Improving coverage and parsing quality of a large-scale LFG for German

Christian Rohrer, Martin Forst

Institute for Natural Language Processing (IMS)
University of Stuttgart
Azenbergstr. 12
70174 Stuttgart, Germany
{rohrer,forst}@ims.uni-stuttgart.de

Abstract

We describe experiments in parsing the German TIGER Treebank. In parsing the complete treebank, 86.44% of the sentences receive full parses; 13.56% receive fragment parses. We discuss the methods used to enhance coverage and parsing quality and we present an evaluation on a gold standard, to our knowledge the first one for a deep grammar of German. Considering the selection performed by our current version of a stochastic disambiguation component, we achieve an f-score of 84.2%, the upper and lower bounds being 87.4% and 82.3% respectively.

1. Introduction

For realistic applications we need grammars with broad coverage. The broader the coverage, however, the greater the number of possible readings per sentence and the lower the performance. When increasing coverage, we tried to include the most frequent constructions (based on a corpus study) and at the same time to restrict the grammar rules in order to avoid overgeneration. The restrictions are sometimes too heavy, and we loose certain sentences, but the gain in performance clearly justifies the restrictions. Besides quantity of analyses, one also wants quality. Quality can only be measured by evaluating against a gold standard. Once substantial coverage with high quality has been reached, the problem is to chose the 'intended' reading. Disambiguation of competing syntactic analyses is one of the greatest challenges for computational linguistics. We present first results of experiments with a stochastic disambiguation model.

2. A Broad-Coverage LFG for German

The grammar was developed in the ParGram project (Butt et al., 2002). Besides achieving 50% coverage (Dipper, 2003), the grammar writers concentrated on phenomena discussed in theoretical syntax. With the advent of treebanks and successful attempts to induce grammars from treebanks, we shifted our focus. In a new project (DLFG[1]), we are concentrating on coverage.

The grammar now has 274 LFG style rules, which compile into an automaton with 6,584 states and 22,241 arcs. The grammar uses several lexicons and and a guessing mechanism for default lexical entries. The lexicons record mainly subcategorization information. As a form of preprocessing, the grammar uses a cascade of finite-state transducers (Kaplan et al., 2004), mainly for tokenization and morphological analysis. The input sentences are thus processed by a tokenizer, a multi-word transducer, a morphology and a guesser before they are actually parsed. Later we will also include a named entity recognizer (NER). In the current experiments with the gold standard we simulate the NER by manual marking.

[1] *Disambiguierung einer Lexikalisch-Funktionalen Grammatik für das Deutsche* ('Disambiguation of a Lexical Functional Grammar for German') – research project financed by the DFG (*Deutsche Forschungsgemeinschaft* 'German Research Foundation'), grant *Ro 245/18-1*

3. Enhancing grammar coverage

3.1. Corpus-based enlargement of grammar coverage

In order to increase coverage of the grammar we first had to find out where the grammar was incomplete. We systematically created testsuites extracted from the TIGER Treebank. For instance we extracted all NPs up to the head or all NPs which are modified by a (subcategorized) subordinate clause or a verbphrase. We also extracted the trees associated with the corrsponding strings in order to determine the frequency of a construction. Most of the examples where our grammar failed involved constructions with very limited frequency. Hence, once a grammar has achieved broad coverage progress is slow. There were, however, a few areas where adding new rules really helped to increase coverage:

3.1.1. Coordination

Coordination was one phenomenon of which only the basic instances were covered by the original grammar. We thus introduced new rules for several subtypes of asymmetric or otherwise 'special' coordination.

Coordination of adverbs with PPs

In analogy to predicative constituents like in *he is a Republican and proud of it*, which can be handled with a special coordination rule for predicative constituents that allows, e.g., DPs and APs to be coordinated, we account for the coordination of ADVPs and PPs that function as modifiers with a special coordination rule[2], namely

ADVP → ADVP: ↓∈↑; CONJco PP: ↓∈↑.

(1) hier und in Berlin
here and in Berlin
'here and in Berlin'

[2] For simplicity of presentation, we only present simplified versions of the newly introduced grammar rules.

Subject gap in finite constructions (SGF)

(2) Hierhin kam Hans und hielt seinen Vortrag.
Here came Hans and gave his talk.

'Hans came here and gave his talk.'

In these constructions, which have received a lot of linguistic attention since Höhle (1983), the shared subject is in the Mittelfeld of the first conjunct instead of being in the Vorfeld. This means that it is not distributed automatically into the second conjunct. We have implemented an analysis following Frank (2002), who treats SGF coordination as a marked case of CP coordination that can only occur given a very particular information structure. Unlike Frank (2002), we formulate the rule as a coordination of a CP and a Cbar, but this is a detail motivated by efficiency considerations:

```
CP →  CP:     ↓∈↑ (↑SUBJ) = (↓SUBJ);
      CONJco
      Cbar:   ↓∈↑.
```

Adverbs and PPs between conjunction and conjunct
Coordinated structures where an ADVP or a PP occurs left of the last conjunct, as illustrated in (3), have received much less attention in theoretical linguistics nor are they accounted for in most deep grammars, to our knowledge.

(3) Von Monat zu Monat wächst das
From month to month grows the offer
Angebot und mit ihm auch die Nachfrage.
and with it also the demand.

'The offer grows from month to month and so does the demand.'

However, they are relatively frequent in text corpora, so that coverage can be noticeably improved by the introduction of a rule for these constructions. We therefore formulated coordination rules of the following type, where, in the f-annotations, → refers to the f-structure of the right sister:

```
DP →  DP:          ↓∈↑;
      CONJco
      ( { ADVP:    ↓∈ (→ ADJUNCT);
        | PP:      ↓∈ (→ ADJUNCT) } )
      DP:          ↓∈↑.
```

Parentheticals 4-5% of the sentences in the TIGER Corpus contain constituents marked as parentheticals. We introduce parenthetical constructions via a metarule macro. It allows insertion of a parenthetical between any two constituents on the right hand side of a phrase-structure rule.

Reported speech without 'real' verbum dicendum
In German newspaper text, sentences like the following occur relatively frequently:

(4) "Die Fans waren zunächst irritiert", bewertet
"The fans were at first irritated", Hans die Veränderung der Band.
evaluates Hans the change of the band.

"'The fans were confused at first", says Hans, evaluating the change of the band.'

The first clause, *Die Fans waren zunächst irritiert*, represents reported speech, but the clause which introduces the reported speech does not contain a verb of saying. *Bewerten* does not subcategorize for a sentential complement. In our example, it takes a subject (*Hans*) and an object (*die Veränderung der Band*).
The distribution of this kind of construction is the same as the distribution of reportive parentheticals headed by verbs that subcategorize for a COMP. Hence, in addition to COMP, we allow the reported speech before or around a reportive parenthetical to be projected to the semantic function REPORTEDSPEECH. The f-structure associated to (4) is illustrated in figure 1.

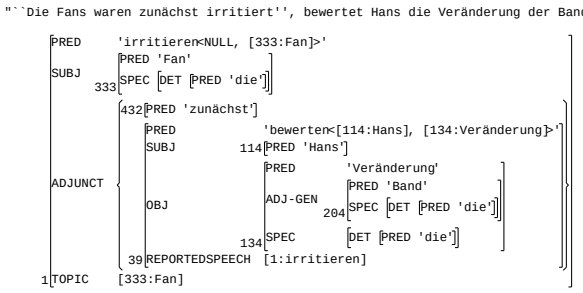

Figure 1: f-structure of (4)

3.2. Corpus-based restriction of grammar rules

3.2.1. Rule specialization

In the original version of our grammar we tried to write rules as general as possible. For instance, a VP can function as an AP if the head verb is transformed into a participle. Instead of an unrestricted rule `AP[+infl]` → `VP[+infl]`, with very negative effects on efficiency, we wrote a special rule `VP-as-AP`, where we limit the number and function of possible constituents and where we exclude recursion in the verbal complex. This is motivated by the fact that, in the TIGER Corpus, there is not a single occurrence of an AP with a participle head dominating a VP.
The exclusion of recursion in deverbal attributive APs has a very positive impact on the efficiency of the grammar because there are numerous forms that can be both an inflected past participle and a past tense form. Consider the following subordinate clause:

(5) Weil er die Frau die Aktien zu verkaufen überredete,
Because he the woman the shares to sell convinced,
...
...
'Because he convinced the woman to sell the shares ...'

The form *überredete* can be both a past tense form and a past participle. As the original grammar allows infinitival VPs to be embedded in attributive deverbal APs, it can analyze the string *die Aktien zu verkaufen überredete* as an inflected AP, and this inflected AP can then be analyzed as a headless DP. This means that a large number of undesired c-structures is built which are only ruled out during the solution of the f-structure constraints. Of course, with respect to efficiency, it is a very attractive feature of the revised grammar that these erroneous c-structures are not built at all in the first place.

3.2.2. Restricting long distance dependencies

Solving the equations which account for long distance dependencies can be very time-consuming. We therefore simplified these equations based on a corpus study, e.g. for extraposed relative clauses.

3.2.3. Restricting rules by 'number of tokens'

We restrict certain rules by limiting the number of tokens covered by the rule. E.g., subjectless insertions like *wie früher berichtet* ('as previously reported') have only very few words between *as* and *reported*.

3.3. Generality of the steps taken to enhance grammar coverage

Our section on corpus-based improvement of grammar coverage may create the impression that we tailored the grammar too closely to the TIGER Corpus. We therefore parsed the 20,614 sentences of the NEGRA Corpus. 81.5% of the sentences obtained a full parse and 18.5%, a partial parse. These results on the NEGRA Corpus are clearly not as good as the results on the TIGER Corpus, but with a grammar coverage of more than 80%, they show that coverage does not drop dramatically on unseen corpora and that at least most of the measures taken to improve coverage carry over to the unseen data.

4. Robustness

We augmented the standard grammar with a FRAGMENT grammar to collect as much information as possible in cases where a sentence does not get a full parse. The parser returns well-formed chunks like NPs, PPs, VPs, Ss, etc. The grammar has a fewest-chunk method for determining the least fragmented parse. It turned out that the quality of fragment parses can be improved by restricting complex rules (e.g. the S-rule) in the fragment grammar wrt. the standard grammar.

In order to cope with timeouts and memory problems, we use the SKIMMING technique (Riezler et al., 2002). When the amount of time or memory spent on a sentence exceeds a given threshold, XLE 'skims' the constituents whose processing has not yet been completed, i.e. XLE does only a bounded amount of work per subtree. When skimming, we use a restricted version of our grammar. This is achieved with the help of special OT marks (Frank et al., 2001), so-called SKIMMING_NOGOOD marks, which turn off expensive rules like headless NPs, 'free' datives, etc. during skimming.

5. Testing

5.1. Gold standard

We evaluated parse quality on manually validated dependency annotations for 1602 sentences from the TiGer Dependency Bank (Forst et al., 2004) The annotation from the TIGER Treebank were semi-automatically transformed into dependency triples which were then corrected and extended by human annotators. It encodes the same type of dependency triples as the PARC 700 Dependency Bank (King et al., 2003). The grammatical relations and morphosyntactic features are the ones annotated in the TIGER Treebank, except for systematic changes meant to make the TiGer DB more suitable for parser evaluation.

5.2. Parsing quality

In tables 1 and 2, we give the results of two types of parse selection: (1) lower bound: In the lower bound a parse from the set of parses is chosen randomly. (2) upper bound: In the case of the upper bound the best F-score according to the annotation schema is chosen. F-score is defined as the harmonic mean of precision and recall ($f = \frac{2pr}{p+r}$). We use the triple encoding and evaluation software of (Crouch et al., 2002).

Table 1 shows that full parses achieve a noticeably higher f-score than partial parses; this shows that it is crucial to improve coverage to, say, at least 80% in order to parse free text with a reasonable quality. Table 2 gives the upper bound and the lower bound figures for the 1602 gold standard sentences broken down according to the grammatical relations and morphosyntactic features encoded.

5.3. Disambiguation

Table 3, finally, gives preliminary results for our stochastic disambiguation component. Two versions of the component are compared with each other and with the upper and lower bound. Both versions are based on maximum entropy models that are trained in a supervised manner on partially labelled data. The training material for both models were the parses of 3,817 sentences from the TIGER Corpus (except of sentences 8,001 through 10,000). The *all properties* version uses both the kind of property described in Riezler et al. (2002) and a series of new properties that mainly encode information on the linear order of grammatical functions. The *only original properties* version only makes use of the former.

relation	upper bound	all properties for disamb.	only original properties	lower bound
all	87.39	84.20	82.95	82.28
preds only	81.91	77.19	76.17	75.11
da	67	64	63	59
gr	88	83	82	79
oa	81	77	69	67
op	58	57	57	54
op_loc	63	54	52	45
quant	80	79	79	76
sb	80	77	73	72
sbp	68	62	61	56

Table 3: F-scores for selected grammatical relations in the 1602 TiGer DB examples broken down according to parse selection method

6. Discussion

6.1. Coverage

In order to get a full parse, the input sentence has to be well-formed. At least 1% of the sentences in the testsuite contain spelling mistakes, punctuation errors or grammatical errors. Furthermore the TIGER annotators sometimes assign full structures to elliptical sentences that lack a clear syntactic head.

In order to match the analyses annotated for them, our parser would have to do a lot of structure building, which would lead to overgeneration and inefficiency.

	all	full	full and non-skimmed fragments	non-skimmed fragments	non-skimmed fragments	skimmed fragments
% of test set	100	88.6	96.6	11.4	8.0	3.4
upper bound	87.4	88.9	88.0	76.0	78.7	69.7
lower bound	81.7	83.6	82.9	72.2	74.2	66.3
avg. sentence length	16.2	14.9	15.2	24.6	17.6	41.7
avg. parse time in sec.	3.91	1.52	2.64	18.56	6.00	56.78

Table 1: Upper bound and lower bound f-scores for grammatical relations and morphosyntactic features in the 1602 TiGer DB examples broken down according to parse quality

Among the well-formed sentences which receive a partial parse we have to distinguish three types: (1) constructions for which our grammar contains rules, which, however, are turned off for efficiency reasons (e.g. coordination without an explicit conjunction), (2) constructions for which we do not have rules (e.g., special types of non-constituent coordination, certain parenthetical constructions, heavy ellipsis), (3) sentences which contain lexical material that is not in the lexicon and which our guesser cannot handle (e.g., problems of subcategorization, idioms and collocations). Subcategorization poses problems especially if a MWE as a whole subcategorizes for a sentential function like COMP despite the fact that none of its parts subcatgorizes for a COMP. This is the case with the MWE *zu Protokoll geben* which subcategorizes for a COMP but neither *geben* nor *Protokoll* subcategorize for a COMP.

6.2. Parsing quality

As Table 1 shows, the results for the complete testsuite are quite good. Breaking them down according to parse quality shows that our upper bound for full parses is roughly identical to Riezler et al. (2002). Our values for the complete test set are better (87.4% vs. 84.1%) because more sentences of our testsuite receive a full parse. If we subtract the 55 sentences with an average length of 41.7 words that get a partial parse after skimming, we obtain for 96.6% of our testsuite an upper bound of 88.0% and a lower bound of 82.9%.

The F-score of our non-skimmed fragment parses is surprisingly high. Only highly elliptical sentences get really bad values. One explanation for our good values are our detailed subcategorization lexicons.

The figures in table 2 are more informative than overall F-score. They illustrate that the f-scores for grammatical relations are not as good as those for morphosyntactic features. The lower values for *case* are due to syntactic ambiguity and are therefore not a purely morphological problem; to a limited extent this is also true for the feature *num* (number). In the preds-only evaluation the values for arguments *sb* (subject) and *oa* (accusative object) are better than those for *da* (dative object) and *og* (genitive object). So-called 'free datives' are quite frequent in German, and as the name indicates, difficult to predict and to specify in the subcategorization lexicon. We guess free datives and, apparently, we go wrong sometimes. For genitive objects we get bad values because, for efficiency reasons, we require that the genitive be morphologically marked. Furthermore, genitive NPs may be attached to preceding NPs. The figures for *sbp* (logical subject in passives) are worse than those for grammatical subjects because the PP denoting the logical subject is introduced by *von*, which has many different functions.

Subcategorized PPs (and ADVPs) are annotated as *op* (oblique), *op_dir* (directional argument), *op_loc* (locative argument) and *op_manner* (modal argument). The low f-score for subcategorized PPs indicates gaps in the subcategorization lexicon. In addition, this low score has a negative effect on the f-score of *mo* (modifiers or ajduncts).

*pd*s (predicative complements) with the copula *sein* can be confused with stative passives. E.g., *Er ist ihm übergeordnet* is analyzed as stative passive by our grammar and as *pd* by the annotators.

The values for the subcategorized functions *oc_fin* (finite complement clauses) and *oc_inf* (non-finite argument VPs) differ. The figures for clauses with the function *oc_fin* are lower because clauses introduced by interrogative or relative pronouns in adverbial function can be interpreted as *oc_fin*s if the embedding clause contains a word which subcategorizes for such a clause. Furthermore there is interference with *rs* (reported speech) and *app_cl* (appositive clauses).

gl (genitive left) denotes possessives and *gr* (genitive right) denotes genitive adjuncts and *von* PPs with genitive function. *gl* constructions are easy to identify because they always precede their head, whereas the analysis of *gr* ultimately is a semantic problem, at least when it is realized by a *von* PP.

Comparative complements (*cc*) and relative clauses (*rc*), which are often extraposed, are difficult to attach to the corresponding head. Coordination (*cj*) is also notoriously difficult and achieves fairly low values.

6.3. Disambiguation

The figures in table 3 show that a selection performed by one of the versions of the stochastic disambiguation component clearly performs better than a random selection (lower bound). We also observe that the *all properties* version of the disambiguation component performs noticeably better than the *only original properties* version. In terms of overall f-score, the gain with respect to the lower bound doubles with the help of the additional properties; for the core grammatical functions, such as *oa*, *sb* etc., which are particularly important for the potential construction of a semantic representation on the basis of f-structures, this gain is even far more important. For many of the grammatical functions, the additional properties allow the *all properties* f-score to be closer to the upper bound f-score than to the lower bound f-score. As this is not the case of the *only original properties* f-scores, we believe that property design will be partic-

relation or	upper bound			lower bound		
feature	precision	recall	f-score	precision	recall	f-score
all	61213/69508 = 88.1	61213/70577 = 86.7	87.4	57646/69636 = 82.8	57646/70482 = 81.8	82.3
preds only	22050/26475 = 83.29	22050/27363 = 80.6	81.9	20236/26554 = 76.2	20236/27328 = 74.0	75.1
ams		0/2 = 0	0		0/2 = 0	0
app	185/268 = 69	185/337 = 55	61	186/282 = 66	186/336 = 55	60
app_cl	23/27 = 85	23/77 = 30	44	22/26 = 85	22/77 = 29	43
cc	17/23 = 74	17/46 = 37	49	14/20 = 70	14/45 = 31	43
cj	1183/1412 = 84	1183/1806 = 66	74	1106/1412 = 78	1106/1806 = 61	69
da	118/190 = 62	118/162 = 73	67	114/226 = 50	114/162 = 70	59
det	3655/3816 = 96	3655/3938 = 93	94	3582/3822 = 94	3582/3930 = 91	92
gl	292/316 = 92	292/317 = 92	92	280/305 = 92	280/316 = 89	90
gr	804/928 = 87	804/902 = 89	88	708/897 = 79	708/899 = 79	79
measured	9/20 = 45	9/24 = 38	41	9/20 = 45	9/24 = 38	41
mo	4997/6878 = 73	4997/6610 = 76	74	4244/6946 = 61	4244/6601 = 64	63
mod	2087/2219 = 94	2087/2228 = 94	94	1967/2226 = 88	1967/2227 = 88	88
name_mod	336/420 = 80	336/385 = 87	83	331/424 = 78	331/385 = 86	82
number	370/469 = 79	370/424 = 87	83	357/456 = 78	357/423 = 84	81
oa	923/1098 = 84	923/1191 = 77	81	764/1104 = 69	764/1189 = 64	67
oa2				0/1 = 0		0
obj	2916/3213 = 91	2916/3180 = 92	91	2805/3227 = 87	2805/3174 = 88	88
oc_fin	151/212 = 71	151/226 = 67	69	147/211 = 70	147/226 = 65	67
oc_inf	340/379 = 90	340/411 = 83	86	339/387 = 88	339/411 = 82	85
og	5/5 = 100	5/9 = 56	71	3/5 = 60	3/9 = 33	43
op	267/389 = 69	267/526 = 51	58	244/377 = 65	244/526 = 46	54
op_dir	29/38 = 76	29/140 = 21	33	20/38 = 53	20/140 = 14	22
op_loc	35/52 = 67	35/59 = 59	63	23/44 = 52	23/59 = 39	45
op_manner	6/8 = 75	6/16 = 38	50	2/4 = 50	2/16 = 12	20
pd	258/358 = 72	258/403 = 64	68	239/358 = 67	239/403 = 59	63
pred_restr	110/121 = 91	110/122 = 90	91	103/123 = 84	103/122 = 84	84
quant	172/195 = 88	172/234 = 74	80	159/184 = 86	159/234 = 68	76
rc	175/212 = 83	175/250 = 70	76	141/209 = 67	141/250 = 56	61
rs	2/19 = 11	2/4 = 50	17	2/19 = 11	2/4 = 50	17
sb	2549/3128 = 81	2549/3274 = 78	80	2297/3140 = 73	2297/3272 = 70	72
sbp	35/46 = 76	35/57 = 61	68	28/43 = 65	28/57 = 49	56
topic_disloc	1/16 = 6	1/3 = 33	11	0/18 = 0	0/3 = 0	0
case	7941/9004 = 88	7941/9098 = 87	88	7205/8991 = 80	7205/9085 = 79	80
circ_form	5/8 = 62	5/6 = 83	71	5/8 = 62	5/6 = 83	71
comp_form	99/115 = 86	99/160 = 62	72	96/111 = 86	96/160 = 60	71
coord_form	557/613 = 91	557/648 = 86	88	550/615 = 89	550/648 = 85	87
degree	2346/2640 = 89	2346/2488 = 94	91	2313/2668 = 87	2313/2486 = 93	90
det_type	3628/3780 = 96	3628/3779 = 96	96	3619/3772 = 96	3619/3771 = 96	96
fut	61/63 = 97	61/71 = 86	91	61/65 = 94	61/71 = 86	90
gend	7207/7829 = 92	7207/7875 = 92	92	6880/7850 = 88	6880/7864 = 87	88
mood	2129/2254 = 94	2129/2366 = 90	92	2117/2253 = 94	2117/2364 = 90	92
num	8739/9495 = 92	8739/9333 = 94	93	8349/9510 = 88	8349/9319 = 90	89
pass_asp	258/287 = 90	258/324 = 80	84	257/287 = 90	257/324 = 79	84
perf	296/301 = 98	296/355 = 83	90	292/299 = 98	292/355 = 82	89
pers	2392/2621 = 91	2392/2800 = 85	88	2192/2617 = 84	2192/2796 = 78	81
precoord_form	7/8 = 88	7/9 = 78	82	6/7 = 86	6/9 = 67	75
pron_form	71/74 = 96	71/72 = 99	97	71/74 = 96	71/72 = 99	97
pron_type	1282/1689 = 76	1282/1482 = 87	81	1261/1700 = 74	1261/1482 = 85	79
tense	2145/2240 = 96	2145/2360 = 91	93	2136/2239 = 95	2136/2358 = 91	93

Table 2: Upper bound and lower bound precisions, recalls and F-scores for grammatical relations and morphosyntactic features in the 1602 TiGer DB examples

ularly important for the further improvement of the stochastic disambiguation component.

A further step that we plan to take and that, as we hope, will improve the results of the stochastic disambiguation, regardless of the properties that are used for it, is the acquisition of more training data.

6.4. Comparison with previous work

Our results are comparable to those reported by Riezler et al. (2002) and Cahill et al. (2005) for English. Our score is improved by the fact that we check some morphological information like gender, number or tense, which a good chunker could also identify correctly. In a preds-only evaluation, the figures are lower, but the same tendency is observed with other parsers that are evaluated on dependency-based gold standards.

Dubey and Keller (2003) induce a grammar from the NEGRA Treebank, a predecessor of TIGER. They report a labelled precision and recall of up to 74%. The results for induced grammars seem to be worse for German with its free word order than for English. This also holds for the German LFG induced from the TIGER Corpus (Cahill et al., 2005). The authors report an f-score of 71%. The evaluation is equivalent to ours, i.e. based on dependency triples obtained via conversion from TIGER graphs. The testsuite which functions as a gold standard, however, is fairly small. One of the reasons for the low f-score seems to be the lack of morphological information and the very flat structure of the TIGER graphs. Integrating morphological information would certainly improve the score. The flat structure of the NEGRA and TIGER Treebanks may also have a negative influence on the quality of the induced grammars.

Foth et al. (2005) describe a parsing system for unrestricted German text. Total coverage is achieved by means of defeasible, graded constraints. The authors report an f-score of 87% in an evaluation with the NEGRA Corpus. These are clearly the best results for German so far. They are also better than those reported by Schiehlen (2003), who achieves an f-score of 81.7% on the NEGRA data. In support of our approach, we would like to mention that our grammar is fully reversible and comes with a fullfledged generator.

7. Conclusion

We have shown that a hand-crafted 'deep' grammar can achieve good results on free text. The next step will be to refine our stochastic disambiguation component. Our grammar can also be used in generation, unlike other large-scale grammars of German.

8. References

Miriam Butt, Helge Dyvik, Tracy H. King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In *Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation*, pages 1–7.

Aoife Cahill, Michael Burke, Martin Forst, Ruth O'Donovan, Christian Rohrer, Josef van Genabith, and Andy Way. 2005. Treebank-Based Multilingual Unification-Grammar Resources. *Research in Language and Computation*.

Richard Crouch, Ronald M. Kaplan, Tracy H. King, and Stefan Riezler. 2002. A comparison of evaluation metrics for a broad-coverage parser. In *Proceedings of the LREC Workshop 'Beyond PARSEVAL—Towards improved evaluation mesures for parsing systems'*, pages 67–74, Las Palmas, Spain.

Stefanie Dipper. 2003. *Implementing and Documenting Large-scale Grammars – German LFG*. Ph.D. thesis, IMS, University of Stuttgart. Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (AIMS), Volume 9, Number 1.

Amit Dubey and Frank Keller. 2003. Probabilistic Parsing for German using Sister-Head Dependencies. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 96–103, Sapporo, Japan.

Martin Forst, Núria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra, and Valia Kordoni. 2004. Towards a dependency-based gold standard for German parsers – The TiGer Dependency Bank. In *Proceedings of the COLING Workshop on Linguistically Interpreted Corpora (LINC '04)*, Geneva.

Kilian Foth, Wolfgang Menzel, and Ingo Schröder. 2005. Robust parsing with weighted constraints. *Natural Language Engineering*, 11(1):1–25.

Anette Frank, Tracy Holloway King, Jonas Kuhn, and John T. Maxwell III. 2001. Optimality Theory Style Constraint Ranking in Large-Scale LFG Grammars. In Peter Sells, editor, *Formal and Empirical Issues in Optimality Theoretic Syntax*.

Anette Frank. 2002. A (Discourse) Functional Analysis of Asymmetric Coordination. In *Proceedings of the 7th International LFG Conference (LFG'05)*, Athens, Greece. CSLI Publications.

Tilmann Höhle. 1983. *Topologische Felder*. Ph.D. thesis, University of Cologne.

Ronald M. Kaplan, John T. Maxwell, Tracy H. King, and Richard Crouch. 2004. Integrating Finite-state Technology with Deep LFG Grammars. In *Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP*, Nancy, France.

Tracy Holloway King, Richard Crouch, Stefan Riezler, Mary Dalrymple, and Ronald M. Kaplan. 2003. The PARC 700 Dependency Bank. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora (LINC '03)*, Budapest.

Stefan Riezler, Tracy Holloway King, Ronald M. Kaplan, Richard Crouch, John T. Maxwell III, and Mark Johnson. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2002*, Philadelphia.

Michael Schiehlen. 2003. Combining Deep and Shallow Approaches in Parsing German. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan.