Cultivating a Swedish Treebank

Joakim Nivre
Beáta Megyesi
Sofia Gustafson-Čapková
Filip Salomonsson
Bengt Dahlqvist

Uppsala University
Department of Linguistics and Philology

1 Introduction

Treebanks, or syntactically annotated corpora, are an invaluable resource for the development and evaluation of syntactic parsers, as well as for empirical research on natural language syntax. Treebanks for Swedish have a long and venerable history, represented by the pioneering work on Talbanken (Einarsson, 1976a,b) and SynTag (Järborg, 1986). In recent years, however, treebank development for Swedish has mostly been limited to smaller projects, such as the reconstruction of Talbanken into Talbanken05 (Nivre et al., 2006), efforts to create a Swedish treebank with texts from the medical domain (Kokkinakis, 2006), and development of small parallel treebanks for Swedish-English-German (Gustafson-Čapková et al., 2007), and Swedish-English (Ahrenberg, 2007). As a consequence, there is still no Swedish treebank of the same scale as the largest available treebanks, such as the Penn Treebank for English (Marcus et al., 1993) or the Prague Dependency Treebank for Czech (Hajič et al., 2001).

Given the high cost of manual annotation and post-editing in treebank development, the possibility to reuse existing annotated resources is potentially of great importance. Often the efficient reuse of such resources is hampered by the fact that different resources, even for the same language, have been developed with different annotation guidelines or encoding standards. In many cases, however, it is possible to overcome these obstacles through a process of cross-corpus harmonization and annotation projection.

In this paper, we describe an ongoing project with the aim of bootstrapping a large Swedish treebank, ultimately with a size of about 1.5 million tokens, by reusing two existing annotated corpora: the previously mentioned treebank Talbanken, consisting of about 350,000 tokens, and the more recent Stockholm-Umeå Corpus (Ejerhed and Källgren, 1997), a part-of-speech-tagged corpus of about 1.2 million words. This treebanking effort is part of the

project *Methods and Tools for Automatic Grammar Extraction*, supported by the Swedish Research Council and initiated by Anna Sågvall Hein at Uppsala University. Besides treebank development, this project involves research on grammar induction with evaluation in the context of machine translation.

The paper is structured as follows. We first give an overview of the project and the different steps needed to develop a new treebank from existing resources and then focus on the two most interesting steps: the harmonization of tokenization and sentence segmentation, and the projection of annotation from one corpus to the other using data-driven taggers and parsers. The latter step also involves annotation refinement, where the syntactic annotation in Talbanken is extended and modified to better suit present-day requirements Finally, we briefly discuss how much is gained by reusing existing corpora and annotation, as opposed to creating a new treebank from scratch.

2 Bootstrapping a Large Swedish Treebank

Our ultimate goal is to produce a Swedish treebank containing 1.5 million tokens by reusing two existing annotated corpora. In section 2.1, we describe important properties of the two corpora; in section 2.2 we describe the major steps that need to be taken to reuse them as part of a single, consistently annotated treebank.

P10835069001	0000	«	GM	074
P10835069002	*DEN	PODPHH	SS	074
P10835069003	1000	RC	SSET	074
P1083506900410002	SOM	PORPHH	SS	074
P1083506900510002	VÄNTAR	VVPS	FV	074
P1083506900610002	MED	PR	OAPR	074
P1083506900710002	1100	IF	OA	074
P1083506900811003	ATT	IM	IM	074
P1083506900911003	TA	VVIV	IV	074
P1083506901011003	UT	ABZA	PL	074
P1083506901111003	ÅLDERSPENSIONEN	NNDDSS	OO	074
P1083506901210002	TILL	PR	TAPR	074
P1083506901310002	EFTER	PR	TATAPR	074
P1083506901410002	67-ÅRSMÅNADEN	NNDDSS	TATA	074
P10835069015	FÅR	FVPS	FV	074
P10835069016	HÖGRE	AJKP	OOAT	074
P10835069017	PENSION	NN	OO	074
P10835069018	.	IP	IP	074

Figure 1: Annotated sentence from Talbanken: *Den som väntar med att ta ut ålderspensionen till efter 67-årsmånaden får högre pension* (Those who do not claim their old age pension until after the 67-year month get a higher pension).

2.1 Component Corpora

Talbanken (Einarsson, 1976a,b) is a syntactically annotated corpus, containing both written and spoken Swedish, produced in the 1970s at the Department of Scandinavian Languages, Lund University, by a group led by Ulf Teleman. In total, the corpus contains about 350,000 tokens, divided into 200,000 tokens of written text (professional prose and high school essays) and 150,000 tokens of spoken language (interviews, debates, and informal conversations). The annotation consists of two layers: a lexical layer, with parts of speech and morphosyntactic features, and a syntactic layer, with a relatively flat phrase structure and grammatical functions (or dependencies). The annotation scheme, known as MAMBA, is described in Teleman (1974) and illustrated in figure 1, which shows a small extract from Talbanken.

The main asset of Talbanken, from our point of view, resides in the syntactic annotation, which contains enough information to support the extraction of both phrase structure and dependency structure representations, as shown in Nilsson et al. (2005) and Nivre et al. (2006), and therefore provides a good base representation for a treebank. Moreover, since Talbanken is by far the largest available corpus of Swedish with manually validated syntactic annotation, including it in the new treebank not only lets us reuse a manually validated syntactic annotation of 350,000 tokens, but also gives us a good basis for training parsers that can be used in the annotation of additional data.

The Stockholm-Umeå Corpus (SUC) (Ejerhed and Källgren, 1997) is a balanced corpus of written Swedish, modeled after the Brown Corpus and similar corpora for English, developed at Stockholm University and at Umeå University in a project led by Gunnel Källgren and Eva Ejerhed. The corpus consists of 1.2 million tokens of text from a variety of different genres, the corpus encoding follows the guidelines of the Text Encoding Initiative (TEI), and the annotation includes lemmatization, parts of speech, morphosyntactic features, and named entities. Since SUC was first released in the 1990s, its annotation scheme has become a de facto standard for Swedish, especially in research on part-of-speech tagging, where SUC data is standardly used for training and evaluation (see, e.g., Carlberger and Kann, 1999; Nivre, 2000; Megyesi, 2002). The annotation scheme is illustrated in figure 2, which shows a small extract from the corpus.

Given that SUC is a larger and more recently developed corpus, which has been extensively used to train taggers and other tools for Swedish, it makes sense to use SUC as a model for the new treebank wherever possible, thus minimizing the need for (new) manual validation and maximizing the conformance with current practice in Swedish language technology. This means, among other things, that principles of tokenization and sentence segmentation should be kept intact in SUC but modified for Talbanken in cases of conflict. We will refer to this as the *harmonization* of tokenization and sentence segmentation. The same holds for the annotation of parts of speech and mor-

```
<s id=fh06-089>
<w n=1487>Senare<ana><ps>AB<m>KOM<b>sen</w>
<w n=1488>på<ana><ps>PP<b>på</w>
<w n=1489>1940-talet<ana><ps>NN<m>NEU SIN DEF NOM<b>1940-tal</w>
<w n=1490>byggde<ana><ps>VB<m>PRT AKT<b>bygga</w>
<NAME TYPE=PERSON>
<w n=1491>John<ana><ps>PM<m>NOM<b>John</w>
<w n=1492>von<ana><ps>PM<m>NOM<b>von</w>
<w n=1493>Neumann<ana><ps>PM<m>NOM<b>Neumann</w>
</NAME>
<w n=1494>i<ana><ps>PP<b>i</w>
<NAME TYPE=PLACE>
<w n=1495>Princeton<ana><ps>PM<m>NOM<b>Princeton</w>
</NAME>
<w n=1496>i<ana><ps>PP<b>i</w>
<NAME TYPE=PLACE>
<ABBR>
<w n=1497>USA<ana><ps>PM<m>NOM<b>USA</w>
</ABBR>
</NAME>
<w n=1498>sina<ana><ps>PS<m>UTR/NEU PLU DEF<b>sin</w>
<num>
<w n=1499>första<ana><ps>RO<m>NOM<b>första</w>
</num>
<w n=1500>datamaskiner<ana><ps>NN<m>UTR PLU IND NOM<b>datamaskin</w>
<d n=1501>.<ana><ps>MAD<b>.</d>
</s>
```

Figure 2: Annotated sentence from the Stockholm-Umeå Corpus: *Senare på 1940-talet byggde John von Neumann i Princeton i USA sina första datamaskiner* (Later in the 1940s John von Neumann at Princeton in the USA built his first computers).

phosyntactic features, where the kind of annotation used in SUC has to be *projected* to Talbanken, which unfortunately uses a different scheme.[1] Since no simple mapping exists from the Talbanken scheme to the SUC scheme (nor in the other direction), this projection will have to be induced by training a tagger on the SUC corpus, using it to reannotate Talbanken, and finally correcting the errors performed by the tagger in a manual post-editing phase. In the following, we will use the term *morphological annotation* (in contrast to *syntactic annotation*) to include both basic parts of speech and morphosyntactic features.

[1]Other kinds of annotation found in SUC, such as lemmatization and named entities, are outside the scope of the current project but should in principle be projected in the same way from SUC to Talbanken.

2.2 Treebank Development

Given the considerations so far, we propose the following overall plan for the production of a new treebank based on Talbanken and SUC:

1. Convert both corpora with their existing annotation into a common standard for corpus encoding (XCES with standoff annotation).
2. Harmonize tokenization and sentence segmentation in Talbanken, applying as far as possible the principles adopted in SUC.
3. Project morphological annotation from SUC to Talbanken, using a data-driven tagger trained on SUC with manual post-editing.
4. Refine the syntactic annotation in Talbanken by automatic inference.
5. Project syntactic annotation from Talbanken to SUC, using a data-driven parser trained on Talbanken with manual post-editing.

In the following two sections, we describe the problems involved in harmonization, annotation refinement and annotation projection in a little more detail.

3 Harmonization

To harmonize the two corpora, we convert the tokenization and sentence segmentation of Talbanken according to the principles of SUC.

3.1 Tokenization

In the tokenization of SUC, abbreviations are always represented as single tokens. This means that when abbreviations in the original text contain spaces, the different elements are concatenated into one token where spaces in the original text are represented by underscores. Moreover, different variants of the same abbreviations are normalized to one form. Thus, the following variants of the abbreviation of *till exempel* (for example):

t. ex.
t ex

are all tokenized as one token:

t_ex

In Talbanken, on the other hand, abbreviations consisting of several elements are annotated as multi-word expressions. Each element of abbreviation is treated as a separate token, but only the first token is assigned proper lexical and syntactic annotation, while the subsequent token are assigned the dummy

tag ID (in both the lexical and the syntactic annotation). To find the abbreviations in Talbanken, we automatically extract tokens annotated with ID tags together with the preceding token, and convert these into a single token with the form prescribed by SUC's tokenization standards.

Certain numerical expressions, such as *3–5*, are also tokenized differently in the two corpora, such that SUC often has a single token where Talbanken splits the expression into several tokens (which may or may not be annotated as a multi-word expression). For certain types of numerical expressions, it is again possible to perform the harmonization automatically, but most cases here need to be checked manually.

3.2 Sentence Segmentation

The sentence segmentation also differs in the two corpora. Above all, lists have a different structural annotation. In SUC, items in lists are handled as different sentence units, while in Talbanken the entire list consisting of several items is treated as one sentence if there are clear syntactic relations between the items. This could lead to errors when we use a data-driven parser to project the syntactic annotation from Talbanken to SUC, since the sentences in Talbanken that would serve as training data would have a different structure compared to the sentences in SUC that need to be parsed. Therefore, we treat each list item in Talbanken as a separate sentence unit as far as possible.

4 Annotation Projection and Refinement

4.1 Morphological Annotation

In order to harmonize the morphological annotation of the two corpora, we project the part-of-speech tags and morphological features from SUC to Talbanken. We do this by training the data-driven TnT tagger (Brants, 2000) on SUC, bootstrapping the tagger by training it on a considerably larger automatically tagged corpus (Forsbom, 2005), and then applying the trained model to Talbanken. Finally, we correct the automatic annotation manually following SUC's annotation principles. The result is a merged corpus with consistent morphological annotation.

At the time of writing, all closed class words in Talbanken have been checked and we predict that the work on morphological annotation will be completed during the spring of 2008. One of the advantages of reusing a previously annotated corpus, even if the annotation is inconsistent, is that checking can be speeded up by pattern matching on the combined old and new annotation. This is convenient especially for closed word classes, where the lack of morphological features often makes the two annotation systems equivalent so that some words need to be checked only if the new and the old annotations are inconsistent.

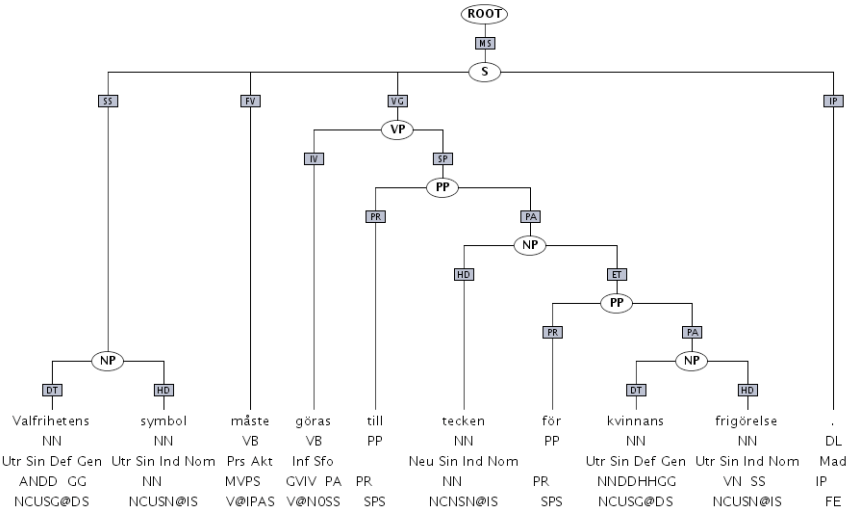

Figure 3: Refined syntactic annotation for the sentence *Valfrihetens symbol måste göras till tecken för kvinnans frigörelse* (The symbol of free choice must be turned into a sign of women's liberation).

4.2 Syntactic Annotation

For the syntactic annotation, we have to project the syntactic analysis of Talbanken to SUC, since SUC lacks syntactic annotation. This is achieved in the first place by training the data-driven MaltParser (Nivre et al., 2006) on Talbanken and using the trained model to parse SUC. At a later stage, we may complement this annotation with the output of different parsers, which will enable us to use an ensemble of classifiers to facilitate the manual correction of the syntactic annotation.

However, before we can train parsers on Talbanken, the syntactic annotation needs to be refined to meet the requirements of present-day language technology. The original annotation has very flat structures and lacks phrase categories such as NP, VP, PP, etc. In order to obtain a hierarchical analysis involving both phrase structure and dependency structure (grammatical functions), we therefore need to do two things:

- Infer phrase categories using information about grammatical functions (both internal and external to a given phrase) and about the parts of speech of constituent words.
- Infer additional structure, such as NPs within PPs, VPs within VPs, based on information about grammatical functions, parts of speech, and inferred phrase categories.

The methodology for automatic annotation refinement is described in Nilsson et al. (2005). Figure 3 shows an example of the current version of the refined syntactic annotation. Our goal is to complete this work during the spring of 2008, which means that a first version of the entire treebank, with purely automatic syntactic annotation in the SUC part, could be released in the fall of 2008. A version where all the annotation has been checked manually remains as a long-term goal for our efforts.

5 How Much Is Gained?

A reasonable question to ask is how much is actually gained by reusing existing corpora, as opposed to building a new treebank from scratch, given the considerable amount of work involved in the harmonization and projection processes. Let us therefore make an attempt at quantifying the gains and balancing them against the disadvantages.

By reusing all the annotation in SUC and the syntactic annotation in Talbanken, we save all the work needed to manually correct tokenization, sentence segmentation, and morphological annotation of 1.2 million tokens, and syntactic annotation of 350,000 tokens. In addition, we save the work needed to check tokenization and sentence segmentation for 350,000 tokens in Talbanken, minus a few person weeks spent on harmonization. Finally, although the morphological annotation of 350,000 tokens in Talbanken still has to be checked manually, both the efficiency and the accuracy of this process can be improved by making use of the old morphological annotation for consistency checking.

To give just one illustrative example, the string *men* in Swedish can be either a coordinating conjunction (but) or a noun (injury). After projecting the new morphological annotation from SUC to Talbanken, it was found that one occurrence of *men* was tagged as a noun in the old annotation and as a conjunction in the new annotation, whereas the remaining 364 occurrences were tagged as conjunctions in both cases. Unsurprisingly, the single occurrence with inconsistent annotation turned out to be a tagging error, which in this way could be detected and corrected. With very high probability, the remaining 364 occurrences are correctly tagged as conjunctions (since the old annotation has been checked manually) and therefore do not need to be checked.[2]

To sum up, we see that cross-corpus harmonization and annotation projection can lead to substantial gains in the manual work needed to validate segmentation and annotation. This of course has to be weighed against a number of other factors, in particular that the new treebank has to be based on old data (in the case of Talbanken, texts from the 1970s) and that the annotation

[2]Other examples are *man*, which is ambiguous between a pronoun (one) with 699 occurrences and a noun (man) with 67 occurrences, and *Vi*, which has a single occurrence as the name of a magazine and 328 occurrences as a capitalized pronoun (we).

schemes have to be inherited from at least one of the old corpora. Still, in situations where manual effort has to be minimized, the approach taken appears to be a viable methodology for producing a large-scale treebank from existing resources.

6 Conclusion

In this paper, we have presented ongoing work to produce a large treebank of Swedish by reusing two existing annotated corpora, Talbanken and SUC. A key component in the bootstrapping methodology is the use of cross-corpus harmonization and annotation projection, supported by automatic conversion procedures and data-driven linguistic analyzers, with a minimum of manual validation. In this way, we hope to be able to create a large-scale, high-quality Swedish treebank, a resource that is badly needed for research and development in language technology, as well as for empirical linguistic research.

References

Ahrenberg, L. (2007). LinES: An English-Swedish parallel treebank. In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA)*, pp. 270–273.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the 6th Conference on Applied Natural Language Processing (ANLP)*.

Carlberger, J. and V. Kann (1999). Implementing an efficient part-of-speech tagger. *Software Practice and Experience 29*, 815–832.

Einarsson, J. (1976a). Talbankens skriftspråkskonkordans. Lund University, Department of Scandinavian Languages.

Einarsson, J. (1976b). Talbankens talspråkskonkordans. Lund University, Department of Scandinavian Languages.

Ejerhed, E. and G. Källgren (1997). Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University. ISBN 91-7191-348-3.

Forsbom, E. (2005). Big is beautiful: Bootstrapping a pos tagger for swedish. Poster presentation at the GSLT retreat. Gullmarsstrand, January 27-29.

Gustafson-Čapková, S., Y. Samuelsson, and M. Volk (2007). SMULTRON (version 1.0) – The Stockholm MULtilingual parallel TReebank. http://www.ling.su.se/dali/research/ smultron/index.htm. An English-German-Swedish parallel treebank with sub-sentential alignments.

Hajič, J., B. Vidova Hladka, J. Panevová, E. Hajičová, P. Sgall, and P. Pajas (2001). Prague Dependency Treebank 1.0. LDC, 2001T10.

Järborg, J. (1986). Manual för syntaggning. Technical report, Göteborg University, Department of Swedish.

Kokkinakis, D. (2006). Towards a swedish medical treebank. In J. Hajič and J. Nivre (Eds.), *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories*, pp. 199–210.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics 19*, 313–330.

Megyesi, B. (2002). *Data-Driven Syntactic Analysis: Methods and Applications for Swedish*. Ph. D. thesis, KTH: Department of Speech, Music and Hearing.

Nilsson, J., J. Hall, and J. Nivre (2005). MAMBA meets TIGER: Reconstructing a Swedish treebank from Antiquity. In P. J. Henrichsen (Ed.), *Proceedings of the NODALIDA Special Session on Treebanks*.

Nivre, J. (2000). Sparse data and smoothing in statistical part-of-speech tagging. *Journal of Quantitative Linguistics 7*, 1–17.

Nivre, J., J. Hall, and J. Nilsson (2006). Maltparser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 2216–2219.

Nivre, J., J. Nilsson, and J. Hall (2006). Talbanken05: A swedish treebank with phrase structure and dependency annotation. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pp. 1392–1395.

Teleman, U. (1974). *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur.