Robust Parsing: More with Less

Kilian Foth, Wolfgang Menzel
Fachbereich Informatik, Universität Hamburg, Germany
foth|menzel@informatik.uni-hamburg.de

Abstract

Covering as many phenomena as possible is a traditional goal of parser development, but the broader a grammar is made, the blunter it may become, as rare constructions influence the behaviour on simple sentences that were already solved correctly. We observe the effects of intentionally removing support for specific constructions from a broad-coverage grammar of German. We show that accuracy of analysing sentences from the NEGRA corpus can be improved not only for sentences that do not need the extra coverage, but even when including those that do.

1 Introduction

Traditionally, broad coverage has always been considered to be a desirable property of a grammar: the more linguistic phenomena are treated properly by the grammar, the better results can be expected when applying it to unrestricted text (c.f. (Grover et al., 1993; Doran et al., 1994)). With the advent of empirical methods and the corresponding evaluation metrics, however, this view changed considerably. (Abney, 1996) was among the first who noted that the relationship between coverage and statistical parsing quality is a more complex one. Adding new rules to the grammar, i.e. increasing its coverage, does not only allow the parser to deal with more phenomena, hence more sentences; at the same time it opens up new possibilities for abusing the newly introduced rules to mis-analyse constructions which were already treated properly before. As a consequence, a net reduction in parsing quality might be observed for simple statistical reasons, since the gain usually is obtained for relatively rare phenomena, while the adverse effects might well affect frequent ones.

(Abney, 1996) uses this observation to argue in favour of stochastic models which attempt to choose the optimal structural interpretation instead of only providing a list of equally probable alternatives. However, using such an optimization procedure is not necessarily a sufficient precondition to completely rule out the effect. Compared to traditional handwritten grammars, successful stochastic models like (Collins, 1999; Charniak, 2000) open up an even greater space of alternatives for the parser and accordingly offer a great deal of opportunities to construct odd structural descriptions from them. Whether the guidance of the stochastic model can really prevent the parser from making use of these unwanted opportunities so far remains unclear.

In the following we make a first attempt to quantify the consequences that different degrees of coverage have for the output quality of a wide-coverage parser. For this purpose we use a Weighted Constraint Dependency Grammar (WCDG), which covers even relatively rare syntactic phenomena of German and performs reliably across a wide variety of different text genres (Foth et al., 2005). By combining hand-written rules with an optimization procedure for hypothesis selection, such a parser makes it possible to successively exclude certain rare phenomena from the coverage of the grammar and to study the impact of these modifications on its output quality

2 Some rare phenomena of German

What are good candidates of ‘rare’ phenomena that might be intentionally removed from the coverage of our grammar? One possibility is to remove coverage for constructions that are already slightly dispreferred. For instance, apposition and coordination of noun phrases often violate the principle of projectivity:

“I got a sled for Christmas, a parrot and a motor-bike.”

This is quite a common construction, but still ‘rare’ in the sense that the great majority of appositions does respect projectivity, so that the example seems at least slightly unusual. But there are also syntactic relations that are quite rare but nevertheless appear perfectly normal when they do occur, such as direct appellations:

“James, please open the door.”

This might be because their frequency varies considerably between text types; everyone is familiar with personal appellation from everyday conversation, but it would be surprising to hear it from the mouth of a television news reader.

Finally, some constructions form variants e.g. by omitting certain words:

“I bought a new broom [in order] to clean the drive-

No.	Phenomenon	Example	$f/1000$
1	*Mittelfeld* extraposition	“Es strahlt über DVB-T neben dem Fernsehprogramm auch **seinen Digitext** aus, **einen Videotext-ähnlichen Informationsdienst**.”	32.5
2	ethical dative	“Noch erobere **sich** der PC neue Käuferschichten, heißt es weiter.”	18.5
3	Nominalization	“Täglich kommen rund **1000 neue** hinzu.”	13.4
4	Vocative	“So nicht, **ICANN**!”	9.7
5	Parenthetical matrix clause	“Bis zum Jahresende 2002, **prognostiziert Roland Berger**, werden die am Neuen Markt gelisteten Unternehmen 200.000 Mitarbeiter beschäftigen.”	8.8
6	verb-first subclause	“**Erfüllt ein Mitgliedstaat keines oder nur eines dieser Kriterien**, so erstellt die Kommission einen Bericht.”	8.3
7	Headline phrase	“**Lehrer** kaum auf Computer **vorbereitet**”	3.9
8	coordination cluster	“Auf den Webseiten der Initiative können **Spender PCs anbieten und Schulen ihren Bedarf anmelden**.”	3.1
9	Adverbial pronoun	“Ihre Sprachen sollen **alle** gleichberechtigt sein.”	2.6
10	*um* omission	“Und Dina ging aus, **die Töchter des Landes zu sehen**.”	2.1
11	Metagrammatical usage	“Die Bezugnahmen auf die gemeinsame Agrarpolitik oder auf die Landwirtschaft und die Verwendung des Wortes **“landwirtschaftlich”** sind in dem Sinne zu verstehen, dass damit unter Berücksichtigung der besonderen Merkmale des Fischereisektors auch die Fischerei gemeint ist.”	1.8
12	Auxiliary flip	“Die Geschädigten werfen Ricardo nun eine erhebliche Mitschuld vor, da größerer Schaden **hätte verhindert werden können**, wenn der Anbieter sofort gesperrt worden wäre.”	1.1
13	Adjectival subclause	“Die Union unterhält ferner, **soweit zweckdienlich**, Beziehungen zu anderen internationalen Organisationen.”	0.9
14	Suffix drop	“Ein **freundlich** Wort, das Maslo intervenieren ließ:”	0.5
15	Elliptical genitive	“**Martins** war auch nicht besser.”	0.3
16	Adverbial noun	“Sie stehen sich **Auge in Auge** gegenüber.”	0.1
17	Verb/particle mismatch	“Außer Windows 9x selbst **können** auch andere Hard- und Softwarekomponenten eines PC mit zu viel Hauptspeicher manchmal nicht **zurecht**.”	0.1
18	*Vorfeld* extraposition	“**Der Verdacht** liegt nahe, **daß hier Schwarzarbeit betrieben wird**.”	0.1
19	double relative subject	“Ich bin der Herr, **der ich** dich aus Ägyptenland herausgeführt habe.”	0.02
20	Relative subject clause	“**Die dir fluchen**, seien verflucht, und **die dich segnen**, seien gesegnet!”	0.04
21	NP extraposition	“Die Verpflichtungen und die Zusammenarbeit in diesem Bereich bleiben im Einklang mit den im Rahmen **der Nordatlantikvertrags-Organisation** eingegangenen Verpflichtungen, **die für die ihr angehörenden Staaten weiterhin das Fundament ihrer kollektiven Verteidigung und das Instrument für deren Verwirklichung ist**.”	0.01

Table 1: Some rare phenomena in modern German.

way.”

Here the longer variant is unambiguously a subclause expressing purpose, while the shorter might be mistaken for a prepositional phrase, so it could be regarded as misleading for the parser.

The selection is necessarily subjective, not only because the delimitation of a phenomenon is subjective (are all kinds of ellipsis fundamentally the same phenomenon or not?) but also because we can remove only those phenomena that are already covered in the first place. Therefore we have selected phenomena

- that were explicitly added to the grammar at some point in order to deal with actually occurring unforeseen constructions,
- that can easily be removed from the grammar without affecting other phenomena,
- and that are relatively rare in all the texts we have investigated.

Table 1 shows the 21 phenomena that we consider in this paper. (Note that the three earlier example sentences correspond to lines 1, 4, and 10 in this table, but that not all lines have exact counterparts in English.) The last column gives the overall frequency per 1,000 sentences of each phenomenon when measured across all trees in our collection.

The collection contains sections of Bible text (Genesis 1–50), law text (the constitutions of Federal Germany and of the European Union), online technical newscasts (`www.heise.de`), novel text, and sentences from the NEGRA corpus of newspaper articles. Table 2 shows the sentence counts of the different sections and the frequency per 1000 of all 21 phenomena in each text type. It can be seen that most of the constructions remain quite rare overall, but often the frequency depends heavily on the text type, so that a high influence of the corpus can be expected for our experiments.

f/1000 Phen.	Bible (2,709)	Law (3,722)	Online (55,327)	Novel (20,253)	News (4,000)	overall (86,011)
1	93.6	24.6	29.0	36.7	28.2	32.6
2	59.6	17.5	12.2	31.3	16.2	18.6
3	21.0	22.7	12.3	12.4	19.5	13.4
4	18.4	0.0	0.1	38.2	1.2	9.7
5	1.1	0.0	5.8	18.2	15.8	8.8
6	3.4	51.4	7.8	2.6	6.8	8.3
7	0.7	3.6	4.8	1.3	7.2	3.9
8	7.1	4.4	3.3	2.4	1.8	3.1
9	7.1	0.5	1.6	5.0	3.5	2.6
10	12.7	1.9	1.9	1.2	1.2	2.0
11	0.4	0.3	2.2	0.5	4.8	1.8
12	1.5	0.0	0.9	1.8	1.5	1.1
13	2.2	0.8	1.0	0.5	0.2	0.9
14	0.7	0.0	0.6	1.2	0.2	0.7
15	1.9	0.0	0.7	0.0	1.0	0.5
16	0.4	0.3	0.2	0.0	0.0	0.1
17	1.1	0.0	0.1	0.0	0.0	0.1
18	0.0	0.0	0.1	0.0	0.2	0.1
19	0.7	0.0	0.0	0.0	0.2	0.0
20	0.7	0.0	0.0	0.0	0.0	0.0
21	0.0	0.3	0.0	0.0	0.0	0.0

Table 2: Frequency of phenomena by text type.

3 Weighted Constraint Dependency Grammar

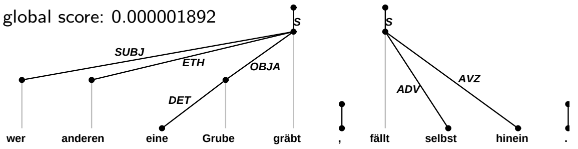
In WCDG (Schröder, 2002), natural language is modelled as labelled *dependency trees*, in which each word is assigned exactly one other word as its regent (only the root of the syntax tree remains unsubordinated) and a label that describes the nature of their relation. The set of acceptable trees is defined not by way of *generative* rules, but only through *constraints* on well-formed structures. Every possible dependency tree is considered correct unless one of its edges or edge pairs violates a constraint. This permissiveness extends to many properties that other grammar formalisms consider non-negotiable; for instance, a WCDG can allow non-projective (or, indeed, cyclical) dependencies simply by not forbidding them. Since the constraints can be arbitrary logical formulas, a grammar rule can also allow some types of non-projective relations and forbid others, and in fact the grammar in question does precisely that.

Weighted constraints can be written to express the fact that a construction is considered acceptable but not fully so. This mechanism is used extensively to achieve robustness against proper errors such as wrong inflection, ellipsis or mis-ordering; all of these are in fact expressed through defeasible constraints. But it can also express more subtle dispreferences against a specific phenomenon by writing only a weak constraint that forbids it; most of the phenomena listed in Table 1 are associated with such constraints to ensure that the parser assumes a rare construction only when this is necessary.

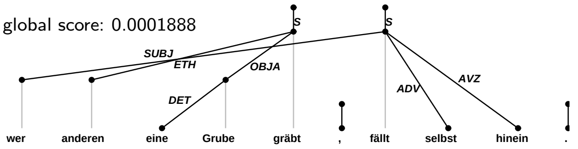
We employ a previously existing wide-coverage WCDG of modern German (Foth et al., 2005) that covers all of the presented rare phenomena. It comprises about 1,000 constraints, 370 of which are hard constraints. The entire parser and the grammar of German are publicly available at `http://nats-www.informatik.uni-hamburg.de/Papa/PapaDownloads`.

The optimal structure could be defined as the tree that violates the least important constraint (as in Optimality Theory), or the tree that violates the fewest constraints; in fact a multiplicative measure is used that combines both aspects by minimizing the collective dispreference for all phenomena in a sentence. Unfortunately, the resulting combinatorial problem is $\mathcal{NP}$-complete and admits of no efficient exact solution algorithm. However, variants of a heuristic *local search* can be used, which try to find the optimal tree by constructing a complete tree and then changing it in those places that violate important constraints. This involves a trade-off between parsing accuracy and processing time, because the correct structure is more likely to be found if there is more time to try out more alternatives. Given enough time, the method works well enough that the overall system exhibits a competitive accuracy even though the theoretical accuracy of the language model may be compromised by search errors.

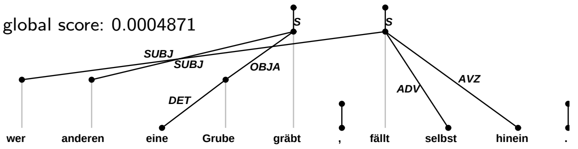
As an example of the process, consider the following analysis of the German proverb “Wer anderen eine Grube gräbt, fällt selbst hinein.” *(He who digs a hole for others, will fall into it himself.)* The transformation starts with the following initial assumption



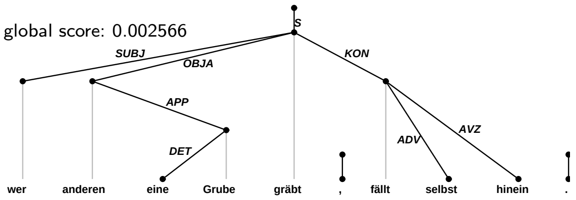
which, besides producing two isolated fragments instead of a spanning tree, also lacks a subject for the second clause.



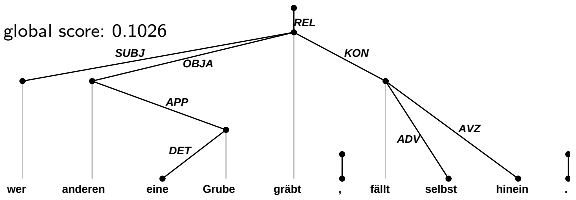
To mend this problem the relative pronoun from the first clause has been taken as a subject for the second one, with the result that the conflict has simply been moved to the first part of the sentence. Nevertheless, the global score improved considerably, since the verb-second condition for German main clauses is violated less often.



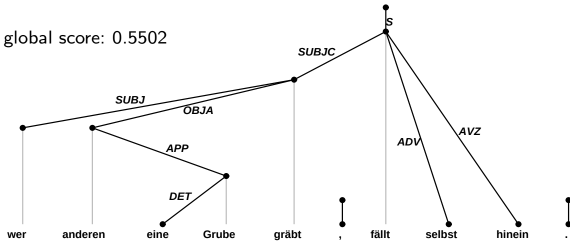
Here, the indefinite plural pronoun ‘anderen’ is taken as the subject for the second clause, creating, however, an agreement error with the finite verb, which is singular. Both subclauses have still not been integrated into a single spanning tree.



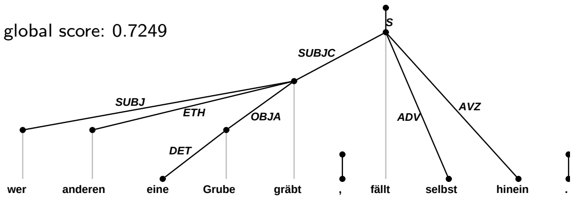
The integration is then achieved, but unfortunately as a coordination without an appropriate conjunction being available. Moreover there is a problem with the hypothesized main clause, since it again does not obey the verb-second condition of German.



Therefore the interpretation is changed to a relative clause, which however cannot appear in isolation. The valency requirements of the verb ‘gräbt’ are satisfied by taking the indefinite pronoun ‘anderen’ as a direct object with the true object (‘eine Grube’) as a (mal-formed) apposition.



Finally, the analysis switches to an interpretation which accepts the second part of the sentence as the main clause and subordinates the first part as a subject clause. The problem with the apposition reading persists.



By interpreting the indefinite pronoun as an ethical dative, the direct object valence is freed for the NP ‘eine Grube’. Although this structure still violates some constraints (e.g. the ethical dative is slightly penalized for being somewhat unusual) a better one cannot be found. Note that the algorithm does not take the shortest possible transformation sequence; in fact, the first analysis could have been transformed directly into the last by only one exchange. Because the algorithm is *greedy*, it chooses a different repair at that point, but it still finds the solution in about three seconds on a 3 GHz Pentium machine.

In contrast to stochastic parsing approaches, a WCDG can be modified in a specifically targeted manner. It therefore provides us with a grammar formalism which is particularly well suited to precisely measure the contributions of different linguistic knowledge sources to the overall parsing quality. In particular it allows us to

1. switch off constraints, i.e. increase the space of acceptable constructions and/or syntactic structures,
2. weaken constraints, by changing the weight in a way that it makes the violation of the constraint condition more easily acceptable,
3. introduce additional dependency labels into the model,
4. remove existing dependency labels from the model
5. reinforce constraints, by removing guards for exceptional cases from them,
6. reinforce constraints, by strengthening their weights or making the constraint non-defeasible in the extreme case, and
7. introducing new constraints, to prohibit certain constructions and/or syntactic structures.

Since for the purpose of our experiments, we start with a fairly broad-coverage grammar of German, from which certain rare phenomena will be removed, options 4 to 7 are most important for us.

4 Robust behaviour under limited coverage

In general, it is not easy to predict the possible outcome of a parsing run when using a grammar with a reduced coverage. Whether a sentence can be analysed at all solely depends on the available alternatives for structuring it. Which structural description it can receive, however, is influenced by the scores resulting from rule applications or constraint violations. Moreover, the transformation-based solution method used for the WCDG-experiments introduces yet another condition: since it is based on a limited heuristics for candidate generation, the grammar must license not only the final parsing result for a sentence, but also all the intermediate transformation steps with a sufficiently high score. This might exclude some structural interpretations from being considered at all if the grammar is not tolerant enough to accommodate highly deviant structures.

Thus, the ability to deal with extragrammatical input in a robust manner is a crucial property if we are going to use a grammar with coverage limitations. Unfortunately, robust behaviour is usually achieved by *extending* instead of *reducing* the coverage of the model and compensating the resulting increase in ambiguity by an appropriately designed scoring scheme together with an optimization procedure.

To deal with these opposing tendencies, it is obviously important to determine which parts of the model need to be relaxed to achieve a sufficient degree of robustness, and which ones can be reinforced to limit the space of alternatives in a sensible way. Excluding phenomena from the grammar which never occur in a corpus should always give an advantage, since this reduces the number of alternatives to consider at each step without forbidding any of the correct ones.

On the other hand, removing support for a construction that is actually needed forces the parser to choose an incorrect solution for at least some part of a sentence, so that a deterioration might occur instead. But even if coverage is reduced below the strictly necessary amount, a net gain in accuracy could occur for two reasons:

1. Leaking: The grammar overgenerates the construction in question, so that forbidding it prevents errors occurring on 'normal' sentences.

2. Focussing: Due to a more restricted search space, the parser is not led astray by rare hypotheses, thus saving processing time which can be used to come closer to the optimum.

4.1 Experiment 1: More with less

In our first experiment, we analysed 10,000 sentences of online newscast texts both with the normal grammar and with the 21 rare phenomena explicitly excluded. As usual for dependency parsers, we measure the parsing quality by computing the structural accuracy (the ratio of correct subordinations to all subordinations) and labelled accuracy (the ratio of all correct subordinations that also bear the correct label to all subordinations). Note that the WCDG parser always establishes exactly one subordination for each word of a sentence, so that no distinction between precision and recall arises. Also, the grammar is written in such a way that even if a necessary phenomenon is removed, the parser will at least find *some* analysis, so that the coverage is always 100%.

As expected, those 'rare' sentences in which at least one of these constructions does actually occur are analyzed less accurately than before: structural and labelled accuracy drop by about 2 percent points (see Table 3). However, the other sentences receive slightly better analyses, and since they are in the great majority, the overall effect is an increase in parsing quality. Note also that the 'rare' sentences appear to be more difficult to analyze in the first place.

Grammar:	Normal	Reduced
Online newscasts		
rare (717)	87.6%/85.2%	85.8%/85.8%
normal (9,283)	91.0%/89.8%	91.4%/90.4%
overall (10,000)	91.0%/89.4%	91.3%/89.7%
NEGRA corpus		
rare (91)	85.5%/83.7%	84.0%/81.4%
normal (909)	91.2%/89.3%	91.5%/89.7%
overall (1,000)	90.5%/88.6%	90.6%/88.7%

Table 3: Structural and labelled accuracy when parsing the same text with reduced coverage.

The net gain in accuracy might be due to plugged leaks (misleading structures that used to be found are rejected in favor of correct structures) or to focussing (structures that were preferred but missed through search errors are now found). A point in case of the latter explanation is the fact that the average runtime decreases by 10% with the reduced grammar. Also, if we consider only those sentences on which the local search originally exceeded the time limit of 500 s and therefore had to be interrupted, the accuracy rises from 85.2%/83.0% to 86.5%/84.4%, i.e. even more pronounced than overall.

4.2 Experiment 2: Stepwise refinement

For comparison with previous work and to investigate corpus-specific effects, we repeated the experiment with the test set of the NEGRA corpus as defined by (Dubey and Keller, 2003). For that purpose the NEGRA annotations were automatically transformed to dependency trees with the freely available tool DEPSY (Daum et al., 2004). Some manual corrections were made to its output to conform to the annotation guidelines of the WCDG of German; altogether, 1% of all words had their regents changed for this purpose.

Table 3 shows that the proportion of sentences with rare phenomena is somewhat higher in the NEGRA sentences, and consequently the net gain in parsing accuracy is smaller; apparently the advantage of reducing the problem size is almost cancelled by the disadvantage of losing necessary coverage.

To test this theory, we then reduced the coverage of the grammar in smaller steps. Since constraints allow us to switch off each of the 21 rare phenomena individually, we can test whether the effects of reducing coverage are merely due to the smaller number of alternatives to consider or whether some constructions affect the parser more than others, if allowed.

We first took the first 3,000 sentences of the NEGRA corpus as a training set and counted how often each construction actually occurs there and in the test set. Table 4 shows that the two parts of the corpus, while different, seem similar enough that statistics obtained

Nr	Phenomenon	Frequency per 1000 on training set	test set
1	*Mittelfeld* extraposition	33.3	13.0
2	ethical dative	16.7	15.0
3	Nominalization	20.3	17.0
4	Vocative	1.0	2.0
5	Parenthetical matrix clause	13.3	23.0
6	verb-first subclause	8.0	3.0
7	Headline phrase	6.7	9.0
8	coordination cluster	1.7	2.0
9	Adverbial pronoun	4.0	2.0
10	*um* omission	1.3	1.0
11	Metagrammatical usage	5.7	2.0
12	Auxiliary flip	2.0	0.0
13	Adjectival subclause	0.0	1.0
14	Suffix drop	1.0	1.0
15	Elliptical genitive	0.0	1.0
16	Adverbial noun	0.0	0.0
17	Verb/particle mismatch	0.0	0.0
18	*Vorfeld* extraposition	0.0	1.0
19	double relative subject	0.0	0.0
20	Relative subject clause	0.3	0.0
21	NP extraposition	0.0	0.0

Table 4: Comparison of training and test set.

on the one could be useful for processing the other. The test set was then parsed again with the coverage successively reduced in several steps: first, all constructions were removed that *never* occur in the training set, then those which occur less than 10 times or 100 times respectively were also removed. We also performed the opposite experiment, first removing support for the least rare phenomena and only then for the really rare ones.

Phenomena removed	structural accuracy	labelled accuracy
none	90.5%	88.6%
$= 0$	90.5%	88.7%
< 10	90.6%	88.8%
< 100	90.7%	88.6%
$>= 100$	90.5%	88.6%
$>= 10$	90.4%	88.5%
> 0	90.5%	88.6%
all	90.6%	88.7%

Table 5: Parsing with coverage reduced stepwise.

Table 5 shows the results of parsing the test set in this way (the first and last lines are repetitions from Table 3). The resulting effects are very small, but they do suggest that removing coverage for the very rare constructions is somewhat more profitable: the first three new experiments tend to yield better accuracy than the original grammar, while in the last three it tends to drop.

4.3 Experiment 3: Plugging known leaks

The previous experiment used only counts from the treebank annotations to determine how rare a phenomenon is supposed to be, but it might also be important how rare the parser actually assumes it to be. The fact that a particular construction never occurs in a corpus does not prevent the parser from using it in its analyses, perhaps more often than another construction that is much more common in the annotations. In other words, we should measure how much each construction actually leaks. To this end, we parsed the training set with the original grammar and grouped all 21 phenomena into three classes:

A: Phenomena that are predicted much more often than they are annotated

B: Phenomena that are predicted roughly the right number of times

C: Phenomena that are predicted less often than annotated (or in fact not at all).

'Much more often' here means 'by a factor of two or more'; constructions which were never predicted *or* annotated at all were grouped into class C.

There are different reasons why a phenomenon might leak more or less. Some constructions depend on particular combinations of word forms in the input; for instance, an auxiliary flip can only be predicted when the finite verb does in fact precede the full verb (phenomenon 12 in Table 1), so that covering it should not change the behaviour of the system much. But most sentences contain more than one noun phrase which the parser might possibly misrepresent as a non-projective extraposition (phenomenon 1). Also, some rare phenomena are dispreferred more than others even when they are allowed. We did not investigate these reasons in detail.

Phenomena removed	structural accuracy	labelled accuracy
none	90.5%	88.6%
A (1,3,4,6–10,13,16,18–21)	90.9%	89.0%
B (2,5,11,12)	90.4%	88.5%
C (14,15,17)	90.4%	88.6%
1–21	90.6%	88.7%

Table 6: Parsing with coverage reduced by increasing leakage.

Table 6 shows an interesting asymmetry: of our 21 constructions, 14 regularly leak into sentences where they have no place, while 4 work more or less as designed. Only 3 are predicted too seldom. This is consistent with our earlier interpretation that most added coverage is in fact unhelpful when judging a parser solely by its empirical accuracy on a corpus.

Accordingly, it is in fact more helpful to judge constructions by their observed tendency to leak than just by their annotated frequency: the first experiment (A) yields the highest accuracy for the newspaper text. Conversely, removing those constructions which actually work largely as intended (B) reduces even the overall accuracy, and not just the accuracy on 'rare' sentences. The third class contains only three very rare phenomena, and removing them from the grammar does not influence parsing very much at all.

Note that this result was obtained although the distribution of the phenomena differs between parser predictions on the training set and the test set; had we classified them according to their behaviour on the test set itself, the class A would have contained only 9 items (of which 7 overlap with the classification actually used).

5 Related work

The fact that leaking is an ubiquitous property of natural language grammars has been noted as early as 80 years ago by (Sapir, 1921). Since no precise definition was given, the notion offers room for interpretation. In general linguistics, leaking is usually understood as the underlying reason for the apparent impossibility to write a grammar which is complete, in the sense that it covers all sentences of a language, while maintaining a precise distinction between correct an incorrect word form sequences (see e.g. (Sampson, forthcoming)). In Computational Linguistics, attention was first drawn to the resulting consequences for obtaining parse trees when it became obvious that all attempts to build wide-coverage grammars led to an increase in output ambiguity, and that even more fine-grained feature-based descriptions were not able solve the problem. Stochastic approaches are usually considered to provide a powerful countermeasure (Manning and Schütze, 1999). However, as (Steedman, 2004) already noted, stochastic models do not address the problem of overgeneration directly.

Disregarding rare phenomena is something that can be achieved in a stochastic framework by putting a threshold on the minimum number of occurrences to be considered. Such an approach is mainly used to either exclude rare phenomena in grammar induction (c.f. (Solsona et al., 2002)) or to prune the search space by adjusting a beam width during parsing itself (Goodman, 1997). The direct use of thresholding techniques at the level of the stochastic model, however, has not been investigated extensively so far. Stochastic models of syntax suffer to such a degree from data sparseness that in effect strong efforts in the opposite direction become necessary: instead of ignoring rare events in the training data, even unseen events are included by smoothing techniques. The only experimental investigation of the impact of rare events we are aware of is (Bod, 2003), where heuristics are explored to constrain the model in the DOP framework by ignoring certain tree fragments. Contrary to the results of our experiments, very few constraints have been found that do not decrease the parse accuracy. In particular, no improvement by disregarding selected observations was possible.

The tradeoff between processing time and output quality which our transformation-based problem solving strategy exhibits, is also a fundamental property of all beam-search procedures. While a limited beam width might cause search errors, widening the beam in order to improve the quality requires investing more computational resources (see e.g. (Collins, 1999)). In contrast to our transformation-based procedure, however, the commonly used Viterbi search is not interruptible and therefore not in a position to really profit from the tradeoff. Thus, focussing as a possibility to increase output quality to our knowledge has never been investigated elsewhere.

6 Conclusions and future work

We have investigated the effect of systematically reducing the coverage of a general grammar of German. By removing support for 21 rare phenomena, the overall parsing accuracy could be improved. We confirmed the initial assumption about the effects that broad coverage has on the parser: while it allows some special sentences to be analysed more accurately, it also causes a slight decrease on the much more numerous normal sentences.

This result shows that at least with respect to this particular grammar, *more* coverage can indeed lead to *less* parsing accuracy. In the first experiment we measured the overall loss through adding coverage where it is not needed as about 0.4% of structural accuracy on newscast text, and 0.1% on NEGRA sentences. This figure can be interpreted as the result of overgenerating or 'leaking' of rare constructions into sentences where they are not wanted.

Although we found that it makes little difference whether to remove support for very rare or for somewhat rare phenomena, judging constructions by how many leaks they actually cause leads to a greater improvement. On the NEGRA test set, removing the 'known troublemakers' leads to a greater increase of in accuracy of 0.4%, reducing the error rate for structural attachment by 4.2%.

Of course, removing rare phenomena is not a viable technique to substantially improve parser accuracy, if only for the simple fact that it does not scale up. However, it confirms that as soon as a certain level of coverage has been reached, robustness, i.e. the ability to deal with unexpected data, is more crucial than coverage itself to achieve high quality results on unrestricted input.

On the other hand, the improvement we obtained is not very large, compared to the already rather high overall performance of the parser. This may be due to the consistent use of weighted constraints in the original grammar, which slightly disprefer many of the 21 phenomena even when they are allowed, and we assume that the original grammar is already reasonably effective at preventing leaks. This claim might be confirmed by reversing the experiment: if all phenomena were allowed *and* all dispreferences switched off, we would expect even more leaks to occur.

To carry out comparable experiments on generative stochastic models presents us with the difficulty that it would first be necessary to determine which of its parameters are responsible for covering a specific phenomenon, and whether they can be modified as to remove the construction from the coverage without affecting others as well. Even in WCDG it is difficult to quantify how much of the observed improvement results from plugged leaks, and how much from focussing. This could only be done by observing all intermediate steps in the solution algorithm, and counting how many trees that were used as intermediate results or considered as alternatives exhibit each phenomenon.

The most promising result from the last experiment is that it is possible to detect particularly detracting phenomena, which are prime candidates for exclusion, in one part of a corpus and use them on another. This suggests itself to be exploited as a method to automatically adapt a broad-coverage grammar more closely to the characteristics of a particular corpus.

References

Steven Abney. 1996. Statistical Methods and Linguistics. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 1–26. The MIT Press, Cambridge, Massachusetts.

Rens Bod. 2003. Do all fragments count? *Natural Language Engineering*, 9(4):307–323.

Eugene Charniak. 2000. A Maximum-Entropy-Inspired Parser. In *Proc. 1st Meeting of the North American Chapter of the ACL, NAACL-2000*, Seattle, WA.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania, Philadephia, PA.

Michael Daum, Kilian Foth, and Wolfgang Menzel. 2004. Automatic transformation of phrase treebanks to dependency trees. In *Proc. 4th Int. Conf. on Language Resources and Evaluation*, pages 99–106, Lisbon, Portugal.

Christy Doran, Dania Egedi, Beth Ann Hockey, B. Srinivas, and Martin Zaidel. 1994. XTAG system - A Wide Coverage Grammar for English. In *Proc. 15th Int. Conf. on Computational Linguistics, COLING-1994*, pages 922 – 928, Kyoto, Japan.

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

Kilian Foth, Michael Daum, and Wolfgang Menzel. 2005. Parsing unrestricted German text with defeasible constraints. In H. Christiansen, P. R. Skadhauge, and J. Villadsen, editors, *Constraint Solving and Language Processing*, volume 3438 of *Lecture Notes in Artificial Intelligence*, pages 88–101, Berlin. Springer-Verlag.

Joshua Goodman. 1997. Global thresholding and multiple-pass parsing. In *Proc. 2nd Int. Conf. on Emprical Methods in NLP, EMNLP-1997*, Boston, MA.

C. Grover, J. Carroll, and E. Briscoe. 1993. The Alvey natural language tools grammar (4th release). Technical Report 284, Computer Laboratory, University of Cambridge.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Natural Language Processing*. MIT Press, Cambridge etc.

Geoffrey Sampson. forthcoming. Grammar without grammaticality. *Corpus Linguistics and Linguistic Theory*.

Edward Sapir. 1921. *Language: An Introduction to the Study of Speech*. Harcourt Brace, New York.

Ingo Schröder. 2002. *Natural Language Parsing with Graded Constraints*. Ph.D. thesis, Department of Informatics, Hamburg University, Hamburg, Germany.

Roger Argiles Solsona, Eric Fosler-Lussier, Hong-Kwang J. Kuo, Alexandros Potamianos, and Imed Zitouni. 2002. Adaptive language models for spoken ddialogue systems. In *Proc. Int. Conf. on Acoustics, Seech, and Signal Processing, ICASSP-2002*, Orlando, FL.

Mark Steedman. 2004. Wide Coverage Parsing with Combinatory Grammars. Slides of a seminar presentation, Melbourne University, Australia. `http://www.cs.mu.oz.au/research/lt/seminars/steedman.pdf`. Last time visited: 2006-01-06.