Treebanking in Northern Europe:
A White Paper

by

Joakim Nivre, Koenraad de Smedt and Martin Volk

Abstract

We present the case for an extensive scientific effort to build up large treebanks for the Nordic and Baltic languages, as a step towards developing advanced multilingual communication technologies for these languages in the future.

Nordic language technology is urgent

Language and speech processing is rapidly becoming a priority area for Northern Europe (in which we include the Baltic states). Two recent, broad developments are causes of this urgency. First, the speed at which digital information technologies have penetrated our society has accelerated particularly strongly in Northern Europe. Secondly, the expansion of the European Union has had an impact on the interaction between these countries, in particular between the Nordic states and the Baltic states.

The challenges for multilingual text and speech processing are enormous. The Northern European area comprises about 31.5 million people (not including Northern Russia) speaking 8 official languages (9 including Russian) and several minority languages. Roughly 45% of all EU inhabitants are able to converse in at least two different languages. The language that is most often the common denominator in Europe, also in Northern Europe, is English, although in the Baltic states and Russia, English language competency is lower, for historical reasons. From a political viewpoint, it is not acceptable to promote English at the expense of the languages of the region.

Most of the information on the Internet consists of natural language, of which less than 50% is currently in English, compared to nearly 100% just ten years ago. Consequently, there is a tremendous and increasing need for language processing tools that make this information accessible to users of different languages. Among these linguistic tools we mention information search and retrieval; filtering, indexing and classification; summarization; translation; text-to-speech and dictation, etc. Recent advances in computational linguistic research are making the development of efficient tools feasible, but it is important to remember that these tools cannot be made independent of the particular language to be treated. English language tools will simply not work for Finnish.

Advanced Nordic language resources are needed

The development of language-specific tools typically requires research on very extensive language resources. These comprise large text and speech collections commonly called *corpora*. Adequately coded and quality-controlled corpora provide the empirical basis for nearly every stage in research and development of language technology products: (i) compiling linguistic requirements and specifications for new systems, (ii) extracting linguistic knowledge in the form of word lists, grammar rules, etc. (iii) repeated testing of research stage prototypes against real language data, and (iv) final evaluation of systems and applications.

Not only must corpora be very large in order to be representative, they must also be carefully encoded and enriched with linguistic descriptions. Given the massive ambiguity in natural language, raw text corpora are of limited use. To take a very simple example, in the Swedish sentence (1) it is impossible, based on word order alone, to determine what is the subject and what is the object of the verb *hittade*. It is therefore essential that corpora are *annotated* so as to reflect the underlying linguistic structure.

(1) *Den första gruppen hittade nästan alla.*
    the   first   group   found   almost   all/everyone
    "The first group found almost everyone" or
    "Almost everyone found the first group"

Annotated corpora are very valuable assets for linguistic research and can be used to answer a multitude of questions concerning linguistic usage and to test hypotheses about linguistic structure. For the developer of language technology, annotated corpora are useful in even more ways. Morphological and syntactic analyzers, which are essential components of translation systems and many other language processing applications, can and should be tested on large annotated corpora. In practice, it is almost impossible to manually construct a system that analyzes an entire language. Besides using corpora for testing, there is also the possibility for automatically inducing linguistic rules and statistical models from corpora, as evidenced by the increasingly successful data-driven approaches to language technology (cf. Manning and Schütze 1999).

However, richly annotated corpora of substantial size are currently lacking for the Nordic languages. Good corpora are just as essential to the linguist as large health databases are to the epidemiologist, or big telescopes to the astronomer. Without corpora, our field of vision is too limited. The situation is especially urgent when we consider syntactically annotated corpora, or *treebanks*.

What is a treebank?

A treebank can be defined as a linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech level. The term 'treebank' appears to have been coined by Geoffrey Leech (Sampson 2003) and obviously alludes to the fact that the most common way of representing the grammatical analysis is by means of a tree structure, as illustrated in Figure 1. However, in current usage, the term is in no way restricted to corpora containing tree-shaped representations, but applies to corpora with all kinds of structural analysis, including even semantic analysis. A related term is 'parsed corpus', which is used more or less interchangeably with 'treebank' in most contexts (cf. Abeillé 2003).
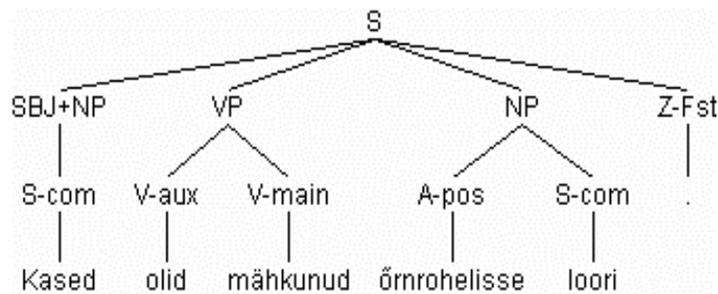

Figure 1 A tree diagram representing the grammatical structure of an Estonian sentence ("The birches were already in pale green leaf.")

Ideally, the design of a treebank should be motivated by its intended usage, whether linguistic research or language technology development. However, in actual practice, there are a number of other factors that influence the design, such as the availability of data and analysis tools. Moreover, given that the development of a treebank is a very labor-intensive task, there is usually also a desire to design the treebank in such a way that it can serve several purposes simultaneously. It is a matter of ongoing debate to what extent it is possible to cater for different needs without compromising the usefulness for each individual use, and different design choices can to some extent be seen to represent different standpoints in this debate. One of the most central design decisions concerns the annotation scheme, i.e. which linguistic categories and representations to use for the grammatical annotation.

Annotation schemes

The choice of annotation scheme for a large-scale treebank is influenced by many different factors. One of the most central considerations is the relation to linguistic theory. Should the annotation scheme be theory-specific, theory-neutral, or even atheoretical? If the first of these options is chosen, which theoretical framework should be adopted? The answers to these questions interact with other factors, in particular the grammatical characteristics of the language that is being analyzed, and the tradition of descriptive grammar that exists for this language. But also the relation to annotation schemes used for other languages is relevant, from the point of view of comparative studies or development of parallel treebank corpora. To this we may add the preferences of different potential user groups, ranging from linguistic researchers and language technology developers to language teachers and students at various levels of education. Finally, when embarking on a large-scale treebank project, researchers usually cannot afford to disregard the resources and tools for automatic and interactive annotation that exist for different candidate annotation schemes.

The number of treebanks available for different languages is growing steadily and with them the number of different annotation schemes. Broadly speaking we may distinguish three main kinds of annotation in current practice:

- Annotation of constituent structure
- Annotation of functional structure
- Annotation of semantic structure

In addition, we can distinguish between (more or less) theory-neutral and theory-specific annotation schemes, a dimension that cuts across the three types of annotation. 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. We will treat the categories in the order in which they are listed above, which also roughly corresponds to the historical development of treebank annotation schemes.

Constituent structure and phrase structure grammar

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. 1992) and the original Penn Treebank (Marcus et al. 1993). 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 2 shows a representative example, taken from the IBM Paris Treebank using a variant of the Lancaster annotation scheme.

```
[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 2 Constituency annotation of a French sentence by means of bracketing in the IBM Paris Treebank (*Vous accédez à cette session à partir de la fenêtre Gestionnaire de tâches.*)

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 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. The drawback of this is that the number of distinct expansions of the same phrase category can become very high. For example, Charniak (1996) was able to extract 10,605 distinct context-free rules from a 300,000 word sample of the Penn Treebank. Of these, only 3,943 occurred more than once in the sample.

Functional annotation and dependency grammar

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 and related theories since Chomsky (1965) has been that grammatical functions are derivable from constituent structure, proponents of dependency syntax such as Mel'čuk (1988) have argued that *functional structure* is more fundamental than constituent structure. Other theories, such as Lexical-Functional Grammar, 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 the annotation schemes based on dependency syntax, exemplified by the Prague Dependency Treebank of Czech (Hajič 1998) and the METU Treebank of Turkish (Oflazer et al. 2000), where the annotation of dependency structure is added directly on top of the morphological annotation without any layer of constituent structure. Figure 3 shows a simple example of dependency annotation from the Prague Dependency Treebank.
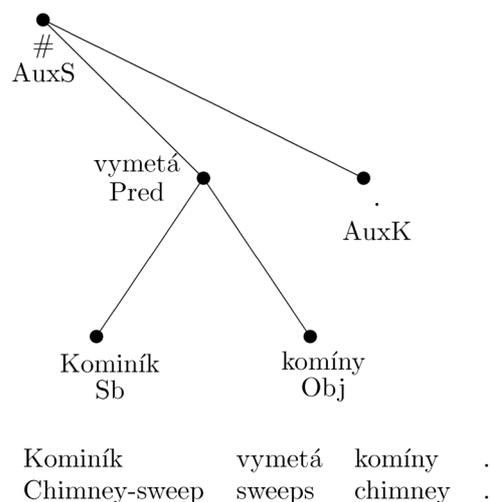

Kominík	vymetá	komíny	.
Chimney-sweep	sweeps	chimney	.

Figure 3 Functional annotation of a Czech sentence 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. A case in point is the Penn Treebank II (Marcus et al. 1994), which adds functional tags to the original phrase structure annotation. Another interesting example in this respect is the annotation scheme adopted in the TIGER Treebank of German (Brants et al. 2002) which integrates the annotation of constituency and dependency in a graph where node labels represent phrasal categories while edge labels represent syntactic functions.

Semantic annotation

From functional annotation, it is only a small step to shallow semantic analysis, such as the annotation of *predicate-argument* structure found in the Proposition Bank (Kingsbury and Palmer 2003). The Proposition Bank is based on the Penn Treebank and adds a layer of annotation, where predicates and their arguments are analyzed in terms of a frame-based lexicon.

Other examples of semantic annotation are the annotation of word senses in a Greek treebank

(Stamou et al. 2003) and the annotation of co-reference relations in the TIGER treebank (Kunz and Hansen-Schirra 2003). Most of the work in this area is still at a pioneering stage, but it can be expected that semantic annotation will be much more prominent in the future. Semantically annotated corpora will open up a world of new research by providing rich empirical material for the study of meaning.

Advances in grammar development are unlocking the potential for constructing treebanks with semantic structures semi-automatically. In the context of the Norwegian LOGON project, the Lexical-Functional Grammar developed in the NorGram project at Bergen has been extended with a semantic projection based on Minimal Recursion Semantics. Consequently, every Norwegian sentence analyzed by the grammar receives a meaning representation in addition to constituent and functional structures. This opens up for the possibility of MRS-banking, i.e. annotating corpora semantically in terms of predicate-argument structures. Moreover, the NorGram grammar was developed parallel to grammars for other languages, in the context of the Parallel Grammar Project (ParGram; Rosén and Zaenen 1999; Butt et al. 2002, Dyvik 2003), which envisages a corpus-based study of grammar and meaning across languages.

Annotations grounded in linguistic theory

Regardless of whether the annotation concerns constituent structure, functional structure or semantic structure, there is a growing interest in annotation schemes that adhere to a specific linguistic theory and use representations from that theory to annotate sentences. Thus, Head-Driven Phrase Structure Grammar (HPSG) has been used as the basis for treebanks of English (Oepen et al. 2002) and Bulgarian (Simov et al. 2003), and the Prague Dependency Treebank is based on the theory of Functional Generative Description (Sgall et al. 1986). CCG-bank is a version of the Penn Treebank annotated within the framework of Combinatory Categorial Grammar (Hockenmaier and Steedman 2002), and there has also been work done on automatic f-structure annotation in the theoretical framework of Lexical-Functional Grammar (see, e.g., Sadler et al. 2000).

In sum, we may say that there has been an overall trend towards more functionally oriented annotation schemes in recent years, and that theory-specific annotation schemes have become more common, but that it is 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.

Tools supporting manual and automatic annotation

A very important methodological issue in the development of treebanks is the division of labor between automatic annotation performed by computational analyzers and human annotation or post-checking. A good example is the methodology for interactive corpus annotation developed by Thorsten Brants and colleagues in the German Negra project, using a cascade of data-driven computational analyzers, which are integrated into the Annotate tool to support efficient interaction with human annotators (Brants and Plaehn 2000). One advantage of using data-driven analyzers is that they can be used to bootstrap the process, since their performance will steadily improve as the size of the treebank grows. Another possibility is the use of ensemble methods, i.e. the combination of several data-driven analyzers in order to improve accuracy (Megyesi 2002).

Exploiting Treebanks

The first treebanks were built as resources for linguistic investigation. This was a time when natural language parsers were still based on hand-crafted rules. But the availability of the Penn Treebank in the early 90s enabled a whole new approach to parsing. Machine Learning algorithms were used to train probabilistic parsers on treebanks. These parsers turned out to be more robust and had the additional advantage of ranking the syntax structures that they found for a given input sentence. Ground work was done by Charniak (1996) and Collins (1996) while Klein and Manning (2001) provide a critical assessment. Nowadays it is widely accepted that wide-coverage parsing—a prerequisite for realistic applications—is impossible without probabilistic components.

Furthermore, the development of translation tools and other multilingual language aids would benefit from the existence of parallel treebanks containing structural annotation at various syntactic and semantic levels, so that structural correspondences between languages could be induced and exploited.

Treebanking activities in Northern Europe

Corpus work has a long and strong tradition in Northern Europe, represented for instance by the well-known Lancaster-Oslo-Bergen (LOB) Corpus for English, begun in 1970 and completed in 1978 with support from the Norwegian Research Council for Science and the Humanities (Johansson et al. 1987). This corpus is now part of the ICAME collection, which has been the basis for an enormous amount of research resulting in well over a thousand scientific articles up to 1998. Another early example of corpus work in the Nordic countries is the Swedish newspaper corpus Press-65, collected at Göteborg University.

Between these pioneering efforts and the year 2000, there has been an increasing production of corpora in Northern Europe. To name but a few examples for the Nordic languages, there is the Stockholm-Umeå Corpus for Swedish, the first version of which was released in 1992, a full version in 1997 (Ejerhed et al. 1992), the Oslo corpus of tagged Norwegian texts, completed in 1999 (Hagen et al. 2000), the Tartu University Corpus of Written Estonian (Hennoste et al. 1998), and the Gothenburg Spoken Language Corpus, a corpus of transcribed spoken Swedish (Allwood et al. 2000).

In view of this long tradition, it comes as no surprise that one of the first treebanks ever was created in Lund by Ulf Teleman and his colleagues in the early 1970's, a treebank comprising some 250,000 words, both written and spoken Swedish (Teleman 1974; Einarsson 1976a, 1976b). This was followed by work at Göteborg University, resulting in an English treebank covering 128,000 words by Ellegård 1978, and by a second treebank of Swedish based on a subset of 100,000 words taken from Press-65 (Järborg 1986). However, after these early treebank projects, which were very impressive at the time but rather small by modern standards, and which suffered to some extent from the lack of appropriate tools for treebank editing, very little has happened on the Swedish scene. (For a discussion of early Swedish annotation schemes, see Nivre 2002.)

Since the turn of the millennium, progress has instead been made in Denmark, where two Danish treebanks have been established, one compiled by Eckhard Bick as part of the VISL project, using a hybrid annotation scheme that combines word-based shallow dependency tags with constituent trees (Bick 2003), and one developed by Matthias Trautner Kromann, using dependency-based annotation (Kromann 2003). However, no treebank currently exists for any of the Nordic languages that has reached the critical size and quality where it becomes usable for developing realistic applications.

The Nordic Treebank Network

This was the situation at the start of the Nordic Treebank Network, an initiative started in 2003 and financed by the Nordic Council of Ministers through its Language Technology research program. The network currently consists of fifteen research groups in six Nordic countries (Denmark, Estonia, Finland, Iceland, Norway and Sweden) and has as its goal to promote research related to treebanks in the Nordic countries. (More information about the network can be found at its website http://w3.msi.vxu.se/~nivre/research/nt.html.)

Besides organizing workshops and graduate courses, the activities of the network have been focused on two areas. The first is the use of common tools, methods and standards for annotation. Although several of the groups in the network have developed their own tools and standards, the network has decided to adopt the TIGER-XML format as the basic interchange format, and has worked out a number of recommendations for the use of TIGER-XML. For example, although TIGER-XML was originally designed to represent constituent trees, recommendations now exist for representing dependency trees in this format. One of the advantages of adopting TIGER-XML

as an interchange format is that it is supported by TIGER-Search, a powerful treebank search tool that imports grammatical structures from various formats, makes them searchable and displays them graphically (König et al. 2002).

A second decision by the Nordic Treebank Network was to build a small parallel treebank as a testing ground and showcase. In order to be able to compare work on different languages, it was decided to build a parallel corpus consisting of the first two chapters of the novel *Sofies verden* (Sophie's World; Gaarder 1991), originally published in Norwegian and translated into all Nordic and many other languages. After considerable negotiations, permission was obtained to use the text in all translations that were needed within the network, and member groups are now working to annotate the same text in different languages, using their own tools and grammatical theories. The separate treebanks are collected and aligned at the sentence level by the Text Laboratory at the University of Oslo. The Text Laboratory also maintains the parallel corpus and provides access for research purposes on request. Currently, the parallel treebank contains data in seven different languages (Danish, Estonian, Faroese, German, Icelandic, Norwegian, Swedish). Figure 1 is taken from this treebank. Some language versions are analyzed with more than one annotation scheme, and several other languages are in the pipeline to be added (notably Dutch, English, Finnish, Greenlandic and Turkish).

The *Sophie treebank* is one of the first parallel treebanks around and, to our knowledge, the biggest one with respect to the number of languages. Building this treebank has created a fruitful common ground for discussions that have already led to important recommendations for annotation standards and has promoted the exchange of information on tools and methods within the network. Moreover, the unique experience of working with a parallel treebank has led to innovative ideas about exploiting translational relations between grammatical structures in different languages (Volk and Samuelsson 2004). This could turn into an important breeding ground for new impulses in machine translation between the Nordic languages. However, the amount of material for each language is still very limited (about 500 sentences when the first two chapters of Sophie's World have been analyzed in their entirety), which is a major obstacle for further development of such methods. It is estimated that a treebank for developing realistic applications needs to be at least two orders of magnitude larger.

Current research in Northern Europe

Despite the absence of treebanks of adequate size, there is nevertheless considerable research going on in the Nordic countries, covering nearly all aspects of treebanking: annotation standards, methods and tools, evaluation, and applications. Below we highlight some research activities found within the Nordic Treebank Network.

At the University of Southern Denmark, constraint grammar parsers are used for treebank annotation. Treebanks are instrumental in applications including e-learning, where they provide a selection of grammar learning exercises.

The Danish Dependency treebank has been established at the Copenhagen Business School, which is now embarking on a parallel treebank for English-Danish, aimed at machine translation applications.

A data-driven dependency parser has been built at Växjö University, and the old Swedish treebank from Lund has been recycled in order to train and test the parser for Swedish (Nivre et al. 2004). Parsers have also been developed for English (using the Penn Treebank; Nivre and Scholz 2004), Danish (using the Danish Dependency Treebank), and Czech (using the Prague Dependency Treebank).

At Stockholm University, research has focused on advanced annotation methods for treebanking, including reuse of tools for one language (in this case a German chunker) for the annotation of another language (Swedish constituent trees), treebank transformations (deepening flat tree structures by automatically adding new nodes), and visualization of parallel treebanks (Volk and Samuelsson 2004).

The University of Bergen is just starting up TREPIL, a pilot project aimed at the semi-automatic construction of a Norwegian treebank. The project will explore a variety of annotation levels including a semantic level and will reuse a large-coverage Norwegian grammar developed in the above-mentioned NorGram project.

The Text Laboratory at Oslo University maintains a server for parallel treebanks and is further developing the search possibilities via a web interface (Johannessen and Nygaard 2004).

The University of Tartu is building a treebank of Estonian, starting from a corpus with shallow syntactic annotation, consisting of 200,000 words, and reusing technology and experience from the VISL project in order to build trees semi-automatically on top of the shallow syntactic annotation (Bick et al. 2004).

Prospects and needs: size matters

The annotation of large knowledge resources has proven a very successful basis for progress in language technology. The Princeton ontology WordNet and the Penn Treebank are two of the most prominent examples. The Stockholm-Umeå Corpus may serve as an example of a widely used resource of a Nordic language. Such resources have become important for measuring the capability of language technology systems, thus enabling objective comparisons. But they also serve as training material for the induction of linguistic knowledge. New methods using data mining and machine learning approaches on large-scale corpora have proven to lead to robust natural language understanding systems.

In this context, it is crucial to realize that size matters. For building realistic systems, it is of limited use to train and test language processing systems on small treebanks. A treebank only pays off when its size reaches millions of words. For English, such large-scale corpora, including treebanks, currently exist. None of the Nordic languages has sufficiently large treebanks today, nor are there current projects that will achieve sufficiently large treebanks in the near future. Therefore there is an immediate need to build full-scale treebanks for the Nordic languages.

The Nordic Treebank Network has been instrumental in promoting research on advanced treebanking methods. However, the network has not been given the financial resources to start building large-scale treebanks, even if the competence to do so is now present. The financial preconditions for large-scale treebanking projects are considerable. The construction of good, useful treebanks cannot be expected to fit within a typical linguistic research project budget. It is more suitable to think of a treebank as a large national or international research facility, requiring an investment of the same order as building an advanced medical, astronomical or nuclear physics research facility.

Based on the experience of the Nordic Treebank Network we want to promote multilingual language technology and work with parallel treebanks. One advantage is that the experience in working with one language can speed up the work for the other languages. At the same time consistency and quality of annotation can be assured over a range of languages rather than individual languages. Finally, having a parallel treebank is enormously more valuable than having isolated monolingual treebanks, due to the fact that translational correspondences of structural information can be exploited for new applications.

Extrapolating from the above-mentioned trends we conjecture that future treebanks will have deeper linguistic annotation than past and current treebanks, including a richer semantic annotation, possibly with links to ontologies for word sense disambiguation. In addition, we will probably see the emergence of rich annotation schemes with built-in conversions between different theory-specific annotation schemes (Nivre 2003). Projects that are going in these directions have already been launched for other languages, and it is important that the Nordic languages follow soon, so that they do not lose their cutting edge profile in corpus research and the full potential for multilingual text and processing can be realized for the Nordic languages.

R e f e r e n c e s

Abeillé, A. (ed.) 2003. *Treebanks: Building and Using Parsed Corpora*. Kluwer.

Allwood, J., Björnberg, M. and Grönqvist, L. 2000. The spoken language corpus at the Linguistics Department, Göteborg University. *Forum Qualitative Social Research*, Vol. 1.

Bick, E. 2003. Arboretum, a hybrid treebank for Danish. In Nivre, J. and Hinrichs, E. (eds.) *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö University Press, pp. 9–20.

Bick, E., Uibo, H. and Müürisep, K. 2004. Arborest – a VISL style treebank derived from an Estonian constraint grammar corpus. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany.

Brants, T. and Plaehn, O. 2000. Interactive Corpus Annotation. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S. and Stainhaouer, G. (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Brants, S., Dipper, S., Hansen, S., Lezius, W. and Smith, G. 2002. The TIGER Treebank. In Hinrichs, E. and Simov, K. (eds.) *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, Sozopol, pp. 24–42.

Butt, M., Dyvik, H., King, T., Masuichi, H., and Rohrer, C. 2002. The Parallel Grammar Project. *Proceedings of COLING 2002 Workshop on Grammar Engineering and Evaluation*.

Charniak, E. 1996. Tree-Bank Grammars. In AAAI/IAAI, Vol. 2, pp. 1031–1036.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Collins, M. 1996. A new statistical parser based on bigram lexical dependencies. In *Proceedings of the Annual Meeting of the ACL*. pp. 184-191.

Dyvik, H. 2003. ParGram: Developing Parallel Grammars. Feature article, *Elsnews* 12.2, pp. 12-14.

Ejerhed, E., Källgren, G., Wennstedt, O. and Åström, M. 1992. The linguistic annotation system of the Stockholm-Umeå corpus project. Umeå universitet, Institutionen för lingvistik, Rapport No. 33.

Einarsson, J. 1976a. Talbankens skriftspråkskonkordans. Lunds universitet: Institutionen för nordiska språk.

Einarsson, J. 1976b. Talbankens talspråkskonkordans. Lunds universitet: Institutionen för nordiska språk.

Gaarder, J. 1991. *Sofies Verden. Roman om filosofiens historie*. Aschehoug & Co.

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

Hagen, K., Johannessen, J. B. and Nøklestad, A. 2000. A web-based advanced and user friendly system: The Oslo corpus of tagged Norwegian texts. In Gavrilidou, M., Carayannis, G., Markantonatou, S., Piperidis, S. and Stainhaouer, G. (eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Hajič, J. 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning*, Karolinum, pp. 106–132.

Hennoste, T., Koit, M., Roosmaa, T. and Saluveer, M. 1998. Structure and Usage of the Tartu University Corpus of Written Estonian. *International Journal of Corpus Linguistics*. Amsterdam, John Benjamins. 3(2), pp. 1-26.

Hockenmaier, J. and Steedman, M. 2002. Acquiring Compact Lexicalized Grammars from a Cleaner Treebank. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pp. 1974–1981.

Johannessen, J. B. and Nygaard, L. 2004. A user-friendly treebank search interface. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*, Tübingen, Germany.

Johansson, S., Leech, G. and Goodluck, H. 1978. Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers. http://helmer.aksis.uib.no/icame/lob/lob-dir.htm.

Järborg, J. 1986. Manual för syntaggning. Göteborgs universitet: Institutionen för språkvetenskaplig databehandling.

Kingsbury, P. and Palmer, M. 2003. PropBank: the next level of TreeBank. In Nivre, J. and Hinrichs, E. (eds.) *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö University Press, pp. 105–116.

Klein, D. and Manning, C.D. 2001. Parsing with Treebank Grammars: Empirical Bounds, Theoretical Models, and the Structure of the Penn Treebank. In *Proceedings of the Annual Meeting of the ACL*.

König, E., Lezius, W., and Voormann, H. 2003. TIGERSearch User's Manual. IMS, University of Stuttgart, Stuttgart.

Kromann, M. T. 2003. The Danish Dependency Treebank and the DTAG treebank tool. In Nivre, J. and Hinrichs, E. (eds.) *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö University Press, pp. 217–220.

Kunz, K. and Hansen-Schirra, S. 2003. Coreference annotation of the TIGER treebank. In Nivre, J. and Hinrichs, E. (eds.) *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö University Press, pp. 221–224.

Manning, C. D. And Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

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

Marcus, M. P., Kim, G., Marcinkiewics, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K. and Schasberger, B. 1994. The Penn Trebank: Annotating Predicate Argument Structure. In ARPA Human Language Technology Workshop.

Megyesi, B. 2002. Data-Driven Syntactic Analysis. Doctoral dissertation, KTH.

Mel'čuk, I. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.

Nivre, J. 2002. What kinds of trees grow in Swedish soil? A comparison of four annotation schemes for Swedish. In Hinrichs, E. and Simov, K. (eds.) *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, Sozopol, pp. 123–138.

Nivre, J. 2003. Theory-supporting treebanks. In Nivre, J. and Hinrichs, E. (eds.) *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö University Press, pp. 117–128.

Nivre, J., Hall, J. and Nilsson, J. 2004. Memory-based dependency parsing. In Ng, H. T. and Riloff, E. (eds.) *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*, Boston, Massachusetts, pp. 49–56.

Nivre, J. and Scholz, M. 2004. Deterministic dependency parsing of English text. In *Proceedings of COLING 2004*, Geneva, Switzerland, pp. 64–70.

Oepen, S., Flickinger, D., Toutanova, K. and Manning, C. D. 2002. LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. In Hinrichs, E. and Simov, K. (eds.) *Proceedings of the First Workshop on Treebanks and Linguistic Theories*, Sozopol, pp. 139–149.

Oflazer, K., Say, B. and Hakkani Tur, D. 2000. A syntactic annotation scheme for Turkish. *In Proceedings of the 10th International Conference on Turkish Linguistics* (ICTL-2000).

Rosén, V. and Zaenen, A. 1999. Grammar Writing in LFG. In Butt, M. and King, T. H. (eds.) *Proceedings of the Fifth International Conference on Lexical-Functional Grammar*. CSLI Publications.

Sadler, L., van Genabith, J., Way, A. 2000. Automatic f-structure annotation from the AP treebank. In Butt, M. and King, T. H. (eds.) *Proceedings of the Fifth International Conference on Lexical-Functional Grammar*. CSLI Publications.

Sampson, G. 2003. Thoughts on two decades of drawing trees. In Abeillé, A. (ed.) *Treebanks: Building and Using Parsed Corpora*, Kluwer, 23–41.

Sgall, P., Hajičová, E. and Panevová, J. 1986. *The Meaning of the Sentence in Its Pragmatic Aspects*. Reidel.

Simov, K., Popova, G., Osenova, P. 2003. HPSG-based syntactic treebank of Bulgarian (BulTreeBank). In Wilson, A., Rayson, P. and McEnery, T. (eds.) *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*, Lincom-Europa.

Stamou, S., Andrikopoulos, V. and Christodoulakis, D. 2003. Towards developing a semantically annotated treebank corpus for Greek. In Nivre, J. and Hinrichs, E. (eds.) *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*, Växjö University Press, pp. 225–228.

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

Volk, M. and Samuelsson, Y. 2004. Bootstrapping parallel treebanks. In *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora (COLING 2004)*, pp. 63–70.