The TIGER Treebank

Sabine Brants*, Stefanie Dipper†, Silvia Hansen*, Wolfgang Lezius†, George Smith‡

* Computational Linguistics, Saarland University, Postfach 15 11 50 Saarbrücken, Germany
{sabine, hansen}@coli.uni-sb.de

† Institute of Natural Language Processing (IMS), Stuttgart University, Germany
{dipper, lezius}@ims.uni-stuttgart.de

‡ Institut für Germanistik, Potsdam University, Germany
smithg@rz.uni-potsdam.de

Abstract

This paper reports on the TIGER Treebank, a corpus of currently 35.000 syntactically annotated German newspaper sentences. We describe what kind of information is encoded in the treebank and introduce the different representation formats that are used for the annotation and exploitation of the treebank. We explain the different methods used for the annotation: interactive annotation, using the tool *Annotate*, and LFG parsing. Furthermore, we give an account of the annotation scheme used for the TIGER treebank. This scheme is an extended and improved version of the NEGRA annotation scheme and we illustrate in detail the linguistic extensions that were made concerning the annotation in the TIGER project. The main differences are concerned with coordination, verb-subcategorization, expletives as well as proper nouns. In addition, the paper also presents the query tool TIGERSearch that was developed in the project to exploit the treebank in an adequate way. We describe the query language which was designed to facilitate a simple formulation of complex queries; furthermore, we shortly introduce TIGERin, a graphical user interface for query input. The paper concludes with a summary and some directions for future work.

1 Introduction

Corpus-based methods play an important role in empirical linguistics as well as in machine learning methods in NLP. In these two areas of research, large natural language corpora, enriched with syntactic information, are needed. Thus, in recent years, there has been an increasing interest in the construction of these syntactically annotated corpora, commonly called ‘treebanks’ (Lezius, 2001).

For German, the first initiative in the field of treebanks was the NEGRA Corpus ((Skut, Brants, Krenn, & Uszkoreit, 1998) and (Brants, Skut, & Uszkoreit, 1999)), which contains syntactically interpreted newspaper texts. Furthermore, there is the Verbmobil Corpus (Wahlster, 2000), which covers the area of spoken language.

This paper reports on the TIGER Treebank project, which aims at building the largest and most exhaustively annotated treebank for German. The annotation format and scheme are based on the NEGRA corpus; however, the TIGER Treebank exceeds the NEGRA corpus in size as well as in detail of annotation. Since the NEGRA Corpus is rather restricted in its size (20,000 syntactically annotated sentences) and the Verbmobil Corpus in its domains (i.e. spontaneous speech for the appointment negotiation domain), the construction of the TIGER Treebank as a comprehensive resource for the German language was a necessary step to overcome these drawbacks.

This paper is structured in the following way: Section 2 describes the annotation format and provides general information on the annotation scheme. Furthermore, it contains a short overview of treebank initiatives for languages other than German. In Section 3, the different methods used for the annotation of the treebank are presented. The linguistic extensions that were made in the TIGER project concerning the annotation scheme are covered in Section 4. Section 5 gives an overview of the query language and query tool that were developed in the project for the exploitation of the treebank. Finally, Section 6 summarizes the paper and sketches some ideas for future work.

2 The TIGER Corpus

The basis of the TIGER Treebank are texts from the German newspaper *Frankfurter Rundschau*. Only complete articles were used, which were taken from all kind of domains[1] so as to cover a broader range of language variations. At the current stage, the corpus contains 35.000 syntactically annotated sentences. In the second project phase, this amount is to be extended to approximately 45.000 sentences.

2.1 Levels of annotation

In the NEGRA as well as in the TIGER corpus, a hybrid framework is used which combines advantages of dependency grammar and phrase structure grammar. The syntactic structure is represented by a tree. The branches of a tree may cross, allowing the encoding of local and non-local dependencies and eliminating the need for traces. This approach has considerable advantages for free-word order languages such as German, which show a large variety of discontinuous constituency types (Skut, Krenn, Brants, & Uszkoreit, 1997).

The linguistic annotation of each sentence in the TIGER Treebank is represented on a number of different levels (see figure 1): Part-of-speech information is encoded in terminal nodes (on the word level). Non-terminal nodes are labeled with phrase categories. The edges of a tree represent syntactic functions. Furthermore, a supplementary annotation on the word level facilitates the encoding of information on lemmata and morphology[2]. For part-of-speech tagging, the Stuttgart-Tübingen-Tagset (Schiller, Teufel, & Stöckert, 1999) is used in a slightly modified version ((Kramp & Preis, 2000) and (Smith & Eisenberg, 2000)). Information on lemmata and morphology was not annotated in the NEGRA corpus; this is a new feature that was added to the annotation in the TIGER project.

Syntactic structures are rather flat and simple in order to reduce the potential for attachment ambiguities. The distinction between arguments and adjuncts, for instance, is not expressed in the constituent structure, but is instead encoded by means of syntactic functions.

Apart from the annotation of morphology and lemmata, another annotation level was added to the TIGER corpus: Secondary edges, i.e. labelled directed arcs between arbitrary nodes, are used to encode coordination information. Currently, these secondary edges are only employed for the annotation of coordinated sentences and verb phrases; another potential use might be the systematic annotation of attachment ambiguities.

[1] We only excluded regional news and sports news because experiences from the past showed that these texts often contained tables, enumerations, etc. instead of complete sentences.

[2] The example in Figure 1, however, contains no lemmata annotation, but a literal translation instead.

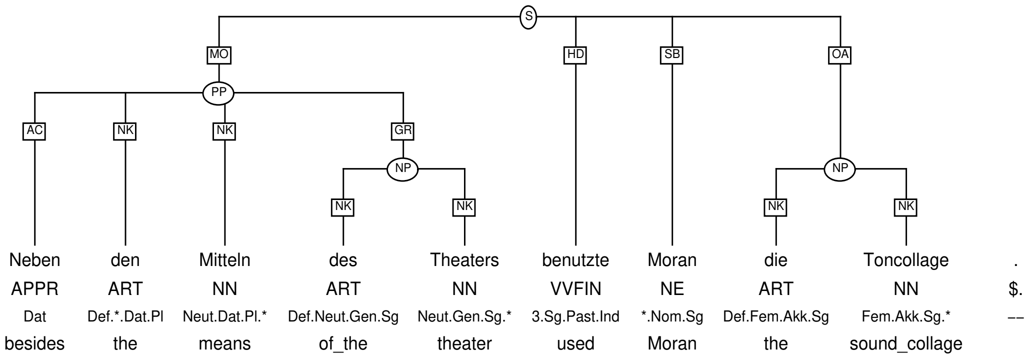

Figure 1: Different levels of annotation

2.2 Annotation formats

In the TIGER project, we use several annotation formats for corpus storage, export and querying. There exist scripts that enable the transformation from one format to another.

First of all, the annotated sentences are stored and maintained in a MySQL database; information about the annotation is contained in tables. An additional output format is used for the export of the sentences. The database entries (words, morphological tags, terminal nodes, non-terminal nodes and edges) can be exported to a table stored in a line-oriented and ASCII-based format (Brants, 1997). The major advantage of this export format is that it is easily readable for humans as well as easily processable for machines. Sentence boundaries are identified through sentence start and end tags. Furthermore, information on sentence origins, editors and used tags is stored at the beginning of each export file.

Based on this export format, the TIGER corpus can be transferred into a third format, namely TIGER XML (Lezius, Biesinger, & Gerstenberger, 2002a). A TIGER XML file is typically split up into header and body. The corpus header contains meta-information on the corpus (such as corpus name, date, author, etc.) and a declaration of the tags that are used for morphology, part-of-speech, non-terminal nodes and edges. In the corpus body, directed acyclic graphs are used as the underlying data model to encode the linguistic annotation. Words, part-of-speech tags, morphological tags and lemmata occur as attributes of the element ‘terminal’, whereas non-terminals are represented in an additional element called ‘nonterminal’ refering to the corresponding terminal ID. Secondary edges are encoded explicitly as well. By using an XML format, the TIGER Treebank is exchangable and usable with a large range of tools. The XML format is also the basis for the use of the corpus query tool TIGERSearch (Lezius & König, 2000).

2.3 Comparable treebank initiatives

One of the first and best known treebanks is the Penn Treebank for the English language (Marcus et al., 1994), which consists of about 1 million words of newspaper text. It contains part-of-speech tagging and rough syntactic and semantic annotation. A bracketing format is used to encode predicate-argument structure and trace-filler mechanisms are used to represent discontinuous phenomena. Other comparable treebanks for English are, for instance, the Susanne Corpus (Sampson, 1995) (containing detailed part-of-speech information and phrase structure annotation), the Lancaster Parsed Corpus (Leech, 1992) (representing phrase structure annotation by means of labelled bracketing) and the

British part of the International Corpus of English (Greenbaum, 1996) (about 1 million words of British English that were tagged, parsed and afterwards checked).

For languages other than English, a fairly well-known treebank is the Prague Dependency Treebank for Czech (Hajic, 1999). It contains about 450.000 tokens and is annotated on three levels: on the morphological level (tags, lemmata, word forms), on the syntactic level (using dependency syntax) and on the tectogrammatical level (encoding functions such as Actor, Patient, etc.). Recently, treebank projects for other languages have come to life as well, e.g. for French (Abeillé, Clément, & Kinyon, 2000), Italian (Bosco, Lombardo, Vassallo, & Lesmo, 2000), Spanish (Moreno, Grishman, López, Sánchez, & Sekine, 2000), Turkish (Oflazer, Hakkani-Tür, & Tür, 1999), Russian (Boguslavsky, Grigorieva, Grigoriev, Kreidlin, & Frid, 2000) and Bulgarian (Simov et al., 2002). More initiatives for linguistically interpreted corpora can be found in (Uszkoreit, Brants, & Krenn, 1999) and (Abeillé, Brants, & Uszkoreit, 2000).

3 Annotation methods

We use two different methods for the syntactic annotation of the TIGER corpus: Interactive annotation and LFG parsing. The first one is a combination of probabilistic parsing and human intervention (Section 3.1). After the parsing is completed, morphological annotation is performed semi-automatically, using the given syntactic annotation for disambiguation. For the second method, a symbolic LFG grammar is used to parse large parts of the corpus; the output is disambiguated by a human annotator (Section 3.2).

3.1 Interactive Tagging and Parsing

Interactive annotation is an efficient combination of automatic parsing and human annotation. Instead of having an automatic parser as preprocessor and a human annotator as postprocessor, the two steps are interwoven in our approach. The parser generates a small part of the annotation, which is immediately presented visually to the human annotator, who can either accept, correct or reject it. Based on the annotator's decision, the parser proposes the next part of the annotation, which is again submitted to the annotator's judgement. This process is repeated until the annotation of the sentence is complete.

The advantage of this interactive method is that the human decisions can be used by the automatic parser. Thus, errors made by the automatic parser at lower levels are corrected instantly and do not 'shine through' on higher levels. The chances grow that the automatic parser proposes correct analyses on higher levels.

The interactive annotation works on several layers. The lowest one is the part-of-speech layer. Higher layers are defined by the depth of the syntactic structure. Each layer is represented by a different Markov Model, hence the name *Cascaded Markov Models* (Brants, 1999). The first step in the annotation process is the generation of part-of-speech tags. This step is performed using the statistical tagger TnT (Brants, 2000b). In addition to the tags, TnT also generates probabilities that help to decide on the reliability of a proposed tag. The lower the probability of alternative tags, the higher the reliability of the best tag for a word (Brants & Skut, 1998). Approximately 84% of all tag assignments are classified as reliable by the tagger. The remaining 16% need to be proof-read by human annotators.

Once the part-of-speech tagging is done, Markov models for higher layers start processing. Hypothetical phrases are generated, and the one with the highest probability is displayed to the annotator. The structure can be accepted, rejected or manually corrected by the annotator. Intervention by the human

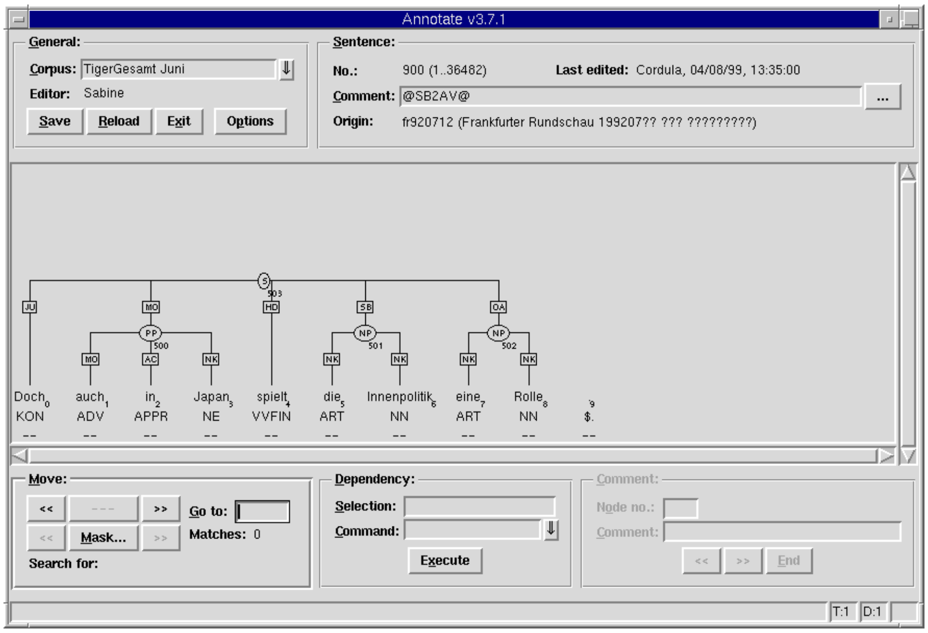

Figure 2: The annotation tool *Annotate*

annotator immediately changes the set of hypotheses used by the parser. The syntactic structure is built phrase by phrase, bottom up. About 71% of the phrases suggested by the parser are correct, 17% need minor intervention (i.e., at most one constituent needs to be added or deleted). The remaining 12% require major intervention by the human annotator.

Both tagger and parser are entirely trained on previously (manually) annotated data. No manual grammar or lexicon development are necessary. The annotation scheme is learnt automatically by tagger and parser. In case of changes in the annotation scheme, only a small amount of data needs to be changed manually. The tagger and parser are then trained on the changed data and are immediately ready for annotation with the new scheme.

The annotation is performed with the help of the tool *Annotate* (Figure 2), a graphical user interface with a comprehensive set of tree manipulation functions and database access (Plaehn & Brants, 2000). *Annotate* runs the TnT tagger and the Cascaded Markov Models in the background.

In order to achieve a high level of consistency and to avoid mistakes, we use a very thorough approach to the annotation: First, each sentence is annotated independently by two annotators. With the support of scripts, they afterwards compare their annotations and correct obvious mistakes. Remaining differences are submitted to a discussion between the annotators. Although this process is rather time-consuming, it has proven to be highly beneficial for the accuracy of the annotation (Brants, 2000a). Furthermore, it also supports the continuous improvement of the annotation scheme: It is in the discussion between the annotators that discrepancies between the annotation scheme and the data become obvious. If this happens to be the case, new rules and better tests for operationalization are added to the annotation scheme. Thus, there is a cross-fertilization between the corpus and the annotation scheme.

Morphology and Lemmata For the analysis of lemmata and morphological tags, we use an additional graphical user interface. This is interleaved with a tool called *TigerMorph* which was developed

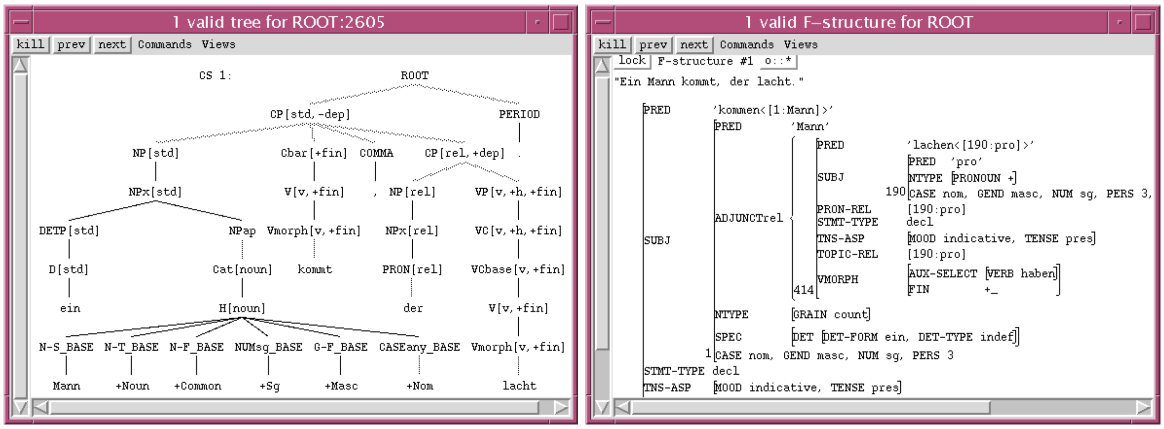

Figure 3: LFG c- and f-structure

by Berthold Crysmann. *TigerMorph* disambiguates the output of a morphological analyser on the basis of the already existing syntactic structure. It proposes lemmata and morphological tags for the words of a sentence, proceeding from left to right. The annotation of morphology and lemmata resembles the interactive annotation of the syntactic structure described above. Ambiguous tags are presented to the annotator, who then decides which one is the correct alternative. This information is returned to *TigerMorph* and used for the disambiguation of the morphological analysis for the remaining words.

3.2 Annotation by LFG Parsing

As an alternative to interactive tagging and parsing, a broad coverage symbolic LFG grammar (Lexical Functional Grammar (Bresnan, 1982)) is used to parse parts of the corpus (Dipper, 2000). Usually, the LFG grammar outputs several analyses for a corpus sentence. The output is first filtered by a grammar internal ranking mechanism and then disambiguated by a human annotator. A transfer component converts the selected analysis into the TIGER format (Zinsmeister, Kuhn, & Dipper, 2002).

One advantage of this approach is the accuracy of the grammar's output. An LFG analysis is always syntactically consistent. It does not contain inconsistencies such as, e.g., missing subject-verb agreement, in case of which the parse would have failed. On the other hand, the grammar certainly is not error-free. But those errors which do occur are systematic and hence easier to correct than errors that occur with manual annotation.

Parsing The LFG grammar applied in parsing has been developed in the ParGram project at the University of Stuttgart, using the Xerox Linguistic Environment (XLE) (ParGram, 2002). The output of an LFG grammar basically consists of two representations, the constituent structure (c-structure) of the sentence being parsed, and its functional structure (f-structure). C-structure encodes information about morphology, constituency, and linear ordering. F-structure represents information about predicate argument structure, about modification, and about tense, mood etc. An example of c- and f-structure is given in Figure 3 for the sentence *Ein Mann kommt, der lacht* ('a man is coming who laughs').

Disambiguation Almost every sentence of a newspaper corpus is syntactically ambiguous. Hence the output of a purely symbolic grammar has to be disambiguated, i.e. a human annotator has to select the correct analysis. This task is supported by XLE which allows for ‘packing’ the different readings into one complex f-structure representation.[3]

On average, however, a sentence of the TIGER corpus receives several thousands of LFG analyses. Clearly it is impossible to disambiguate those analyses manually. Therefore XLE provides a (non-statistical) mechanism for suppressing certain ambiguities automatically (Frank, King, Kuhn, & Maxwell, 1998). By means of this mechanism, the average number of solutions drops down to 17, the median being 2.

Conversion into TIGER format All information that is required by the TIGER annotation scheme is contained in c- and f-structure representations of LFG. Compare the LFG representation (Figure 3) with the TIGER representation (Figure 4) of the sentence *Ein Mann kommt, der lacht.*

(i) LFG c-structure contains categorial information (e.g., NP, CP), lemmas (Mann), part-of-speech tags (+Noun +Common), and morphological tags (+Sg +Masc +Nom); in Figure 3, lemma and tags are only shown for the terminal *Mann*. In the TIGER scheme, this information is encoded in a slightly different terminology: the nodes and tags mentioned above correspond to TIGER nodes NP, S, the part-of-speech tag NN, and the morphological tag Masc.Nom.Sg, respectively.

(ii) LFG f-structure represents dependency relations such as the head argument relation SUBJ and the head modifier relation ADJUNCTrel (‘relative clause’). Note that in c-structure, NP and CP (the relative clause) do not form a constituent; however, their f-structures, SUBJ and ADJUNCTrel, are linked. This linking information is encoded by a crossing branch in TIGER, cf. Figure 4.

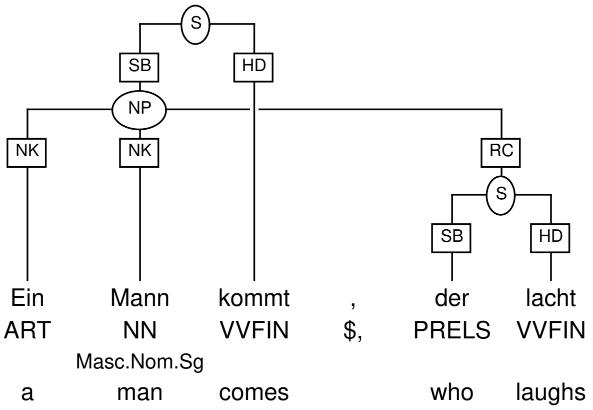

Figure 4: TIGER representation

Often, there is a one-to-one correspondence between LFG and TIGER representations. In these cases the transfer component simply converts one format into another, e.g. +Sg +Masc +Nom is mapped to Masc.Nom.Sg, CP to S, SUBJ to SB, etc. However, in other cases the transfer has to combine information both from c- and f-structure, as in the case of the extraposed relative clause in Figure 3/4. Here the transfer component makes use of the f-structure link between SUBJ and ADJUNCTrel to form a (discontinuous) constituent. The categorial label (NP) is derived from c-structure.

[3] Furthermore XLE provides various browsing tools applying to c-structure as well as to f-structure which can be used for manual disambiguation (cf. (King, Dipper, Frank, Kuhn, & Maxwell, To Appear) where these tools are described in detail).

Results and Outlook When parsed with the current grammar version, 50% of the corpus sentences receive at least one analysis; approx. 70% of the parsed sentences receive the correct analysis (possibly among others).[4] About 2,000 sentences of the TIGER corpus has been annotated this way.

To enlarge coverage, XLE allows for partial parses, providing, e.g., N or P chunks. In first experiments, N chunks were found with a precision of 89% and a recall of 67%; for P chunks, the precision was 96%, and the recall was 79% (Schrader, 2001). Furthermore, to minimize manual effort, a statistical disambiguation tool can be integrated (Riezler et al., 2002).

4 Extensions in the TIGER annotation scheme

The annotation in TIGER is based on the annotation scheme that was used for the NEGRA corpus (Brants et al., 1999). This annotation scheme covered a broad variety of phenomena. However, there was still room for improvement in its linguistic adequacy. A vital part of the work in the TIGER project is the linguistic extension of this annotation scheme. In the following, the major changes that were made in the TIGER project are presented. A more detailed account of these changes and an evaluation of the improved annotation scheme can be found in (Brants & Hansen, 2002).

4.1 Coordination

An essential linguistic extension in the TIGER annotation scheme was made concerning the annotation of coordinated sentences and verb phrases. In the NEGRA corpus, arguments that are shared by both verb conjuncts of a coordination, but that are only mentioned once, were structurally linked only to the nearest part of the coordination. Thus, the NEGRA annotation is in many cases not suitable for the extraction of subcategorization information. In the TIGER treebank, these shared arguments are provided with secondary edges in order to represent their syntactic relation to the more distant verb conjuncts.

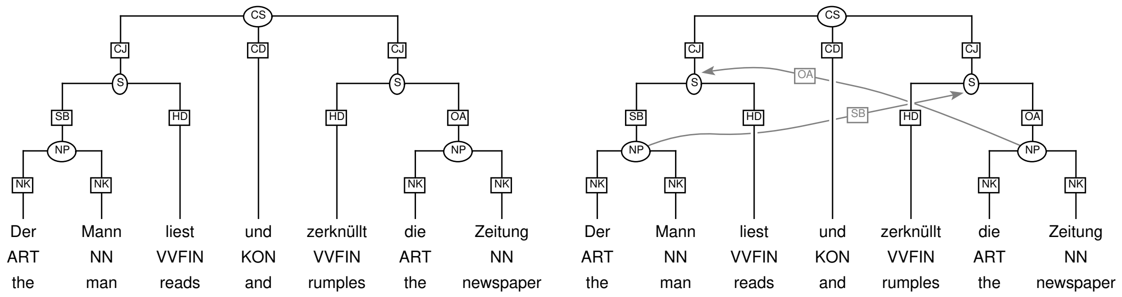

Figure 5: Coordination with shared arguments in NEGRA (left) and TIGER (right)

Figure 5 illustrates the difference between the NEGRA and the TIGER treatment of these cases. In the example sentence *Der Mann liest und zerknüllt die Zeitung* ('the man read and rumpled the newspaper'), the common subject of both verbs is the NP *der Mann*, the common object is the NP *die Zeitung*. However, in the NEGRA annotation scheme, shared arguments are linked only to the nearest verb (cf. left part of Figure 5). The structure of the tree would be exactly the same if the first

[4] 10% of the sentences failed because of gaps in the morphological analyser; 6% failed because of storage overflow or timeouts (with limits set to 100 MB storage and 100 seconds parsing time). 10% of the parsed sentences were not evaluated wrt. the correct analysis because they received more than 20 analyses.

verb were intransitive and did not have *die Zeitung* as its object (e.g. *Der Mann lacht und zerknüllt die Zeitung* ('the man laughs and rumples the newspaper')). In contrast, the right part of Figure 5 shows the annotation of the sentence according to the TIGER annotation scheme, making use of the secondary edges that were introduced. Thus, the TIGER scheme allows the differentiation between transitive and intransitive verbs in coordinations.

4.2 Verb-subcategorization

The NEGRA corpus provides no distinctions between prepositional phrases with respect to their syntactic functions; all PPs occuring in sentences or verb phrases are unexceptionally marked with the label MO (modifier). In the TIGER project, two additional function labels for PPs were introduced: prepositional objects (OP) and collocational verb constructions (CVC). The label OP is applied to constructions like *auf jemanden warten* ('to wait for somebody'). These phenomena are marked by the fact that the preposition *auf* ('on') has lost its lexical meaning.

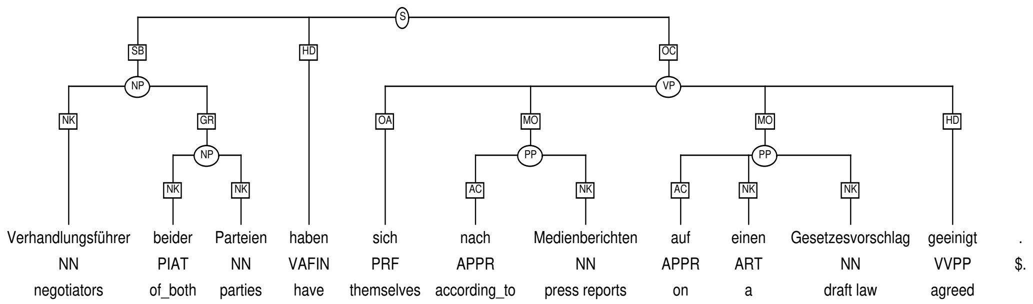

Figure 6: Annotation of PPs in NEGRA: MO

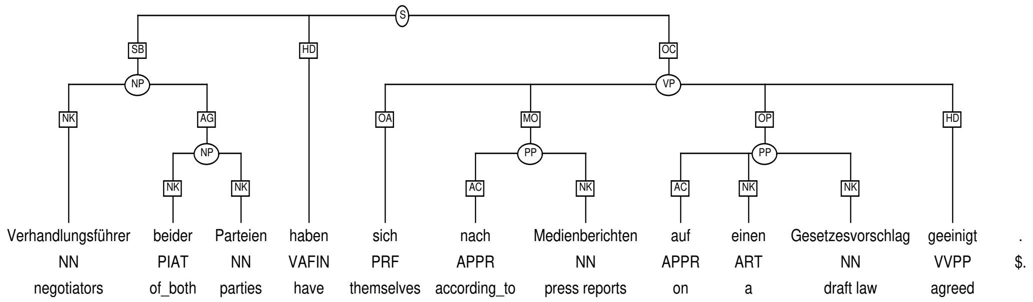

Figure 7: Annotation of PPs in TIGER: MO and OP

Figure 6 exemplifies the fact that NEGRA did not allow the distinction between complements and adjuncts on the level of edge labels[5]. In contrast, the TIGER annotation (Figure 7) mirrors the functional difference between the first PP and the second PP in the use of different edge labels: the first PP is functionally independent of the verb and serves as an adverbial; it still receives the label MO. The second PP represents a typical example for a prepositional object (OP) in German: the preposition *auf* ('on') has completely lost its lexical meaning and is purely functional.

[5]A correct English translation of the example sentence is the following: 'According to press reports, negotiators from both parties have agreed on a draft law'.

The other newly introduced edge label for prepositional phrases, CVC (collocational verb construction), serves to label verb + PP constructions in which the main semantic information is contained in the noun of the PP, not in the verb. This label can only be used with a very limited class of verbs (usually a semantically weak verb with an originally directional or local meaning, e.g. *stellen, kommen*, etc. ('to put', 'to come')) that occur in connection with an equally limited class of prepositions (mostly *zu* and *in* ('to' and 'in')). A typical example for this is the German collocational expression *in Kraft treten* (literal translation: 'to step into force', meaning: 'to take effect').

4.3 Expletives

In the TIGER corpus, finer distinctions with regard to the usage of *es*, the German expletive, have been introduced. In the NEGRA annotation scheme, only one label (PH, meaning place-holder) was used for the non-semantic usage of *es*; in the TIGER scheme, we distinguish three types:

- ***Vorfeld es***: This type of *es* is used to fill the first position of a sentence, called the *Vorfeld* slot. It is marked with the label PH (place-holder). Example: *Es naht ein Gewitter* (literal translation: 'it approaches a thunderstorm'; meaning: 'a thunderstorm is approaching'). As soon as another component occupies the *Vorfeld* slot, the *es* disappears: *Ein Gewitter naht.*
- **Correlative *es***: This second type of *es* is always correlated to some propositional argument in the sentence. It is usually optional. Example: *Mich freut es, dass ...* ('it makes me happy that ...'). This type is also labeled as PH but can be easily distinguished from the first type because it always occurs in connection with a propositional sister node functioning as RE (repeated element) (cf. Figure 8).
- **Expletive *es***: The last type of *es* functions as a non-thematic argument, e.g. in connection with weather verbs: *Heute regnet es* ('today, it is raining'). It receives the label EP (expletive).

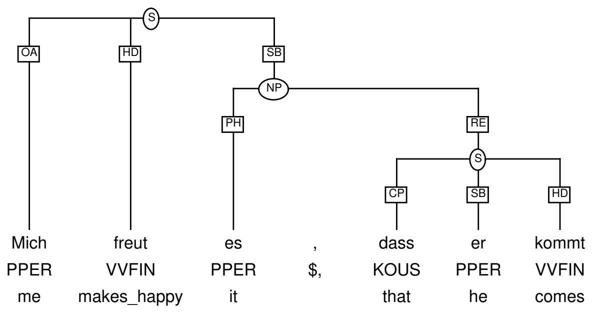

Figure 8: Correlative *es*

4.4 Proper nouns

In the NEGRA as well as in the TIGER corpus, the parent label PN is used to mark ordinary proper nouns, such as 'Gerhard Schröder'. The single components receive the edge label PNC (proper noun component). Furthermore, the label PN is also used for multi-token company names, newspaper names (e.g. 'The San Francisco Chronicle') etc. In the TIGER annotation scheme, the usage of this label was extended to cover titles of films, books, exhibitions etc. that have a complex, sometimes sentence-like structure. Occurences of these phenomena are first annotated structurally and then receive an additional unary parent label PN. The examples in Figure 9 illustrate the different annotations in NEGRA and TIGER.

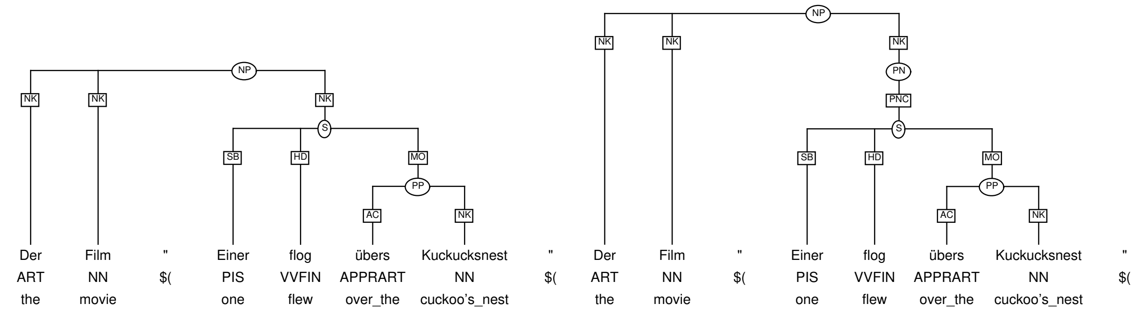

Figure 9: Treatment of structured proper nouns in NEGRA (left) and TIGER (right)

Thus, the TIGER annotation permits the identification of structures that function as names, but do not feature the corresponding part-of-speech tag NE (proper noun) in one of their terminal nodes.

5 TIGERSearch

Syntactically annotated corpora such as the TIGER treebank provide a wealth of information which can only be exploited with an adequate query tool and query language. Thus, a powerful search engine for treebanks has been developed within the TIGER project (Lezius, 2002).

The search engine is freely available for research purposes and can be downloaded from the TIGER web site. To support all popular platforms, the tool is implemented in Java. Input filters are provided for many popular treebank formats. A complete list of supported formats is given in (Lezius, Biesinger, & Gerstenberger, 2002b). In addition, an interface format, TIGER-XML (Mengel & Lezius, 2000; Lezius et al., 2002a), has been defined. This format can be used to import treebanks in other formats, including formats which have not yet been developed. The search engine is thus a tool which can be used by the entire community.

5.1 Query Language

The query language (König & Lezius, 2002a, 2002b) has been designed to fulfill two conflicting requirements: On the one hand, it is close to grammar formalisms, thus easy to learn. It allows modular, understandable code, even for complex queries. A user can pose queries intuitively, mapping linguistic descriptions directly into the query language. On the other hand, its expressiveness has been constrained to guarantee efficient query processing.

The query language consists of three levels. On the *node* level, nodes can be described by Boolean expressions over feature-value pairs. The following query is matched by the terminal node *lacht* ('laughs') in Figure 4:

```
[word="lacht" & pos="VVFIN"]
```

On the *node relation* level, descriptions of two or more nodes are combined by a relation. Since graphs are two-dimensional objects, we need one basic relation for each dimension. These are immediate precedence (“`.`”) for the horizontal dimension and immediate dominance (“>”) for the vertical dimension.[6] There are also derived node relations such as underspecified dominance or siblings:

`>*`	dominance (minimum path length 1)
`>n`	dominance in n steps ($n > 0$)
`>m,n`	dominance in d steps ($m \leq d \leq n$)
`>L`	labeled dominance (edge label `L`)
`>@l`	leftmost terminal successor (‘left corner’)
`.*`	precedence (min. number of intervals: 1)
`$`	siblings

For example, the following query is matched by a subgraph of Figure 4 (the NP node):

```
[cat="NP"] >RC [cat="S"]
```

Finally, on the *graph description* level we allow Boolean expressions over node relations, without negation. For example, a subgraph of Figure 4 (the second S node) satisfies the following query:

```
([cat="S"] > [pos="PRELS"]) &
([cat="S"] > [pos="VVFIN"])
```

Variables can be used to express coreference of nodes or feature values. For example, the two node descriptions `[cat="S"]` in the above query could refer to different nodes. A reformulation of the query using variables prevents this:

```
(#n:[cat="S"] > [pos="PRELS"]) &
(#n > [pos="VVFIN"])
```

The following query describes a clause that comprises a personal pronoun and a finite verb (cf. Figure 5).

```
#v:[cat="S"] &
#t1:[pos="PPER"] & #t2:[pos="VVFIN"] &
(#v > #t1) & (#v > #t2) & arity(#v,2)
```

In addition, the user can define type hierarchies. Subtypes may also be constants, e.g. in the case of part-of-speech symbols. Here is a part of a type hierarchy for the STTS tag set:

```
nominal := noun,properNoun,pronoun.
noun := "NN".
properNoun := "NE".
pronoun := "PPER","PDS","PRELS", ...
```

[6]The precedence of two inner nodes is defined as the precedence of their leftmost terminal successors (König & Lezius, 2000).

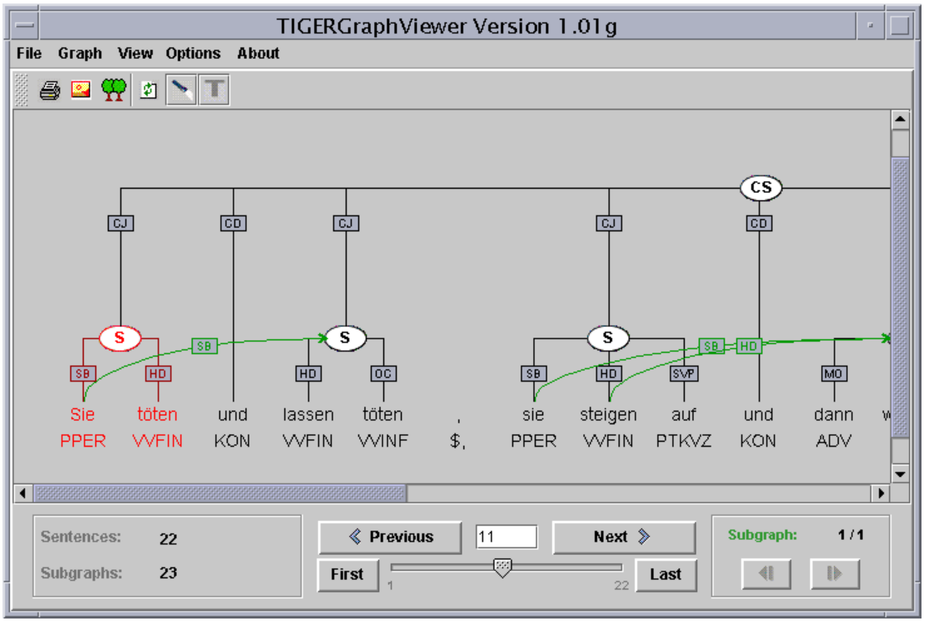

Figure 10: Visualization of query results

This hierarchy can be used to formulate queries in a more adequate way:

```
[pos=nominal] .* [pos="VVFIN"]
```

There are also several useful predicates such as `discontinuous(#n)`,`continuous(#n)` (phrase does/does not contain crossing branches) or `arity(#n,num)` (phrase comprises *num* children). The following example query determines extraposed relative clauses in the TIGER treebank (cf. Figure 4):

```
(#n:[cat="NP"] >RC [cat="S"]) &
discontinuous(#n)
```

To simplify the formulation of more complex queries, templates can be defined. These are described in (König & Lezius, 2002a, 2002b).

5.2 Query Tool

To ensure efficient query processing we have chosen an indexed-based approach. A corpus, encoded in a variety of external formats, e.g., NEGRA/TIGER format, bracketing format or an XML treebank format (Mengel & Lezius, 2000; Lezius et al., 2002a) is imported and indexed. Many partial searches are performed during indexing in order to save processing time during query processing. The indexing of a corpus is realized in a tool called TIGERRegistry, the corpus query tool is called TIGERSearch. To increase performance we have also implemented query optimization strategies and search space filters. The query processing strategy is described in detail in (Lezius, 2002).

In order to facilitate corpus exploration, a combination of searching for phenomena and browsing through a corpus, we have developed a sophisticated graphical user interface for the both the TIGER-

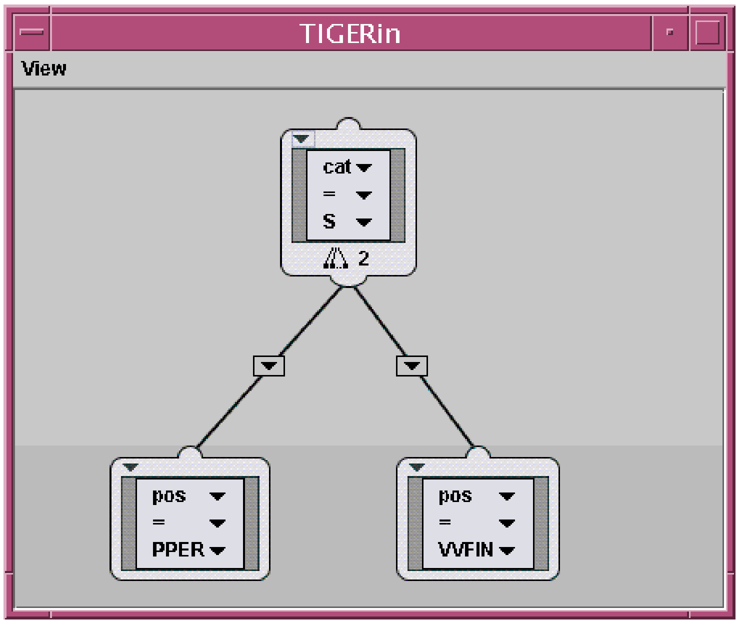

Figure 11: Graphical query input

Registry and the TIGERSearch tool. We have also developed special corpus visualization and query input strategies.

The TIGERSearch GUI comprises a graph viewer to view the matching sentences of a query. Figure 10 illustrates the visualization of a corpus graph that matches the example query above. Users can browse through the forest of matching sentences using a navigation bar and export their favourite matches. Matches can be exported in the TIGER XML format, but also as an interactive SVG image. Thus, match forests can be viewed in a format that does not depend on the TIGERSearch software suite.

We have also developed a graphical query input front-end which enables users to 'draw' queries in a very intuitive way (Voorman, 2002). Queries are expressed by combining nodes and node relations. Figure 11 illustrates how the example query above can be expressed using the graphical query editor.

6 Summary and outlook

In this paper, we presented the TIGER treebank, the largest and most comprehensive treebank for the German language. We explained the different levels of annotation: part-of-speech tags, phrase categories and syntactic functions. Furthermore, information about lemmata and morphology is also encoded in the corpus. The methods of annotation - interactive annotation with *Annotate* and LFG parsing - as well as the different representation formats used for the TIGER treebank were demonstrated. We also gave a short overview of related work in comparable treebank projects.

Moreover, we also described the TIGER annotation scheme, which is based on the annotation scheme used for the NEGRA corpus. The paper also outlined the most important extensions in the TIGER annotation scheme, which concern the use of secondary edges in coordinations, verb-subcategorization, finer distinctions concerning the German expletive *es* and a different treatment of structured proper nouns.

The last section presented TIGERSearch, a query tool that was developed in the project and which can be used to exploit the TIGER treebank and several other treebank formats. We explained the query language, that was designed to pose intuitive queries to the treebank. We also shortly introduced the graphical query input TIGERin.

Future work will be concerned with the extension of the TIGER treebank to approximately 80.000 sentences altogether and additional improvements in the annotation scheme. For instance, we invision the introduction of further distinctions concerning verbal arguments in order to facilitate the identification of thematic roles (Smith, 2000). For the beginning of 2003, a first release is planned which will contain 10.000 sentences completely annotated (part-of-speech tagging, syntactic structure, morphology, lemmata) according to the new TIGER annotation scheme and thoroughly checked for consistency.

All the tools presented in this paper are freely available for research purposes. For further information on the corpus, the corpus tools and how to obtain them, please refer to the project web page:

`http://www.coli.uni-sb.de/cl/projects/tiger`.

References

Abeillé, A., Brants, T., & Uszkoreit, H. (Eds.). (2000). *Proceedings of the COLING-2000 Post-Conference Workshop on Linguistically Interpreted Corpora LINC-2000.* Luxembourg.

Abeillé, A., Clément, L., & Kinyon, A. (2000). Building a treebank for French. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000* (pp. 87 – 94). Athens, Greece.

Boguslavsky, I., Grigorieva, S., Grigoriev, N., Kreidlin, L., & Frid, N. (2000). Dependency treebank for Russian: Concept, tools, types of information. In *18th International Conference on Computational Linguistics COLING-2000.* Saarbrücken, Germany.

Bosco, C., Lombardo, V., Vassallo, D., & Lesmo, L. (2000). Building a treebank for Italian: A data-driven annotation schema. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000* (pp. 99 – 106). Athens, Greece.

Brants, S., & Hansen, S. (2002). Developments in the TIGER annotation scheme and their realization in the corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation LREC-02.* Las Palmas de Gran Canaria, Spain.

Brants, T. (1997). *The NEGRA Export Format* (CLAUS Report No. 98). Saarbrücken, Germany: Dept. of Computational Linguistics, Saarland University.

Brants, T. (1999). *Tagging and parsing with Cascaded Markov Models - automation of corpus annotation.* Saarbrücken, Germany: German Research Center for Artificial Intelligence and Saarland University: Saarbrücken Dissertations in Computational Linguistics and Language Technology (Bd. 6).

Brants, T. (2000a). Inter-annotator agreement for a German newspaper corpus. In *Proceedings of Second International Conference on Language Resources and Evaluation LREC-2000.* Athens, Greece.

Brants, T. (2000b). TnT – A Statistical Part-of-Speech Tagger. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000.* Seattle, WA.

Brants, T., Hendriks, R., Kramp, S., Krenn, B., Preis, C., Skut, W., & Uszkoreit, H. (1999). *Das NEGRA-Annotationsschema* (Tech. Rep.). Saarbrücken, Germany: Dept. of Computational Linguistics, Saarland University.

Brants, T., & Skut, W. (1998). Automation of treebank annotation. In *Proceedings of New Methods in Language Processing NeMLaP-98.* Sydney, Australia.

Brants, T., Skut, W., & Uszkoreit, H. (1999). Syntactic annotation of a German newspaper corpus. In *Proceedings of the ATALA Treebank Workshop* (pp. 69–76). Paris, France.

Bresnan, J. (Ed.). (1982). *The Mental Representation of Grammatical Relations.* MIT Press.

Dipper, S. (2000). Grammar-based Corpus Annotation. In *Proceedings of LINC-2000* (pp. 56–64). Luxembourg.

Frank, A., King, T. H., Kuhn, J., & Maxwell, J. (1998). Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In *Proceedings of the LFG98 Conference.* Brisbane, Australia: CSLI Online Publications, http://www-csli.stanford.edu/publications.

Greenbaum, S. (Ed.). (1996). *Comparing English worldwide: The International Corpus of English.* Oxford, UK: Clarendon Press.

Hajic, J. (1999). Building a syntactically annotated corpus: The Prague Dependency Treebank. In E. Hajicova (Ed.), *Issues of valency and meaning. Studies in honour of Jarmila Panevova.* Prague, Czech Republic: Charles University Press.

King, T. H., Dipper, S., Frank, A., Kuhn, J., & Maxwell, J. (To Appear). Ambiguity Management in Grammar Writing. *Journal of Language and Computation.* (Special issue)

König, E., & Lezius, W. (2000). A description language for syntactically annotated corpora. In *Proceedings of COLING-2000* (pp. 1056–1060). Saarbrücken, Germany.

König, E., & Lezius, W. (2002a). *The TIGER language - A Description Language for Syntax Graphs. Part 1: User's Guidelines.* (Tech. Rep.). IMS, University of Stuttgart. (http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/papers/tigerLanguage.ps.gz)

König, E., & Lezius, W. (2002b). *The TIGER language - A Description Language for Syntax Graphs. Part 2: Formal Definition.* (Tech. Rep.). IMS, University of Stuttgart. (http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/papers/tigerLangForm.ps.gz)

Kramp, S., & Preis, C. (2000). *Konventionen für die Verwendung des STTS im NEGRA-Korpus* (Tech. Rep.). Saarbrücken, Germany: Dept. of Computational Linguistics, Saarland University.

Leech, G. (1992). The Lancaster Parsed Corpus. *ICAME Journal*, *16*(124).

Lezius, W. (2001). Baumbanken. In K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde, & H. Langer (Eds.), *Computerlinguistik und Sprachtechnologie - eine Einführung* (pp. 377 – 385). Heidelberg, Germany: Spektrum Akademischer Verlag.

Lezius, W. (2002). *Ein Werkzeug zur Suche auf syntaktisch annotierten Textkorpora.* IMS, University of Stuttgart. (PhD thesis, in preparation)

Lezius, W., Biesinger, H., & Gerstenberger, C. (2002a). *TIGER-XML Quick Reference Guide* (Tech. Rep.). IMS, University of Stuttgart.

Lezius, W., Biesinger, H., & Gerstenberger, C. (2002b). *TIGERRegistry Manual* (Tech. Rep.). IMS, University of Stuttgart.

Lezius, W., & König, E. (2000). Towards a search engine for syntactically annotated corpora. In *Proceedings of the Fifth KONVENS Conference.* Ilmenau, Germany.

Marcus, M., Kim, G., Marcinkiewicz, M., MacIntyre, R., Bies, A., Gerguson, M., Katz, K., & Schasberger, B. (1994). The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop.* San Francisco, CA: Morgan Kaufman.

Mengel, A., & Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of LREC-2000* (pp. 121–126). Athens, Greece.

Moreno, A., Grishman, R., López, S., Sánchez, F., & Sekine, S. (2000). A treebank of Spanish and its application to parsing. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC-2000* (pp. 107 – 112). Athens, Greece.

Oflazer, K., Hakkani-Tür, D., & Tür, G. (1999). Design for a Turkish treebank. In *Proceedings of the Workshop on Linguistically Interpreted Corpora LINC-99.* Bergen, Norway.

ParGram. (2002). *The ParGram Project.* (URL: http://www2.parc.com/istl/groups/nltt/pargram/)

Plaehn, O., & Brants, T. (2000). Annotate - an efficient interactive annotation tool. In *Proceedings of the Sixth Conference on Applied Natural Language Processing ANLP-2000.* Seattle, WA.

Riezler, S., King, T. H., Kaplan, R., Crouch, R., Maxwell, J., & Johnson, M. (2002). Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the ACL-02,.* Philadephia, PA.

Sampson, G. (1995). *English for the computer. The SUSANNE corpus and analytic scheme.* Oxford, UK: Clarendon Press.

Schiller, A., Teufel, S., & Stöckert, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS* (Tech. Rep.). University of Stuttgart, University of Tübingen.

Schrader, B. (2001). *Modifikation einer deutschen LFG-Grammatik für Partial Parsing.* Studienarbeit, University of Stuttgart.

Simov, K., Osenova, P., Slavcheva, M., 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 Third International Conference on Language Resources and Evaluation LREC-2002* (pp. 1729–1736). Las Palmas de Gran Canaria, Spain.

Skut, W., Brants, T., Krenn, B., & Uszkoreit, H. (1998). A linguistically interpreted corpus of German newspaper text. In *Proceedings of the Conference on Language Resources and Evaluation LREC-98* (pp. 705–711). Granade, Spain.

Skut, W., Krenn, B., Brants, T., & Uszkoreit, H. (1997). An annotation scheme for free word order languages. In *Proceedings of ANLP-97.* Washington, D.C.

Smith, G. (2000). Encoding thematic roles via syntactic categories in a German treebank. In *Proceedings of the Workshop on Syntactic Annotation of Electronic Corpora.* Tübingen, Germany.

Smith, G., & Eisenberg, P. (2000). *Kommentare zur Verwendung des STTS im NEGRA-Korpus* (Tech. Rep.). University of Potsdam.

Uszkoreit, H., Brants, T., & Krenn, B. (Eds.). (1999). *Proceedings of the Workshop on Linguistically Interpreted Corpora LINC-99.* Bergen, Norway.

Voorman, H. (2002). *TIGERin – Graphische Eingabe von Suchanfragen in TIGERSearch.* IMS, University of Stuttgart. (Diploma Thesis, in preparation)

Wahlster, W. (Ed.). (2000). *Verbmobil: Foundations of Speech-to-Speech Translation.* Heidelberg, Germany: Springer.

Zinsmeister, H., Kuhn, J., & Dipper, S. (2002). Utilizing LFG Parses for Treebank Annotation. In *Proceedings of the LFG02 Conference.* Athens, Greece.