13. Treebanks

1. Introduction

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. However, in current usage, the term is in no way restricted to corpora containing tree-shaped representations, but applies to all kinds of grammatically analyzed corpora.

It is customary to restrict the application of the term 'treebank' to corpora where the grammatical analysis is the result of manual annotation or post-editing. This is in contrast to the term 'parsed corpus', which is more often used about automatically analyzed corpora, whether the analysis has been manually corrected or not. This is also the usage that will be adopted here, although it is worth pointing out that the two terms are sometimes used interchangeably in the literature (cf. Abeillé 2003b).

Treebanks have been around in some shape or form at least since the 1970's. One of the earliest efforts to produce a syntactically annotated corpus was performed by Ulf Teleman and colleagues at Lund University, resulting in close to 300,000 words of both written and spoken Swedish, manually annotated with both phrase structure and grammatical functions, an impressive achievement at the time but unfortunately documented only in Swedish (cf. Teleman 1974; Nivre 2002). However, it is only in the last ten to fifteen years that treebanks have appeared on a large scale for a wide range of languages, mostly developed using a combination of automatic processing and manual annotation or post-editing. In this article, we will not attempt to give a comprehensive inventory of available treebanks but focus on theoretical and methodological issues, referring to specific treebanks only to exemplify the points made. A fairly representative overview of available treebanks for a number of languages can be found in Abeillé (2003a), together with a discussion of certain methodological issues. In addition, proceedings from the annual workshops on *Treebanks and Linguistic Theories* (TLT) contain many useful references (Hinrichs/Simov 2002; Nivre/Hinrichs 2003; Kübler et al. 2004). Cf. also article 20 for some of the more well-known and influential treebanks.

The rest of this article is structured as follows. We begin, in section 2, by discussing design issues for treebanks, in particular the choice of annotation scheme. We move on, in section 3, to the development of treebanks, discussing the division of labor between manual and automatic analysis, as well as tools to be used in the development process.

In section 4, we briefly discuss the usage of treebanks, focusing on linguistic research and natural language processing. We conclude, in section 5, with a brief outlook on the future.

2. Treebank design

Ideally, the design of a treebank should be motivated by its intended usage, whether linguistic research or language technology development (cf. section 4 below), in the same way that any software design should be informed by a requirements analysis (cf. article 9 on design strategies). 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. Thus, as observed by Abeillé (2003b), the majority of large treebank projects have emerged as the result of a convergence between computational linguistics and corpus linguistics, with only partly overlapping goals. It is still 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. We will return to this problem in relation to annotation schemes in section 2.2. But first we will consider the choice of corpus material.

2.1. Corpus material

The considerations involved in selecting the data to include in a treebank are essentially the same as for any (annotated) corpus (cf. article 9). Therefore, we will limit the discussion here to a few observations concerning current practice.

One basic design choice is whether to include written or spoken language, or both, in the treebank. For linguistic corpora in general, written language is much more widely represented than spoken language, and this tendency is even stronger with respect to treebanks, partly because theories of syntactic representation have focused more on written language data, which makes the grammatical annotation of spoken language an even more challenging task. Nevertheless, there now exist quite a few treebanks involving spoken language data, especially for English, such as the CHRISTINE Corpus (Sampson 2003), the Switchboard section of the Penn Treebank (Taylor et al. 2003), and the better part of the ICE-GB Corpus (Nelson et al. 2002). In addition, we have the Tübingen Treebanks of spoken German, English and Japanese (Hinrichs et al. 2000), and the Spoken Dutch Corpus (CGN) (Wouden et al. 2002). It can be expected that the number of spoken language treebanks will increase considerably in the future.

Another basic consideration that any corpus project has to face is whether to construct a balanced sample of different text genres (whether written or spoken) or to concentrate on a specific text type or domain. Historically speaking, treebanks have often been based on previously established corpora, which means that they inherit the design choices of the original corpus. Thus, the SUSANNE Corpus (Sampson 1995) is based

on a subset of the Brown Corpus of American English (Kučera/Francis 1967), which is a typical balanced corpus. By and large, however, the majority of available treebanks for written language are based on contemporary newspaper text, which has the practical advantage of being relatively easily accessible. An important case in point is the Wall Street Journal section of the Penn Treebank (Marcus et al. 1993), which has been very influential as a model for treebanks across a wide range of languages.

Although most treebanks developed so far have been based on more or less contemporary data from a single language, there are also exceptions to this pattern. On the one hand, there are historical treebanks, based on data from earlier periods of a language under development, such as the Penn-Helsinki Parsed Corpus of Middle English (Kroch/Taylor 2000) and the Partially Parsed Corpus of Medieval Portuguese (Rocio et al. 2003). On the other hand, there are parallel treebanks based on texts in one language and their translations in other languages. The Czech-English Penn Treebank has been developed for the specific purpose of machine translation at Charles University in Prague (Čmejrek et al. 2004), and several other projects are emerging in this area (cf. Cyrus et al. 2003; Volk/Samuelsson 2004).

Finally, we have to consider the issue of corpus size. Despite recent advances in automating the annotation process, linguistic annotation is still a very labor-intensive activity. Consequently, there is an inevitable tradeoff in corpus design between the amount of data that can be included and the amount of annotation that can be applied to the data. Depending on the intended usage, it may be preferable to build a smaller treebank with a more detailed annotation, such as the SUSANNE corpus (Sampson 1995), or a larger treebank with a less detailed annotation, such as the original bracketed version of the Penn Treebank (Marcus et al. 1993). Because the annotation of grammatical structure is even more expensive than annotation at lower levels, treebanks in general tend to be one or two orders of magnitude smaller than corresponding corpora without syntactic annotation. Thus, whereas an ordinary corpus of one million running words is not considered very big today, there are only a few treebanks that reach this size, and most of them are considerably smaller.

2.2. Annotation scheme

When discussing the annotation format for a treebank, there are at least two different levels that need to be distinguished. On the one hand, we have the level of linguistic analysis, with certain assumptions about the nature of syntactic structure, a specific choice of linguistic categories, and guidelines for the annotation of particular linguistic phenomena. This level, which is what is normally referred to as an *annotation scheme*, is the level that concerns us in this section (although the discussion of guidelines will be postponed until section 3.1.). On the other hand, we have the level of formal representation, or *encoding*, which is where we decide whether the annotation should be represented using a special markup language or ordinary text, whether it should be stored in one file or several files, etc. The encoding of syntactic annotation will be discussed briefly in section 3.2., and for the time being we will assume that annotation schemes are independent of encoding schemes, although this is strictly speaking not true. (For a general discussion of annotation schemes and standards, cf. article 22.)

Most treebank annotation schemes are organized into a number of layers, where the lower layers contain word-level annotations, such as part-of-speech, often supplemented with morpho-syntactic features, lemmatization or morphological analysis. Figure 13.1 shows a representative example taken from the SUSANNE Corpus (Sampson 1995), where each token is represented by one line, with part-of-speech (including morpho-syntactic features) in the first column, the actual token in the second column, and the lemma in the third column.

AT	The	the
JJ	grand	grand
NN1c	jury	jury
VVDv	took	take
AT1	a	a
NN1c	swipe	swipe
II	at	at
AT	the	the
NNL1n	State	state
NN1u	Welfare	welfare
NNJ1c	Department	department
GG	+<apos>s	–
VVGt	handling	handle
IO	of	of
JJ	federal	federal
NN2	funds	fund
YG	–	–
VVNt	granted	grant
IF	for	for
NN1c	child	child
NN1u	welfare	welfare
NN2	services	service
II	in	in
VV0t	foster	foster
NN2	homes	home
YF	+.	–

Fig. 13.1: Word-level annotation in the SUSANNE Corpus

In the following, we will not discuss word-level annotation but concentrate on the annotation of syntactic (and to some extent semantic) structure, since this is what distinguishes treebanks from other annotated corpora. Moreover, word-level annotation tends to be rather similar across different treebank 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 or theory-neutral? If the first of these options is chosen, which theoretical framework should be adopted? If we opt for the second, how do we achieve broad consensus, given that truly theory-neutral annotation is impossible? 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. In addition, the relation to annotation schemes used for other languages is relevant, from the point of view of comparative studies or development of parallel treebanks. 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 can distinguish three main kinds of annotation in current practice:

- Constituency annotation
- Functional annotation
- Semantic annotation

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 noted 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.

The annotation of constituent structure, often referred to as bracketing, is the main kind of annotation found in early large-scale 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 13.2 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]
```

Fig. 13.2: Constituency annotation in the IBM Paris Treebank

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 3943 occurred more than once in the sample.

A variation on the basic constituency analysis is to annotate syntactic chunks (Abney 1991) rather than a complete phrase structure tree. This kind of annotation is found in the French treebanks described in Abeillé et al. (2003) and Vilnat et al. (2003), respectively. As a further variation, the Tübingen Treebanks of German introduces a layer of topological fields on top of the basic constituent structure (Hinrichs et al. 2000).

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; Böhmová et al. 2003), where the annotation of dependency structure is added directly on top of the morphological annotation without any layer of constituent structure, as illustrated in Figure 13.3.

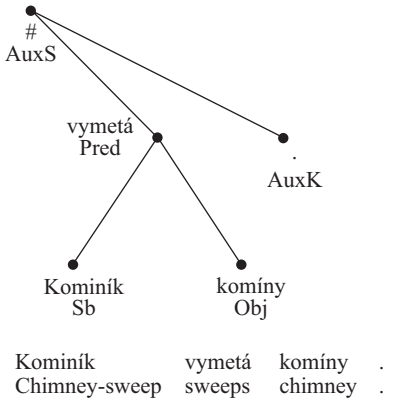

Fig. 13.3: Functional annotation in the Prague Dependency Treebank

Other examples of treebanks based primarily on dependency analysis is the METU Treebank of Turkish (Oflazer et al. 2003), the Danish Dependency Treebank (Kromann 2003), the Eus3LB Corpus of Basque (Aduriz et al. 2003), the Turin University Treebank of Italian (Bosco/Lombardo 2004), and the parsed corpus of Japanese described in Kurohashi/Nagao (2003).

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 1995), which is a development of the Lancaster annotation scheme mentioned above, and Penn Treebank II (Marcus et al. 1994), which adds functional tags to the original phrase structure annotation. A combination of constituent structure and grammatical functions along these lines is currently the dominant paradigm in treebank annotation and exists in many different variations. Adapted versions of the Penn Treebank II scheme are found in the Penn Chinese Treebank (Xue et al. 2004), in the Penn Korean Treebank (Han et al. 2002) and in the Penn Arabic Treebank (Maamouri/Bies 2004), as well as in a treebank of Spanish (Moreno et al. 2003). A similar combination of constituency and grammatical functions is also used in the ICE-GB Corpus of British English (Nelson et al. 2002).

A different way of combining constituency and functional annotation is represented by the TIGER annotation scheme for German (Brants et al. 2002), developed from the earlier NEGRA scheme, which integrates the annotation of constituency and dependency in a graph where node labels represent phrasal categories while edge labels represent syntactic functions, and which allows crossing branches in order to model discontinuous constituents. Another scheme that combines constituent structure with functional annotation while allowing discontinuous constituents is the VISL (Visual Interactive Syntax Learning) scheme, originally developed for pedagogical purposes and applied to 22 languages on a small scale, subsequently used in developing larger treebanks in Portuguese (Afonso et al. 2002) and Danish (Bick 2003). Yet another variation is found in the Italian Syntactic-Semantic Treebank (Montemagni et al. 2003), which employs two independent layers of annotation, one for constituent structure, one for dependency structure.

From functional annotation, it is only a small step to a shallow semantic analysis, such as the annotation of predicate-argument structure found in the Proposition Bank (Kingsbury/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. The Prague Dependency Treebank, in addition to the surface-oriented dependency structure exemplified in Figure 13.3, also provides a layer of tectogrammatical analysis involving case roles, which can be described as a semantically oriented deep syntactic analysis (cf. Hajičová 1998). The Turin University Treebank also adds annotation of semantic roles to the dependency-based annotation of grammatical functions (Bosco/Lombardo 2004), and the Sinica treebank of Chinese uses a combination of constituent structure and functional annotation involving semantic roles (Chen et al. 2003).

Other examples of semantic annotation are the annotation of word senses in the Italian Syntactic-Semantic Treebank (Montemagni et al. 2003) and in the Hellenic National Treebank of Greek (Stamou et al. 2003). Discourse semantic phenomena are annotated in the RST Discourse Treebank (Carlson et al. 2002), the German TIGER Treebank (Kunz/Hansen-Schirra 2003), and the Penn Discourse Treebank (Miltsakaki et al. 2004). Despite these examples, semantic annotation has so far played a rather marginal role in the development of treebanks, but it can be expected to become much more important in the future.

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. 2002), 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/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., Cahill et al. 2002).

Whereas theory-neutral annotation caters for a larger group of users, it runs the risk of not being informative enough or containing too many compromises to be useful for special applications. On the other hand, theory-specific treebanks are clearly more useful for people working within the selected theoretical framework but naturally have a more restricted user group. Recently, there have been attempts at combining the best of both worlds and maximize overall utility in the research community through the use of rich annotation schemes with well-defined conversions to more specific schemes (Nivre 2003; Sasaki et al. 2003). In addition to minimizing the effort required to produce a set of theory-specific treebanks based on the same language data, such a scheme has the advantage of allowing systematic comparisons between different frameworks.

The discussion throughout this section has been focused on the annotation of written language data, as exemplified in the majority of available treebanks across the world. The annotation of spoken language data poses special difficulties that call for an extension of existing annotation schemes. One example is the annotation of so-called disfluencies, which is included in the Switchboard section of the Penn Treebank (cf. Taylor et al. 2003). But more generally, it remains an open question to what extent the annotation schemes developed for written language are adequate for the annotation of spoken language, where interactively defined notions such as turns or dialogue acts may be more central than the syntactic notion of sentence inherited from traditional syntactic theory. Nevertheless, the currently available treebanks of spoken language are all annotated using relatively minor adaptations of schemes originally developed for written language.

3. Treebank development

The methods and tools for treebank development have evolved considerably from the very first treebank projects, where all annotation was done manually, to the present-day situation, which is characterized by a more or less elaborate combination of manual work and automatic processing, supported by emerging standards and customized software tools. In section 3.1., we will discuss basic methodological issues in treebank development, including the division of labor between manual work and automatic processing. In section 3.2., we will then give a brief overview of available tools and standards in the area. We will focus on the process of syntactic annotation, since this is what distinguishes treebank development from corpus development in general.

3.1. Methodology

One of the most important considerations in the annotation of a treebank is to ensure consistency, i. e. to ensure that the same (or similar) linguistic phenomena are annotated

in the same (or similar) ways throughout the corpus, since this is a critical requirement in many applications of treebanks, be it frequency-based linguistic studies, parser evaluation or induction of grammars (cf. section 4 below). This in turn requires explicit and well- documented annotation guidelines, which can be used in the training of human annotators, but which can also serve as a source of information for future users of the treebank. Besides documenting the general principles of annotation, including the annotation scheme as described in section 2.2., the guidelines need to contain detailed examples of linguistic phenomena and their correct annotation. Among linguistic phenomena that are problematic for any annotation scheme, we can mention coordination structures, discontinuous constituents, and different kinds of multi-word expressions. The need to have a rich inventory of examples means that the annotation guidelines for a large treebank project will usually amount to several hundred pages (cf. Sampson 2003).

Another important methodological issue in treebank development is the division of labor between automatic annotation performed by computational analyzers and human annotation or post-editing. Human annotation was the only feasible solution in early treebank projects, such as Teleman (1974) and Järborg (1986) for Swedish, but has the drawback of being labor-intensive and therefore expensive for large volumes of data. In addition, there is the problem of ensuring consistency across annotators if several people are involved. Fully automatic annotation has the advantage of being both inexpensive and consistent but currently cannot be used without introducing a considerable proportion of errors, which typically increases with the complexity of the annotation scheme. Hence, fully automatic annotation is the preferred choice only when the amount of data to be annotated makes manual annotation or post-editing prohibitively expensive, as in the 200 million word corpus of the Bank of English (Järvinen 2003). In addition, fully automatic analysis of a larger section of the treebank can be combined with manual post-correction for smaller sections, as in the Danish Arboretum, which contains a "botanical garden" of a few hundred thousand words completely corrected, a "forest" of one million words partially corrected, and a "jungle" of nine million words with automatic analysis only (Bick 2003).

Given the complementary advantages and drawbacks of human and automated annotation, most treebank projects today use a combination of automatic analysis and manual work in order to make the process as efficient as possible while maintaining the highest possible accuracy. The traditional way of combining automated and manual processing is to perform syntactic parsing (complete or partial) followed by human post-editing to correct errors in the parser output. This methodology was used, for example, in the development of the Penn Treebank (Taylor et al. 2003) and the Prague Dependency Treebank (Böhmová et al. 2003). One variation on this theme is to use human disambiguation instead of human post-correction, i. e. to let the human annotator choose the correct analysis from a set of possible analyses produced by a nondeterministic parser. This approach is used in the TreeBanker (Carter 1997) and in the development of the LinGO Redwood Treebanks (Oepen et al. 2002).

Regardless of the exact implementation of this methodology, human post-editing or parse selection runs the risk of biasing the annotation towards the output of the automatic analyzer, since human editors have a tendency of accepting the proposed analysis even in doubtful cases. The desire to reduce this risk was one of the motivating factors behind the methodology for interactive corpus annotation developed by Thorsten Brants and colleagues in the German NEGRA project (Brants et al. 2003), which uses a cascade

of data-driven computational analyzers and gives the human annotator the opportunity to correct the output of one analyzer before it is fed as input to the next (Brants/Plaehn 2000). Data-driven analyzers also have an advantage in that they can be used to bootstrap the process, since their performance will steadily improve as the size of the treebank grows.

Another issue to consider is the order in which data are fed to human annotators for post-correction. Wallis (2003) argues that transverse correction, i. e. checking all instances of a particular construction together, can improve the consistency of the annotation, as compared to traditional longitudinal correction (sentence-by-sentence). On the other hand, transverse correction is harder to implement and manage. A related issue is the order in which different layers of a multi-layered annotation scheme should be processed and whether different layers should be annotated together or separately. In many cases, there are dependencies between layers that dictate a particular order, but it may also be possible to annotate layers in parallel (cf. Taylor et al. 2003). Whether the work is done in sequence or in parallel, it is usually considered best to let each annotator work with a single layer at a time.

Finally, it is worth mentioning that consistency in treebank annotation can be improved by letting several people annotate or correct the same sentences and compare their work. However, this procedure is very expensive and can therefore normally be used only for a small subpart of the treebank, often with the specific purpose of investigating inter-annotator agreement. A less expensive method is to use automated analysis to detect potential errors or inconsistencies in the annotation, as proposed by Dickinson/Meurers (2003) and Ule/Simov (2004), among others.

3.2. Tools and standards

Many of the software tools that are used in treebank development are tools that are needed in the development of any annotated corpus, such as tokenizers and part-of-speech taggers (cf. article 24). Tools that are specific to treebank development are primarily tools for syntactic preprocessing (cf. article 28) and specialized annotation tools.

Well-known examples of syntactic parsers used in treebank development are the deterministic Fidditch parser (Hindle 1994), used in the development of the Penn Treebank, and the statistical parser of Collins et al. (1999), used for the Prague Dependency Treebank. It is also common to use partial parsers (or chunkers) for syntactic preprocessing, since partial parsing can be performed with higher accuracy than full parsing.

Breaking down the parsing process into several steps has the advantage that it allows human intervention between each step, as discussed in connection with interactive corpus annotation above. This is one of the motivations behind the Annotate tool (Brants/Plaehn 2000), which is a tool for interactive corpus annotation incorporating a cascade of data-driven analyzers for tagging and chunking. Another annotation tool developed especially for treebank annotation is the graphical editor TrEd, developed in the Prague Dependency Treebank project, but it is also quite common to use more or less sophisticated extensions to existing editors such as Emacs (cf. Taylor et al. 2003; Abeillé et al. 2003).

As a general assessment of the state of the art in treebank development, it seems fair to say that there is a lack of standardized tools and that most projects tend to develop

their own tools suited to their own needs. To some extent this can be explained by the fact that different projects use different annotation schemes, motivated by properties of the particular language analyzed and the purpose of the annotation, and that not all tools are compatible with all annotation schemes (or software platforms). However, it probably also reflects the lack of maturity of the field and the absence of a widely accepted standard for the encoding of treebank annotation. While there have been several initiatives to standardize corpus encoding in general (cf. article 22), these recommendations have either not extended to the level of syntactic annotation or have not gained widespread acceptance in the field. Instead, there exist several *de facto* standards set by the most influential treebank projects, in particular the Penn Treebank, but also the Prague Dependency Treebank for dependency representations. Another popular encoding standard is TIGER-XML (König/Lezius 2003), originally developed within the German TIGER project, which can be used as a general interchange format although it imposes certain restrictions on the form of the annotation.

As observed by Ide/Romary (2003), there is a widely recognized need for a general framework that can accommodate different annotation schemes and facilitate the sharing of resources as well as the development of reusable tools. It is part of the objective of the ISO/TC 37/SC 4 Committee on Language Resource Management to develop such a framework, building on the model presented in Ide/Romary (2003), but at the time of writing there is no definite proposal available.

4. Treebank usage

Empirical linguistic research provided most of the early motivation for developing treebanks, and linguistic research continues to be one of the most important usage areas for parsed corpora. We discuss linguistic research in section 4.1. below. In recent years, however, the use of treebanks in natural language processing, including research as well as technical development, has increased dramatically and has become the primary driving force behind the development of new treebanks. This usage is the topic of section 4.2. (cf. also article 35).

The use of treebanks is not limited to linguistic research and natural language processing, although these have so far been the dominant areas. In particular, there is a great potential for pedagogical uses of treebanks, both in language teaching and in the teaching of linguistic theory. A good example is the Visual Interactive Syntax Learning (VISL) project at the University of Southern Denmark, which has developed teaching treebanks for 22 languages with a number of different teaching tools including interactive games such as Syntris, based on the well-known computer game Tetris (see http://visl.edu.dk).

4.1. Linguistic research

Treebanks are in principle a useful resource for any kind of corpus-based linguistic research that is related to syntax. This includes not only syntactic research in a narrow sense but research on any linguistic phenomenon that is dependent on syntactic proper-

ties. One of the main advantages of using a treebank, rather than an ordinary corpus, is that it enables more precise queries and thereby reduces the noise in the answer set. To take one concrete example, in a recent corpus-based study of English quantifiers, the use of *all* as a so-called floating quantifier (*they all laughed*) had to be excluded from the study, simply because there was no way of constructing the query precisely enough to extract the relevant examples from the much more numerous examples of other uses of *all* (Vannestål 2004). Given a properly annotated treebank, this methodological problem should not arise. However, it is important to remember that an efficient use of treebanks in corpus-based research requires adequate tools for searching and browsing treebanks. We refer to article 34 for a discussion of this topic.

Treebank data, like other corpus data, can be used in a variety of ways in linguistic research. Some of them are qualitative, such as finding an authentic example of a certain linguistic construction, or a counter-example to an empirical claim about syntactic structure, but arguably the most important uses of treebank data are found in quantitative studies of different kinds, where treebanks provide an invaluable source of information about frequencies, cooccurences, etc. For a long time, frequency information has by a majority of linguists been considered as complementary to, but not directly relevant for, theoretical accounts of linguistic structure. However, this is a position that is increasingly called into question, and there are now a number of proposals that incorporate frequency, or probability, into the theoretical description of linguistic categories and rules (see, e. g., Bod et al. 2003). Since corpus-based syntactic research and its relation to syntactic theory is treated in depth in other articles, in particular articles 42 and 43, we will not pursue these issues further here.

4.2. Natural language processing

Broadly speaking, we can distinguish two main uses of treebanks in natural language processing. The first is the use of treebank data in the evaluation of natural language processing systems, in particular syntactic parsers. The second is the induction of linguistic resources from treebanks, especially the use of machine learning to develop or optimize linguistic analysers (cf. article 39).

Empirical evaluation of systems and components for natural language processing is currently a very active field. With respect to syntactic parsing there are essentially two types of data that are used for evaluation. On the one hand, we have so-called test suites, i. e. collections of sentences that are compiled in order to cover a particular range of syntactic phenomena without consideration of their frequency of occurrence (cf. Lehmann et al. 1996). On the other hand, we have treebank samples, which are extracted to be representative with respect to the frequency of different phenomena. Both types of data are clearly relevant for the evaluation of syntactic parsers, but it is also clear that the resulting evaluation will focus on different properties. Test suite evaluation measures the coverage of a syntactic parser in terms of the number of constructions that it can handle, without considering the relative frequency of these constructions. Treebank evaluation, on the other hand, measures the average performance that we can expect from the parser when applied to naturally distributed data from the same source as the evaluation corpus.

An important methodological issue in treebank evaluation is the way in which performance of a parser is measured relative to a manually annotated treebank sample (a so-called *gold standard*). An obvious metric to use is the proportion of sentences where the parser output completely matches the gold standard annotation (the *exact match* criterion). However, it can be argued that this is a relatively crude evaluation metric, since an error in the analysis of a single word or constituent will have the same impact on the result as the failure to produce any analysis whatsoever. Consequently, the most widely used evaluation metrics measure various kinds of partial correspondence between the parser output and the gold standard parse.

The most well-known evaluation metrics are the PARSEVAL measures (Black et al. 1991), which are based on the number of matching constituents between the parser output and the gold standard, and which have been widely used in parser evaluation using data from the Penn Treebank. As an alternative to the constituency-based PARSEVAL measures, several researchers have proposed evaluation schemes based on dependency relations and argued that these provide a better way of comparing parsers that use different representations (Lin 1998; Carroll et al. 1998).

A very successful use of treebanks during the last decade has been the induction of probabilistic grammars for parsing, with lexicalised probabilistic models like those of Collins (1999) and Charniak (2000) representing the current state of the art. An even more radical approach is Data-Oriented Parsing (Bod 1998), which eliminates the traditional notion of grammar completely and uses a probabilistic model defined directly on the treebank. But there has also been great progress in broad-coverage parsing using so-called deep grammars, where treebanks are mainly used to induce statistical models for parse selection (see, e. g., Riezler et al. 2002; Toutanova et al. 2002). In fact, one of the most significant results in research on syntactic parsing during the last decade is arguably the conclusion that treebanks are indispensable in order to achieve robust broad-coverage parsing, regardless of which basic parsing methodology is assumed.

Besides using treebanks to induce grammars or optimize syntactic parsers, it is possible to induce other linguistic resources that are relevant for natural language processing. One important example is the extraction of subcategorization frames (cf. Briscoe/ Carroll 1997). Cf. also articles 35 and 39.

5. Conclusion

Treebanks have already been established as a very valuable resource both in linguistic research and in natural language processing. In the future, we can expect their usefulness to increase even more, with improved methods for treebank development and usage, with more advanced tools built on universal standards, and with new kinds of annotation being added. Treebanks with semantic-pragmatic annotation have only begun to emerge and will play an important role in the development of natural language understanding. Parallel treebanks, which hardly exist at the moment, will provide an invaluable resource for research on translation as well as the development of better methods for machine translation. Spoken language treebanks, although already in existence, will be developed further to increase our understanding of the structure of spoken discourse and lead to enhanced methods in speech technology.

5. Literature

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

Abeillé, A. (2003b), Introduction. In: Abeillé 2003a, xiii−xxvi.

Abeillé, A./Clément, L./Toussenel, F. (2003), Building a Treebank for French. In: Abeillé 2003a, 165−187.

Abney, S. (1991), Parsing by Chunks. In: Berwick, R./Abney, S./Tenny, C. (eds.), *Corpus-based Methods in Language and Speech*. Dordrecht: Kluwer, 257−278.

Aduriz, I./Aranzabe, M. J./Arriola, J. M./Atutxa, A./Díaz de Ilarraza, A./Garmendia, A./Oronoz, M. (2003), Construction of a Basque Dependency Treebank. In: Nivre/Hinrichs 2003, 201−204.

Afonso, S./Bick, E./Haber, R./Santos, D. (2002), Floresta Sintá(c)tica, a Treebank for Portuguese. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 1698−1703.

Bick, E. (2003), Arboretum, a Hybrid Treebank for Danish. In: Nivre/Hinrichs 2003, 9−20.

Black, E./Abney, S./Flickinger, D./Gdaniec, C./Grishman, R./Harrison, P./Hindle, D./Ingria, R./Jelinek, F./Klavans, J./Liberman, M./Marcus, M./Roukos, S./Santorini, B./Strzalkowski, T. (1991), A Procedure for Quantitatively Comparing the Syntactic Coverage of English Grammars. In: *Proceedings of the DARPA Speech and Natural Language Workshop.* Pacific Grove, CA, 306−311.

Bod, R. (1998), *Beyond Grammar.* Chicago: CSLI Publications.

Böhmová, A./Hajič, J./Hajičová, E./Hladká, B. (2003), The PDT: A 3-level Annotation Scenario. In: Abeillé 2003a, 103−127.

Bosco, C./Lombardo, V. (2004), Dependency and Relational Structure in Treebank Annotation. In: *Proceedings of the Workshop Recent Advances in Dependency Grammar*. Geneva, Switzerland, 9−16.

Brants, T./Plaehn, O. (2000), Interactive Corpus Annotation. In: *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens, Greece, 453−459.

Brants, S./Dipper, S./Hansen, S./Lezius, W./Smith, G. (2002), The TIGER Treebank. In: Hinrichs/Simov 2002, 24−42.

Brants, T./Skut, W./Uszkoreit, H. (2003), Syntactic Annotation of a German Newspaper Corpus. In: Abeillé 2003a, 73−87.

Briscoe, E./Carroll, J. (1997), Automatic Extraction of Subcategorization from Corpora. In: *Proceedings of the Fifth Conference on Applied Natural Language Processing.* Washington, DC, 356−363.

Cahill, A./McCarthy, M./Van Genabith, J./Way, A. (2002), Evaluating F-structure Annotation for the Penn-II Treebank. In: Hinrichs/Simov 2002, 43−60.

Carlson, L./Marcu, D./Okurowski, M. E. (2002), *RST Discourse Treebank*. Philadelphia, PA: Linguistic Data Consortium.

Carroll, J./Briscoe, E./Sanfilippo, A. (1998), Parser Evaluation: A Survey and a New Proposal. In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC).* Granada, Spain, 447−454.

Carter, D. (1997), The TreeBanker: A Tool for Supervised Training of Parsed Corpora. In: *Proceedings of the ACL Workshop on Computational Environments for Grammar Development and Linguistic Engineering*. Madrid, Spain, 9−15.

Charniak, E. (1996), Tree-bank Grammars. In: *Proceedings of AAAI/IAAI*, 1031−1036.

Charniak, E. (2000), A Maximum-Entropy-Inspired Parser. In: *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics.* Seattle, WA, 132−139.

Chen, K./Luo, C./Chang, M./Chen, F./Chen, C./Huang, C./Gao, Z. (2003), Sinica Treebank. In: Abeillé 2003a, 231−248.

Chomsky, N. (1965), *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Čmejrek, M./Cuřín, J./Havelka, J./Hajič, J./Kuboň, V. (2004), Prague Czech-English Dependency Treebank: Syntactically Annotated Resources for Machine Translation. In: *Proceedings of the IV International Conference on Language Resources and Evaluation.* Lisbon, Portugal, 1597−1600.

Collins, M. (1999), Head-driven Statistical Models for Natural Language Parsing. PhD Thesis, University of Pennsylvania.

Collins, M./Hajič, J./Brill, E./Ramshaw, L./Tillmann, C. (1999), A Statistical Parser of Czech. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics,* 505−512.

Cyrus, L./Feddes, H./Schumacher, F. (2003), FuSe − A Multi-layered Parallel Treebank. In: Nivre/Hinrichs 2003, 213−216.

Dickinson, M./Meurers, W. D. (2003), Detecting Inconsistencies in Treebanks. In: Nivre/Hinrichs 2003, 45−56.

Garside, R./Leech, G./Varadi, T. (compilers) (1992), Lancaster Parsed Corpus. A Machine-readable Syntactically Analyzed Corpus of 144,000 Words. Available for Distribution through ICAME. Bergen: The Norwegian Computing Centre for the Humanities.

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

Hajičová, E. (1998), Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In: *Proceedings of the First Workshop on Text, Speech, Dialogue*. Brno, Czech Republic, 45−50.

Han, C./Han, N./Ko, S. (2002), Development and Evaluation of a Korean Treebank and its Application to NLP. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 1635−1642.

Hindle, D. (1994), A Parser for Text Corpora. In: Zampolli, A. (ed.), *Computational Approaches to the Lexicon*. New York: Oxford University Press, 103−151.

Hinrichs, E./Simov, K. (eds.) (2002), *Proceedings of the First Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

Hinrichs, E. W./Bartels, J./Kawata, Y./Kordoni, V./Telljohann, H. (2000), The Tübingen Treebanks for Spoken German, English and Japanese. In: Wahlster, W. (ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer, 552−576.

Hockenmaier, J./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, 1974−1981.

Ide, N./Romary, L. (2003), Encoding Syntactic Annotation. In: Abeillé 2003a, 281−296.

Järborg, J. (1986), *Manual för syntaggning*. Göteborg University: Department of Swedish.

Järvinen, T. (2003), Bank of English and Beyond. In: Abeillé 2003a, 43−59.

Kingsbury, P./Palmer, M. (2003), PropBank: The Next Level of TreeBank. In: Nivre/Hinrichs 2003, 105−116.

König, E./Lezius, W. (2003), *The TIGER Language − A Description Language for Syntax Graphs. Formal Definition.* Technical Report, IMS, University of Stuttgart.

Kroch, A./Taylor, A. (2000), Penn-Helsinki Parsed Corpus of Middle English. URL: ⟨http://www.ling.upenn.edu/mideng/.⟩

Kromann, M. T. (2003), The Danish Dependency Treebank and the DTAG Treebank Tool. In: Nivre/Hinrichs 2003, 217−220.

Kübler, S./Nivre, J./Hinrichs, E./Wunsch, H. (eds.) (2004), *Proceedings of the Third Workshop on Treebanks and Linguistic Theories*. Tübingen, Germany.

Kučera, H./Francis, W. (1967), *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Kunz, K./Hansen-Schirra, S. (2003), Coreference annotation of the TIGER treebank. In: Nivre/Hinrichs 2003, 221−224.

Kurohashi, S./Nagao, M. (2003), Building a Japanese Parsed Corpus. In: Abeillé 2003a, 249−260.

Lehmann, S./Oepen, S./Regnier-Prost, S./Netter, K./Lux, V./Klein, J./Falkedal, K./Fouvry, F./Estival, D./Dauphin, E./Compagnion, H./Baur, J./Balkan, L./Arnold, D. (1996), TSNLP − Test Suites for Natural Language Processing. In: *Proceedings of the 16th International Conference on Computational Linguistics*. Copenhagen, Denmark, 711−716.

Lin, D. (1998), A Dependency-based Method for Evaluating Broad-coverage Parser. In: *Journal of Natural Language Engineering* 4, 97−114.

Maamouri, M./Bies, A. (2004), Developing an Arabic Treebank: Methods, Guidelines, Procedures, and Tools. In: *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.* Geneva, Switzerland, 2−9.

Marcus, M. P./Santorini, B./Marcinkiewics, M. A. (1993), Building a Large Annotated Corpus of English: The Penn Treebank. In: *Computational Linguistics* 19, 313−330.

Marcus, M. P./Kim, G./Marcinkiewics, M. A./MacIntyre, R./Bies, A./Ferguson, M./Katz, K./Schasberger, B. (1994), The Penn Treebank: Annotating Predicate Argument Structure. In: *Proceedings of the Human Language Technology Workshop*. Plainsboro, NJ, 114−119.

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

Miltsakaki, E./Prasad, R./Joshi, A./Webber, B. (2004), The Penn Discourse Treebank. In: *Proceedings of the IV International Conference on Language Resources and Evaluation.* Lisbon, Portugal, 2237−2240.

Montemagni, S./Barsotti, F./Battista, M./Calzolari, N./Corazzari, O./Lenci, A./Zampolli, A./Fanciulli, F./Massetani, M./Raffaelli, R./Basili, R./Pazienza, M. T./Saracino, D./Zanzotto, F./Nana, N./Pianesi, F./Delmonte, R. (2003), Building the Italian Syntactic-semantic Treebank. In: Abeillé 2003a, 189−210.

Moreno, A./López, S./Sánchez, F./Grishman, R. (2003), Developing a Spanish Treebank. In: Abeillé 2003a, 149−163.

Nelson, G./Wallis, S./Aarts, B. (2002), *Exploring Natural Language: Working with the British Component of the International Corpus of English.* Amsterdam: John Benjamins.

Nivre, J. (2002), What Kinds of Trees Grow in Swedish Soil? A Comparison of Four Annotation Schemes for Swedish. In: Hinrichs/Simov 2002, 123−138.

Nivre, J. (2003), Theory-supporting treebanks. In: Nivre/Hinrichs 2003, 117−128.

Nivre, J./Hinrichs, E. (eds.) (2003), *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*. Växjö, Sweden: Växjö University Press.

Oepen, S./Flickinger, D./Toutanova, K./Manning, C. D. (2002), LinGO Redwoods: A Rich and Dynamic Treebank for HPSG. In: Hinrichs/Simov 2002, 139−149.

Oflazer, K./Say, B./Hakkani-Tür, D. Z./Tür, G. (2003), Building a Turkish Treebank. In: Abeillé 2003a, 261−277.

Riezler, S./King, M./Kaplan, R./Crouch, R./Maxwell, J. (2002), Parsing the Wall Street Journal Using a Lexical-functional Grammar and Discriminative Estimation Techniques. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 271−278.

Rocio, V./Alves, M. A./Lopes, J. G./Xavier, M. F./Vicenter, G. (2003), Automated Creation of a Medieval Portuguese Treebank. In: Abeillé 2003a, 211−227.

Sampson, G. (1995), *English for the Computer. The SUSANNE Corpus and Analytic Scheme*. Oxford: Oxford University Press.

Sampson, G. (2003), Thoughts on Two Decades of Drawing Trees. In: Abeillé 2003, 23−41.

Sasaki, F./Witt, A./Metzing, D. (2003), Declarations of Relations, Differences and Transformations between Theory-specific Treebanks: A New Methodology. In: Nivre/Hinrichs 2003, 141−152.

Sgall, P./Hajičová, E./Panevová, J. (1986), *The Meaning of the Sentence in Its Pragmatic Aspects*. Dordrecht: Reidel.

Simov, K./Osenova, P./Kolkovska, S./Balabanova, E./Doikoff, D./Ivanova, K./Simov, A./Kouylekov, M. (2002), Building a Linguistically Interpreted Corpus of Bulgarian: The BulTreeBank. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 1729−1736.

Stamou, S./Andrikopoulos, V./Christodoulakis, D. (2003), Towards Developing a Semantically Annotated Treebank Corpus for Greek. In: Nivre/Hinrichs 2003, 225–228.

Taylor, A./Marcus, M./Santorini, B. (2003), The Penn Treebank: An Overview. In: Abeillé 2003a, 5–22.

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

Toutanova, K./Manning, C. D./Shieber, S. M./Flickinger, D./Oepen, S. (2002), Parse Disambiguation for a Rich HPSG Grammar. In: Hinrichs/Simov 2002, 253–263.

Ule, T./Simov, K. (2004), Unexpected Productions May Well Be Errors. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal, 1795–1798.

Vilnat, A./Paroubek, P./Monceaux, L./Robba, I./Gendner, V./Illouz, G./Jardino, M. (2003), EASY or How Difficult Can It Be to Define a Reference Treebank for French. In: Nivre/Hinrichs 2003, 229–232.

Volk, M./Samuelsson, Y. (2004), Bootstrapping Parallel Treebanks. In: *Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora*. Geneva, Switzerland, 63–70.

Wallis, S. (2003), Completing Parsed Corpora. In: Abeillé 2003a, 61–71.

Wouden, T. van der/Hoekstra, H./Moortgat, M./Renmans, B./Schuurman, I. (2002), Syntactic Analysis in the Spoken Dutch Corpus. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas, Spain, 768–773.

Xue, N./Xia, F./Chiou, F.-D./Palmer, M. (2004), The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. In: *Natural Language Engineering* 11, 207–238.

Joakim Nivre, Växjö (Sweden)