Gefördert vom

bmb+f

Bundesministerium für
Bildung, Wissenschaft,
Forschung und Technologie

Verbmobil

Verbundvorhaben

Towards Robust Understanding of Speech by Combination of Partial Analyses

Karsten L. Worm, C. J. Rupp

Universität des Saarlandes

Vm

Report 222
Mai 1998

Mai 1998

Karsten L. Worm, C. J. Rupp

Computerlinguistik
Universität des Saarlandes
D-66041 Saarbrücken

Tel.: (0681) 302 - 4503
Fax: (0681) 302 - 4351
e-mail: {worm,cj}@coli.uni-sb.de

Gehört zum Antragsabschnitt: 2.5.3 Robuste semantische Verfahren

Die vorliegende Arbeit wurde im Rahmen des Verbundvorhabens Verbmobil vom Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) unter dem Förderkennzeichen 01 IV 701 R 4 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

ISSN 0949-6084

Abstract

We discuss the issue of how a speech understanding system can be made robust against spontaneous speech phenomena and speech recognition errors. We argue that treating these problems before or during parsing raises certain problems which can be circumvented by an approach handling them after parsing, and we describe how this approach is currently being investigated in the Verbmobil speech translation project.

This paper will appear in the proceedings of ECAI 1998 (Worm and Rupp 1998).

Contents

1 INTRODUCTION

Processing spontaneous speech by machine is difficult as the input to the natural language processing component is perturbed by two influences. First, speakers make mistakes, correct themselves during speaking, produce false starts and use ungrammatical constructions. Second, the acoustic signal produced by a human speaker is mapped by a speech recogniser onto a written form; this mapping is rarely completely correct. This introduces two levels of uncertainty into the processing of speech, which make the task of linguistically analysing a spoken utterance in a speech processing system hard.

In this paper, we present an architecture for a speech processing system which is intended to be a step forward in coping with these problems. While in many cases the input will be too deformed to achieve a complete and spanning analysis, it is usually possible to arrive at a set of partial analyses which together cover the whole input. The main idea of the approach presented here is to take these partial analyses if nothing better can be found and combine them using heuristic rules and linguistic knowledge.

2 THE VERBMOBIL PROJECT

Verbmobil (Wahlster 1997) is a large scale research project in the area of spoken language translation. Its goal is to develop a system that translates negotiation dialogues between speakers of German, English and Japanese in face-to-face or video conferencing situations. The integrated system developed during the first project phase (1993-96), the *Research Prototype*, was successfully demonstrated in autumn 1996 (Bub, Wahlster, and Waibel 1997). The final *Verbmobil Prototype* is due in 2000.

2.1 The system architecture

Verbmobil employs different approaches to machine translation. A semantic transfer approach (Dorna and Emele 1996) based on a deep linguistic analysis of the input utterance competes with statistical, example-based and dialogue act-based translation approaches.

The spoken input is mapped onto a *word hypothesis graph* (WHG) by a speech recognizer (see Section 3.2). A prosody component divides the input into segments and annotates the WHGs with prosodic features. Within the semantic transfer line of processing, three different parsers (an HPSG-based chart parser, a chunk parser using cascaded finite state automata, and a statistical parser) attempt to analyse the WHG syntactically and semantically. All three deliver their analyses in the VIT format (see Section 2.2). The parsers are coordinated by an *integrated processing* component which chooses paths through the WHG to be analysed in parallel by the parsers until an analysis spanning the whole input is found or the system runs out of time.

Since in many cases no complete analysis spanning the whole input can be found, the parsers produce partial analyses along the way and send them to the *robust semantic processing* component, which stores and combines them to yield larger analyses. We describe this component in Section 5. The relevant part of the system's architecture is shown in Figure 1.

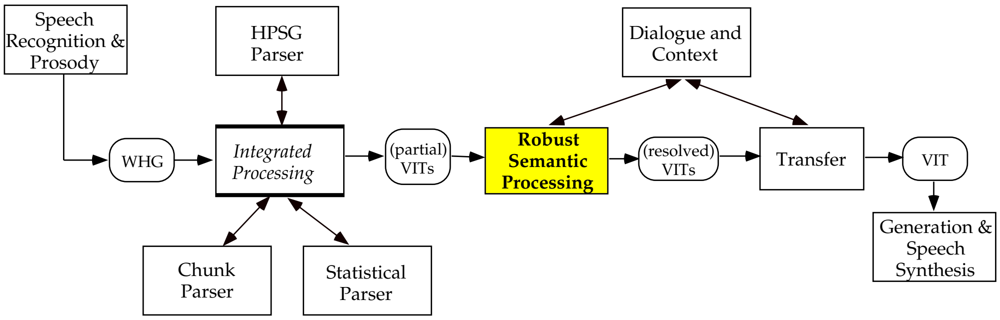

Figure 1: Part of the system architecture.

2.2 The VIT format

The VIT (*Verbmobil Interface Term*) was designed as a common output format for two alternative and independently developed syntactic-semantic analysis components of the first project phase (Bos et al. 1998). Their internal semantic formalisms differed, but both were to be attached to a single transfer module. The need for a common output format is still present, since there are three alternative syntactic-semantic parsing modules in the new Verbmobil system, all of which produce output for just one transfer module.

The VIT can be viewed as a theory-independent representation for underspecified semantic representations (Bos et al. 1996). It specifies a set of discourse representation structures, DRSs, (Kamp and Reyle 1993). If an utterance is structurally ambiguous, it will be represented by one VIT, which specifies the set of DRSs corresponding to the different readings of the utterance.

Formally, a VIT is a nine-place PROLOG term. There are slots for an identifier for the input segment to which the VIT corresponds, a list of the core semantic predicates, a list of scopal constraints, syntactic, prosodic and pragmatic information as well as tense and aspect and sortal information. For an example of a VIT, see (Bos et al. 1998).

3 PROBLEMS IN SPEECH PROCESSING

3.1 Spontaneous speech phenomena

Spontaneously uttered speech deviates considerably from written language (Levelt 1983; Block and Schachtl 1995). Speakers make mistakes, correct themselves, start a sentence and do not finish it, and use ungramatical constructions. The following examples are taken from the Verbmobil corpus of appointment scheduling dialogues.

(1) Das paßt mir auch. Ach nee, das paßt mir schon wieder nicht. (*That suits me, too. Oh, sorry, that does not suit me either.*)

(2) Sagen wir am ersten Juni, äh, am ersten Juli. (*Let's say on the first of June, uh, on the first of July.*)

(3) Das ist . . . freut mich. (*That is . . . makes me happy.*)

In the first example, the speaker utters a sentence, followed by interjections marking a self correction, and then the opposite of what was expressed in the first sentence. In the second example, the speaker notices that the date uttered was incorrect and corrects it by appending the intended date. In the third example, the sentence is not finished in the way originally intended; the speaker utters a different verb after the copula *ist* and finishes the sentence, still referring to the originally uttered subject.

3.2 Speech recognition errors

Although there has been rapid progress in the development of speech recognizers during recent years, the mapping they perform from an acoustic signal to a written representation of what has been said is not always correct — words actually uttered do not appear in the transcription, are mistaken for different words, or words not uttered are hallucinated. The sum of the numbers of these errors divided by the number of words actually uttered is the *word error rate*, a measure for the performance of the speech recognizer.

Since the strings of words a speech recognizer considers with the highest probability to be the one uttered often is not the one actually spoken, the recognizer's output usually consists of an ordered list of the n-best hypotheses. To avoid redundancy, they are often represented as a weighted graph, a *word hypothesis graph* (WHG, cf. (Oerder and Ney 1993)). The word error rate of a WHG is defined as the minimum of the word error rates of all paths through the graph. The current Verbmobil speech recognizers achieve a word error rate of about 13% (Bub, Wahlster, and Waibel 1997). Though this is quite a good result, it means that only for about 30% of the utterances in the corpus, the corresponding WHG contains the sequence of words actually uttered.

Figure 2 gives an example of a WHG. The edges correspond to the individual word hypotheses and are annotated with scores[1]. The speaker uttered *Mir wäre es am liebsten in den nächsten zwei Wochen* ('I would prefer it during the next two weeks'), but no path from the start node to the end node contains exactly that sequence of words. The one coming closest contains *Mir wäre es am liebsten den nächsten zwei Wochen*, leaving out the preposition *in*. This path and all the others are ungrammatical.

Because of such recognition errors, the linguistic analysis component cannot expect to find the correct string of words in each WHG, nor even to find a different string which is analysable. In the evaluation of the Verbmobil-1 system, the linguistic analysis managed to deal with slightly more than half of the input WHGs (Bub, Wahlster, and Waibel 1997).

4 APPROACHES TO ROBUST SPEECH PROCESSING

What are the options for making a speech understanding system more robust to spoken language phenomena and recognizer errors, given that one wants to stick to the idea of a deep linguistic analysis of the input? Basically, there are three different places in the processing where one could apply new mechanisms: *before*, *during*, or *after* parsing. Each alternative has its advantages and drawbacks.

[1]The scores are negative logarithmic probabilities, i.e. the lower the score, the higher the probability.

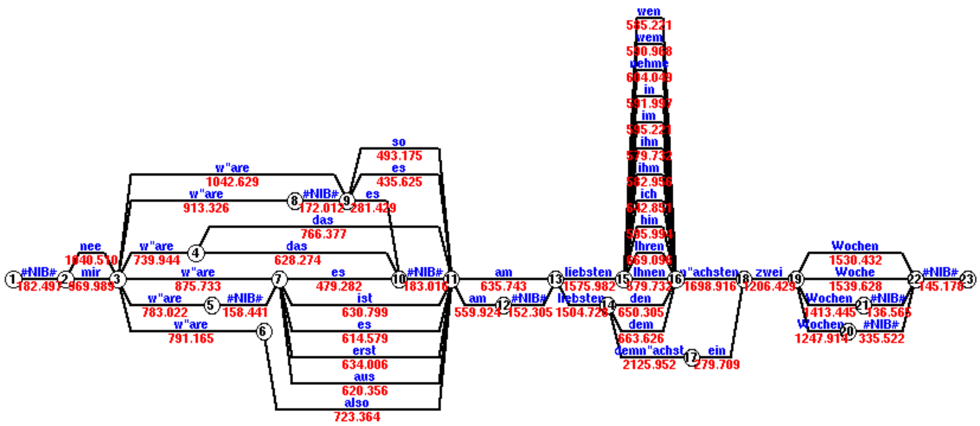

Figure 2: A word hypothesis graph for the utterance *Mir wäre es am liebsten in den nächsten zwei Wochen*.

4.1 Before parsing

Several approaches in the literature explore the idea of handling spontaneous speech phenomena before parsing. E. g, they attempt to recognize self correction on the basis of pattern matching techniques (Bear, Dowding, and Shriberg 1992; Heeman and Allen 1994), sometimes including prosodic features (Nakatani and Hirschberg 1993). However, so far these techniques have only been applied to human-made transliterations of speech. I. e, the additional level of uncertainty and complexity introduced by looking at speech recognizer output in the form of WHGs has not been considered.

WHGs contain a very large number of possible paths which have to be considered when applying such repair operations, which increases the complexity of the task considerably. The complexity of the parser's task is also greater, since these techniques should be applied conservatively, only adding new 'shortcut' paths to the WHGs in case the presence of a spontaneous speech phenomenon has been hypothesized, in order to avoid pruning away a correct path. Adding edges to the WHG is critical especially in a system with real time requirements like Verbmobil.

Nonetheless, the applicability of this kind of techniques to WHGs is currently being explored by our project partners at the Univ. of Erlangen-Nürnberg in the Verbmobil project. First results, however, suggest that they will have to be restricted to handling rather short self corrections because of the computational complexity. How a division of labour between this approach and the one suggested in this paper can be achieved, is a very interesting question for our research.

4.2 During parsing

Treating spontaneous speech phenomena and recognizer errors during parsing (e. g. (Hindle 1983; Langer 1990; Hipp 1993)) entails making the analysis procedure more liberal — the grammar or the parser has to be changed so that strings which are not grammatical still get some sort of analysis, on the assumption that one of the problems mentioned is present and will somehow be taken care of.

While this may be a reasonable alternative for processing written language, it becomes infeasible for speech processing on the basis of WHGs. In this context, the parser has two intertwined tasks:

1. find an analysable path through the WHG, and
2. provide a syntactic-semantic analysis for it.

Here one has a trade off — if the analysis procedure is too strict and only accepts well-formed paths, there will be no result if no well-formed path is present in the WHG. If it is too liberal, it may accidentally accept some ill-formed path before having seen a well-formed one with a lower score.

4.3 After parsing

Handling the phenomena mentioned after parsing has certain advantages. First, the grammar employed for parsing can be restrictive, thus limiting the search space of the parser and avoiding analysing an ill-formed WHG path before a well-formed, lower-ranked one has been considered. This means that such an approach is more efficient (Rosé 1997), since the flexibility required for robust behaviour is only applied when needed.

Second, the information accumulated about analysable parts of the input is available for further processing. The fragmentary analyses built up from the input often provide enough valuable information about the utterance that a better analysis spanning large parts of the input can be constructed from them.

Third, in a system like Verbmobil with several parsers delivering output in a common format, it becomes possible to combine partial results from different sources. This would not be possible using the internal data structures of the parsers, since they are completely incompatible.

A model for fallback processing of partial analyses was proposed by Rosé (Rosé 1997). However, in contrast to the approach suggested here, she employs interlingual representations of utterance meanings as the units which are combined, which makes all linguistic constraints on the combinatorics of fragments unavailable.

5 ROBUST SEMANTIC PROCESSING

The robust semantic processing component we are developing in the Verbmobil project has three main tasks: *Store* partial analyses from the parsers, *combine* them on the basis of heuristic rules, and *choose* a result.

5.1 The chart

The first task of the robust semantic processing is to manage a possibly large number of partial analyses, each spanning a certain sub-interval of the input utterance.

The basic mode of processing, store competing analyses and combine them to larger analyses, while avoiding unnecessary redundancy, resembles that of a chart parser. Indeed we use a chart-like data structure to store the competing partial analyses delivered by the parsers and new analyses obtained by combining analyses already there. All the advantages of the chart in chart parsing are preserved: The chart allows storing competing hypotheses, even from different sources, without redundancy.

Since the input to the parsers consists of WHGs rather than strings, the analyses entered cannot refer to the string positions they span. Rather they have to refer to a time interval. This means also that the chart cannot be indexed by string positions, but is indexed by the time frames the speech recognizer uses. This makes necessary slight modifications to the chart handling algorithms. Figure 3 shows an example chart.[2]

Every time a new VIT is delivered by a parser, a corresponding passive edge is constructed, the active edges induced by the rules whose conditions are met by the VIT are computed, as are new active or passive edges resulting from the VIT satisfying an active edge's requirements.

All new edges are stored in an agenda before being entered into the chart. Agenda items are processed one by one; in between, the system checks whether new VITs from a parser have arrived and deals with them first. Since a parser may at first deliver analyses for small substrings and only later for larger parts of a path, agenda items might be already outdated when they are about to be entered.

We capture this by classifying the edges as *in* or *out*.[3] By default, an edge is *in*. It becomes *out* when an edge from the same source (the same parser or the robust processing component) is entered into the chart which covers a larger substring of the same WHG path. Edges that are *out* are not considered further when applying rules or selecting a result.

Thus, the processing is incremental: Small analyses are already dealt with while the parsers are still aiming at bigger ones. If bigger and better analyses arrive, they are prefered. The algorithm also has the anytime property (Dean and Boddy 1988), since as soon as the first parser results have been entered, a result can be retrieved. The more processing time is available, the more parser results have been processed and the more rules have been applied, thus the better the result will be.

Once no further parser results are to be expected (since there are no further paths to consider, or the time bound for parsing is reached), gaps in the chart are bridged. We do this by extending active edges over passive edges not contributing to any other edge.

Processing an incoming VIT requires 30-50 ms in our SICStus PROLOG implementation on a SUN UltraSparc2. This includes entering the corresponding edge into the chart, the application of rules and creation of active and passive edges induced by the firing of rules.

5.2 The rules

To describe the conditions under which two or more partial analyses should be combined, an analysis should be left out or modified, we use a set of heuristic rules. Each

[2]The edges are numbered according to the order in which they are entered into the chart. Each edge is annotated with its number, the substring(s) of the input it is built from, and a list of the identifiers of the edges which are its immediate constituents.

[3]In Figure 3, the edges that are *in* are drawn in the upper half, the ones that are *out* in the lower half of the diagram.

rule specifies the conditions under which it should be applied, the operations to be performed, and what the result of the rule application is. Rules have the following format (the notation is in PROLOG style):

```
[Cond1,...,CondN] ---> [Op1,...,OpN] & Result.
```

The left hand side consists of a list of conditions on partial analyses, `Cond1` being a condition (or a list of conditions) on the first partial analysis (VIT), etc., where the order of conditions parallels the expected temporal order of the analyses. When these conditions are met, the rule fires and the operations `Op1`, etc. are performed on the input VITs. One VIT, `Result`, is designated as the result of the rule. As a result of applying the rules, it is entered into the chart, spanning the union of the spans of the analyses on the left hand side. Examples of rules are given below.

Generally, all applicable rules will be applied. To aid scoring of hypotheses, each operation performed returns a confidence value. If e. g. a verbal functor is combined with a possible argument, the confidence value will depend on whether the sortal restrictions on the argument role in question are met by the argument, and whether syntactic constraints such as case requirements or agreement are violated.

We split the rule set into four subsets: Three language-specific ones (for German, English, and Japanese) and a general one. The general rule set currently contains 11 rules, e. g. basic rules for applying a verbal functor to a possible argument, or for applying a modifier to a proposition. The language-specific rules deal with the different structures spontaneous speech phenomena exhibit in the three languages. These rules are still under development.

5.3 Choosing a result

The last step in processing is to choose the 'best' sequence of analyses from the chart which covers the whole input and deliver it to the transfer module. In the ideal case, there will be an analysis spanning the whole input.

We assign each edge k a score $\sigma(k)$ as defined in (4).[4]

(4) $$\sigma(k) \stackrel{\text{def}}{=} (|k|^2 - C(k)) \cdot VC(k) \cdot C(k)$$

The score for a sequence $k_1, \ldots, k_n$ of edges is defined in (5).

(5) $$\sigma(k_1, \ldots, k_n) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \sigma(k_i)$$

The scoring function σ satisfies the following important property. If we have an edge k_1 delivered by a parser covering an interval completely, two edges k_2 and k_3 from a parser which together cover the interval completely, and an edge k_4 produced by the robust semantic processing out of k_2 and k_3, we have:

(6) $$\sigma(k_1) > \sigma(k_4) > \sigma(k_2, k_3)$$

[4] $|k|$ is the length of the corresponding time interval, $C(k)$ the number of component edges k has been built from, $VC(k)$ is the sum of the lengths of the words which are covered by the VIT of k ($0 < VC(k) \leq 1$; $VC(k) < 1$ iff there is an unalysed gap), and $C(k)$ is a confidence value returned by the operations that resulted in k ($0 < C(k) \leq 1$).

Thus, σ favors larger edges, edges delivered by a parser, edges with less gaps, edges with less constituents, edges being the result of operations with a high confidence. The sequence of edges with the highest score is delivered.

5.4 An Example

Consider the analysis of the WHG from figure 2. Since function words like prepositions are usually short, speech recognizers often have trouble recognizing them. In this example, the speaker uttered *Mir wäre es am liebsten in den nächsten zwei Wochen* ('I would prefer it during the next two weeks'), while in all paths through the WHG the preposition *in* was missing. The parsers delivered analyses for the segments *Mir wäre es am liebsten* and *den nächsten zwei Wochen*.

These fragments are handled by two rules. The first turns a temporal NP like the second fragment into a temporal modifier, expressing that something is standing in an underspecified temporal relation to the temporal entity the NP denotes:

```
[temporal_np(V1)] --->
          [typeraise_to_mod(V1,V2)] & V2.
```

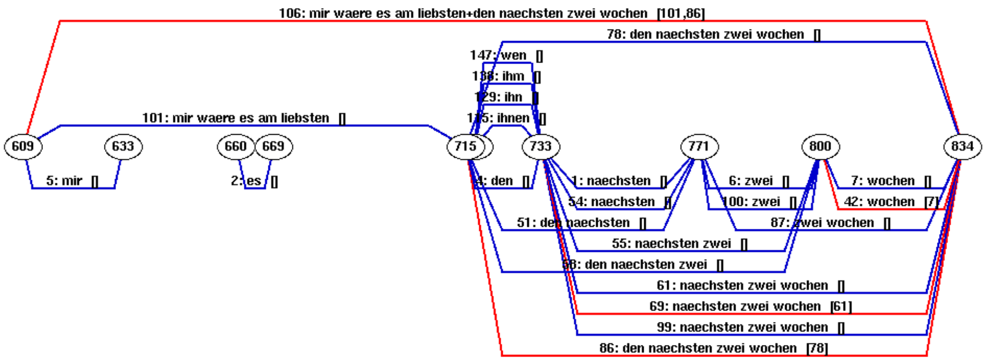

Figure 3: The chart for *Mir wäre es am liebsten in den nächsten zwei Wochen*.

Then a very general rule can apply that modifier to the proposition expressed by the first fragment:

```
[type(V1,prop),type(V2,mod)] --->
           [apply(V2,V1,V3)] & V3.
```

This sequence of rule applications produces an analysis spanning the entire utterance and making use of all accessible information. The chart corresponding to this example is shown in Figure 3.

6 CONCLUSION AND OUTLOOK

We have discussed two important problems for building a speech understanding system: *spontaneous speech phenomena* and *speech recognition errors*, and argued that the treatment of these before or during parsing multiplies the uncertainty and complexity inherent in the task. As a possible solution, we described a model which aims at treating these problems after parsing by combining partial analyses if no spanning analysis can be found. The strategy is currently being integrated in the 1998 version of the Verbmobil Prototype.

ACKNOWLEDGEMENTS

The authors wish to thank their colleagues Johan Bos, Aljoscha Burchardt, Björn Gambäck, Walter Kasper, Bernd Kiefer, Uli Krieger, Manfred Pinkal, Tobias Ruland, Jörg Spilker, Hans Weber for their collaboration, and Markus Egg and the two anonymous reviewers for their comments which helped improve this paper.

This research was supported by the German Federal Ministry for Education, Science, Research and Technology under grant no. 01 IV 701 R4.

References

Bear, J., J. Dowding, and E. Shriberg (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proc. of the 30 th ACL*, Newark, DE, pp. 56–63.

Block, H.-U. and S. Schachtl (1995). What a grammar of spoken dialogues has to deal with. In G. Heyer and H. Haugeneder (Eds.), *Language Engineering*, Advanced Studies in Computer Science, pp. 101–126. Braunschweig: Vieweg.

Bos, J., B. Buschbeck-Wolf, M. Dorna, and C. J. Rupp (1998). Managing information at linguistic interfaces. In *Proc. of the 17th COLING/36th ACL*, Montréal, Canada.

Bos, J., B. Gambäck, C. Lieske, Y. Mori, M. Pinkal, and K. Worm (1996). Compositional semantics in Verbmobil. In *Proc. of the 16th COLING*, Copenhagen, Denmark, pp. 131–136.

Bub, T., W. Wahlster, and A. Waibel (1997). Verbmobil: The combination of deep and shallow processing for spontaneous speech translation. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, München, Germany, pp. 71–74. IEEE Signal Processing Society.

Dean, T. and M. Boddy (1988). An analysis of time-dependent planning. In *Proc. of AAAI-88*, St. Paul, MN, pp. 49–54.

Dorna, M. and M. C. Emele (1996). Semantic-based transfer. In *Proc. of the 16th COLING*, Copenhagen, Denmark, pp. 316–321.

Heeman, P. and J. Allen (1994). Detecting and correcting speech repairs. In *Proc. of the 32 th ACL*, Las Cruces, NM, pp. 295–302.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proc. of the 21 th ACL*, Cambridge, MA, pp. 123–128.

Hipp, D. R. (1993). *Design and Development of Spoken Natural-Language Dialog Parsing Systems*. Ph. D. thesis, Department of Computer Science, Duke University, Durham, NC.

Kamp, H. and U. Reyle (1993). *From Discourse to Logic*. Dordrecht: Kluwer.

Langer, H. (1990). Syntactic normalization of spontaneous speech. In *Proc. of the 13 th COLING*, Helsinki, Finnland, pp. 180–183.

Levelt, W. J. M. (1983, July). Monitoring and self-repair in speech. *Cognition 14*(1), 41–104.

Nakatani, C. and J. Hirschberg (1993). A speech-first model for repair detection and correction. In *Proc. of the 31 th ACL*, Columbus, OH, pp. 46–53.

Oerder, M. and H. Ney (1993). Word graphs: An efficient interface between continuous-speech recognition and language understanding. In *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Minneapolis, MN. IEEE Signal Processing Society.

Rosé, C. P. (1997). *Robust Interactive Dialogue Interpretation*. Ph. D. thesis, Carnegie Mellon University, Pittsburgh, PA. Language Technologies Institute.

Wahlster, W. (1997, June). Verbmobil: Erkennung, Analyse, Transfer, Generierung und Synthese von Spontansprache. Verbmobil-Report 198, DFKI GmbH, Saarbrücken. Available from http://www.dfki.de/verbmobil/.

Worm, K. L. and C. J. Rupp (1998). Towards robust understanding of speech by combination of partial analyses. In *Proc. of the 13 th ECAI*, Brighton, UK, pp. 190–194.