What kinds of trees grow in Swedish soil?

xxx

yyy
uuu
E-mail: email@email

1 Introduction

One of the issues brought up in this workshop concerns the relationship between the syntactic properties of a given language and the choice of linguistic theory for annotation purposes. Our Swedish treebank consortium, consisting of researchers from Växjö University, KTH and Stockholm University, is currently facing a specific instance of this issue in trying to define an annotation standard for a large-scale treebank of Swedish written and spoken language.

In this paper, I will discuss and compare four different annotation schemes that have been proposed for Swedish in terms of their suitability for Swedish syntax as well as their relationship to linguistic theory and annotation schemes proposed for other languages. Other aspects that will be touched upon are the availability of parsers and/or annotated training data for developing parsers, the different requirements for annotation of spoken and written language, and the different needs of different user groups.

By way of background, I will start by reviewing some basic facts about the syntax of Swedish, a Germanic verb second language with moderately fixed word order. In doing this I will also introduce the Scandinavian tradition of descriptive grammar, in particular the influential field model due to Diderichsen [5]. The background section also contains a brief discussion of existing annotation schemes for other languages and their relation to current linguistic theory.

The main part of the paper will be devoted to a discussion and comparison of the following four annotation schemes for Swedish:

- MAMBA (Teleman [19])
- SynTag (Järborg [6])

- SWECG (Birn [1])
- S-CLE (Gambäck [7])

The four schemes fall naturally into two groups, MAMBA and SynTag being standards designed for manual annotation of corpus material, while SWECG and S-CLE are primarily general purpose parsing systems which have corpus annotation as one of their (potential) applications.

2 Treebanks and Linguistic Theory

The number of treebanks available for different languages is growing steadily and with them the number of different annotation schemes. This makes it very difficult to say something general about the relation between annotation schemes and linguistic theory, but broadly speaking I think we may distinguish three main kinds of annotation in current practice:

- Annotation of constituent structure
- Annotation of functional structure
- Theory-specific annotation

This is obviously not a proper taxonomy, since theory-specific annotation may concern both constituent structure and functional structure. Rather, the first two categories are meant to cover more or less theory-neutral annotation schemes, focusing on constituent structure or functional structure, respectively. It should also be pointed out immediately that the annotation found in many if not most of the existing treebanks actually combines two or even all three of these categories. Still, I believe that the categories may be useful in discussing existing annotation schemes and their relation to linguistic theory. I will treat the categories in the order in which they are listed above, which I think roughly corresponds to the historical development of treebank annotation schemes.

The annotation of *constituent structure*, often referred to as *bracketing*, is the main kind of annotation found in pioneering projects such as the Lancaster Parsed Corpus (Garside et al. [8]) and the original Penn Treebank (Marcus et al. [10]). Normally, this kind of annotation consists of part-of-speech tagging for individual word tokens and annotation of major phrase structure categories such as NP, VP, etc. Figure 1 shows a representative example, taken from the IBM Paris Treebank using a variant of the Lancaster annotation scheme.

Annotation schemes of this kind are usually intended to be theory-neutral and therefore try to use mostly uncontroversial categories that are recognized in all or

```
[N Vous_PPSA5MS N]
[V accedez_VINIP5
   [P a_PREPA
      [N cette_DDEMFS session_NCOFS N]
   P]
   [Pv a_PREP31 partir_PREP32 de_PREP33
      [N la_DARDFS fenetre_NCOFS
         [A Gestionnaire_AJQFS
            [P de_PREPD
               [N taches_NCOFP N]
            P]
         A]
      N]
   Pv]
V]
```

Figure 1: Constituency annotation in the IBM Paris Treebank

most syntactic theories that assume some notion of constituent structure. Moreover, the structures produced tend to be rather flat, since intermediate phrase level categories are usually avoided, as well as complex structures such as Chomsky adjunction. The drawback of this is that the number of distinct expansions of the same phrase category can become very high. For example, Charniak [3] was able to extract 10,605 distinct context-free rules from a 300,000 word sample of the Penn Treebank. Of these, only 3943 occurred more than once in the sample.

The status of grammatical functions and their relation to constituent structure has long been a controversial issue in linguistic theory. Thus, whereas the standard view in transformational syntax since Chomsky [4] has been that grammatical functions are derivable from constituent structure, proponents of dependency syntax such as Mel'čuk [13] have argued that functional structure is more fundamental than constituent structure. Other theories, such as LFG, steer a middle course by assuming both notions as primitive.

When it comes to treebank annotation, the annotation of *functional structure* has become increasingly important in recent years. The most radical examples are perhaps the annotation schemes based on dependency syntax, exemplified by the Prague Dependency Treebank of Czech (Hajic [9]) and the METU Treebank of Turkish (Oflazer et al. [14]), where the annotation of dependency structure is added directly on top of the morphological annotation without any layer of constituent

structure. Figure 2 shows a simple example of dependency annotation from the Prague Dependency Treebank.

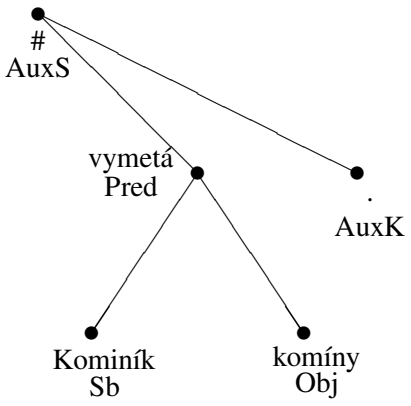

Figure 2: Functional annotation in the Prague Dependency Treebank

The trend towards more functionally oriented annotation schemes is also reflected in the extension of constituency-based schemes with annotation of grammatical functions. Cases in point are SUSANNE (Sampson [17]), which is a development of the Lancaster annotation scheme mentioned above, and Penn Treebank II (Marcus et al. [11]), which adds functional tags to the original phrase structure annotation. One of the most interesting examples in this respect is the annotation scheme adopted in the TIGER Treebank of German (Brants and Hansen [2]), developed from the earlier NEGRA treebank and annotation scheme, which integrates the annotation of constituency and dependency in a graph where node labels represent phrasal categories while edge labels represent syntactic functions.

The third kind of annotation scheme that is found in available treebanks is the kind that adheres to a specific linguistic theory and uses representations from that theory to annotate sentences. Thus, HPSG has been used as the basis for treebanks of Bulgarian (Simov et al. [18]) and Polish (Marciniak et al. [12]), and the Prague Dependency Treebank mentioned earlier is based on the theory of Functional Generative Description (Sgall et al. [16]). There has also been work done on automatic f-structure annotation in the theoretical framework of LFG (see, e.g., Sadler et al. [15]).

In conclusion, we may perhaps say that there has been a trend towards more functionally oriented annotation schemes in recent years, and that theory-specific annotation schemes have become more common, but that it is probably still true to say that the dominant paradigm in treebank annotation is the kind of theory-neutral annotation of constituent structure with added functional tags represented by schemes such as the Penn Treebank II standard.

3 Conclusion

In conclusion, MAMBA and SWECG emerge as the strongest candidates for use in the annotation of a Swedish treebank. The other two schemes considered, SynTag and S-CLE, are interesting in their own right but are on the whole less suitable for adoption in a large-scale treebank project.

MAMBA and SWECG have the advantage of being firmly based in the Swedish tradition of descriptive grammar and can therefore be expected to have good descriptive adequacy and coverage. This is true especially for MAMBA, which has been designed especially to handle spoken language as well as written language. Moreover, the fact that these schemes are based on notions of traditional grammar means that they provide an annotation which may be more accessible to non-expert treebank users.

The main weakness of SWECG is that the annotation contains little or no information about phrase structure and is therefore difficult to relate to many current linguistic theories. However, this situation has clearly improved with the development of FDG, which establishes a more direct connection to dependency-based theories of syntax and also provides a better basis for the reconstruction of phrase structure from dependency structure if this is required.

For MAMBA the biggest problem is instead the lack of resources for automatic annotation, although it may be possible to improve the situation by using the available annotated corpora for bootstrapping a parsing system.

References

[1] Birn, Juhani (1998) Swedish Constraint Grammar. Lingsoft Inc. (URL: http://www.lingsoft.fi/ doc/swecg/intro/).

[2] Brants, Sabine and Hansen, Silvia (2002) Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)*, pp. 1643–1649, Las Palmas.

[3] Charniak, Eugene (1996) Tree-Bank Grammars. In *AAAI/IAAI*, Vol. 2, pp. 1031–1036.

[4] Chomsky, Noam (1965) *Aspects of the Theory of Syntax.* MIT Press.

[5] Diderichsen, Paul (1946) *Elementær dansk grammatik.* Copenhagen: Gyldendal.

[6] Järborg, Jerker (1986) Manual för syntaggning [Manual for syntagging]. Göteborgs universitet: Institutionen för språkvetenskaplig databehandling.

[7] Gambäck, Björn and Rayner, Manny (1992) The Swedish Core Language Engine. In *Papers from the 3rd Nordic Conference on Text Comprehension in Man and Machine*, Linköping University, Linköping, Sweden, pp. 71–85.

[8] Garside, R., Leech, G. and Varadi, T. (compilers) (1992) *Lancaster Parsed Corpus*. A machine-readable syntactically-analysed corpus of 144,000 words, available for distribution through ICAME, The Norwegian Computing Centre for the Humanities, Bergen.

[9] Hajic, Jan (1998) Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning*, pp. 106–132. Prague: Karolinum.

[10] Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19, 313–330. [Reprinted in Armstrong, Susan (ed.) (1994) *Using large corpora*, pp. 273–290. Cambridge, MA: MIT Press.]

[11] Marcus, Mitchell P., Kim, Grace, Marcinkiewicz, Mary Ann, MacIntyre, Robert, Bies, Ann, Ferguson, Mark, Katz, Karen and Schasberger, Britta (1994) The Penn Treebank: Annotating Predicate Argument Structure", In *ARPA Human Language Technology Workshop*.

[12] Marciniak, Malgorzata, Mykowiecka, Agniezska, Kupść, Anna and Przepiórkowski, Adam (2000) An HPSG-Annotated Test Suite for Polish. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000).*

[13] Mel'čuk, Igor (1988) *Dependency Syntax: Theory and Practice*. State University of New York Press.

[14] Oflazer, Kemal, Say, Bilge and Hakkani Tur, Dilep (2000) A Syntactic Annotation Scheme for Turkish. In *Proceedings of 10th International Conference on Turkish Linguistics (ICTL-2000).*

[15] Sadler, Louisa, von Genabith, Josef and Way, Andy (2000) Automatic F-Structure Annotation from the AP Treebank. In Butt, Miriam and Holloway King, Tracy (eds.) *Proceedings of the Fifth International Conference on Lexical-Functional Grammar*, The University of California at Berkeley, 19 July – 20 July 2000. Stanford, CA: CSLI Publications.

[16] Sgall, Petr, Hajicova, Eva and Panevova, Jarmila (1986) *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel.

[17] Sampson, Geoffrey (1995) *English for the Computer*. Oxford University Press.

[18] Simov, Kiril, Popova, Gergana, Osenova, Petya (forthcoming) HPSG-Based Syntactic Treebank of Bulgarian (BulTreeBank). In Wilson, Andrew, Rayson, Paul, McEnery, Tony (eds.) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, pp. 135-142. Munich: Lincom-Europa.

[19] Teleman, Ulf (1974) *Manual för grammatisk beskrivning av talad och skriven svenska [Manual for grammatical description of spoken and written Swedish]*. Lund: Studentlitteratur.