SYNTACTIC ANNOTATION OF A GERMAN NEWSPAPER CORPUS

Thorsten Brants **Wojciech Skut** **Hans Uszkoreit***

Résumé - Abstract

We report on the syntactic annotation of a German newspaper corpus. The annotations consists of context-free structures, additionally allowing crossing branches, with labeled nodes (phrases) and edges (grammatical functions). Furthermore, we present a new, interactive semi-automatic annotation process that allows efficient and reliable annotations.

Mots Clefs - Keywords

Corpus Annotation, German Newspaper Corpus, Annotation Tools

INTRODUCTION

Data-oriented and corpus-based methods have become one of the most important areas of applied as well as theoretical NLP. Currently, the methods prevailingly belong to the *supervised learning* paradigm, i.e., they require as training material large corpora annotated with linguistic information. Since the preparation of such corpora usually involves manual human work, a lot of effort is put into the optimisation of the annotation process in order to make it less time-consuming and labour-intensive.

The amount of annotation work depends on how much information is associated with the raw language data. The present paper deals with the construction of a *treebank*, which is particularly labour-intensive. There are two main optimisation issues. Firstly, automatic processing methods can be used, so that the annotator only supervises, corrects and completes the analyses proposed by some tool. Secondly, the speed and consistency of annotation may depend on the annotation scheme employed.

Several annotation procedures are described in the literature. The Penn Treebank (Marcus M. *et al.* 1993) used the Fidditch parser (Hindle D. 1983) as pre-processing. The structures were subsequently presented to human annotators and manually corrected. The *Treebanker* (Carter D. 1997) is used for syntactical annotation in the ATIS (air travel information system) domain. It

*Saarland University, FR 8.7 Computational Linguistics, P.O.Box 171150, D-66041 Saarbrücken, Germany, Email: {brants,skut,uszkoreit}@coli.uni-sb.de

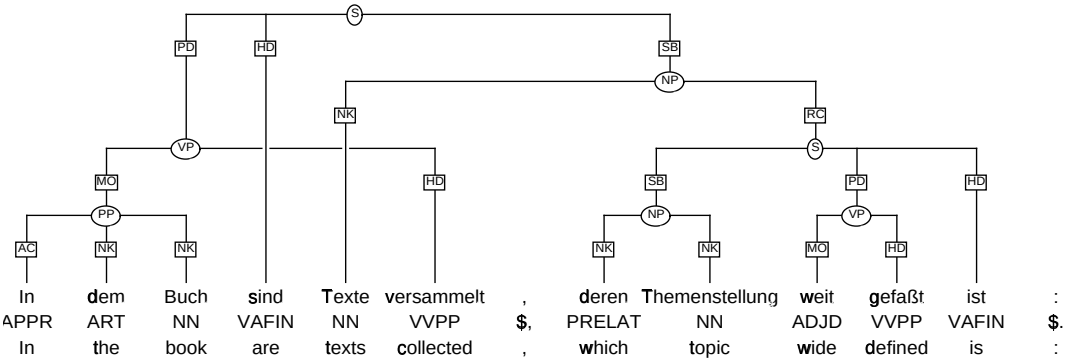

Figure 1: Example sentence (#2412 in the corpus). It contains two crossing branches. One is caused by the PP in the Vorfeld, the other one by the extraposed relative clause. Edges are labeled with grammatical functions.

presents alternative readings in parallel and the annotator selects correct or rejects wrong readings. This mode of annotation requires a parser that that generates the correct complete parse for most of the sentences. (Brants T. & Skut W. 1998) present methods to automate the labeling of annotated structures as well as partial structures for NPs and PPs. Structures and labeling are generated incrementally.

In this paper, we present an interactive annotation mode that suggests new phrases to the annotator who can accept or reject the suggestion. Probabilities of suggested phrases are used to determine if confirmation by the annotator is necessary. It is also shown how to optimise the annotations scheme as far as the consistency and efficiency of annotation are concerned.

1. TREEBANK DEVELOPMENT

The NEGRA treebank has been developed since 1996 in the NEGRA project at the Saarland University in Saarbrücken. To our knowledge, it was the first attempt to build a large-scale corpus of German text at that time. Over the past three years, more than 20,000 sentences have been annotated with syntactic structures, syntactic categories and grammatical functions.

NEGRA was one of the first treebanks for languages other than English. The only documented large-scale treebanks available to us were the Penn Treebank (Marcus M. *et al.* 1993) and the Susanne corpus (Sampson G. 1995), which were both based on context-free syntactic representations with additional trace-filler annotations for discontinuous constituents. This seemed very likely to become a standard in the corpora community, which suggested adopting this kind of annotations for our corpus.

On the other hand, it was clear from the beginning that the experience made during the construction of English corpora would not always be useful in our case. In particular, we expected to encounter problems specific to the

language type represented by German.

In fact, it turned out that the considerable degree of free word order exhibited by German makes context-free annotation of syntactic structures difficult. This is particularly due to the frequency of discontinuous constituents, which can be annotated by trace-filler coreferences, but only at the cost of lower transparency of annotations (the reader may imagine a sentence containing three non-local dependencies, each annotated by a separate co-reference).

These considerations suggested that we should adopt a different representation format, preferably one specially developed for free word order languages. There is one such formalism, the Dependency Grammar, which interprets the natural language in terms of head-modifier relations (*dependencies*) between words. If the words are interpreted as tree nodes and the dependencies as branches connecting the words, the structure of a sentence can be drawn as a tree. In the original, non-projective version of Dependency Grammar, there are no restrictions on attaching words to each other. There may be crossing branches, which makes the formalism particularly suitable for free word order languages. The approach of Dependency Grammar annotations was chosen for the Prague Dependency Treebank (Hajic J. 1998).

We used Dependency Grammar as the starting point for the development of our annotation scheme. It soon turned out that the sharp distinction between heads and modifiers stipulated by the formalism causes difficulties in practice. In particular, all kinds of constructions without a clear syntactic head are difficult to analyse: ellipses, sentences without a verb (e.g., copula-less predicatives), and coordinations.

Linguistic theories may well have solutions to these problems, but such solutions tend to be highly theory-specific, while our annotation effort aims at descriptive and theory-neutral annotation.

As a result, we abandoned the idea of pure dependency-based annotation scheme and adopted a hybrid framework that combines the advantages of phrase-structure and dependency grammars. We do employ phrasal nodes, but try to keep the structures flat: a phrasal node mostly corresponds to exactly one lexical head. The branches of such a tree may cross, so there is no need for a special treetment of non-local dependencies, which occupy the same structural status as local ones. In addition to syntactic categories and structures proper, *grammatical functions* (subject, object, etc.) are annotated more explicitly than, e.g., in the Penn Treebank. This is no surprise since German does not express functional information by position as does English.

The hybrid representation format adopted in NEGRA permits fast and consistent annotation. Consisteny is supported by the following features:

- flat syntactic structures exhibit a low potential for attachment ambiguities;
- trees are mostly quite close to semantics (predicate-argument structures), and the annotators disambiguate sentences on the basis of how they understand their meaning;
- remaining spurious ambiguities are dealt with by “default attachment”;

Step	Type	Description
1.	automatic	part-of-speech tagging
2.	automatic	determine reliable and unreliable tag assignments
3.	manual	confirm unreliable and alter wrong tags
4.	automatic	Suggest new phrase that has highest probability
5.	manual	accept $\rightarrow$ step 7, retry $\rightarrow$ step 4, reject $\rightarrow$ step 6
6.	manual	insert or alter a phrase
7.	automatic	insert labels into new structure
8.	automatic	determine reliable and unrelibale label assignments
9.	manual	confirm unreliable and alter wrong labels
10.	manual	finished $\rightarrow$ end, continue $\rightarrow$ step 4

Figure 2: Processing steps during corpus annotation, which combine automatic processing and manual intervention/manual annotation based on the annotator's decisions.

- only clear-cut distinctions between grammatical functions are made.

Figure 1 shows an example annotation. As was presented in (Skut W. *et al.* 1997), the crossing branches occuring in the annotations can be automatically converted to context-free structures with traces if required.

2. CORPUS ANNOTATION

We employ a new interactive annotation process for the annotation of our corpus data. The process incrementally guides a human annotator through the structure that is to be generated. The annotator can stop the process at any point and correct or alter the structure or the labeling manually. Furthermore, the automatic process includes a reliability measure. Using this automatic measure, the process classifies its decisions either as reliable or as unreliable. The latter require additional actions by the annotator in order to ensure a high accuracy of the annotated corpus. The process is integrated in a flexible graphical user interface.

2.1. The Interactive Annotation Process

The annotation of our corpus includes part-of-speech tags, syntactic structures (including discontinuous constituents), and labeling of nodes and edges. A tagger and a parser are running in the background, the annotation is visualized by a graphical user interface (see next section).

The process is outlined in figure 2. Annotation starts with a raw sequence of words. As a first step, a statistical tagger is run on this sequence. The tagger implements Markov Models. Unknown words are handled by suffix analysis (Samuelsson C. 1993). We not only exploit the tags that are assigned, but also their probabilities. We look at the distance of the best and second best tag's probabilities, $P(t_{best})$ and $P(t_{second})$, expressed by their quotient. If their distance is large, the best tag is simply added to the annotation. If their distance

is small, the annotator is asked to confirm the tag. So we choose a threshold θ and classify:

$$\text{reliable, if } \frac{P(t_{best})}{P(t_{second})} \geq \theta, \qquad \text{unreliable, if } \frac{P(t_{best})}{P(t_{second})} < \theta$$

We empirically choose θ such that the expected accuracy of all reliable cases is above 99%.

Next, the syntactic structure is built incrementally. The best phrase hypothesis given the current annotation is determined by Cascaded Markov Models (Brants T. 1999). Each layer of the structure is represented by a separate Markov Model, which assigns probabilities to each hypothetical phrase. The phrase with the highest probability is presented to the annotator, who can either accept or reject it. In the first case, annotation proceeds and the system builds the next phrase. In the latter case, the phrase is removed from the internal structures of the parser, probabilities are re-calculated, and the next-best phrase is presented. At any point, the annotator can manually change the structure if the suggested phrase needs a small change or if the parser is not able to generate the correct structure.

If a structural element is confirmed, the parser tests the labels in the nodes and edges for reliability (as it was done for the part-of-speech tags). Labeling is again performed with Markov Models. The system highlights unreliable assignments and asks the annotator for confirmation. Again, thresholds are set such that reliable assignments have an accuracy above 99%. Nevertheless, annotators can change all labels if an error occurs, or change the structure at a later point.

The tagger and parser need some minimum amount of annotated training to enable automation. The exact amount depends on the language and the annotation scheme. It is typically around a few thousand tokens. Accuracy is lower with small amounts of training data, but the problem is allevated by using the reliability measure.

2.2. Graphical User Interface

The process of selecting the best phrase hypothesis is integrated in a graphical user interface, which is shown in figure 3. The tool allows a variety of manual tree manipulations (grouping, ungrouping, re-attachment, labeling, etc.). It additionally runs taggers and parsers in the background. These are used for interactive annotation.

The tool can handle virtually any annotation scheme that is based on context-free structures (with crossing branches) and labelings of nodes and edges. For instance, it can be also used for the Penn Treebank and the Susanne corpus. The graphical user interface in combination with cascaded Markov models allows very efficient annotation. A trained annotator needs in average 50 seconds per sentence of newspaper text with an average sentence

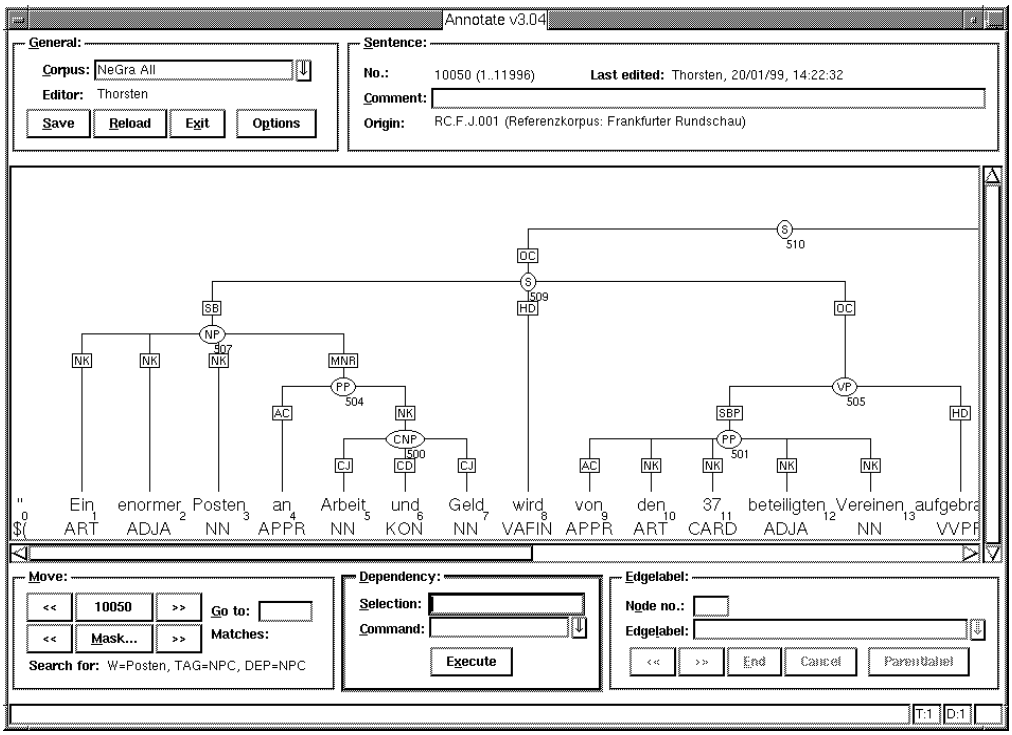

Figure 3: Graphical user interface for corpus annotation. The GUI uses an interface to cascaded Markov models that generate the next phrase hypothesis presented to the annotator. Additionally, all necessary manual tree manipulations are supported.

length of 17.5 tokens (approx. 1,300 tokens/minute)[1]. The annotation includes part-of-speech tagging, syntactic structure, and labeling of nodes and edges.

2.3. Comparison of Annotations

In order to ensure a high degree of accuracy, each sentence is independently annotated by two annotators. The annotations are compared. Differences in the structure and the labeling are automatically detected and presented to the annotators. Their task is to agree on one annotation. Before comparison, approx. 52% of all sentences receive the same annotation by both annotators, i.e., both sentences receive the same structure, same part-of-speech tags, same phrase categories and same grammatical functions. If only the structure is tested, 68% of all sentences are identical. This is equivalent to approx. 93% recall/precision. After comparison, differences remain in about 6% of the sentences, which is equivalent to 99% recall/precision.

[1]The annotation time was measured for two annotators and averaged over 2,000 sentences.

3. APPLICATIONS

The design of the corpus and its annotation scheme aims at re-usability for applications in language technology and for linguistic investigations. The corpus has been used to train part-of-speech taggers and chunkers (Skut W. & Brants T. 1998; Brants T. 1999) with very good results. It also has been used for an investigation on the extraposition of relative clauses (Uszkoreit H. *et al.* 1998). The distribution of material between the NP and the relative clause exhibits length effects as proposed by Hawkins principle of early immediate constituents (Hawkins J. A. 1994), but it shows a systematic difference to results of a psycholinguistic experiment. Furthermore, the corpus provided data for an investigation on collocations (Krenn B. 1998) and for an investigation on word order in the German Mittelfeld (Kurz D. ming). The corpus is freely available to non-profit organizations for research purposes[2]. This enabled other investigations on rule-based and statistical processing as well as linguistic investigations, which have recently started.

4. CONCLUSIONS

The presented corpus is annotated with a hybrid scheme combining the advantages of phrase-structure and dependency grammars. In addition to context-free trees, crossing branches are allowed and predicate-argument information is encoded. This approach was motivated by free word order phenomena in German and by “incomplete” structures frequently found in corpora. The resulting flat structures are very well suited to human annotation.

We presented an efficient interactive annotation mode that guides the annotator through the annotation and asks for confirmation if the process detects unreliable assignments. The process is incremental, each increment consists of one new phrase. This is probably a point for further optimization. If the complete increment is classified as reliable, the process could enlarge the increment in order to speed up annotation.

Currently, the annotated part of the corpus consists of approx. 20,000 sentences (350,000 tokens).

RÉFÉRENCES

BRANTS, Thorsten ; SKUT, Wojciech (1998) : “Automation of treebank annotation”, *in Proceedings of New Methods in Language Processing*, Sydney, Australia.

BRANTS, Thorsten (1999) : “Cascaded markov models”, *in Proceedings of 9th Conference of the European Chapter of the Association for Computational Linguistics EACL-99*, Bergen, Norway.

CARTER, David (1997) : “The TreeBanker: a tool for supervised training of parsed

[2] see `http://www.coli.uni-sb.de/sfb378/negra-corpus/`

corpora", *in ACL Workshop: Computational Environments for Grammar Development and Linguistic Engineering*, Madrid, Spain.

HAJIC, Jan (1998) : "Building a syntactically annotated corpus: The prague dependency treebank", *in Issues of Valency and Meaning*, Praha, Karolinum.

HAWKINS, John A. (1994) : *A performance theory of order and constituency*, Cambridge Studies in Linguistics 73, Cambridge Univ. Press.

HINDLE, Don (1983) : "Deterministic parsing of syntactic non-fluencies", *in Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics ACL-83*, Cambridge, MA.

KRENN, Brigitte (1998) : "A competence-performance integrated representation model for collocations", *in Proceedings of AMLaP-98*, Freiburg.

KURZ, Daniela (forthcoming) : *Wortstellungsphänomene im Deutschen Mittelfeld*, Master thesis, Universität des Saarlandes, Computational Linguistics.

MARCUS, Mitchell ; SANTORINI, Beatrice ; MARCINKIEWICZ, Mary Ann (1993) : "Building a large annotated corpus of English: The Penn Treebank", *Computational Linguistics*, vol. 19, n° 2, pp. 313–330.

SAMPSON, Geoffrey (1995) : *English for the Computer*, Oxford University Press.

SAMUELSSON, Christer (1993) : "Morphological tagging based entirely on Bayesian inference", *in 9th Nordic Conference on Computational Linguistics NODALIDA-93*, Stockholm University, Stockholm, Sweden.

SKUT, Wojciech ; BRANTS, Thorsten (1998) : "A maximum-entropy partial parser for unrestricted text", *in Sixth Workshop on Very Large Corpora*, Montreal, Canada.

SKUT, Wojciech ; BRANTS, Thorsten ; KRENN, Brigitte ; USZKOREIT, Hans (1997) : "Annotating unrestricted german text", *in Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft fr Sprachwissenschaft*, Heidelberg, Germany.

USZKOREIT, Hans ; BRANTS, Thorsten ; DUCHIER, Denys ; KRENN, Brigitte ; KONIECZNY, Lars ; OEPEN, Stephan ; SKUT, Wojciech (1998) : "Studien zur performanzorientierten Linguisitk. Aspekte der Relativsatzextraposition im Deutschen", *Kognitionswissenschaft*, vol. 7, n° 3.