Talbanken05: A Swedish Treebank with Phrase Structure and Dependency Annotation

Joakim Nivre, Jens Nilsson, Johan Hall

Växjö University, Sweden
School of Mathematics and Systems Engineering
{nivre, jni, jha}@msi.vxu.se

Abstract

We introduce Talbanken05, a Swedish treebank based on a syntactically annotated corpus from the 1970s, Talbanken76, converted to modern formats. The treebank is available in three different formats, besides the original one: two versions of phrase structure annotation and one dependency-based annotation, all of which are encoded in XML. In this paper, we describe the conversion process and exemplify the available formats. The treebank is freely available for research and educational purposes.

1. Introduction

Treebanks have become an essential resource for the development, optimization and evaluation of broad-coverage syntactic parsers, and treebanks have therefore been developed for a wide range of languages on a smaller or larger scale. For Swedish there has until now been no large-scale treebank generally available, which is somewhat surprising, since some of the earliest examples of syntactically annotated corpora, Talbanken in the 70s (Einarsson, 1976a; Einarsson, 1976b) and SynTag in the 80s (Järborg, 1986) were based on Swedish data. Talbanken was created in Lund and contains close to 300,000 words of both written and spoken Swedish, manually annotated with partial phrase structure and grammatical functions according to the MAMBA scheme (Teleman, 1974), and was a very impressive achievement at the time of its creation. However, by modern standards it is probably best characterized as a "proto-treebank", since the annotation format makes it rather difficult to use with contemporary parsers and treebank tools.

In order to facilitate the reuse of Talbanken, we have converted the data to a modern format, using TIGER-XML and Malt-XML as the formal representation languages. A conversion program has been created in order to do this automatically. In this process, we have converted the original annotation to three different formats, two phrase structure representations together with grammatical functions, similar to the annotation in the German TIGER Treebank (Brants et al., 2002), and one using pure dependency representations, as used e.g. in the Prague Dependency Treebank (Hajič et al., 2001). The existence of two parallel and consistent annotations makes the newly created treebank rather unusual, although conversions of treebanks to dependency structure have been conducted before. The Spanish part of the constituency-based treebank 3LB (Gelbukh et al., 2005) is a recent example, where heuristic rules have been used in order to infer the dependency structure. Our approach is also based on heuristic rules. Although no large-scale evaluation of the conversion process has been performed, preliminary studies indicate that the conversion is very reliable.

In this paper, we present the recently released treebank, referred to as Talbanken05 to distinguish it from the original annotated corpus, which we call Talbanken76. We begin with a brief description of Talbanken76 and move on to the conversion process leading to Talbanken05 and the different formats in which it is available. Talbanken05 is freely available for research and educational purposes.[1]

2. Talbanken76

Talbanken76 consists of a written language part (Einarsson, 1976a) and a spoken language part (Einarsson, 1976b) of roughly equal size. The written language part in turn consists of two sections, the so-called professional prose section (P), with data from textbooks, brochures, newspapers, etc., and a collection of high school students' essays (G). The spoken language part also has two sections, interviews (IB) and conversations and debates (SD). Altogether, the corpus contains close to 300,000 running tokens.

The MAMBA annotation scheme (Teleman, 1974) consists of two layers, the first being a lexical analysis, consisting of part-of-speech information including morphological features, and the second being a syntactic analysis, in terms of grammatical functions. Both layers are flat in the sense that they consist of tags assigned to individual word tokens, but the syntactic layer also gives information about constituent structure, as exemplified in the annotation of the sentence *Genom skattereformen införs individuell beskattning av arbetsinkomster* (Through the tax reform, individual taxation of work income is introduced):

```
*GENOM              PR          AAPR
SKATTEREFORMEN      NNDDSS      AA
INFÖRS              VVPSSMPA    FV
INDIVIDUELL         AJ          SSAT
BESKATTNING         VN          SS
AV                  PR          SSETPR
ARBETSINKOMSTER     NN  SS      SSET
.                   IP          IP
```

The first column of annotation is the lexical analysis, while the second column is the syntactic analysis. The grammatical subject of the sentence is the phrase *individuell beskattning av arbetsinkomster* (individual taxation of work income), where the head word *beskattning* (taxation) is assigned the simple tag SS for subject, while the

[1] URL: http://www.msi.vxu.se/users/nivre/research/Talbanken05.html

pre-modifying adjective *individuell* (individual) is tagged SS and AT for adjectival modifier; in the post-modifying prepositional phrase, the noun *arbetsinkomster* (work income) is tagged SS and ET for post-modifier, while the preposition *av* (of) is tagged SS, ET and PR for preposition.

3. Conversion

The syntactic analysis in Talbanken76 is described by its creators as an eclectic combination of dependency grammar, topological field analysis and immediate constituent analysis (Teleman, 1974). This makes it very suitable for conversion to both phrase structure and dependency annotation. The conversion has proceeded in four steps:

1. The original flat but multi-layered annotation is converted to a bare phrase structure annotation, i.e. a phrase structure with unlabeled nonterminal nodes, and edges labeled with grammatical functions. This conversion is rather straightforward given the partially hierarchical annotation exemplified above.
2. The bare phrase structure annotation is extended to a full phrase structure representation by labeling nonterminal nodes with syntactic categories. These categories are not part of the original annotation and have to be inferred from other parts of the annotation.
3. The full phrase structure annotation is deepened by inserting extra non-terminal nodes, which are not directly warranted by the original flat annotation but by theoretical considerations, such as NP nodes within PPs, S nodes within SBARs, etc.
4. The deepened phrase structure annotation is converted to a dependency annotation using the standard technique with head-finding rules (Magerman, 1995; Collins, 1996) and preserving grammatical functions as edge labels. Head-finding rules are not part of the original annotation scheme and have to be constructed manually.

The full phrase structure annotation, the deepened phrase structure annotation, and the dependency annotation are the currently available formats for Talbanken05.

4. Phrase Structure Annotation

The full phrase structure annotation, which is the outcome of the second conversion step, uses a conventional set of phrase types (S, NP, VP, etc.) in combination with the grammatical functions of the original MAMBA annotation. The representation allows discontinuous phrases, as in the German TIGER annotation scheme (Brants et al., 2002), although discontinuous constituents are relatively rare in the treebank.

The conversion of MAMBA to TIGER-XML gives rise to a bare phrase structure, i.e. a phrase structure without nonterminal node labels. Nonterminal node labels have been inferred by considering:

1. Grammatical functions of the node's children
2. Grammatical function of the node
3. Lexical categories of the node's children

An ordered set of labeling rules has been created manually, which is applied to all nodes in the phrase structure trees. A labeling rule is a quadruple (C, P, L, N), where:

1. C and P are lists of grammatical functions,
2. L is a list of lexical categories,
3. N is a nonterminal node label.

A labeling rule (C, P, L, N) assigns the label N to a node n if the following conditions are satisfied:

1. n has a child with a grammatical function $g \in C$, or C = *,
2. n has a grammatical function $g \in P$, or P = *,
3. n has a child with a lexical category $l \in L$, or L = *.

Below we show the set of rules needed to label the example sentence in figure 1,[2] ordered by decreasing priority.

MS	*	*	ROOT
PR	*	*	PP
SS, FV	*	*	S
*	+F	*	S
IV, IM	*	*	VP
*	VS, VO	*	VP
DT, AT, ET	*	*	NP
HD	*	AJ, TP, SP	AP
*	*	*	XP

The first rule is only applied on the nonterminal acting as the root of each sentence, since only the root can have children with the grammatical function *MS* (first column). The second rule is used for assigning the label *PP* to prepositional phrases, because children with the grammatical function *PR* mark prepositions. This is the situation for the nonterminal having the preposition *vid* as one of its children. In a similar fashion, when the conversion process encounters a finite verb or a subject among the children, having the grammatical function *SS* or *FV*, we infer that the label of the nonterminal is *S*. Consequently, the words *Man* (SS) and *fäster* (FV) trigger the assignment of the label *S* to their mutual mother node.

The deepened phrase structure is constructed from the full phrase structure annotation by inserting, e.g., NPs within PPs and VPs within (larger) VPs. The deepening has also been performed automatically. Figure 2 shows the same sentence with the deepened phrase structure annotation.

5. Dependency Annotation

The dependency annotation, which is the outcome of the third conversion step, consists of terminal nodes connected by edges labeled with grammatical functions of MAMBA and is encoded in Malt-XML, a representation defined for the data-driven parser-generator MaltParser (Nivre and

[2] English translation: "One gives greater weight to the pupils' spontaneous ability to express themselves orally and in writing".

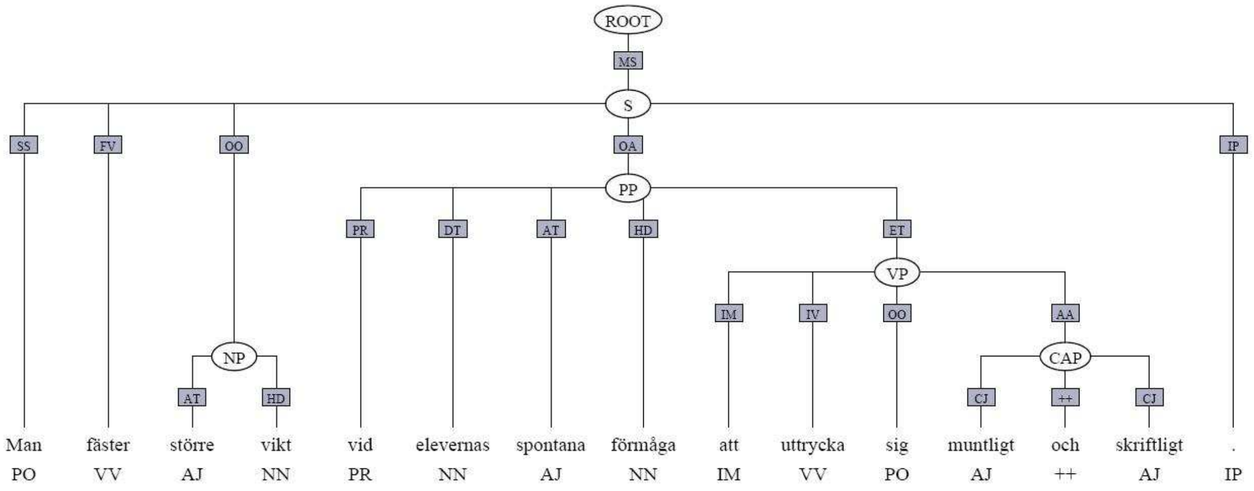

Figure 1: Phrase structure annotation in Talbanken05 (flat version)

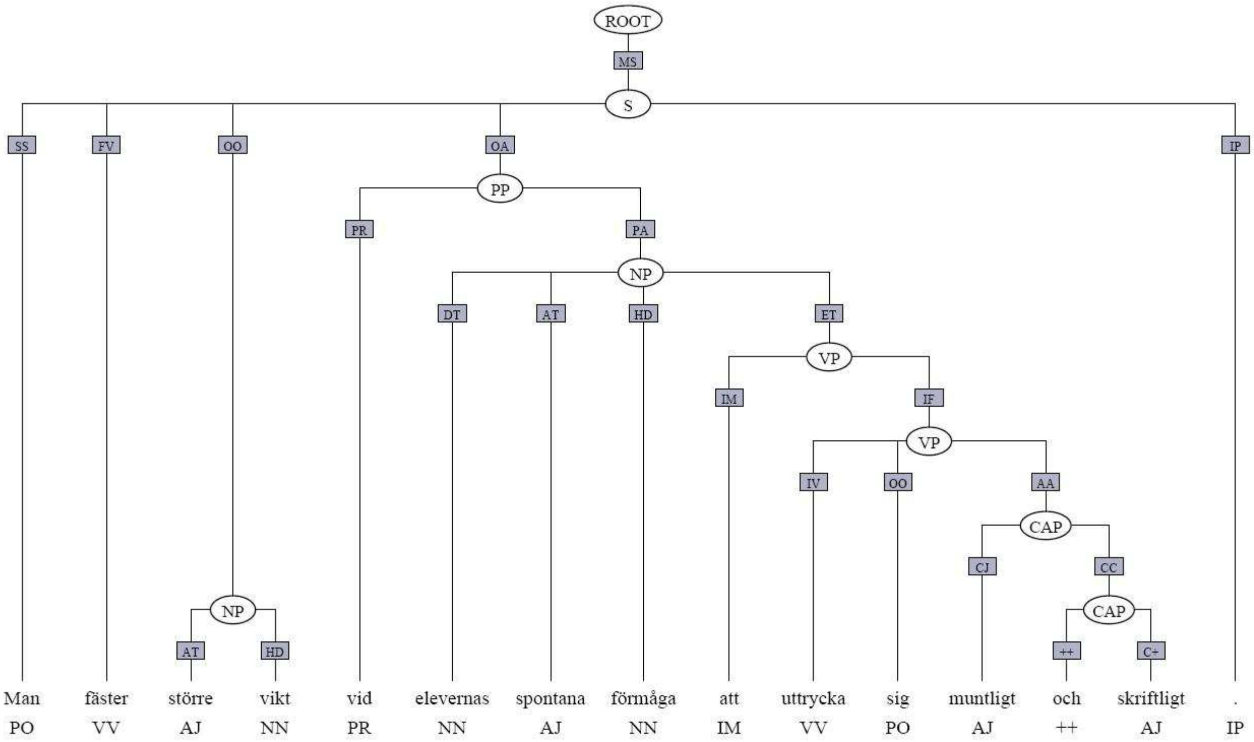

Figure 2: Phrase structure annotation in Talbanken05 (deepened version)

Hall, 2005). The representation allows non-projective dependency structures, which are needed to capture discontinuous constituents. The dependency representation is extracted from the deepened phrase structure.

The conversion process traverses all nonterminals in each sentence. Provided that every nonterminal node n has a unique head child c_h, the constituent structure can be converted to a dependency structure by recursively letting the head d of each non-head child c_d of n be a dependent to the head h of the head child c_h (where a terminal node $c_h = h$ is its own head). MAMBA provides explicit annotation of head children for many constituent types, such as noun phrases. However, there are two problematic cases:

1. A nonterminal node n has more than one head child.
2. A nonterminal node n has no head child.

The first case is found in coordinate structures and is resolved by letting the leftmost head child be the head in the dependency structure. The second case is found at the clause level, where the MAMBA annotation encodes a topological field analysis, and is resolved by having a priority list of grammatical functions to identify the head. In table 1 we show the priority list of grammatical functions. In principle, the node with an edge label with the highest priority becomes the head, breaking ties from left to right. When a head has been identified for a nonterminal, the next

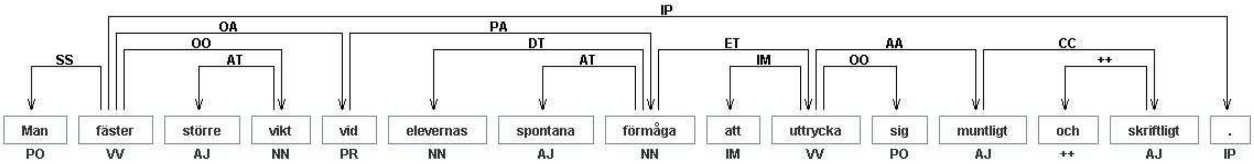

Figure 3: Dependency annotation in Talbanken05

EDGE LABEL
Head (HD)
Finite verb (FV)
Non-finite verb (IV)
Predicative complement (SP)
Subjects and objects (e.g. SS, ES, FS, OO, EO, FO)
Clause adverbials (AA, KA, RA, OA, TA)
Phrase adverbials (+A, CA, MA, NA, VA, XA)
Nominal pre-modifier (AT)
Nominal post-modifier (ET)
Other noun dependents (DT, XT)
Unclassifiable dependent (XX)
Others not involved in coordination (e.g. +F, IM, AN)
Conjunct (CJ)
Label to the right of conjuncts (CC)
Label to the right of conjunctions (C+)
Coordinator, conjunction (++)
Punctuation (e.g. I?, IC, IG, IK, IP, IQ, IR, IS, IT, ST)

Table 1: The priority list of head-finding rules

step is to make all other words dependents of the head. In the simpler case, when the head is an individual word (terminal node), that word be the head of all its child nodes. The verb *fäster* is a terminal that becomes the head of the other terminals and nonterminals of the phrase it belongs to. When the identified head instead is a non-terminal node, the terminal identified as the head inside that nonterminal will also be the head of all the child nodes. (No such case exists in the phrase structure example.)

The corresponding procedure holds for the children too. If a child is an individual word, then that word is the dependent of the head. For example, this is the case for the subject *Man* and its head *fäster*. Also, the dependency relation between the head word and the dependent is set to the edge label of the child, which in this case is *SS*. If the child on the other hand is a nonterminal, the head word of that phrase is the dependent to the head. This is the kind of relation between the words *vikt* and *fäster*, since the former word is the head of the NP constituting the object of the verb. The dependency relation between them is OO, since that is the edge label of the phrase that is dependent on *fäster*. Figure 3 shows our running example sentence annotated with the pure dependency representation.

6. Conclusion

In this paper, we have presented Talbanken05, a recently released Swedish treebank with annotation of both phrase structure and dependency structure, derived from an older syntactically annotated corpus, Talbanken76. In the absence of a large-scale treebank based on contemporary language data, we hope that Talbanken05 can serve as a useful resource for Swedish language technology. The treebank comes with no guarantee but is freely available for research and educational purposes as long as proper credit is given for the work done to produce the material.

7. References

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. TIGER treebank. In Erhard Hinrichs and Kiril Simov, editors, *Proceedings of the First Workshop on Treebanks and Linguistic Theories (TLT)*, pages 24–42.

Michael Collins. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 184–191.

Jan Einarsson. 1976a. Talbankens skriftspråkskonkordans. Lund University, Department of Scandinavian Languages.

Jan Einarsson. 1976b. Talbankens talspråkskonkordans. Lund University, Department of Scandinavian Languages.

A. Gelbukh, H. Calvo, and S. Torres. 2005. Transforming a constituency treebank into a dependency treebank. In *Procesamiento Lenguaje Natural*, Spain.

Jan Hajič, Barbora Vidova Hladka, Jarmila Panevová, Eva Hajičová, Petr Sgall, and Petr Pajas. 2001. Prague Dependency Treebank 1.0. LDC, 2001T10.

Jerker Järborg. 1986. Manual för syntaggning. Technical report, Göteborg University, Department of Swedish.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 276–283.

Joakim Nivre and Johan Hall. 2005. MaltParser: A language-independent system for data-driven dependency parsing. In *Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories (TLT)*.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Studentlitteratur.