Using Chunk Based Partial Parsing of Spontaneous Speech in Unrestricted Domains for Reducing Word Error Rate in Speech Recognition

Klaus Zechner and **Alex Waibel**
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213, USA
{zechner,ahw}@cs.cmu.edu

Abstract

In this paper, we present a chunk based partial parsing system for spontaneous, conversational speech in unrestricted domains. We show that the chunk parses produced by this parsing system can be usefully applied to the task of reranking Nbest lists from a speech recognizer, using a combination of chunk-based n-gram model scores and chunk coverage scores.
The input for the system is Nbest lists generated from speech recognizer lattices. The hypotheses from the Nbest lists are tagged for part of speech, "cleaned up" by a preprocessing pipe, parsed by a part of speech based chunk parser, and rescored using a backpropagation neural net trained on the chunk based scores. Finally, the reranked Nbest lists are generated.
The results of a system evaluation are promising in that a chunk accuracy of 87.4% is achieved and the best performance on a randomly selected test set is a decrease in word error rate of 0.3 percent (absolute), measured on the new first hypotheses in the reranked Nbest lists.

1 Introduction

In the area of parsing spontaneous speech, most work so far has primarily focused on dealing with texts within a narrow, well-defined domain. Full scale parsers for spontaneous speech face severe difficulties due to the intrinsic nature of spoken language (e.g., false starts, hesitations, ungrammaticalities), in addition to the well-known complexities of large coverage parsing systems in general (Lavie, 1996; Light, 1996).

An even more serious problem is the imperfect word accuracy of speech recognizers, particularly when faced with spontaneous speech over a large vocabulary and over a low bandwidth channel. This is particularly the case for the SWITCHBOARD database (Godfrey et al., 1992) which we mainly used for development, testing, and evaluation of our system. Current state-of-the-art recognizers exhibit word error rates (WER[1]) for this corpus of approximately 30%–40% (Finke et al., 1997). This means that in fact about every third word in an input utterance will be misrecognized. Thus, any parser which is too restrictive with respect to the input it accepts will likely fail to find a parse for most of these utterances.

When the domain is restricted, sufficient coverage can be achieved using semantically guided approaches that allow skipping of unparsable words or segments (Ward, 1991; Lavie, 1996).

Since we cannot build on semantic knowledge for constructing parsers in the way it is done for limited domains when attempting to parse spontaneous speech in *unrestricted domains*, we argue that more shallow approaches have to be employed to reach a sufficient reliability with a reasonable amount of effort.

In this paper, we present a chunk based partial parser, following ideas from (Abney, 1996), which is used to to generate shallow syntactic structures from speech recognizer output. These representations then serve as the basis for scores used in the task of reranking Nbest lists.

The organization of this paper is as follows: In section 2 we introduce the concept of chunk parsing and how we interpret and use it in our system. Section 3 deals with the issue of reranking Nbest lists and the question of why we consider it appropriate to use chunk representations for this task. In section 4, the system architecture is described, and then the results from an evaluation of the system are presented and discussed (sections 5 and 6). Finally, we give the results of a small study with human subjects on an analogous task (section 7), before pointing out directions for future research (section 8) and summarizing our work (section 9).

2 Chunk Parsing

There have been recent developments which encourage the investigation of the possibility of parsing speech in unrestricted domains. It was demonstrated that parsing natural language[2] can be han-

[1] The word error rate (WER in %) is defined as follows: $WER = 100.0 * \frac{substitutions + deletions + insertions}{correct + substitutions + deletions}$

[2] mostly of the written, but also of the spoken type

dled by very simple, even finite-state approaches if one adheres to the principle of "chunking" the input into small and hence easily manageable constituents (Abney, 1996; Light, 1996).

We use the notion of a *chunk* similar to (Abney, 1996), namely a *contiguous, non-recursive phrase*. Chunk phrases mostly correspond to traditional notions of syntactic constituents, such as NPs or PPs, but there are exceptions, e.g. VCs ("verb complex phrases"), which are not used in most traditional linguistic paradigms.[3] Unlike in (Abney, 1996), our goal was not to build a multi-stage, cascaded system to result in full sentence parses, but to confine ourselves to parsing of "basic chunks".

A strong rationale for following this simple approach is the nature of the ill-formed input due to (i) spontaneous speech dysfluencies, and (ii) errors in the hypotheses of the speech recognizer.

To get an intuitive feel about the output of the chunk parser, we present a short example here:[4]

```
[conj BUT] [np HE] [vc DOESN'T REALLY LIKE]
[np HIS HISTORY TEACHER] [advp VERY MUCH]
```

3 Reranking of Speech Recognizer Nbest Lists

State-of-the-art speech recognizers, such as the JANUS recognizer (Waibel et al., 1996) whose output we used for our system, typically generate lattices of word hypotheses. From these lattices, Nbest lists can be computed automatically, such that it is ensured that the ordering of hypotheses in these lists corresponds to the internal ranking of the speech recognizer.

As an example, we present a reference utterance (i.e., "what was actually said") and two hypotheses from the Nbest list, given with their rank:

```
REF: YOU WEREN'T BORN JUST TO SOAK UP SUN
  1: YOU WEREN'T BORN JUSTICE SO CUPS ON
190: YOU WEREN'T BORN JUST TO SOAK UP SUN
```

This is a typical example, in that it is frequently the case that hypotheses which are ranked further down the list, are actually closer to the true (reference) utterance (i.e., the WER would be lower).[5] So, if we had an oracle that could tell the speech recognizer to always pick the hypothesis with the lowest WER from the Nbest list (instead of the top ranked hypothesis), the global performance could be improved significantly.[6]

In the speech recognizer architecture, the search module is guided mostly by very local phenomena, both in the acoustic models (a context of several phones), and in the language models (a context of several words). Also, the recognizer does not make use of any syntactic (or constituent-based) knowledge.

Thus, the intuitive idea is to generate representations that allow for a discriminative judgment between different hypotheses in the Nbest list, so that eventually a more plausible candidate can be identified, if, as it is the case in the following example, the resulting chunk structure is more likely to be well-formed than that of the first ranked hypothesis:

```
  1: [np YOU] [vc WEREN'T BORN] [np JUSTICE]
     [advp SO] [np CUPS] [advp ON]
190: [np YOU] [vc WEREN'T BORN]
     [advp JUST] [vc TO SOAK UP] [np SUN]
```

We use two main scores to assess this plausibility: (i) a *chunk coverage* score (percentage of input string which gets parsed), and (ii) a *chunk language model* score, which is using a standard n-gram model based on the chunk sequences. The latter should give worse scores in cases like hypothesis (1) in our example, where we encounter the `vc-np-advp-np-advp` sequence, as opposed to hypothesis (190) with the more natural `vc-advp-vc-np` sequence.

4 System Architecture

4.1 Overview

Figure 1 shows the global system architecture. The Nbest lists are generated from lattices that are produced by the JANUS speech recognizer (Waibel et al., 1996). First, the hypothesis duplicates with respect to silence and noise words are removed from the Nbest lists[7], next the word stream is tagged with Brill's part of speech (POS) tagger (Brill, 1994), Version 1.14, adapted to the SWITCHBOARD Corpus. Then, the token stream is "cleaned up" in the preprocessing pipe, which then serves as the input of the POS based chunk parser. Finally, the chunk representations generated by the parser are used to compute scores which are the basis of the rescoring component that eventually generates new reranked Nbest lists.

In the following, we describe the major components of the system in more detail.

[3] A VC-chunk is a *contiguous* verbal segment of an utterance, whereas a VP usually comprises this verbal segment *and* its arguments together.

[4] conj=conjunction chunk, np=noun phrase chunk, vc=verb complex chunk, advp=adverbial phrase chunk

[5] In this case, hypothesis 190 is completely correct; generally it is not the case, particularly for longer utterances, to find the correct hypothesis in the lattice.

[6] On our data, from WER=43.5% to WER=30.4%, using the top 300 hypotheses of each utterance (see Table 1).

[7] since we are ignoring these pieces of information in later stages of processing

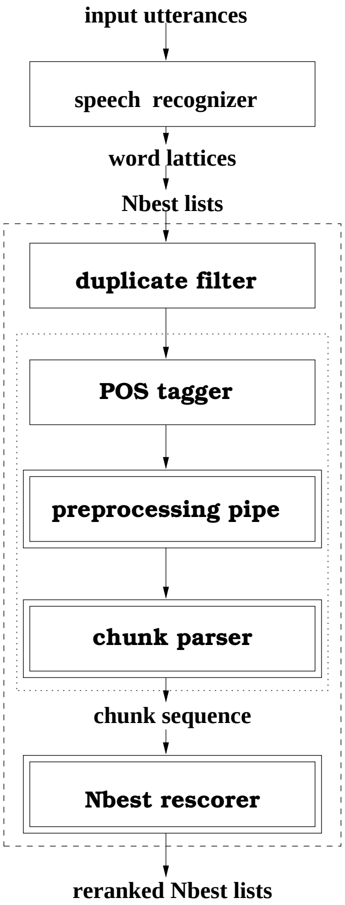

Figure 1: Global system architecture

4.2 Preprocessing Pipe

This preprocessing pipe consists of a number of filter components that serve the purpose of simplifying the input for subsequent components, without loss of essential information. Multiple word repetitions and non-content interjections or adverbs (e.g., "actually") are removed from the input, some short forms are expanded (e.g., "we'll" → "we will"), and frequent word sequences are combined into a single token (e.g., "a lot of" → "a_lot_of"). Longer turns are segmented into *short clauses*, which are defined as consisting of at least a subject and an inflected verbal form.

4.3 Chunk Parser

The chunk parser is a chart based context free parser, originally developed for the purpose of semantic frame parsing (Ward, 1991). For our purposes, we define the *chunks* to be the relevant concepts in the underlying grammar. We use 20 different chunks that consist of part of speech sequences (there are 40 different POS tags in the version of Brill's tagger that we are using). Since the grammar is non-recursive, no attachments of constituents are made, and, also due to its small size, parsing is extremely fast (more than 2000 tokens per second).[8] The parser takes the POS sequence from the tagged input, parses it in chunks, and finally, these POS-chunks are combined again with the words from the input stream.

4.4 Nbest Rescorer

The rescorer's task is to take an Nbest list generated from the speech recognizer and to label each element in this list (=hypothesis) with a *new score* which should correspond to the *true WER* of the respective hypothesis; these new scores are then used for the reranking of the Nbest list. Thus, in the optimal case, the hypothesis with lowest WER would move to the top of the reranked Nbest list.

The three main components of the rescorer are:

1. **Score Calculation:**
   There are three types of scores used:
   (a) *normalized score from the recognizer* (with respect to the acoustic and language models used internally): highest score = lowest rank number in the original Nbest list
   (b) *chunk coverage scores:* derived from the relative coverage of the chunk parser for each hypothesis: highest score = complete coverage, no skipped words in the hypothesis
   (c) *chunk language model score:* this is a standard n-gram score, derived from the sequence of *chunks* in each hypothesis (as opposed to the sequence of *words* in the recognizer): high score = high probability for the chunk sequence; the chunk language model was computed on the chunk parses of the LDC[9] SWITCHBOARD transcripts (about 3 million words total; we computed standard 3-gram and 5-gram backoff models).
2. **Reranking Neural Network:** We are using a standard three layer backpropagation neural network. The input units are the scores described here, the output unit should be a good predictor of the *true* WER of the hypothesis. For training of the neural net, the data was split randomly into a training and a test set.
3. **Cutoff Filter:** Initial experiments and data analysis showed clearly that in short utterances (less than 5–10 words) the *potential* reduction in WER is usually low: many of these utterances are (almost) correctly recognized in the

[8]DEC Alpha, 200MHz
[9]Linguistic Data Consortium

data set	Utts.	true WER	opt. WER
train	271	45.05	30.75
test	103	40.50	29.83
Total	374	43.51	30.41

Table 1: Characteristics of train and test sets (WER in %)

first place. For this reason, this filter prevents application of reranking to these short utterances.

5 Experiment: System Performance

5.1 Data

The data we used for system training, testing, and evaluation were drawn from the SWITCHBOARD and CALLHOME LVCSR[10] evaluation in spring 1996 (Finke and Zeppenfeld, 1996). In total, 374 utterances were used that were randomly split to form a training and test set. For these utterances, Nbest lists of length 300 were created from speech recognizer lattices.[11] The word error rates (WER) of these sets are given in Table 1. While the *true WER* corresponds to the WER of the first hypothesis (=top ranked), the *optimal WER* is computed under the assumption that an oracle would always pick the hypothesis with the lowest WER in every Nbest list. The difference between the average *true* WER and the *optimal* WER is 13.1%; this gives the maximum margin of improvement that reranking can possibly achieve on this data set. Another interesting figure is the *expected WER gain*, when a *random* process would rerank the Nbest lists and just pick any hypothesis to be the (new) top one. For the test set, this expected WER gain is -4.9% (i.e., the WER would *drop* by 4.9%).

5.2 Global System Speed

The system runtime, starting from the POS-tagger through all components up to the final evaluation of WER gain for the 103 utterances of the test set (ca. 8400 hypotheses, 145000 tokens) is less than 10 minutes on a DEC Alpha workstation (200 MHz, 192MB RAM), i.e., the throughput is more than 10 utterances per minute (or 840 hypotheses per minute).

5.3 Part Of Speech Tagger

We are using Brill's part of speech tagger as an important preprocessing component of our system (Brill, 1994). As our evaluations prove, the performance of this component is quite crucial to the whole system's performance, in particular to the segmentation module and to the POS based chunk parser.

Since the original tagger was trained on written corpora (Wall Street Journal, Brown corpus), we had to adapt it and retrain it on SWITCHBOARD data. The tagset was slightly modified and adapted, to accommodate phenomena of spoken language (e.g., hesitation words, fillers), and to facilitate the task of the segmentation module (e.g., by tagging clausal and non-clausal coordinators differently). After the adaptive training, the POS accuracy is 91.2% on general SWITCHBOARD[12] and 88.3% on a manually tagged subset of the training data we used for our experiments.[13]

Fortunately, some of these tagging errors are irrelevant with respect to the POS based chunk grammar: the tagger's performance with respect to this grammar is 92.8% on general SWITCHBOARD, and 90.6% for the manually tagged subset from our training set.

test set	words	miss.	wrong	sup.fl.	error
`20utts`	372	33	13	1	12.6%
`20utts-corr`	372	10	0	1	3.0%

Table 2: Performance of the chunk parser on different test sets

5.4 Chunk Parser

The evaluation of the chunk parser's accuracy was done on the following data sets: (i) 20 utterances (5 references and 15 speech recognizer hypotheses) (`20utts`); (ii) the same data, but with manual corrections of POS tags and short clause segment boundaries (`20utts-corr`).

For each word appearing in the chunk parser's output (including the skipped words[14]), it was determined, whether it belonged to the correct chunk, or whether it had to be classified into one of these three error categories:

- "missing": either not parsed or wrongfully incorporated in another chunk;
- "wrong": belongs to the wrong *type* of chunk;
- "superfluous": parsed as a chunk that should not be there (because it should be a part of another chunk)

[10] Large Vocabulary Continuous Speech Recognition

[11] Short utterances tend to have small lattices and therefore not *all* Nbest lists comprise the maximum of 300 hypotheses.

[12] The original LDC transcripts *not* used in our rescoring evaluations.

[13] These numbers are significantly lower than those achievable by taggers for written language; we conjecture that one reason for this lower performance is due to the more refined tagset we use which causes a higher amount of ambiguity for some frequent words.

[14] Skipped words are words that could not be parsed into any chunks.

data set	best performance	expected WER gain
eval21	+2.0	+0.5
test	**+0.3**	**-4.9**

Table 3: WER gain: best results in neural net experiments for two test sets (in absolute %)

The results of this evaluation are given in Table 2. We see that an optimally preprocessed input is indeed crucial for the accuracy of the parser: it increases from 87.4% to 97.0%.[15]

5.5 Nbest Rescorer

The task of the Nbest list rescorer is performed by a neural net, trained on chunk coverage, chunk language model, and speech recognizer scores, with the true WER as target value. We ran experiments to test various combinations of the following parameters: type of chunk language model (3-gram vs. 5-gram); chunk score parameters (e.g., penalty factors for skipped words, length normalization parameters); hypothesis length cutoffs (for the cutoff filter); number of hidden units; number of training epochs.

The net with the best performance on the test set has one hidden unit, and is trained for 10 epochs. A length cutoff of 8 words is used, i.e., only hypotheses whose average length was ≥ 8 are actually considered as reranking candidates. A 3-gram chunk language model proved to be slightly better than a 5-gram model.

Table 3 gives the results for the entire test set and a subset of 21 hypotheses (`eval21`) which had *at least* a potential gain of three word errors (when comparing the first ranked hypothesis with the hypothesis which has the fewest errors).[16]

We also calculated the cumulative average WER *before* and *after* reranking, over the size of the Nbest list for various hypotheses.[17] Figure 2 shows the plots of these two graphs for the example utterance in section 3 ("you weren't born just to soak up sun"). We see very clearly, that in this example not only has the new first hypothesis a significant WER gain compared to the old one, but that *in general* hypotheses with lower WER moved towards the top of the Nbest list.

[15] (Abney, 1996) reports a comparable per word accuracy of his CASS2 chunk parser (92.1%).

[16] While the latter set was obtained *post hoc* (using the known WER), it is conceivable to approximate this biased selection, when fairly reliable confidence annotations from the speech recognizer are available (Chase, 1997).

[17] Average of the WER from hypotheses 1 to k in the Nbest list.

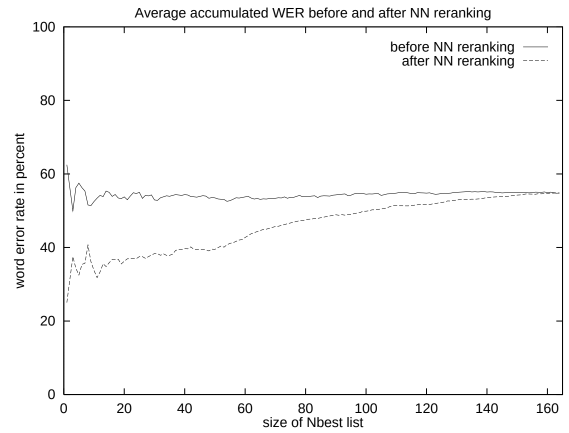

Figure 2: Cumulative average WER before and after reranking for an example utterance

rank/nr.	hypothesis
1/1	you weren't born justice so cups on
2/3	you weren't born just to sew cups on
3/189	you weren't born justice vocal song
4/190	**you weren't born just to soak up sun**
5/214	you weren't foreign just to sew cups on
6/269	you weren't born justice so courts on
7/273	you weren't born just to sew carp song
8/296	you weren't boring just to soak up son

Table 4: Recognizer hypotheses from an example utterance (hypothesis nr. 190 exactly corresponds to the reference)

A more detailed account of 8 hypotheses from the same example utterance is given in tables 4 (which lists the recognizer hypotheses) and 5 (where various scores, WER, and the ranks before and after the reranking procedure are provided). It can be seen that while the new first best hypothesis is *not* the one with the lowest WER, it *does* have a lower WER than the originally first ranked hypothesis (25.0% vs. 62.5%).

6 Discussion

Using the neural net with the characteristics described in the previous section, we were able to get a *positive* effect in WER reduction on a non-biased test set. While this effect is quite small, one has to keep in mind that the (constituent-like) chunk representations were the *only* source of information for our reranking system, in addition to the internal scores of the speech recognizer. It can be expected that including more sources of knowledge, like the plausibility of correct verb-argument structures (the correct match of subcategorization frames), and the likelihood of selectional restrictions between the verbal heads and their head noun arguments would further improve these results.

Hypo-Rank New/Old	True WER in %	Chunk-Cov. Score	Skipped Words	Chunk-LM Score	Norm.SR Score
1/8	**25.0**	0.875	0	0.984	0.93
2/7	37.5	0.625	0	0.865	0.94
3/4	0.0	0.75	0	0.954	0.97
4/3	62.5	0.5	0	0.618	0.98
5/6	62.5	0.625	0.125	0.715	0.95
6/5	50.0	0.75	0.125	1.056	0.96
7/1	**62.5**	0.625	0.125	0.715	1.0
8/2	37.5	0.625	0.125	1.032	0.99

Table 5: Scores, WER, and ranks before and after reranking of 8 hypotheses from an example utterance

The second observation we make when looking at the markedly positive results of the `eval21` set concerns the potential benefit of selecting good *candidates* for reranking in the first place.

7 Comparison: Human Study

One of our motivations for using syntactic representations for the task of Nbest list reranking was the intuition that frequently, by just *reading* through the list of hypotheses, one can eliminate highly implausible candidates or favor more plausible ones.

To put this intuition to test, we conducted a small experiment where human subjects were asked to look at pairs of speech recognizer hypotheses drawn from the Nbest lists and to decide which of these they considered to be "more well-formed". Well-formedness was judged in terms of (i) structure (syntax) and (ii) meaning (semantics). 128 hypothesis pairs were extracted from the training set (the top ranked hypothesis and the hypothesis with lowest WER), and presented in random order to the subjects.

4 subjects participated in the study and table 6 gives the results of its evaluation: WER gain is measured the same way as in our system evaluation — here, it corresponds to the average reduction in WER, when the well-formedness judgements of the human subjects were to be used to rerank the respective hypothesis-pairs.

While the maximum WER gain for these 128 hypothesis-pairs is 15.2%, the expected WER gain (i.e., the WER gain of a random process) is 7.6%.

Whereas the difference between both methods to a random choice is highly significant (syntax: $\alpha = 0.01, t = 9.036, df = 3$; semantics: $\alpha = 0.01, t = 11.753, df = 3$)[18] , the difference between these two methods is *not* ($\alpha = 0.05, t = -1.273, df = 6$)[19]. The latter is most likely due to the fact that there were only few hypotheses that were judged *differently* in terms of syntactic or semantic well-formedness by one subject: on average, only 6% of the hypothesis-pairs received a different judgement by one subject.

Subject	Syntax pref.	Sem. pref.
A	10.0	10.3
B	10.0	10.2
C	9.1	9.7
D	10.2	10.8
Total Avg.	9.8	10.2

Table 6: Human Performance (WER gain in %)

8 Future Work

From our results and experiments, we conclude that there are several directions of future work which are promising to pursue:

- *improvement of the POS tagger:* Since the performance of this component was shown to be of essential importance for later stages of the system, we expect to see benefits from putting efforts into further training.
- *alternative language models:* An idea for improvement here is to integrate skipped words into the LM (similar to the modeling of noise in speech). In this way we get rid of the skipping penalties we were using so far and which blurred the statistical nature of the model.
- *identifying good reranking candidates:* So far, the only and exclusive heuristics we are using for determining when to rerank and when not to, is to use the length-cutoff filter to exclude short utterances from being considered in the final reranking procedure. (Chase, 1997) showed that there are a number of potentially useful "features" from various sources within the recognizer which can predict, at least to a certain extent, the "confidence" that the recognizer has about a particular hypothesis. Hypotheses

[18] These results were obtained using the one-sided t-test.
[19] Two-sided t-test.

which have a higher WER on average also exhibit a higher word gain potential, and therefore these predictions appear to be promising indeed.

- *adding argument structure representations:* The chunk representation in our system only gives an idea about which constituents there are in a clause and what their ordering is. A richer model has to include also the *dependencies* between these chunks. Exploiting statistics about subcategorization frames of verbs and selectional restrictions would be a way to enhance the available representations.

9 Summary

In this paper we have shown that it is feasible to produce chunk based representations for spontaneous speech in unrestricted domains with a high level of accuracy.

The chunk representations are used to generate scores for an Nbest list reranking component.

The results are promising, in that the best performance on a randomly selected test set is an absolute decrease in word error rate of 0.3 percent, measured on the new first hypotheses in the reranked Nbest lists.

10 Acknowledgements

The authors are grateful for valuable discussions and suggestions from many people in the Interactive Systems Laboratories, CMU, in particular to Alon Lavie, Klaus Ries, Marsal Gavaldà, Torsten Zeppenfeld, and Michael Finke. Also, we wish to thank Marsal Gavaldà, Maria Lapata, Alon Lavie, and the three anonymous reviewers for their comments on earlier drafts of this paper.

More details about the work reported here can be found in the first author's master's thesis (Zechner, 1997).

This work was funded in part by grants of the Austrian Ministry for Science and Research (BMWF), the Verbmobil project of the Federal Republic of Germany, ATR – Interpreting Telecommunications Research Laboratories of Japan, and the US Department of Defense.

References

Steven Abney. 1996. Partial parsing via finite-state cascades. In *Workshop on Robust Parsing, 8th European Summer School in Logic, Language and Information, Prague, Czech Republic*, pages 8–15.

Eric Brill. 1994. Some advances in transformation-based part of speech tagging. In *Proceeedings of AAAI-94*.

Lin Chase. 1997. *Error-responsive feedback mechanisms for speech recognizers.* Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Michael Finke, Jürgen Fritsch, Petra Geutner, Klaus Ries and Torsten Zeppenfeld. 1997. The JanusRTk SWITCHBOARD/CALLHOME 1997 Evaluation System. In *Proceedings of LVCSR Hub5-e Workshop, May 13-15, Baltimore, Maryland.*

Michael Finke and Torsten Zeppenfeld. 1996. LVCSR SWITCHBOARD April 1996 Evaluation Report. In *Proceedings of the LVCSR Hub 5 Workshop, April 29 - May 1, 1996 Maritime Institute of Technology, Linthicum Heights, Maryland.*

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: telephone speech corpus for research and development. In *Proceedings of the ICASSP-92*, volume 1, pages 517–520.

Alon Lavie. 1996. *GLR*: A Robust Grammar-Focused Parser for Spontaneously Spoken Language.* Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.

Marc Light. 1996. CHUMP: Partial parsing and underspecified representations. In *Proceedings of the 12th European Conference on Artificial Intelligence (ECAI-96), Budapest, Hungary.*

Alex Waibel, Michael Finke, Donna Gates, Marsal Gavaldà, Thomas Kemp, Alon Lavie, Lori Levin, Martin Maier, Laura Mayfield, Arthur McNair, Ivica Rogina, Kaori Shima, Tilo Sloboda, Monika Woszczyna, Torsten Zeppenfeld, and Puming Zhan. 1996. JANUS-II - advances in speech recognition. In *Proceedings of the ICASSP-96.*

Wayne Ward. 1991. Understanding spontaneous speech: The PHOENIX system. In *Proceedings of ICASSP-91*, pages 365–367.

Klaus Zechner. 1997. Building chunk level representations for spontaneous speech in unrestricted domains: The CHUNKY system and its application to reranking Nbest lists of a speech recognizer. M.S. Project Report, CMU, Department of Philosophy. Available from `http://www.contrib.andrew.cmu.edu/~zechner/publications.html`