Annotation Schemes and their Influence on Parsing Results

Wolfgang Maier
Seminar für Sprachwissenschaft, Universität Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
wmaier@sfs.uni-tuebingen.de

Abstract

Most of the work on treebank-based statistical parsing exclusively uses the Wall-Street-Journal part of the Penn treebank for evaluation purposes. Due to the presence of this quasi-standard, the question of to which degree parsing results depend on the properties of treebanks was often ignored. In this paper, we use two similar German treebanks, TüBa-D/Z and NeGra, and investigate the role that different annotation decisions play for parsing. For these purposes, we approximate the two treebanks by gradually taking out or inserting the corresponding annotation components and test the performance of a standard PCFG parser on all treebank versions. Our results give an indication of which structures are favorable for parsing and which ones are not.

1 Introduction

The Wall-Street-Journal part (WSJ) of the Penn Treebank (Marcus et al., 1994) plays a central role in research on statistical treebank-based parsing. It has not only become a standard for parser evaluation, but also the foundation for the development of new parsing models. For the English WSJ, high accuracy parsing models have been created, some of them using extensions to classical PCFG parsing such as lexicalization and markovization (Collins, 1999; Charniak, 2000; Klein and Manning, 2003). However, since most research has been limited to a single language (English) and to a single treebank (WSJ), the question of how portable the parsers and their extensions are across languages and across treebanks often remained open.

Only recently, there have been attempts to evaluate parsing results with respect to the properties and the language of the treebank that is used. Gildea (2001) investigates the effects that certain treebank characteristics have on parsing results, such as the distribution of verb subcategorization frames. He conducts experiments on the WSJ and the Brown Corpus, parsing one of the treebanks while having trained on the other one. He draws the conclusion that a small amount of matched training data is better than a large amount of unmatched training data. Dubey and Keller (2003) analyze the difficulties that German imposes on parsing. They use the NeGra treebank for their experiments and show that lexicalization, while highly effective for English, has no benefit for German. This result motivates them to create a parsing model for German based on sister-head-dependencies. Corazza et al. (2004) conduct experiments with model 2 of Collins' parser (Collins, 1999) and the Stanford parser (Klein and Manning, 2003) on two Italian treebanks. They report disappointing results which they trace back to the different difficulties of different parsing tasks in Italian and English and to differences in annotation styles across treebanks.

In the present paper, our goal is to determine the effects of different annotation decisions on the results of plain PCFG parsing without extensions. Our motivation is two-fold: first, we want to present research on a language different from English, second, we want to investigate the influences of annotation schemes via a realistic comparison, i.e. use two different annotation schemes. Therefore, we take advantage of the availability of two similar treebanks of German, TüBa-D/Z (Telljohann et al., 2003) and NeGra (Skut et al., 1997). The strategy we adopt extends Kübler

Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 19–24,
Sydney, July 2006. ©2006 Association for Computational Linguistics

(2005). Treebanks and their annotation schemes respectively are compared using a stepwise approximation. Annotation components corresponding to certain annotation decisions are taken out or inserted, submitting each time the resulting modified treebank to the parser. This method allows us to investigate the role of single annotation decisions in two different environments.

In section 2, we describe the annotation of both treebanks in detail. Section 3 introduces the methodology used. In section 4, we describe our experimental setup and discuss the results. Section 5 presents a conclusion and plans for future work.

2 The Treebanks: TüBa-D/Z and NeGra

With respect to treebanks, German is in a privileged position. Various treebanks are available, among them are two similar ones: NeGra (Skut et al., 1997), from Saarland University at Saarbrücken and TüBa-D/Z (Telljohann et al., 2003), from the University of Tübingen. NeGra contains about 20,000 sentences, TüBa-D/Z about 15,000, both consist of newspaper text. In both treebanks, predicate argument structure is annotated, the core principle of the annotation being its theory independence. Terminal nodes are labeled with part-of-speech tags and morphological labels, non-terminal nodes with phrase labels. All edges are labeled with grammatical functions. Annotation was accomplished semi-automatically with the same software tools.

The main difference between the treebanks is rooted in the partial free word order of German sentences: the positions of complements and adjuncts are of great variability. This leads to a high number of discontinuous constituents, even in short sentences. An annotation scheme for German must account for that. NeGra allows for crossing branches, thereby giving up the context-free backbone of the annotation. With crossing branches, discontinuous constituents are not a problem anymore: all children of every constituent, discontinuous or not, can always be grouped under the same node. The inconvenience of this method is that the crossing branches must be resolved before the treebank can be used with a (PCFG) parser. However, this can be accomplished easily by reattaching children of discontinuous constituents to higher nodes.

TüBa-D/Z uses another mechanism to account for the free word order. Above the phrase level, an additional layer of annotation is introduced. It consists of topological fields (Drach, 1937; Höhle, 1986). The concept of topological fields is widely accepted among German grammarians. It reflects the empirical observation that German has three possible sentence configurations with respect to the position of the finite verb. In its five fields (initial field, left sentence bracket, middle field, right sentence bracket, final field), verbal material generally resides in the two sentence brackets, while the initial field and the middle field contain all other elements. The final field contains mostly extraposed material. Since word order variations generally do not cross field boundaries, with the model of topological fields, the free word order of German can be accounted for in a natural way.

On the phrase level, the treebanks show great differences, too. NeGra does not allow for any intermediate ("bar") phrasal projections. Additionally, no unary productions are allowed. This results in very flat phrases: pre- and postmodifiers are attached directly to the phrase, nominal subjects are attached directly to the sentence, nominal material within PPs doesn't project to NPs, complex (non-coordinated) NPs remain flat. TüBa-D/Z, on the contrary, allows for "deep" annotation. Intermediate productions and unary productions are allowed and extensively used.

To illustrate the annotation principles, the figures 1 and 2 show the annotation of the sentences (1) and (2) respectively.

(1) Darüber muß nachgedacht werden.
About-that must tought be
'This must be tought about.'

(2) Schillen wies dies gestern zurück:
Schillen rejected that yesterday VPART
'Schillen rejected that yesterday.'

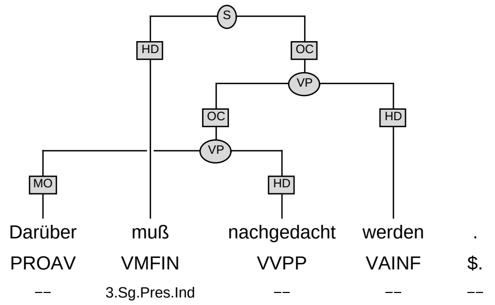

Figure 1: A NeGra tree

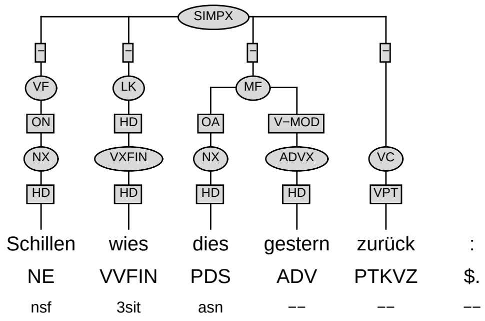

Figure 2: A TüBa-D/Z tree

3 Treebanks, Parsing, and Comparisons

Our goal is to determine which components of the annotation schemes of TüBa-D/Z and NeGra have which influence on parsing results. A direct comparison of the parsing results shows that the TüBa-D/Z annotation scheme is more appropriate for PCFG parsing than NeGra's (see tables 2 and 3). However, this doesn't tell us anything about the role of the subparts of the annotation schemes.

A first idea for a more detailed comparison could be to compare the results for different phrase types. The problem is that this would not give meaningful results. NeGra noun phrases, e.g., cover a different set of constituents than TüBa-D/Z noun phrases, due to NeGra's flat annotation and avoidance of annotation of unary NPs. Furthermore, both annotation schemes contain categories not contained in the other one. There are, e.g., no categories in NeGra that correspond to TüBa-D/Z's field categories, while in TüBa-D/Z, there are no categories equivalent to NeGra's categories for coordinated phrases or verb phrases.

We therefore pursue another approach. We use a method introduced by Kübler (2005) to investigate the usefulness of different annotation components for parsing. We gradually modify the treebank annotations in order to approximate the annotation style of the treebanks to one another. This is accomplished by taking out or inserting certain components of the annotation. For our treebanks, this generally results in reduced structures for TüBa-D/Z and augmented structures for NeGra. Table 1 presents three measures that capture the changes between each of the modifications. The average number of child nodes of nonterminal nodes shows the degree of flatness of the annotation on phrase level. Here, the unmodified NeGra consequently shows the highest values. The average tree height relates directly to the number of annotation hierarchies in the tree. Here, the unmodified TüBa-D/Z has the highest values.

4 Experimental Setup

For our experiments, we use `lopar` (Schmid, 2000), a standard PCFG parser. We read the grammar and the lexicon directly off the trees together with their frequencies. The parser is given the gold POS tagging to avoid parsing errors that are caused by wrong POS tags. Only sentences up to a length of 40 words are considered due to memory limitations.

Traditionally, most of the work on WSJ uses the same section of the treebank for testing. However, for our aims, this method has a shortcoming: since both treebanks consist of text created by different authors, linguistic phenomena are not evenly distributed over the treebank. When using a whole section as test set, some phenomena may only occur there and thus not occur in the grammar. To reduce data sparseness, we use another test/training-set split for the treebanks and their variations. Each 10th sentence is put into the test set, all other sentences go into the training set.

4.1 Preprocessing the Treebanks

Since we want to read the grammars for our parser directly off the treebanks, preprocessing of the treebanks is necessary due to the non-context-free nature of the original annotation. In both treebanks, punctuation is not included in the trees, furthermore, sentence splitting in both treebanks does not always coincide with the linguistic notion of a sentence. This leads to sentences consisting of several unconnected trees. All nodes in a sentence, i.e. the roots and the punctation, are grouped by a virtual root node, which may cause crossing branches. Furthermore, the NeGra annotation scheme allows for crossing branches for linguistic reasons, as described in section 2. All of the crossing branches have to be removed before parsing.

The crossing branches caused by the NeGra annotation scheme are removed with a small program by Thorsten Brants. It attaches some of the children of discontinuous constituents to higher nodes. The virtual root node is made continuous by attaching all punctuation to the highest possible location in the tree. Pairs of parenthesis and quotation marks are preferably attached to

	NeGra	*NE_fi.*	*NE_NP*	*NE_tr.*	TüBa	Tü_NF	Tü_NU	Tü_f	Tü_f_NU	Tü_f_NU_NF
N/T	0.41	0.70	0.50	0.41	1.21	0.89	0.54	1.00	0.42	0.35
μ D/N	2.92	2.22	2.59	2.92	1.61	1.89	2.53	1.83	2.93	3.35
μ H(T)	4.86	5.81	5.16	4.68	6.88	5.68	5.45	5.94	4.72	4.15

Table 1: Properties of the treebank modifications[1]

the same node, to avoid low-frequent productions in the grammar that only differ by the position of parenthesis marks on their right hand side.

4.2 Results of the Comparison

We use the standard parseval measures for the evaluation of parser output. They measure the percentage of correctly parsed constituents, in terms of precision, recall, and F-Measure. The parser output of each modified treebank version is evaluated against the correspondingly modified test set. Unparsed sentences are fully included in the evaluation.

NeGra. Along with the unmodified treebank, two modifications of NeGra are tested. Both of them introduce annotation components present in TüBa-D/Z but not in NeGra. In the first one, *NE_fi*, we add an annotation layer of **topological fields**[2], as existing in TüBa-D/Z. The precision value benefits the most from this modification. When parsing without grammatical functions, it increases about 6,5%. When parsing with grammatical functions, it increases about 14%. Thus, the additional rules provided by a topological field level that groups phrases below the clausal level are favorable for parsing. The average number of crossing brackets per sentence increases, which is due to the fact that there are simply more brackets to create.

A detailed evaluation of the results for node categories shows that the new field categories are easy to recognize (e.g. LF gets 97.79 F-Measure). Nearly all categories have a better precision value. However, the F-Measure for VPs is low (only 26.70 while 59.41 in the unmodified treebank), while verb phrases in the unmodified TüBa-D/Z (see below) are recognized with nearly 100 points F-Measure. The problem here is the following. In the original NeGra annotation, a verb and its complements are grouped under the same VP. To preserve as much of the annotation as possible, the topological fields are inserted *below* the VP (complements are grouped by a middle field node, the verb complex by the right sentence bracket). Since this way, the phrase node VP resides above the field level, it becomes difficult to recognize.

In the second modification, *NE_NP*, we approximate NeGra's PPs to TüBa-D/Z's by grouping all **nominal material below the PPs to separate NPs**. This modification gives us a small benefit in terms of precision and recall (about 2-3%). Although there are more brackets to place, the number of crossing parents increases only slightly, which can be attributed to the fact that below PPs, there is no room to get brackets wrong.

We finally parse a version of NeGra where for each node movement during the resolution of crossing edges, a **trace label** was created in the corresponding edge (*NE_tr*). Although this brings the treebank closer to the format of TüBa-D/Z, the results get even worse than in the version without traces. However, the high number of unparsed sentences indicates that the result is not reliable due to data sparseness.

	NeGra	*NE_fi.*	*NE_NP*	*NE_tr.*
	without grammatical functions			
cross. br.	1.10	1.67	1.14	—
lab. prec.	68.14%	74.96%	70.43%	—
lab. rec.	69.98%	70.37%	72.81%	—
lab. F_1	69.05	72.59	71.60	—
not parsed	1.00%	0.10%	0.15%	—
	with grammatical functions			
cross. br.	1.10	1.21	1.27	1.05
lab. prec.	52.67%	67.90%	59.77%	51.81%
lab. rec.	52.17%	65.18%	60.36%	49.19%
lab. F_1	52.42	66.51	60.06	50.47
not parsed	12.90%	1.66%	9.88%	16.01%

Table 2: Parsing NeGra: Results

TüBa-D/Z. Apart from the original treebank, we test six modifications of TüBa-D/Z. In each of the modifications, annotation material is removed in order to obtain NeGra-like structures. Since they are equally absent in NeGra, we delete the annotation of **topological fields** in the first modification, *Tü_NF*. This results in small losses.

[1] explanation: N/T = node/token ratio, μ D/N = average number of daughters of non-terminal nodes, μ H(T) = average tree height

[2] We are grateful to the DFKI Saarbrücken for providing us with the topological field annotation.

	TüBa	Tü_NF	Tü_NU	Tü_flat	Tü_f_NU	Tü_f_NU_NF
	without grammatical functions					
crossing brackets	2.21	1.82	1.67	1.04	0.80	1.03
labeled precision	87.39%	86.31%	79.97%	86.22%	75.18%	63.05%
labeled recall	83.57%	83.43%	78.52%	85.41%	76.11%	66.86%
labeled F-Measure	85.44	84.85	79.24	85.81	75.64	64.90
not parsed	0.07%	0.07%	2.45%	0.07%	2.99%	6.87%
	with grammatical functions					
crossing brackets	1.84	1.82	1.79	0.98	1.01	1.12
labeled precision	76.99%	68.55%	63.71%	76.93%	58.91%	45.15%
labeled recall	75.30%	68.40%	62.79%	77.21%	58.92%	44.76%
labeled F-Measure	76.14	68.47	63.25	77.07	58.92	44.96
not parsed	0.07%	0.27%	4.49%	0.07%	7.21%	17.76%

Table 3: Parsing TüBa-D/Z: Results

A closer look at category results shows that losses are mainly due to categories on the clausal level; structures within fields do not deteriorate. Field categories are thus especially helpful for the clausal level.

In the second modification of TüBa-D/Z, *Tü_NU*, **unary nodes** are collapsed with the goal to get structures comparable to NeGra's. As the figures show, the unary nodes are very helpful, the F-Measure drops about 6 points without them. The number of crossing brackets also drops, along with the total number of nodes. When parsing with grammatical functions, taking out unary productions has a detrimental effect, F-Measure drops about 13 points. A plausible explanation could be data sparseness. 32.78% of the rules that the parser needs to produce a correct parse don't occur in the training set.

An evaluation of the results for the different categories shows that all major phrase categories loose both in precision and recall. Since field nodes are mostly unary, many of them disappear, but most of the middle field nodes stay because they generally contain more than one element. However, their recall drops about 10%. Supposedly it is more difficult for the parser to annotate the middle field “alone” without the other field categories.

We also test a version of TüBa-D/Z with **flattened phrases** that mimic NeGra's flat phrases, *Tü_flat*. With this treebank version, we get results very similar to those of the unmodified treebank. The F-Measure values are slightly higher and the parser produces less crossing brackets. A single category benefits the most from this treebank modification: EN-ADD, its F-Measure rising about 45 points. It was originally introduced as a marker for named entities, which means that it has no specific syntactic function. In the TüBa-D/Z version with flattened phrases, many of the nominal nodes below EN-ADD are taken out, bringing EN-ADD closer to the lexical level. This way, the category has more meaningful context and therefore produces better results.

Furthermore, we test combinations of the modifications. Apart from the average tree height, the dimensions of TüBa-D/Z with **flattened phrases and without unary productions** (*Tü_f_NU*) resemble those of the unmodified NeGra treebank, which indicates their similarity. Nevertheless, parser results are worse on NeGra. This indicates that TüBa-D/Z still benefits from the remaining field nodes. The number of crossing branches is the lowest in this treebank version.

In the last modification that **combines all modifications** made before (*TÜ_f_NU_NF*), as expected, all values drop dramatically. F-Measure is about 5 points worse than with the unmodified NeGra treebank.

POS tagging. In a second round, we investigate the benefits that gold POS tags have when making them available in the parser input. We repeat all experiments without giving the parser the perfect tagging.

This leads to higher time and space requirements during parsing, caused by the additional tagging step. With TüBa-D/Z, NeGra, and all their modifications, the F-Measure results are about 3-5 points worse when parsing with grammatical functions. When parsing without them, they drop 3-6 points. We can determine two exceptions: TüBa-D/Z with flattened phrases, where the F-Score drops more than 9 points when parsing with grammatical functions, and the TüBa-D/Z version with all modifications combined, where F-Score drops only a little less than 2 points. The behavior of the flattened TüBa-D/Z relates directly to the fact that the categories that loose the most without gold POS tags are phrase categories (particularly infinite VPs and APs). They are directly conditioned on the POS tagging and thus behave accordingly to its quality. For the TüBa-D/Z version with all modifications combined, one could argue that the results are not reliable because of data sparseness, which is confirmed by the high number of unparsed sentences in this treebank version. However, in all cases, less crossing brackets are produced.

To sum up, obviously, it is more difficult for the parser to build a parse tree onto an already existing layer of POS-tagging. This explains the bigger number of unparsed sentences. Nevertheless, in terms of F-Score, the parsing results profit visibly from the gold POS tagging.

5 Conclusions and Outlook

We presented an analysis of the influences of the particularities of annotation schemes on parsing results via a comparison of two German treebanks, NeGra and TüBa-D/Z, based on a stepwise approximation of both treebanks. The experiments show that as treebanks are approximated, the parsing results also get closer. When annotation structure is deleted in TüBa-D/Z, the number of crossing brackets drops, but F-Measure drops, too. When annotation structure is added in NeGra, the contrary happens. We can conclude that, being interested in good F-Measure results, the deep TüBa-D/Z structures are more appropriate for parsing than NeGra's flat structures. Moreover, we have observed that it is beneficial to provide the parser with the gold POS tags at parsing time. However, we see that especially when parsing with grammatical functions, data sparseness becomes a serious problem, making the results less reliable.

Seen in the context of a parse tree, the expansion probability of a PCFG rule just covers a subtree of height 1. This is a clear deficiency of PCFGs since this way, e.g., the expansion probability of a VP is independent of the choice of the verb. Our future work will start at this point. We will conduct further experiments with the Stanford Parser (Klein and Manning, 2003) which considers broader contexts in its probability. It uses markovization to reduce horizontal context (right hand sides of rules are broken up) and add vertical context (rule probabilities are conditioned on (grand-)parent-node information). This way, we expect further insights in NeGra's an TüBa-D/Z's annotation schemes.

References

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proceedings of NAACL 2000*.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Anna Corazza, Alberto Lavelli, Giorgio Satta, and Roberto Zanoli. 2004. Analyzing an Italian treebank with state-of-the-art statistical parsers. In *Proceedings of the 3^{rd} Workshop on Treebanks and Linguistic Theories (TLT 2004)*.

Erich Drach. 1937. *Grundgedanken der deutschen Satzlehre*. Diesterweg, Frankfurt/Main.

Amit Dubey and Frank Keller. 2003. Probabilistic parsing for German using sisterhead dependencies. In *Proceedings of ACL 2003*.

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of EMNLP 2001*.

Tilman Höhle. 1986. Der Begriff "Mittelfeld", Anmerkungen ber die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*, Göttingen, Germany.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL 2003*.

Sandra Kübler. 2005. How do treebank annotation schemes influence parsing results? Or how not to compare apples and oranges. In *Proceedings of RANLP 2005*.

Mitchell P. Marcus, Grace Kim, Marry Ann Marcinkiewicz, Robert MacIntyre, Ann Biew, Mark Freguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *Proceedings of the 1994 Human Language Technology Workshop, HLT 94, Plainsboro, NJ*.

Helmut Schmid. 2000. LoPar: Design and implementation. Technical report, Universität Stuttgart, Germany.

Wojciech Skut, Brigitte Krenn, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. In *Proceedings of ANLP 1997*.

Heike Telljohann, Erhard W. Hinrichs, and Sandra Kübler, 2003. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Germany.