GLR* – An Efficient Noise-skipping Parsing Algorithm For Context Free Grammars

Alon Lavie and Masaru Tomita

School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213
email: lavie@cs.cmu.edu

Abstract

This paper describes GLR*, a parser that can parse *any* input sentence by ignoring unrecognizable parts of the sentence. In case the standard parsing procedure fails to parse an input sentence, the parser nondeterministically skips some word(s) in the sentence, and returns the parse with fewest skipped words. Therefore, the parser will return some parse(s) with any input sentence, unless no part of the sentence can be recognized at all.

The problem can be defined in the following way: Given a context-free grammar G and a sentence S, find and parse S' - the largest subset of words of S, such that $S' \in L(G)$.

The algorithm described in this paper is a modification of the Generalized LR (Tomita) parsing algorithm [Tomita, 1986]. The parser accommodates the skipping of words by allowing shift operations to be performed from inactive state nodes of the Graph Structured Stack. A heuristic similar to beam search makes the algorithm computationally tractable.

There have been several other approaches to the problem of robust parsing, most of which are special purpose algorithms [Carbonell and Hayes, 1984], [Ward, 1991] and others. Because our approach is a modification to a standard context-free parsing algorithm, all the techniques and grammars developed for the standard parser can be applied as they are. Also, in case the input sentence is by itself grammatical, our parser behaves exactly as the standard GLR parser.

The modified parser, GLR*, has been implemented and integrated with the latest version of the Generalized LR Parser/Compiler [Tomita *et al.*, 1988], [Tomita, 1990].

We discuss an application of the GLR* parser to spontaneous speech understanding and present some preliminary tests on the utility of the GLR* parser in such settings.

1 Introduction

In this paper, we introduce a technique for substantially increasing the robustness of syntactic parsers to two particular types of extra- grammaticality: noise in the input, and limited grammar coverage. Both phenomena cause a common situation, where the input contains words or fragments that are unparsable. The distinction between these two types of extra- grammaticality is based to a large extent upon whether or not the unparsable fragment, in its context, can be considered grammatical by a linguistic judgment. This distinction may indeed be vague at times, and practically unimportant.

Our approach to the problem is to enable the parser to overcome these forms of extra-grammaticality by ignoring the unparsable words and fragments and focusing on the maximal subset of the input that is covered by the grammar. Although presented and implemented as an enhancement to the Generalized LR parsing paradigm, our technique is applicable in general to most phrase- structured based parsing formalisms. However, the efficiency of our parser is due in part to several particular properties of GLR parsing, and may thus not be easily trans-

ferred to other syntactic parsing formalisms.

The problem can be formalized in the following way: Given a context-free grammar G and a sentence S, find and parse S' - the largest subset of words of S, such that $S' \in L(G)$.

A naive approach to this problem is to exhaustively list and attempt to parse all possible subsets of the input string. The largest subset can then be selected from among the subsets that are found to be parsable. This algorithm is clearly computationally infeasible, since the number of subsets is exponential in the length of the input string. We thus devise an efficient method for accomplishing the same task, and pair it with an efficient search approximation heuristic that maintains runtime feasibility.

The algorithm described in this paper, which we have named GLR*, is a modification of the Generalized LR (Tomita) parsing algorithm. It has been implemented and integrated with the latest version of the GLR Parser/Compiler [Tomita *et al.*, 1988], [Tomita, 1990].

There have been several other approaches to the problems of robust parsing, most of which have been special purpose algorithms. Some of these approaches have abandoned syntax as a major tool in handling extra-grammaticalities and have focused on domain dependent semantic methods [Carbonell and Hayes, 1984], [Ward, 1991]. Other systems have constructed grammar and domain dependent fall-back components to handle extra-grammatical input that causes the main parser to fail [Stallard and Bobrow, 1992], [Seneff, 1992].

Our approach can be viewed as an attempt to extract from the input the maximal syntactic structure that is possible, within a purely syntactic and domain independent setting. Because the GLR* parsing algorithm is an enhancement to the standard GLR context-free parsing algorithm, all of the techniques and grammars developed for the standard parser can be applied as they are. In particular, the standard LR parsing tables are compiled in advance from the grammar and used "as is" by the parser in runtime. The GLR* parser inherits the benefits of the original parser in terms of ease of grammar development, and, to a large extent, efficiency properties of the parser itself. In the case that the input sentence is by itself grammatical, GLR* behaves exactly as the standard GLR parser.

The remaining sections of the paper are organized in the following way: Section 2 presents an outline of the basic GLR* algorithm itself, followed by a detailed example of the operation of the parser on a simple input string. In section 3 we discuss the search heuristic that is added to the basic GLR* algorithm, in order to ensure its runtime feasibility. We discuss an application of the GLR* algorithm to spontaneous speech understanding, and present some preliminary test results in section 4. Finally, our conclusions and further research directions are presented in section 5.

(1) S → NP VP
(2) NP → det n
(3) NP → n
(4) NP → NP PP
(5) VP → v NP
(6) PP → p NP

Figure 1: A Simple Natural Language Grammar

2 The GLR* Parsing Algorithm

The GLR* parsing algorithm is an extension of the Generalized LR Parser, as implemented in the Universal Parser Architecture developed at CMU [Tomita, 1986]. This implementation incorporates an SLR(0) parsing table.

The parser accommodates skipping words of

the input string by allowing shift operations to be performed from inactive state nodes in the Graph Structured Stack (GSS). Shifting an input symbol from an inactive state is equivalent to skipping the words of the input that were encountered after the parser reached the inactive state and prior to the current word being shifted. Since the parser is LR(0), reduce operations need not be repeated for skipped words (the reductions do not depend on any lookahead). Information about skipped words is maintained in the symbol nodes that represent parse sub-trees.

An initial version of the GLR* parser has been implemented in Lucid Common Lisp, in the integrated environment of the Universal Parser Architecture.

	Reduce	Shift					Goto			
State		det	n	v	p	$	NP	VP	PP	S
0		sh3	sh4				2			1
1						acc				
2				sh7	sh8			5	6	
3			sh9							
4	r3									
5	r1									
6	r4									
7		sh3	sh4				10			
8		sh3	sh4				11			
9	r2									
10	r5				sh8				6	
11	r6				sh8				6	

Table 1: SLR(0) Parsing Table for Grammar in Figure 1

2.1 An Example

To clarify how the proposed GLR* parser actually works, in lieu of a more formal description of the algorithm itself, we present a step by step runtime example. For the purpose of the example, we use a simple natural language grammar that is shown in Figure 1. The terminal symbols of the grammar are depicted in lower-case, while the non-terminals are in upper-case. The grammar is compiled into an SLR(0) parsing table, which is displayed in Table 1. Note that since the table is SLR(0), the reduce actions are independent of any lookahead. The actions on states 10 and 11 include both a shift and a reduce.

To understand the operation of the parser, we now follow some steps of the GLR* parsing algorithm on the input `x = det n v n det p n`. This input is ungrammatical due to the second "`det`" token. The maximal parsable subset of the input in this case is the string that includes all words other than the above mentioned "`det`".

In the figures ahead, which graphically display the GSS of the parser in various stages of the parsing process, we use the following notation:

- An *active* (top level) state node is represented by the symbol "`@`", with the state number indicated above it. Actions that are attached to the node are indicated to the right of the node.
- An *inactive* state node is represented by the symbol "`*`". The state number is indicated above the node and actions that are attached to the node are indicated above the state number.
- Grammar symbol nodes are represented by the symbol "`#`", with the grammar symbol itself displayed above it.

```
0                                  after initialization
@ sh3                              (and empty reduce phase)
```

Figure 2: Initial GSS

```
sh4                                after first shift phase
 0  det  3                         (and empty reduce phase)
 *---#---@ sh9
```

Figure 3: GSS after first shift phase

The parser operates in phases of shifts and reductions. We follow the GSS of the parser following each of these phases, while processing the input string. Reduce actions are distributed to the active nodes during initialization and after each shift phase. Shift actions are distributed after each reduce phase. Note that the GLR* parsing algorithm distributes shift actions to *all* state nodes (both active and inactive), whereas the original parser distributed shift actions only to active nodes. Reduce actions are distributed only to active state nodes.

Figure 2 is the initial GSS, with an active state node of state 0. Since there are no reduce actions from state 0, the first reduce phase is empty. With the first input token being "det", the shift action attached to state node 0 is "sh3".

Figure 3 shows the GSS after the first shift phase. The symbol node labeled "det" has been shifted and connected to the initial state node and to the new active state node of state 3. Since there are no reduce actions from state 3, the next reduce phase is empty. The next input token is "n". Shift actions are distributed by the algorithm to both the active node of state 3 and the inactive node of state 0, as can be seen in Figure 3.

Figure 4 shows the GSS after the next shift phase. The input token "n" was shifted from both state nodes, creating active state nodes of states 9 and 4. The shifting of the input token "n" from state 0 corresponds to a parsing possibility in which the first input token "det" is skipped. Reduce actions are distributed to both of the active nodes.

The following reduce phase reduces both branches into noun phrases. The two "NP"s are packed together by a local ambiguity packing procedure. Using information on skipped words that is maintained within the symbol nodes, the ambiguity packing can detect that one of the noun phrases (the one that was reduced from "det n") is more complete, and the other noun phrase is discarded. The resulting GSS is displayed in Figure 5. Shift actions with the next input token "v" are then distributed to all the state nodes. However, in this case, only state 2 allows a shift of "v" into state 7.

Figure 6 shows the GSS after the third shift phase. The state 7 node is the only active node at this point. Since no reduce actions are specified for this state, the fourth reduce phase is empty. Shift actions with the next input token "n" are distributed to all state nodes, as can be seen in the figure.

Figure 7 shows the GSS after the fourth shift phase and Figure 8 after the fifth reduce phase. Note that there are no active state nodes after the fifth reduce phase. This is due to the fact that none of the state nodes produced by the reduce phase allow the shifting of the next input token "det". The original parser would have thus failed as this point. However, the GLR* parser succeeds in distributing shift actions to two inactive state nodes in this case.

For the sake of brevity we do not continue to further follow the parsing step by step. The final GSS is displayed in Figure 9. Several different parses, with different subsets of skipped words are actually packed into the single "S" node seen at the bottom of the figure. The parse that corre-

sponds to the maximal subset of the input is the one in which the second "det" is the only word skipped.

Figure 4: GSS after second shift phase

2.2 Efficiency of the Parser

Efficiency of the parser is achieved by a number of different techniques. The most important of these is a sophisticated process of local ambiguity packing and pruning. A local ambiguity is a part of the input sentence that corresponds to a phrase (thus, reducible to some non-terminal symbol of the grammar), and is parsable in more than one way. The process of skipping words creates a large number of local ambiguities. For example, the grammar in Figure 1 allows both determined and undetermined noun phrases (rules 2 and 3). As seen in the example presented earlier, this results in the creation of two different noun phrase symbol nodes for the initial fragment "det n". The first node is created for the full phrase after a reduction according to the first rule. A second symbol node is created when the determiner is skipped and a reduction by the second rule takes place.

Locally ambiguous symbol nodes are detected as nodes that are surrounded by common state nodes in the GSS. The original GLR parser detects such local ambiguities and packs them into a single symbol node. This procedure was extended in the GLR* parser. Locally ambiguous symbol nodes are compared in terms of the words skipped within them. In cases such as the example described above, where one phrase has more skipped words than the other, the phrase with more skipped words is discarded in favor of the more complete parsed phrase. This subsuming operation drastically reduces the number of parses being pursued by the parser.

Another technique employed to increase the efficiency of the parser is the merging of state nodes of the same state after a reduce phase and after a shift phase. This allows the parsing through the GSS to continue with fewer state nodes.

2.3 Selecting the Best Maximal Parse

An obvious and unsurprising side effect of the GLR* parser is an explosion in the number of parses found by the parser. In principle, we are only interested in finding the maximal parsable subset of the input string (and its parse). However, in many cases there are several distinct maximal parses, each consisting of a different subset

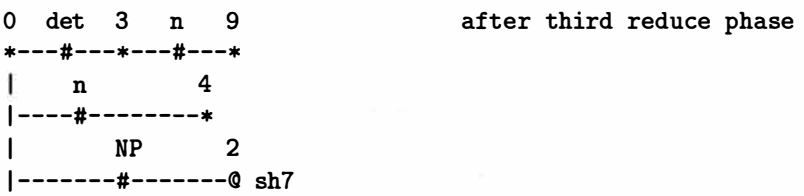

Figure 5: GSS after third reduce phase

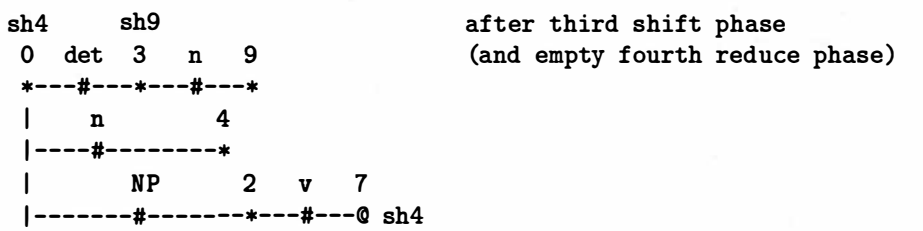

Figure 6: GSS after third shift phase

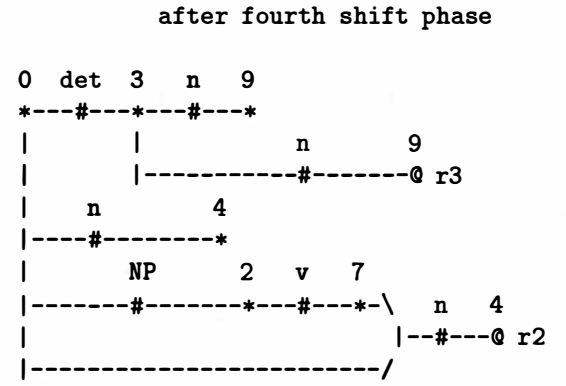

Figure 7: GSS after fourth shift phase

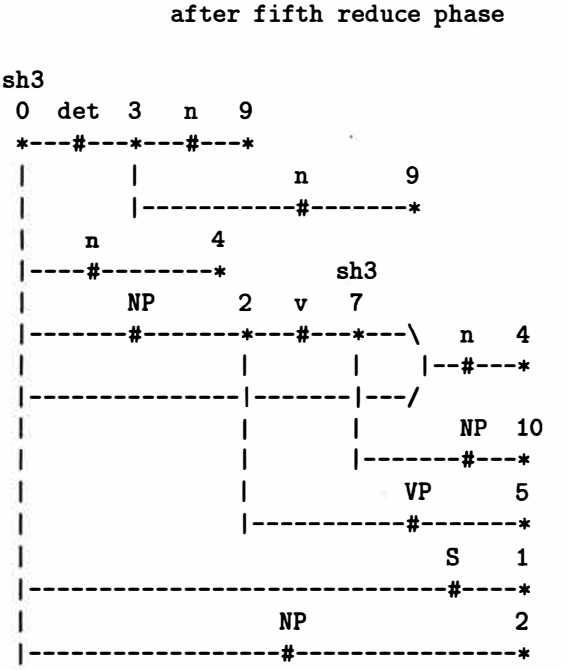

Figure 8: GSS after fifth reduce phase

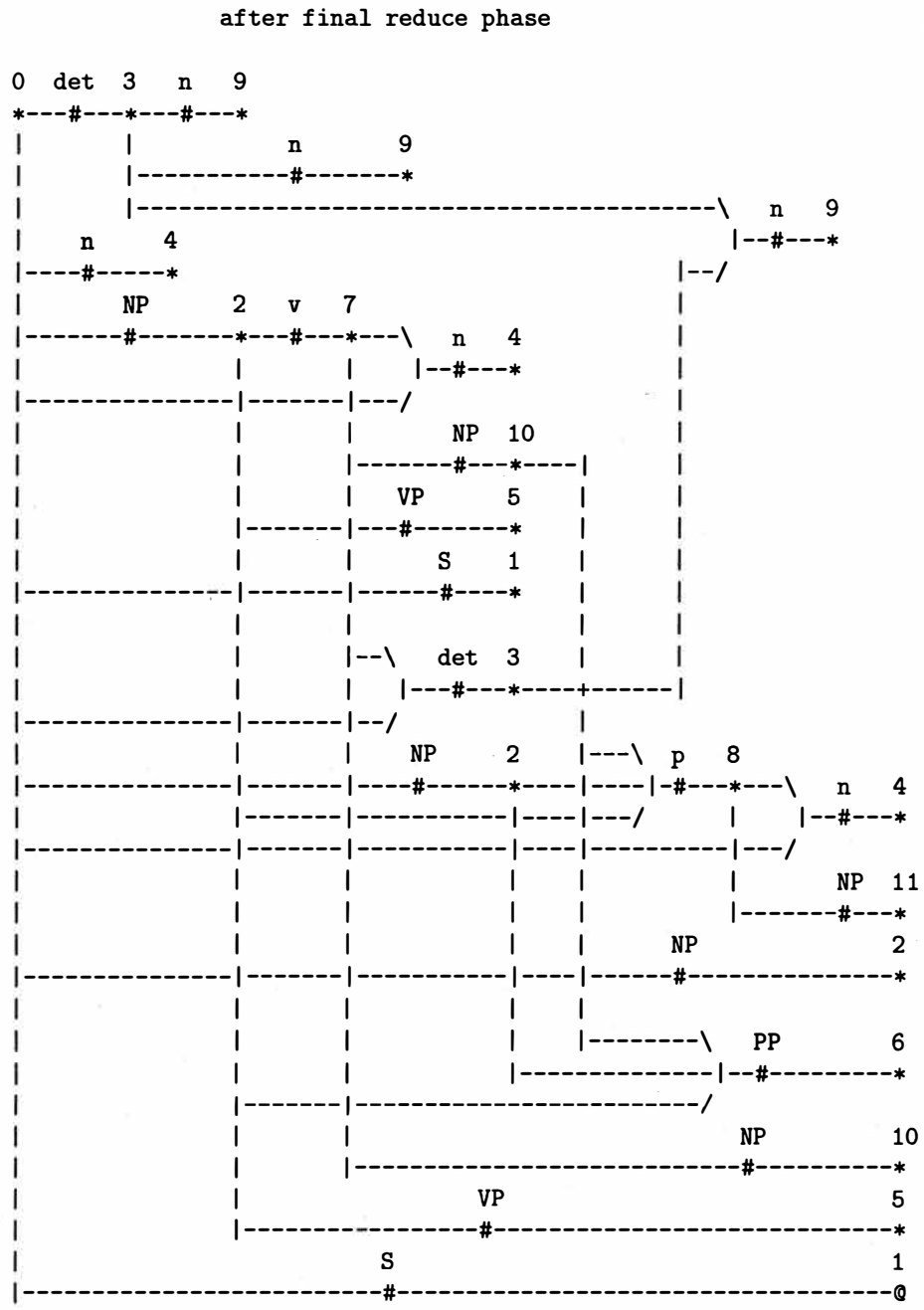

Figure 9: GSS after final reduce phase

of words of the original sentence. Additionally, there are cases where a parse that is not maximal in terms of the number of words skipped may be deemed preferable.

To select the "best" parse from the set of parses returned by the parser, we use a scoring procedure that ranks each of the parses found. We then select the parse that was ranked best. [1] Presently, our scoring procedure is rather simple. It takes into account the number of words skipped and the fragmentation of the parse (i.e. the number of S-nodes that the parsed input sentence was divided into). Both measures are weighed equally. Thus a parse that skipped one word but parsed the remaining input as a single sentence is preferred over a parse that fragments the input into three sentences, without skipping any input word.

On the top of our current research goals is the enhancement of this simple scoring mechanism. We plan on adding to our scoring function several additional heuristic measures that reflect various syntactic and semantic properties of the parse tree. We will measure the effectiveness of our enhanced scoring function in ranking the parse results by their desirability.

3 The Beam Search Heuristic

Although implemented efficiently, the basic GLR* parser is still not guaranteed to have a feasible running time. The basic GLR* algorithm described computes parses of all parsable subsets of the original input string, the number of which is potentially exponential in the length of the input string. Our goal is to find parses of maximal subsets of the input string (or almost maximal subsets). We have therefore developed and added to the parser a heuristic that prunes parsing options that are not likely to produce a maximal parse. This process has been traditionally called "beam search".

A direct way of adding a beam search to the parser would be to limit the number of active state nodes pursued by the parser at each stage, and continue processing only active nodes that are most promising in terms of the number of skipped words associated with them. However, the structure of the GSS makes it difficult to associate information on skipped words directly with the state nodes. [2] We have therefore opted to implement a somewhat different heuristic that has a similar effect.

Since the skipping of words is the result of performing shift operations from inactive state nodes of the GSS, our heuristic limits the number of inactive state nodes from which a input symbol is shifted. At each shift stage, shift actions are first distributed to the active state nodes of the GSS. This corresponds to no additional skipped words at this stage. If the number of state nodes that allow a shift operation at this point is less than a predetermined constant threshold (the "beam-limit"), then shift operations from inactive state nodes are also considered. Inactive states are processed in an ordered fashion, so that shifting from a more recent state node that will result in fewer skipped words is considered first. Shift operations are distributed to inactive state nodes in this way until the number of shifts distributed reaches the threshold.

This beam search heuristic reduces the runtime of the GLR* parser to within a constant factor of the original GLR parser. Although it is not guaranteed to find the desired maximal parsable subset of the input string, our preliminary tests have shown that it works well in practice.

The threshold (beam-limit) itself is a parameter that can be dynamically set to any constant value at runtime. Setting the beam-limit to a value of 0 disallows shifting from inactive states all together, which is equivalent to the original GLR parser. In preliminary experiments that we have conducted (see next section) we have achieved good results with a setting of the beam-limit to values in the range of 5 to 10. There exists a direct tradeoff between the value of the beam-limit and the runtime of the GLR* parser. With a set value of 5, our tests have indicated a runtime that is within a factor of 2-3 times that of the original GLR parser, which amounts to a parse time of only several seconds on sentences that are up to 30 words long.

[1] The system will display the n best parses found, where the parameter n is controlled by the user at runtime. By default, we set n to one, and the highest ranking parse is displayed.

[2] This is due to the fact that state nodes are merged, so that a state node may be common to several different parses, with different skipped words associated with each parse.

	Robust Parser
	number (and percent)
Parsable	99
Unparsable	1
Good/Close Parses	77
Bad Parses	22

Table 2: Performance of the GLR* Parser on Spontaneous Speech

4 Parsing of Spontaneous Speech Using GLR*

4.1 The Problem of Parsing Spontaneous Speech

As a form of input, spontaneous speech is full of noise and irrelevances that surround the meaningful words of the utterance. Some types of noise can be detected and filtered out by speech recognizers that process the speech signal. A parser that is designed to successfully process speech recognized input must however be robust to various forms of noise, and be able to weed out the meaningful words from the rest of the utterance.

When parsing spontaneous spoken input that was recognized by a speech recognition system, the parser must deal with three major types of extra-grammaticality:

- Noise due to the spontaneity of the speaker, such as repeated words, false beginnings, stuttering, and filled pauses (i.e. "ah", "um", etc.).
- Ungrammaticality that is due to the language of the speaker, or to the coverage of the grammar.
- Noise due to errors of the speech recognizer.

We have conducted two preliminary experiments to evaluate the GLR* parser's ability to overcome the first two types of extra-grammaticality. We are in the process of experimenting with the GLR* parser on actual speech recognized output, in order to test its capabilities in handling errors produced by the speech recognizer.

4.2 Parsing of Noisy Spontaneous Speech

The first test we conducted was intended to evaluate the performance of the GLR* parser on noisy sentences typical of spontaneous speech. The parser was tested on a set of 100 sentences of transcribed spontaneous speech dialogues on a conference registration domain. The input is hand-coded transcribed text, not processed through any speech recognizer. The grammar used was an upgraded version of a grammar for the conference registration task, developed and used by the JANUS speech-to-speech translation project at CMU [Waibel et al. 1991]. Since the test sentences were drawn from actual speech transcriptions, they were not guaranteed to be covered by the grammar. However, since the test was meant to focus on spontaneous noise, sentences that included verbs and nouns that were beyond the vocabulary of the system were avoided. Also pruned out of the test set were short opening and closing sentences (such as "hello" and "goodbye"). The transcriptions include a multitude of noise in the input. The following example is one of the sentences from this test set:

```
"fckn2_10 /ls/ /h#/ um okay {comma}
then yeah I am disappointed {comma}
*pause* but uh that is okay {period}"
```

The performance results are presented in Table 2. Note that due to the noise contaminating the input, the original parser is unable to parse a single one of the sentences in this test set. The GLR* parser succeeded to return some parse result in all but one of the test sentences. However, since returning a parse result does not by itself guarantee an analysis that adequately reflects the meaning of the original utterance, we reviewed the parse results by hand, and classified them into the categories of "good/close" and "bad" parses. The results of this classification are included in the table.

4.3 Grammar Coverage

We conducted a second experiment aimed exclusively on evaluating the ability of the GLR* parser to overcome limited grammar coverage. In this experiment, we compared the results of the GLR* parser with those of the original GLR parser on a common set of sentences using the same grammar. We used the grammar from the spontaneous speech experiment for this test as well. The common test set was a set of 117 sentences from the conference registration task of the JANUS project. These sentences are simple synthesized text sentences. They contain no spontaneous speech noise, and are not the result of any speech recognition processing. Once again, to evaluate the quality of the parse results returned by the parser, we classified the parse results of both parsers by hand into two categories: "good/close parses" and "bad parses". The results of the experiment are presented in Table 3.

The results indicate that using the GLR* parser results in a significant improvement in performance. The percentage of sentences, for which the parser returned good or close parses increased from 52% to 70%, an increase of 18%. Fully 97% of the test sentences (all but 3) are parsable by the GLR* parser, an increase of 36% over the original parser. However, this includes a significant increase (from 9% to 27%) in the number of bad parses found. Thus, fully half of the additional parsable sentences of the set return with parses that may be deemed bad.

The results of the two experiments clearly point to the following problem: Compared with the GLR* parser, the original GLR parser, although fragile, returned results of relatively good quality, when it succeeded in parsing the input. The GLR* parser, on the other hand, will succeed in parsing almost any input, but this parse result may be of little or no value in a significant portion of cases. This indicates a strong need in the development of methods for discriminating between good and bad parse results. We intend to try and develop some effective heuristics to deal with this problem. The problem is also due in part to the ineffectiveness of the simple heuristics currently employed for selecting the best parse result from among the large set of parses returned by the parser. As mentioned earlier, we intend to concentrate efforts on developing more sophisticated and effective heuristics for selecting the best parse.

5 Conclusions and Future Research Directions

Motivated by the difficulties that standard syntactic parses have in dealing with extra-grammaticalities, we have developed GLR*, an enhanced version of the standard Generalized LR parser, that can effectively handle two particular problems that are typical of parsing spontaneous speech: noise contamination and limited grammar coverage.

Given a grammar G and and input string S, GLR* finds and parses S', the maximal subset of words of S, such that S' is in the language $L(G)$. The parsing algorithm accommodates the skipping of words and fragments of the input string by allowing shift operations to be performed from inactive states of the GSS (as well as from the active states, as is done by the standard parser). The algorithm is coupled with a beam-search-like heuristic, that controls the process of shifting from inactive states to a limited beam, and

	Original Parser		Robust Parser	
	number	percent	number	percent
Parsable	71	61%	114	97%
Unparsable	46	39%	3	3%
Good/Close Parses	61	52%	82	70%
Bad Parses	10	9%	32	27%

Table 3: Performance of the GLR* Parser vs. the Original Parser

maintains computational tractability.

Most other approaches to robust parsing have suffered to some extent from a lack of generality and from being domain dependent. Our approach, although limited to handling only certain types of extra-grammaticality, is general and domain independent. It attempts to maximize the robustness of the parser within a purely syntactic setting. Because the GLR* parsing algorithm is a modification of the standard GLR context-free parsing algorithm, all of the techniques and grammars developed for the standard parser can be applied as they are. In the case that the input sentence is by itself grammatical, GLR* behaves exactly as the standard GLR parser. The techniques used in the enhancement of the standard GLR parser into the robust GLR* parser are in principle applicable to other phrase-structure based parsers.

Preliminary experiments conducted on the effectiveness of the GLR* parser in handling noise contamination and limited grammar coverage have produced encouraging results. However, they have also pointed out a definite need to develop effective heuristics that can select the best parse result from a potentially large set of possibilities produced by the parser. Since the GLR* parser is likely to succeed in producing some parse in practically all cases, successful parsing by itself can no longer be an indicator to the value and quality of the parse result. Thus, additional heuristics need to be developed for evaluating the quality of the parse found.

We intend to concentrate on developing such effective heuristics that will complement the GLR* parser, and boost its performance in handling spontaneously spoken input. We plan to conduct extensive experiments with speech recognized input to evaluate our system and guide its further development. We also plan to investigate the potential of the GLR* parser in several other application areas and domains.

References

[Carbonell and Hayes, 1984] J. G. Carbonell and P. J. Hayes. Recovery Strategies for Parsing Extragrammatical Language. *Technical Report CMU-CS-84-107*, 1984.

[Seneff, 1992] S. Seneff. A relaxation method for understanding spontaneous speech uttera nces. In *Proceedings of DARPA Speech and Natural Language Workshop*, pages 299–304, February 1992.

[Stallard and Bobrow, 1992] D. Stallard and R. Bobrow. Fragment processing in the DELPHI system. In *Proceedings of DARPA Speech and Natural Language Workshop*, pages 305–310, February 1992.

[Tomita *et al.*, 1988] M. Tomita, T. Mitamura, H. Musha, and M. Kee. The Generalized LR Parser/Compiler - Version 8.1: User's Guide . *Technical Report CMU-CMT-88-MEMO*, 1988.

[Tomita, 1986] M. Tomita. *Efficient Parsing for Natural Language.* Kluwer Academic Publishers, Hingham, Ma., 1986.

[Tomita, 1990] M. Tomita. The Generalized LR Parser/Compiler - Version 8.4. In *Proceedings of International Conference on Computational Lingui stics (COLING-90)*, pages 59–63, Helsinki, Finland, 1990.

[Ward, 1991] W. Ward. Understanding spontaneous speech: The Phoenix system. In *Proceedings of IEEE International Conference on Acoustics, Spee ch and Signal Processing (ICASSP)*, pages 365–367, April 1991.