus want to fully incorporate the resources we have available at Språkbanken, such as the semantic lexicon SALDO and its Swedish morphology, to improve tagging. Another method to improve results is to use multiple taggers (Henrich et al., 2009; Loftsson, 2009) based on different principles and training data. Assigning probabilities to the (possibly multiple) tags for each word, together with weights from other tools such as parsers, can help us find the most probable tag. Some of the tools, however, require adaptation to be fully compatible.

Syntactic annotation is currently done with an off-the-shelf version of the statistical dependency parser for Swedish, Maltparser (Nivre et al., 2007), which offers state-of-the-art accuracy for Swedish. The available model was trained on a modest amount of professionally written Swedish non-fiction published in the 1970s. Parsing quality appears to suffer for some of the diverse text types in Språkbanken. This could be ameliorated by adding knowledge from external sources to the parser, like dictionaries, or by adding statistical information from large corpora of different text types.

In the strict pipeline model, syntactic analysis is the end-point of the annotation workflow as it relies on information coming from POS-tagging and lexical analysis. However, we have identified problems at earlier levels of annotation, such as lexical analysis, that can be addressed with the help of the output of syntactic analysis. Research on the interaction between annotation levels has shown that this may increase annotation quality overall, for instance, Bohnet and Nivre (2012) on POS tagging and dependency parsing. We thus expect that integration of the parser in the new workflow will benefit annotation accuracy. For this, the parser will need to be adapted to efficiently process sets of alternative annotation hypotheses, in contrast to just one fixed annotation as is the case in the traditional NLP pipeline.

5. Conclusions

The Koala project is a major effort aimed at improving the annotation of the large corpus collection freely available through Språkbanken. The results will be better search possibilities in the research infrastructure, better tools for annotating Swedish texts, and new high-quality corpora.

Acknowledgements

The Koala project is funded by a grant from Riksbankens Jubilemsfond to the authors (Yvonne Adesam, PI), grant number In13-0320:1.

References

B. Bohnet and J. Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proc. of EMNLP/CoNLL.*

Lars Borin, Dana Danlls, Markus Forsberg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2010. The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.

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.

C. Chiarcos, J. Ritz, and M. Stede. 2009. By all these lovely tokens... merging conflicting tokenizations. In *Proc. of LAW.*

M. Dickinson and D. Meurers. 2003. Detecting errors in part-of-speech annotation. In *Proc. of EACL.*

M. Dickinson and D. Meurers. 2005. Detecting errors in discontinuous structural annotation. In *Proc. of ACL.*

R. Dridan and S. Oepen. 2012. Tokenization: returning to a long solved problem. a survey, contrastive experiment, recommendations, and toolkit. In *Proc. of ACL.*

E. Ejerhed, G. Källgren, O. Wennstedt, and M. Åström. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project - description and guidelines. Technical report, Department of Linguistics, Umeå University.

P. Halácsy, A. Kornai, and C. Oravecz. 2007. HunPos: an open source trigram tagger. In *Proc. of ACL.*

V. Henrich, T. Reuter, and H. Loftsson. 2009. Combitagger: A system for developing combined taggers. In *Proc. of FLAIRS.*

J. Järborg. 2003. *Semantisk uppmärkning: Metoder, problem och resultat.*, volume GU-ISS-03-2. Research Reports from the Dept of Swedish, Göteborg University.

H. Loftsson, S. Helgadóttir, and E. Rögnvaldsson. 2011. Using a morphological database to increase the accuracy in PoS tagging. In *Proc. of RANLP.*

H. Loftsson. 2009. Correcting a POS-tagged corpus using three complementary methods. In *Proc. of EACL.*

B. Megyesi. 2009. The open source tagger HunPoS for Swedish. In *Proc. of Nodalida.*

J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryiğit, S. Kübler, S. Marinov, and E. Marsi. 2007. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

U. Teleman, S. Hellberg, and E. Andersson. 1999. *Svenska akademiens grammatik.* Norstedts ordbok.