# Parsing Coordinations

**Sandra Kübler**
Indiana University
skuebler@indiana.edu

**Erhard Hinrichs**
Universität Tübingen
eh@sfs.uni-tuebingen.de

**Wolfgang Maier**
Unversität Tübingen
wo.maier@uni-tuebingen.de

**Eva Klett**
Universität Tübingen
eklett@sfs.uni-tuebingen.de

## Abstract

The present paper is concerned with statistical parsing of constituent structures in German. The paper presents four experiments that aim at improving parsing performance of coordinate structure: 1) reranking the $n$-best parses of a PCFG parser, 2) enriching the input to a PCFG parser by gold scopes for any conjunct, 3) reranking the parser output for all possible scopes for conjuncts that are permissible with regard to clause structure. Experiment 4 reranks a combination of parses from experiments 1 and 3.

The experiments presented show that $n$-best parsing combined with reranking improves results by a large margin. Providing the parser with different scope possibilities and reranking the resulting parses results in an increase in F-score from 69.76 for the baseline to 74.69. While the F-score is similar to the one of the first experiment ($n$-best parsing and reranking), the first experiment results in higher recall (75.48% vs. 73.69%) and the third one in higher precision (75.43% vs. 73.26%). Combining the two methods results in the best result with an F-score of 76.69.

## 1 Introduction

The present paper is concerned with statistical parsing of constituent structures in German. German is a language with relatively flexible phrasal ordering, especially of verbal complements and adjuncts. This makes processing complex cases of coordination particularly challenging and error-prone. The paper presents four experiments that aim at improving parsing performance of coordinate structures: the first experiment involves reranking of $n$-best parses produced by a PCFG parser, the second experiment enriches the input to a PCFG parser by offering gold pre-bracketings for any coordinate structures that occur in the sentence. In the third experiment, the reranker is given all possible pre-bracketed candidate structures for coordinated constituents that are permissible with regard to clause macro- and microstructure. The parsed candidates are then reranked. The final experiment combines the parses from the first and the third experiment and reranks them. Improvements in this final experiment corroborate our hypothesis that forcing the parser to work with pre-bracketed conjuncts provides parsing alternatives that are not present in the $n$-best parses.

Coordinate structures have been a central issue in both computational and theoretical linguistics for quite some time. Coordination is one of those phenomena where the simple cases can be accounted for by straightforward empirical generalizations and computational techniques. More specifically, it is the observation that coordination involves two or more constituents of the same categories. However, there are a significant number of more complex cases of coordination that defy this generalization and that make the parsing task of detecting the right scope of individual conjuncts and correctly delineating the correct scope of the coordinate structure as a whole difficult. (1) shows some classical examples of this kind from English.

(1) a. Sandy is a Republican and proud of it.
b. Bob voted, but Sandy did not.
c. Bob supports him and Sandy me.

In (1a), unlike categories (NP and adjective) are conjoined. (1b) and (1c) are instances of ellipsis (VP ellipsis and gapping). Yet another difficult set of examples present cases of non-constituent conjunction, as in (2), where the direct and indirect object of a ditransitive verb are conjoined.

(2) Bob gave a book to Sam and a record to Jo.

## 2 Coordination in German

The above phenomena have direct analogues in German.[1] Due to the flexible ordering of phrases, their variability is even higher. For example, due to constituent fronting to clause-initial position in German verb-second main clauses, cases of non-constituent conjunction can involve any two NPs (including the subject) of a ditransitive verb to the exclusion of the third NP complement that appears in clause-initial position. In addition, German exhibits cases of asymmetric coordination first discussed by Höhle (1983; 1990; 1991) and illustrated in (3).[2]

(3) In den Wald ging ein Jäger und
Into the woods went a hunter and
schoss einen Hasen.
shot a hare.

Such cases of subject gap coordination are frequently found in text corpora (cf. (4) below) and involve conjunction of a full verb-second clause with a VP whose subject is identical to the subject in the first conjunct.

## 3 Experimental Setup and Baseline

### 3.1 The Treebank

The data source used for the experiments is the Tübingen Treebank of Written German (TüBa-D/Z) (Telljohann et al., 2005). TüBa-D/Z uses the newspaper 'die tageszeitung' (taz) as its data source, version 3 comprises approximately 27 000 sentences. The treebank annotation scheme distinguishes four levels of syntactic constituency: the lexical level, the phrasal level, the level of topological fields, and the clausal level. The primary ordering principle of a clause is the inventory of topological fields (VF, LK, MF, VC, and NF), which characterize the word order regularities among different clause types of German. TüBa-D/Z annotation relies on a context-free backbone (i.e. proper trees without crossing branches) of phrase structure combined with edge labels that specify the grammatical function of the phrase in question. Conjuncts are generally marked with the function label KONJ. Figure 1 shows the annotation that sentence (4) received in the treebank. Syntactic categories are displayed as nodes, grammatical functions as edge labels in gray (e.g. OA: direct object, PRED: predicate). This is an example of a subject-gap coordination, in which both conjuncts (FKONJ) share the subject (ON) that is realized in the first conjunct.

(4) Damit hat sich der Bevölkerungs-
So has itself the decline in
rückgang zwar abgeschwächt, ist
population though lessened, is
aber noch doppelt so groß wie 1996.
however still double so big as 1996.
'For this reason, although the decline in population has lessened, it is still twice as big as in 1996.'

The syntactic annotation scheme of the TüBa-D/Z is described in more detail in Telljohann et al. (2004; 2005).

All experiments reported here are based on a data split of 90% training data and 10% test data.

### 3.2 The Parsers and the Reranker

Two parsers were used to investigate the influence of scope information on parser performance on coordinate structures: BitPar (Schmid, 2004) and LoPar (Schmid, 2000). BitPar is an efficient implementation of an Earley style parser that uses bit vectors. However, BitPar cannot handle pre-bracketed input. For this reason, we used LoPar for the experiments where such input was required. LoPar, as it is used here, is a pure PCFG parser, which allows the input to be partially bracketed. We are aware that the results that can be obtained by pure PCFG parsers are not state of the art as reported in the shared task of the ACL 2008 Workshop on Parsing German (Kübler, 2008). While BitPar reaches an F-score of 69.76 (see next section), the best performing parser (Petrov and Klein, 2008) reaches an F-score of 83.97 on TüBa-D/Z (but with a different split of training and test data). However, our experiments require certain features in the parsers, namely the capability to provide $n$-best analyses and to parse pre-bracketed input. To our knowledge, the parsers that took part in the shared task do not provide these features. Should they become available, the methods presented here could be applied to such parsers. We see no reason why our

[1] To avoid having to gloss German examples, they were illustrated for English.

[2] Yet, another case of such asymmetric coordination discussed by Höhle involves cases of conjunction of different clause types: [$_{V-final}$ *Wenn du nach Hause kommst* ] und [$_{V-2nd}$ *da warten Polizeibeamte vor der Tür.* 'If you come home and there are policemen waiting in front of the door ] .'

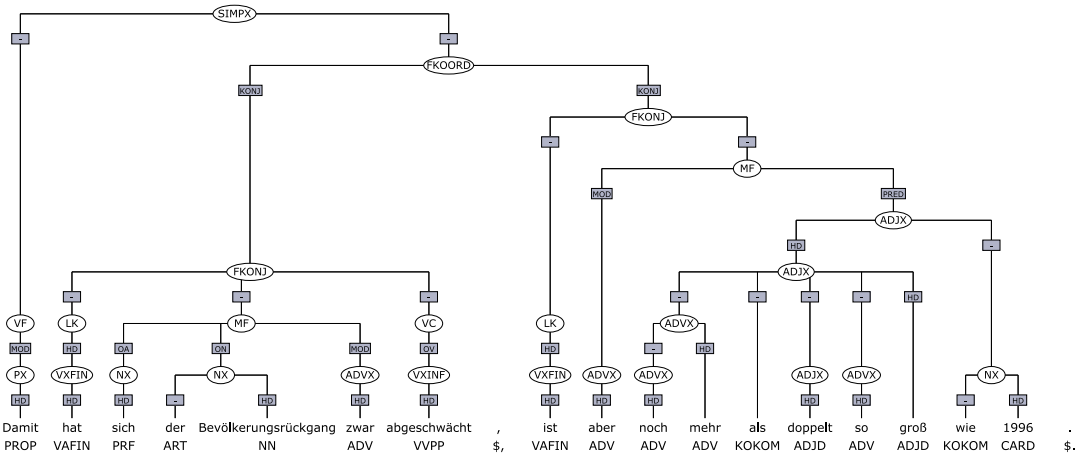


Figure 1: A tree with coordination.

methods should not be able to improve the results of these parsers further.

Since we are interested in parsing coordinations, all experiments are conducted with gold POS tags, so as to abstract away from POS tagging errors. Although the treebank contains morphological information, this type of information is not used in the experiments presented here.

The reranking experiments were conducted using the reranker by Collins and Koo (2005). This reranker uses a set of candidate parses for a sentence and reranks them based on a set of features that are extracted from the trees. The reranker uses a boosting method based on the approach by Freund et al. (1998). We used a similar feature set to the one Collins and Koo used; the following types of features were included: rules, bigrams, grandparent rules, grandparent bigrams, lexical bigrams, two-level rules, two-level bigrams, trigrams, head-modifiers, PPs, and distance for head-modifier relations, as well as all feature types involving rules extended by closed class lexicalization. For a more detailed description of the rules, the interested reader is referred to Collins and Koo (2005). For coordination, these features give a wider context than the original parser has and should thus result in improvements for this phenomenon.

### 3.3 The Baseline

When trained on 90% of the approximately 27,000 sentences of the TüBa-D/Z treebank, BitPar reaches an F-Score of 69.73 (precision: 68.63%, recall: 70.93%) on the full test set of 2611 sentences. These results as well as all further results presented here are labeled results, including grammatical functions. Since German has a relatively free word order, it is impossible to deduce the grammatical function of a noun phrase from the configuration of the sentence. Consequently, an evaluation based solely on syntactic constituent labels would be meaningless (cf. (Kübler, 2008) for a discussion of this point). The inclusion of grammatical labels in the trees, makes the parsing process significantly more complex.

Looking at sentences with coordination (i.e. sentences that contain a conjunction which is not in sentence-initial position), we find that 34.9% of the 2611 test sentences contain coordinations. An evaluation of only sentences with coordination shows that there is a noticeable difference: the F-score reaches 67.28 (precision: 66.36%, recall: 68.23%) as compared to 69.73 for the full test set.

The example of a wrong parse shown below illustrates why parsing of complex coordinations is so hard. Complex coordinations can take up a considerable part of the input string and accordingly of the overall sentence structure. Such global phenomena are particularly hard for pure PCFG parsing, due to the independence assumption inherent in the statistical models for PCFGs.

Sentence (4) has the following Viterbi parse:

```
(VROOT
  (SIMPX
    (VF
      (SIMPX-OS
        (VF (PX-MOD (PROP-HD Damit)))
        (LK
          (VXFIN-HD (VAFIN-HD hat)))
        (MF
```

```
        (NX-OA (PRF-HD sich))
        (NX-ON (ART der)
          (NN-HD Bevölkerungsrückgang))
        (ADVX-MOD (ADV-HD zwar)))
      (VC (VXINF-OV
          (VVPP-HD abgeschwächt)))))
  ($, ,)
  (LK
    (VXFIN-HD (VAFIN-HD ist)))
  (MF
    (ADVX-MOD (ADV-HD aber))
    (ADVX-MOD (ADV-HD noch))
    (ADJX-PRED
      (ADJX-HD (ADVX (ADV-HD mehr))
        (ADJX (KOKOM als)
          (ADJD-HD doppelt))
        (ADVX (ADV-HD so))
        (ADJD-HD groß))
      (NX (KOKOM wie)
        (CARD-HD 1996)))))
($. .))
```

The parse shows that the parser did not recognize the coordination. Instead, the first conjunct including the fronted constituent, `Damit hat sich der Bevölkerungsrückgang zwar abgeschwächt`, is treated as a fronted subordinate clause.

## 4 Experiment 1: n-Best Parsing and Reranking

The first hypothesis for improving coordination parsing is based on the assumption that the correct parse may not be the most probable one in Viterbi parsing but may be recovered by $n$-best parsing and reranking, a technique that has become standard in the last few years. If this hypothesis holds, we should find the correct parse among the $n$-best parses. In order to test this hypothesis, we conducted an experiment with BitPar (Schmid, 2004). We parsed the test sentences in a 50-best setting.

A closer look at the 50-best parses shows that of the 2611 sentences, 195 (7.5%) were assigned the correct parse as the best parse. For 325 more sentences (12.4%), the correct parse could be found under the 50 best analyses. What is more, in 90.2% of these 520 sentences, for which the correct parse was among the 50 best parses, the best parse was among the first 10 parses. Additionally, only in 4 cases were the correct analyses among the 40-best to 50-best parses, an indication that increasing $n$ may not result in improving the results significantly. These findings resulted in the decision not to conduct experiments with higher $n$.

That the 50 best analyses contain valuable information can be seen from an evaluation in which an oracle chooses from the 50 parses. In this case, we reach an F-score of 80.28. However, this F-score is also the upper limit for improvement that can be achieved by reranking the 50-best parses.

For reranking, the features of Collins and Koo (2005) were extended in the following way: Since the German treebank used for our experiments includes grammatical function information on almost all levels in the tree, all feature types were also included with grammatical functions attached: All nodes except the root node of the subtree in question were annotated with their grammatical information. Thus, for the noun phrase (NX) rule with grandparent prepositional phrase (PX) $\texttt{PX}_{GP}\ \texttt{NX} \rightarrow \texttt{ART ADJX NN}$, we add an additional rule $\texttt{PX}_{GP}\ \texttt{NX-HD} \rightarrow \texttt{ART ADJX NN-HD}$.

After pruning all features that occurred in the training data with a frequency lower than 5, the extractions produced more than 5 mio. different features. The reranker was optimized on the training data, the 50-best parses were produced in a 5-fold cross-validation setting. A non-exhaustive search for the best value for the $\alpha$ parameter showed that Collins and Koo's value of 0.0025 produced the best results. The row for exp. 1 in Table 1 shows the results of this experiment. The evaluation of the full data set shows an improvement of 4.77 points in the F-score, which reached 74.53. This is a relative reduction in error rate of 18.73%, which is slightly higher that the error rate reduction reported by Collins and Koo for the Penn Treebank (13%). However, the results for Collins and Koo's original parses were higher, and they did not evaluate on grammatical functions.

The evaluation of coordination sentences shows that such sentences profit from reranking to the same degree. These results prove that while coordination structures profit from reranking, they do not profit more than other phenomena. We thus conclude that reranking is no cure-all for solving the problem of accurate coordination parsing.

## 5 Experiment 2: Gold Scope

The results of experiment 1 lead to the conclusion that reranking the $n$-best parses can only result in restricted improvements on coordinations. The fact that the correct parse often cannot be found in the 50-best analyses suggests that the different possible scopes of a coordination are so different in their probability distribution that not all of the possible scopes are present in the 50-best analyses.

| | | all sentences | | | coord. sentences | | |
|---|---|---|---|---|---|---|---|
| | | precision | recall | F-score | precision | recall | F-score |
| | baseline: | 68.63 | 70.93 | 69.76 | 66.36 | 68.23 | 67.28 |
| exp. 1: | 50-best reranking: | 73.26 | 75.84 | 74.53 | 70.67 | 72.72 | 71.68 |
| exp. 2: | with gold scope: | 76.12 | 72.87 | 74.46 | 75.78 | 72.22 | 73.96 |
| exp. 3: | automatic scope: | 75.43 | 73.96 | 74.69 | 72.88 | 71.42 | 72.14 |
| exp. 4: | comb. 1 and 3: | 76.15 | 77.23 | 76.69 | 73.79 | 74.73 | 74.26 |

Table 1: The results of parsing all sentences and coordinated sentences only

If this hypothesis holds, forcing the parser to consider the different scope readings should increase the accuracy of coordination parsing. In order to force the parser to use the different scope readings, we first extract these scope readings, and then for each of these scope readings generate a new sentence with partial bracketing that represents the corresponding scope (see below for an example). LoPar is equipped to parse partially-bracketed input. Given input sentences with partial brackets, the parser restricts analyses to such cases that do not contradict the brackets in the input.

(5) Was stimmt, weil sie
Which is correct, because they
unterhaltsam sind, aber auch falsche
entertaining are, but also wrong
Assoziationen weckt.
associations wakes.
'Which is correct because they are entertaining, but also triggers wrong associations.'

In order to test the validity of this hypothesis, we conducted an experiment with coordination scopes extracted from the treebank trees. These scopes were translated into partial brackets that were included in the input sentences. For the sentence in (5) from the treebank (sic), the input for LoPar would be the following:

```
Was/PWS stimmt/VVFIN ,/$, weil/
KOUS ( sie/PPER unterhalt-
sam/ADJD sind/VAFIN ) ,/$,
aber/KON ( auch/ADV falsche/ADJA
Assoziationen/NN weckt/VVFIN )
```

The round parentheses delineate the conjuncts. LoPar was then forced to parse sentences containing coordination with the correct scope for the coordination. The results for this experiment are shown in Table 1 as exp. 2.

The introduction of partial brackets that delimit the scope of the coordination improve overall results on the full test set by 4.7 percent points, a rather significant improvement when we consider that only approximately one third of the test sentences were modified. The evaluation of the set of sentences that contain coordination shows that here, the difference is even higher: 6.7 percent points. It is also worth noticing that provided with scope information, the parser parses such sentences with the same accuracy as other sentences. The difference in F-scores between all sentences and only sentences with coordination in this experiment is much lower (0.5 percent points) than for all other experiments (2.5–3.0 percent points).

When comparing the results of experiment 1 ($n$-best parsing) with the present one, it is evident that the F-scores are very similar: 74.53 for the 50-best reranking setting, and 74.46 for the one where we provided the gold scope. However, a comparison of precision and recall shows that there are differences: 50-best reranking results in higher recall, providing gold scope for coordinations in higher precision. The lower recall in the latter experiment indicates that the provided brackets in some cases are not covered by the grammar. This is corroborated by the fact that in $n$-best parsing, only 1 sentence could not be parsed; but in parsing with gold scope, 8 sentences could not be parsed.

## 6 Experiment 3: Extracting Scope

The previous experiment has shown that providing the scope of a coordination drastically improves results for sentences with coordination as well as for the complete test set (although to a lower degree). The question that remains to be answered is whether automatically generated possible scopes can provide enough information for the reranker to improve results.

The first question that needs to be answered is how to find the possible scopes for a coordination. One possibility is to access the parse forest of a chart parser such as LoPar and extract infor-

mation about all the possible scope analyses that the parser found. If the same parser is used for this step and for the final parse, we can be certain that only scopes are extracted that are compatible with the grammar of the final parser. However, parse forests are generally stored in a highly packed format so that an exhaustive search of the structures is very inefficient and proved impossible with present day computing power.

(6) ”Es gibt zwar ein paar
”There are indeed a few
Niederflurbusse, aber das reicht ja
low-floor buses, but that suffices *part.*
nicht”, sagt er.
not”, says he.
’”There are indeed a few low-floor buses, but that isn’t enough”, he says.

Another solution consists of generating all possible scopes around the coordination. Thus, for the sentence in (6), the conjunction is *aber*. The shortest possible left conjunct is *Niederflurbusse*, the next one *paar Niederflurbusse*, etc. Clearly, many of these possibilities, such as the last example, are nonsensical, especially when the proposed conjunct crosses into or out of base phrase boundaries. Another type of boundary that should not be crossed is a clause boundary. Since the conjunction is part of the subordinated clause in the present example, the right conjunct cannot extend beyond the end of the clause, i.e. beyond *nicht*.

For this reason, we used KaRoPars (Müller and Ule, 2002), a partial parser for German, to parse the sentences. From the partial parses, we extracted base phrases and clauses. For (6), the relevant bracketing provided by KaRoPars is the following:

```
( " Es gibt zwar { ein paar
Niederflurbusse } , ) aber ( das
reicht ja nicht ) " , sagt er .
```

The round parentheses mark clause boundaries, the curly braces the one base phrase that is longer than one word. In the creation of possible conjuncts, only such conjuncts are listed that do not cross base phrase or clause boundaries. In order to avoid unreasonably high numbers of pre-bracketed versions, we also use higher level phrases, such as coordinated noun phrases. KaRoPars groups such higher level phrases only in contexts that allow a reliable decision. While a small percentage of such decisions is wrong, the heuristic used turns out to be reliable and efficient.

For each scope, a partially bracketed version of the input sentence is created, in which only the brackets for the suggested conjuncts are inserted. Each pre-bracketed version of the sentence is parsed with LoPar. Then all versions for one sentence are reranked. The reranker was trained on the data from experiment 1 ($n$-best parsing). The results of the reranker show that our restrictions based on the partial parser may have been too restrictive. Only 375 sentences had more than one pre-bracketed version, and only 328 sentence resulted in more than one parse. Only the latter set could then profit from reranking.

The results of this experiment are shown in Table 1 as exp. 3. They show that extracting possible scopes for conjuncts from a partial parse is possible. The difference in F-score between this experiment and the baseline reaches 5.93 percent points. The F-score is also minimally higher than the F-score for experiment 2 (gold scope), and recall is increased by approximately 1 percent point (even though only 12.5% of the sentences were reranked). This can be attributed to two factors: First, we provide different scope possibilities. This means that if the correct scope is not covered by the grammar, the parser may still be able to parse the next closest possibility instead of failing completely. Second, reranking is not specifically geared towards improving coordinated structures. Thus, it is possible that a parse is reranked higher because of some other feature. It is, however, not the case that the improvement results completely from reranking. This can be deduced from two points: First, while the F-score for experiment 1 (50-best analyses plus reranking) and the present experiment are very close (74.53 vs. 74.69), there are again differences in precision and recall: In experiment 1, recall is higher, and in the present experiment precision. Second, a look at the evaluation on only sentences with coordination shows that the F-score for the present experiment is higher than the one for experiment 1 (72.14 vs. 71.68). Additionally, precision for the present experiment is more than 2 percent points higher.

## 7 Experiment 4: Combining $n$-Best Parses and Extracted Scope Parses

As described above, the results for reranking the 50-best analyses and for reranking the versions

with automatically extracted scope readings are very close. This raises the question whether the two methods produce similar improvements in the parse trees. One indicator that this is not the case can be found in the differences in precision and recall. Another possibility of verifying our assumption that the improvements do not overlap lies in the combination of the 50-best parses with the parses resulting from the automatically extracted scopes. This increases the number of parses between which the reranker can choose. In effect, this means a combination of the methods of experiments 1 ($n$-best) and 3 (automatic scope). Consequently, if the results from this experiment are very close to the results from experiment 1 ($n$-best), we can conclude that adding the parses with automatic scope readings does not add new information. If, however, adding these parses improves results, we can conclude that new information was present in the parses with automatic scope that was not covered in the 50-best parses. Note that the combination of the two types of input for the reranker should not be regarded as a parser ensemble but rather as a resampling of the $n$-best search space since both parsers use the same grammar, parsing model, and probability model. The only difference is that LoPar can accept partially bracketed input, and BitPar can list the $n$-best analyses.

The results of this experiment are shown in Table 1 as exp. 4. For all sentences, both precision and recall are higher than for experiment 1 and 3, resulting in an F-score of 76.69. This is more than 2 percent points higher than for the 50-best parses. This is a very clear indication that the parses contributed by the automatically extracted scopes provide parses that were not present in the 50 best parses from experiment 1 ($n$-best). The same trend can be seen in the evaluation of the sentences containing coordination: Here, the improvement in F-score is higher than for the whole set, a clear indication that this method is suitable for improving coordination parsing. A comparison of the results of the present experiment and experiment 3 (with automatic scope only) shows that the gain in precision is rather small, but the combination clearly improves recall, from 73.96% to 77.23%. We can conclude that adding the 50 best parses remedies the lacking coverage that was the problem of experiment 3. More generally, experiment 4 suggests that for the notoriously difficult problem of parsing coordination structures, a hybrid approach that combines parse selection of $n$ best analyses with pre-bracketed scope in the input results in a considerable reduction in error rate compared to each of these methods used in isolation.

## 8 Related Work

Parsing of coordinate structures for English has received considerable attention in computational linguistics. Collins (1999), among many other authors, reports in the error analysis of his WSJ parsing results that coordination is one of the most frequent cases of incorrect parses, particularly if the conjuncts involved are complex. He manages to reduce errors for simple cases of NP coordination by introducing a special phrasal category of base NPs. In the experiments presented above, no explicit distinction is made between simple and complex cases of coordination, and no transformations are performed on the treebank annotations used for training.

Our experiment 1, reranking 50-best parses, is similar to the approaches of Charniak and Johnson (2005) and of Hogan (2007). However, it differs from their experiments in two crucial ways: 1) Compared to Charniak and Johnson, who use 1.1 mio. features, our feature set is appr. five times larger (more than 5 mio. features), with the same threshold of at least five occurrences in the training set. 2) Both Hogan and Charniak and Johnson use special features for coordinate structures, such as a Boolean feature for marking parallelism (Charniak and Johnson) or for distinguishing between coordination of base NPs and coordination of complex conjuncts (Hogan), while our approach refrains from such special-purpose features.

Our experiments using scope information are similar to the approaches of Kurohashi and Nagao (1994) and Agarwal and Bogges (1992) in that they try to identify coordinate structure bracketings. However, the techniques used by Agarwal and Bogges and in the present paper are quite different. Agarwal and Bogges and Kurohashi and Nagao rely on shallow parsing techniques to detect parallelism of conjuncts while we use a partial parser only for suggesting possible scopes of conjuncts. Both of these approaches are limited to coordinate structures with two conjuncts only, while our approach has no such limitation. Moreover, the goal of Agarwal and Bogges is quite different from ours. Their goal is robust detection of coordinate structures only (with the intended ap-

plication of term extraction), while our goal is to improve the performance of a parser that assigns a complete sentence structure to an input sentence.

Finally, our approach at present is restricted to purely syntactic structural properties. This is in contrast to approaches that incorporate semantic information. Hogan (2007) uses bi-lexical head-head co-occurrences in order to identify nominal heads of conjuncts more reliably than by syntactic information alone. Chantree et al. (2005) resolve attachment ambiguities in coordinate structures, as in (7a) and (7b), by using word frequency information obtained from generic corpora as an effective estimate of the semantic compatibility of a modifier vis-à-vis the candidate heads.

(7) a. Project managers and designers
b. Old shoes and boots

We view the work by Hogan and by Chantree et al. as largely complementary to, but at the same time as quite compatible with our approach. We must leave the integration of structural syntactic and lexical semantic information to future research.

## 9 Conclusion and Future Work

We have presented a study on improving the treatment of coordinated structures in PCFG parsing. While we presented experiments for German, the methods are applicable for any language. We have chosen German because it is a language with relatively flexible phrasal ordering (cf. Section 2) which makes parsing coordinations particularly challenging. The experiments presented show that $n$-best parsing combined with reranking improves results by a large margin. However, the number of cases in which the correct parse is present in the $n$-best parses is rather low. This led us to the assumption that the $n$-best analyses often do not cover the whole range of different scope possibilities but rather present minor variations of parses with few differences in coordination scope. The experiments in which the parser was forced to assume predefined scopes show that the scope information is important for parsing quality. Providing the parser with different scope possibilities and reranking the resulting parses results in an increase in F-score from 69.76 for the baseline to 74.69. One of the major challenges for this approach lies in extracting a list of possible conjuncts. Forcing the parser to parse all possible sequences results in a prohibitively large number of possibilities, especially for sentences with 3 or more conjunctions. For this reason, we used chunks above base phases, such as coordinated noun chunks, to restrict the space. However, an inspection of the lists of bracketed versions of the sentences shows that the definition of base phrases is one of the areas that must be refined. As mentioned above, the partial parser groups sequences of "NP KON NP" into a single base phrase. This may be correct in many cases, but there are exceptions such as (8).

(8) Die 31jährige Gewerkschaftsmitarbeiterin und ausgebildete Industriekauffrau
The 31-year-old union staff member and trained industrial clerk
aus Oldenburg bereitet nun ihre erste eigene CD vor.
from Oldenburg is preparing now her first own CD *part.*.

For (8), the partial parser groups *Die 31jährige Gewerkschaftsmitarbeiterin und ausgebildete Industriekauffrau* as one noun chunk. Since our proposed conjuncts cannot cross these boundaries, the correct second conjunct, *ausgebildete Industriekauffrau aus Oldenburg*, cannot be suggested. However, if we remove these chunk boundaries, the number of possible conjuncts increases dramatically, and parsing times become prohibitive. As a consequence, we will need to find a good balance between these two needs. Our plan is to increase flexibility very selectively, for example by enabling the use of wider scopes in cases where the conjunction is preceded and followed by base noun phrases. For the future, we are planning to repeat experiment 3 (automatic scope) with different phrasal boundaries extracted from the partial parser. It will be interesting to see if improvements in this experiment will still improve results in experiment 4 (combining 50-best parses with exp. 3).

Another area of improvement is the list of features used for reranking. At present, we use a feature set that is similar to the one used by Collins and Koo (2005). However, this feature set does not contain any coordination specific features. We are planning to extend the feature set by features on structural parallelism as well as on lexical similarity of the conjunct heads.

## References


Rajeev Agarwal and Lois Boggess. 1992. A simple but useful approach to conjunct identification. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics (ACL-92)*, pages 15–21, Newark, DE.

Francis Chantree, Adam Kilgarriff, Anne de Roeck, and Alistair Willis. 2005. Disambiguating coordinations using word distribution information. In *Proceedings of Recent Advances in NLP (RANLP 2005)*, pages 144–151, Borovets, Bulgaria.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 173–180, Ann Arbor, MI.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–69.

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

Anette Frank. 2002. A (discourse) functional analysis of asymmetric coordination. In *Proceedings of the LFG-02 Conference*, Athens, Greece.

Yoav Freund, Ray Iyer, Robert Shapire, and Yoram Singer. 1998. An efficient boosting algorithm for combining preferences. In *Proceedings of the 15th International Conference on Machine Learning*, Madison, WI.

Deirdre Hogan. 2007. Coordinate noun phrase disambiguation in a generative parsing model. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 680–687, Prague, Czech Republic.

Tilman Höhle. 1983. Subjektlücken in Koordinationen. Universität Tübingen.

Tilman Höhle. 1990. Assumptions about asymmetric coordination in German. In Joan Mascaró and Marina Nespor, editors, *Grammar in Progress. Glow Essays for Henk van Riemsdijk*, pages 221–235. Foris, Dordrecht.

Tilman Höhle. 1991. On reconstruction and coordination. In Hubert Haider and Klaus Netter, editors, *Representation and Derivation in the Theory of Grammar*, volume 22 of *Studies in Natural Language and Linguistic Theory*, pages 139–197. Kluwer, Dordrecht.

Andreas Kathol. 1990. Linearization vs. phrase structure in German coordination constructions. *Cognitive Linguistics*, 10(4):303–342.

Sandra Kübler. 2008. The PaGe 2008 shared task on parsing German. In *Proceedings of the ACL Workshop on Parsing German*, pages 55–63, Columbus, Ohio.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4):507–534.

Frank Henrik Müller and Tylman Ule. 2002. Annotating topological fields and chunks–and revising POS tags at the same time. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING'02*, pages 695–701, Taipei, Taiwan.

Slav Petrov and Dan Klein. 2008. Parsing German with latent variable grammars. In *Proceedings of the ACL Workshop on Parsing German*, pages 33–39, Columbus, Ohio.

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

Helmut Schmid. 2004. Efficient parsing of highly ambiguous context-free grammars with bit vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland.

Heike Telljohann, Erhard Hinrichs, and Sandra Kübler. 2004. The TüBa-D/Z treebank: Annotating German with a context-free backbone. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 2229–2235, Lisbon, Portugal.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister, 2005. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, Germany.

Dieter Wunderlich. 1988. Some problems of coordination in German. In Uwe Reyle and Christian Rohrer, editors, *Natural Language Parsing and Linguistic Theories*, Studies in Linguistics and Philosophy, pages 289–316. Reidel, Dordrecht.