A Study of Human Dialogue Strategies in the Presence of Speech Recognition Errors

Teresa Zollo
Department of Computer Science
University of Rochester
Rochester, New York 14627-0226
zollo@cs.rochester.edu

Abstract

Continuous speech recognition technology has recently matured to the point where it has become feasible to develop spoken dialogue interfaces to computer systems that help people perform tasks in simple domains. Word recognition accuracy for spontaneous dialogue, however, is still a long way from perfect, and any system that uses spoken natural language input has to have some mechanism for dealing with speech recognition errors. In fact, since we cannot expect speech recognition technology ever to surpass human performance, which is itself imperfect, the handling of misrecognition will continue to be an issue that needs to be addressed by spoken dialogue systems.

Many techniques have been incorporated into various components of spoken dialogue systems to try to provide robustness in the presence of speech recognition errors. This paper describes a study that explores how humans go about signaling communication problems caused by speech recognition errors. The study is the first step in designing a conversational agent that has human-like repair initiation and feedback behavior, and can perhaps provide robustness by resolving communication problems within the course of the dialogue itself.

Introduction

Dialogue is, by definition, a collaborative endeavor. A dialogue can not be said to have occurred unless the participants collaborate at least to the point where one participant addresses a message to the other, and the other receives, decodes, and acknowledges the message. Analyses of actual dialogues show that the collaboration involved in maintaining a cooperative dialogue is actually much more involved than this basic behavior.

The general goal of dialogue is to affect the beliefs and actions of the dialogue participants. Cooperative dialogue participants engage in a highly coordinated process that systematically builds up a set of beliefs held mutually by the participants. Dialogue involves taking turns to introduce propositions, provide indications of whether the propositions are understood and agreed to, monitor the conversation for potential misunderstandings, and repair misunderstandings when they occur.

As a dialogue progresses, participants regularly provide feedback to indicate whether they have heard and understood what the other participant has said, a process referred to as *grounding* (Clark & Schaefer 1987; Clark & Brennan 1991). Speakers often make an extra effort to correct themselves if they detect a potential source of misunderstanding in their original utterance (Schegloff, Jefferson, & Sacks 1977), and will correct their conversational partner if the feedback provided by the partner indicates a misunderstanding (Schegloff 1992). A psychological model presented by Clark and Wilkes-Gibbs (1986) predicts that a cooperative dialogue participant will provide more substantial evidence of understanding when the content of an utterance is perceived to be especially important, and when the communication channel between the participants is perceived to be less reliable. Brennan and Hulteen (1995) emphasize the importance of providing both positive and negative evidence of understanding in order to efficiently identify and repair misunderstandings as they occur.

There are several levels at which dialogue participants can claim to understand their interlocutor's utterance:

1. They know that their dialogue partner has said *something* (i.e., they saw the lips moving, but did not hear anything).

2. They know that they are being addressed by their dialogue partner.

3. They heard the words that were said by their dialogue partner (at least well enough for their current purposes).

4. They were able to form a coherent semantic interpretation of their dialogue partner's utterance (in-

cluding the references made, intended sense of the words used, etc.).

5. They understand the illocutionary force of the initiator's utterance.

6. They understand how the propositional content of the utterance relates to the dialogue.

Spoken dialogue systems

Automatic speech recognition (ASR) technology has recently matured to the point that it is feasible to use ASR systems as platforms on which to build limited-domain spoken dialogue systems. The speech recognition component of a spoken dialogue system is responsible for identifying the sequence of words uttered by a speaker.

Word recognition rates on the order of 95% have been reported for speech recognition systems having vocabularies of more than a thousand words. However, when state of-the-art recognizers are applied to the task of recognizing continuous, unconstrained, spontaneous dialogue, word recognition accuracy is closer to 75%. As tasks performed by spoken language systems become more complex, users' speech is likely to become more variable, reducing speech recognition performance further.

A critical challenge to developers of spoken dialogue systems is how to provide robustness in the presence of speech recognition errors. Some techniques that have been applied to this problem are using statistically-based post-processing for speech recognizers (Ringger & Allen 1996) and robust parsing techniques that are capable of identifying meaningful fragments of utterances (Allen *et al.* 1996; Sadek *et al.* 1996).

Current spoken dialogue systems inform the user when they cannot understand an utterance. However, in most cases the user is left to wonder if the problem was due to poor speech recognition, in which case they might be able to correct the problem by repeating the same utterance (although utterances are even *less* likely to be understood the second time (Oviatt *et al.* 1996)), or if the system could not understand the utterance for some other reason (i.e., some unknown word was used or there was some ambiguity that couldn't be resolved), in which case the user needs to rephrase the utterance.

The strategies used by people when they encounter communication problems are much more sophisticated than those used by current dialogue systems. Consider the fragment of TRAINS-93 dialogue 12.4 given below[1].

```
utt3   m:   um [sil] [click] [sil] here's
            the problem we need to transport
            two tankers of orange juice to
            Avon and three boxcars of
            bananas to Elmira [brth] uh the
            + bananas +
utt4   a:   + okay + so wait a second [brth]
utt5   a:   two tankers of
utt6   m:   orange juice
utt7   a:   yep
utt8   m:   to Avon
utt9   a:   to Avon
utt10  a:   okay
utt11  m:   three [sil] boxcars of bananas
utt12  a:   yep
utt13  m:   to Elmira
utt14  a:   okay
utt15  m:   the bananas [sil] must arrive
            at nine p.m.
utt16  a:   nine p.m.
utt17  m:   right
utt18  a:   okay
```

In utterance 3, the manager starts to introduce the goal of the dialogue, before being interrupted by the assistant (in utterance 4). In utterance 5, the assistant signals that he has not understood by repeating a fragment of utterance 3, and drawing-out the final word ("of") to indicate that he wishes his interlocutor to complete the utterance. The manager, now aware that he was not understood, gives the assistant a single bit of information and then waits for an acknowledgment (utt5) before proceeding. The manager changed his strategy after the communication problem occurred, and waits for acknowledgment of each piece of information he gives.

In utterances 7, 12, 14 and 18, the assistant signals that he has understood using simple affirmations. In utterance 9 the assistant uses repetition to provide evidence to the manager that he correctly heard what was said. In utterance 16, however, the assistant uses rep-

[1] The participants in the TRAINS-93 dialogues could not see each other, and the only mode of communication was speech. In the example dialogue, an utterance number is found to the left of each utterance. An **m** indicates that the utterance was spoken by the "manager", and an **a** indicates that the utterance was spoken by the "assistant". In the TRAINS-93 corpus, both dialogue participants are humans and are aware that the other participant is a human, although the "assistant" plays role of a computerized system. Words enclosed between +-signs indicate overlapping speech. Silence markers and non-words are denoted by [sil]. For more information about the TRAINS-93 corpus, see Heeman and Allen (1995).

etition to indicate uncertainty, emphasizing the "P" in P.M., presumably to contrast it with the other possibility: A.M.

In the dialogue fragment above we find examples of two different devices humans use when there is a problem with communication: (1) requesting a repair (in the case of utterance 5, repeating the part of an utterance that *was* understood and prosodically indicating that the rest of the utterance needs to be repeated), and (2) requesting acknowledgment (in the case of utterance 16, by repeating part of an utterance emphasizing some part that is in question).

As a first step in designing a conversational agent capable of generating repair requests and positive feedback in a natural and principled way, we want to identify how humans repair communication failures due to mishearing/misrecognition. The testbed for our conversational agent will be the TRIPS system (Ferguson & Allen 1998), an interactive planning assistant with a spoken natural language interface, developed at the University of Rochester. In support of this goal, we have collected a corpus of dialogues in the TRIPS domain in which the interface has been degraded to stimulate the need for repairs.

Research methodology

Formal models of computation can be conceptually elegant, rigorous, and general, but they are of little use if they cannot be implemented in a usable system. On the other hand, strategies that work for limited domains, but lack a well-founded theoretical basis, are unlikely to be extensible and are thus are also of little use. The current challenge for designers of dialogue systems is to develop strategies to bridge the gap between formally rigorous models that are computationally infeasible on one side, and systems that hand-craft solutions to a specific set of problems encountered in a very limited domain on the other.

The proposed work, an investigation into whether human-like repair initiation strategies are appropriate for spoken language systems, will follow a research and development methodology appropriate for dialogue systems given the current state of the art. The trend in discourse processing, in general, is toward building theory on an empirical basis. Computer systems that accomplish non-trivial tasks collaboratively with human users, using spontaneous spoken language as an input mode, are just becoming available in research environments. These prototype spoken language systems can process rudimentary interactions, such as simple question and answer pairs. However, they are still a long way from being able to engage in a very natural dialogue, even in limited domains.

To advance spoken language systems so that they can interact with human users in a more natural way, we need to observe what "natural" behavior is for humans. Collected corpora provide evidence of both *what* humans do with language and *how* they do it. By studying corpora of a particular genre (e.g., spoken, task-oriented dialogues between humans, such as in the TRAINS-93 corpus), we can identify dialogue phenomena that occur very frequently (and are therefore assumed to be particularly important), and use those observations to focus research and development. In this way, we avoid addressing theoretically interesting problems that rarely occur and so are less critical to the overall success of spoken language systems.[2]

Observations about where current spoken language systems fail, obtained by actually using the systems, further focus research and development by providing concrete evidence about what the current capabilities of the system are, what capabilities are critical to the success of a dialogue, and what the engineering issues are that need to be addressed. Repair subdialogues are an example of a phenomenon that occurs fairly infrequently in human-human dialogue, but are a critical issue in actual spoken language systems given the current state of the art in speech recognition, parsing, reference resolution and speech act interpretation.

Data collection

Researchers from many disciplines have relied on corpora as a source of data for empirical investigation into language use. Unfortunately, some of the repair phenomena that we wanted to observe because of their importance in human-computer dialogues, occur fairly infrequently in human-human dialogues, even in task-oriented dialogues where there may be more significant consequences of unrepaired communication problems.

The data collection performed for this study overcame this data paucity problem by degrading the interface between the dialogue participants, thus stimulating the need for repairs. The dialogue participants did not have visual contact with each other. Furthermore, the participants who played the part of the sys-

[2] Dialogue systems, unlike text understanding systems or machine translation systems, can compensate for incomplete coverage of language when they are capable of requesting repairs or signaling non-understanding to their dialogue partner. For example, if the system encounters a word in the user's utterance that is not in the system's lexicon, it can say something like, "I don't know the word ⟨unknown word⟩. Would you mind rephrasing that?" The user can then make the repair (in this case by finding another word to use), and the dialogue can continue. This capability makes the handling of rarely-used constructs less critical to the success of the system.

tem were only provided with the string of words output by the speech recognizer (they did not *hear* the users' utterances). This restriction was interesting because it allowed us to observe how well humans could perform the language understanding task that spoken dialogue systems must perform, given the same input.

The data collection was done according to the "Wizard of Oz" paradigm, in which one dialogue participant, the system user, believes that the other participant (the wizard) is a computer system. Using the WOZ paradigm gives researchers an indication of how users would interact with computers using spoken natural language as an interface, even when the technology to support the use is not yet developed.

The nature of human-computer dialogues is different from human-human dialogues, even when we consider only task-oriented dialogue. A WOZ data collection provides evidence as to whether this difference is due to the fact that human users actually prefer to interact differently with a computer (it may be the case that humans would like to be impolite or terse) or that it is a result of the current state of the art (the users are merely adapting to the capabilities of the system being used).

Collection environment

To perform the WOZ data collection, a system was implemented that provided wizards with a graphical user interface, enabling them to quickly access requested information, and generate synthesized responses.

The system users, working at Sun SPARCstations, wore headsets with microphones through which they provided input to the speech recognizer.

Software components used in the experiment were:

- A Sphinx-2 speech recognizer developed at CMU, whose language model was tailored for the TRIPS system
- TrueTalk, a commercial off-the-shelf speech generator (available from Entropics, Inc.)

The participants

All the participants in the data collection were native speakers of American English. This restriction was meant to control the performance of the speech recognizer to some degree and avoid idiosyncratic behavior. The system users were computer science undergraduate students and department visitors. Naive users were solicited to be the managers, since naive users presumably have fewer expectations (or at least less realistic expectations) about the capabilities of current spoken language systems. The wizards needed to be proficient at computer keyboarding (they needed to type quickly and accurately to interact effectively); computer science graduate students and faculty were used as wizards in this experiment.

Data Collection Procedure

Each time a dialogue was collected, the same procedure was followed. The collection procedure had the following stages:

1. The wizard was given a hardcopy of general instructions and practiced using the Wizard's GUI.
2. The user was given a hardcopy introduction to the domain and the interface to the system, and practiced using the click-and-hold-protocol.
3. The user and wizard collaborated on solving the given domain task.
4. The user completed a questionnaire about his impressions of the interaction, and his satisfaction with the performance of the "system".
5. The wizard annotated a transcript of the interaction with comments about what knowledge sources they drew on to figure out was said, and what influenced how they responded to particular utterances.

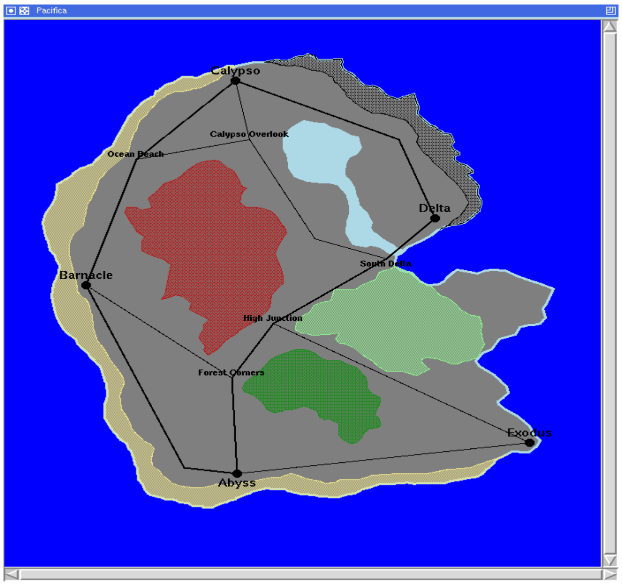

Figure 1: Map of Pacifica

When the users indicated that they were ready to begin the task, the experimenter displayed the map of the fictitious island Pacifica shown in Figure 1, and handed them an index card with the following instructions:

Construct a plan to transport all the civilians on Pacifica to Barnacle by 5 A.M. so that they can be evacuated from the island.

(The plan will be deployed at midnight.)

When the plan was nearly complete, the experimenter handed the users a second index card with the following update:

The road between Calypso and Ocean Beach is impassable.

Results

The results reported here come from a set of 7 dialogues collected in the Spring of 1999. Transcripts of the dialogues were created and annotated to facilitate analysis. Each user utterance was tagged as being either correctly recognized or containing recognition errors. The user utterances also received tags indicating whether they were abandoned or ungrammatical, since abandoned and ungrammatical utterances might be misunderstood for reasons other than recognition errors. There were a total of 348 user utterances in the 7 dialogues. Of these, only 121 were recognized without error. Statistics used to compute word recognition accuracy of each individual dialogue, as well as on the set of dialogues, is given in Table 1.

The word recognition accuracy measurement used in the analysis was computed as follows:

$$WRA = \frac{N - D - S - I}{N}$$

where
WRA = word recognition accuracy
N = the number of words actually spoken,
D = the number of words deleted,
S = the number of word substitutions, and
I = the number of words inserted

Wizard utterances were annotated with tags describing the type of feedback they provided. We developed two taxonomies of feedback for the purpose of annotating our collected dialogues: one for negative feedback and one for positive feedback. For each type of feedback listed, an example from the collected corpus is given. Although the taxonomies presented here were sufficient to describe the utterances encountered in our corpus, they are not comprehensive taxonomies of feedback in general. We were concerned only with verbal feedback to indicate whether the words spoken were understood sufficiently to determine the users' intentions. In our examples of feedback, user utterances are preceded by **u:**, and the wizards responses to those utterances are preceded by **s:**. Speech recognizer output is given in ALL CAPS, followed by what was actually said in parentheses. "." in the speech recognizer output represents silence markers.

Dialog Number	**# User Words Spoken**	**# Words Correctly Recognized**	**Word Recognition Accuracy**
1	434	357	82.3%
2	399	276	69.2%
3	408	266	65.2%
4	182	137	75%
5	529	310	59%
6	216	147	68%
7	263	198	75%
Overall	**2431**	**1691**	**70%**

Table 1: Recognition Accuracy Statistics

Negative feedback

The following list itemizes the types of negative feedback that we encountered in our corpus. When an utterance is tagged as being any of these feedback types, it indicates that the wizard could not identify the words spoken with enough confidence to understand and respond to the user's intention.

1. Request for a full repetition of the user's previous utterance

```
u:   THAT_ARE BY BAD
     (never mind that)
s:   I do not understand
```

2. Wh-term replacement of a missing or erroneous word or phrase.

```
u:  . HOW_MANY PEOPLE . ARE . IN . INCLUDES
    (how many people are in exodus)
s:  how many people are where
```

3. Attempt to salvage a correctly-recognized word or phrase from an otherwise erroneous utterance.

```
u:   . ONE_OF GOALS THE HELICOPTER
     (what about with a helicopter)
s:   What about the helicopter
```

4. Request for verification of what was said.

```
u:  . I_AM RETRIEVE START . .
    (I am ready to start)
s:  are you ready to begin
```

The first type of negative feedback above demonstrates no understanding of what the user had previously said. The other three types of feedback show some confidence in at least a partial understanding of what the user said.

Wizards resorted to asking for full repetitions of the users' previous utterances 33 times. In 24 of the cases where the wizard requested a full repetition, the user complied by repeating or rephrasing their original utterance.

Twelve times in our corpus, wizards used wh-term replacements to indicate that they needed part of the content of utterances repeated. In eight of these cases, the user responded by giving just the information requested. In one case, the user responded with a full repetition. In another case, the system's feedback did not actually correspond to what the user originally said, and the user repeated the original utterance without rejecting the wizards wh-term replacement. In the remaining two cases, the users' utterances were glib remarks that the user did not care to repeat. The users abandoned those utterances.

Twenty times in our corpus, the wizard used the third type of negative feedback, hoping to salvage some of the content of the original utterance. Although this strategy was not always applied effectively by the wizards, it may have had the effect of raising user satisfaction. It may be the case that the users were less frustrated knowing that at least part of what they said was understood. In several cases, the users responded to this type of negative feedback by repeating or rephrasing their original utterance, much like the response given when a full repetition was requested by the wizard.

Fifteen times in our corpus, the wizard requested verification of what it believed the user had said. Of the fifteen requests, ten were responded to by the user with explicit affirmations ("yes", "correct", and "right"). In another case, the user had abandoned the original utterance , and the user responded to the request for verification with a rejection. In one case, the user made a clarification to his original utterance in response to the request for verification. In three cases (2 of which also had explicit verifications), the user followed up with additional information. In one case, the user repeated himself before hearing the request for verification. In the final case, the user changed his mind about how he wanted to continue the dialogue, and so ignored the request for verification.

All of the utterances that the wizards responded to with some form of negative feedback either contained speech recognition errors or had no errors but were abandoned, and so appeared to be erroneous. It is interesting to note that the wizards only provided negative feedback for 80 of the 227 utterances (35%) that were recognized incorrectly or were abandoned. In most cases, the wizards felt confident enough that they had understood the users intentions to continue without requesting a repair. According to the wizards, they compensated for speech recognition errors by ignoring words from the speech recognizer that were not particularly salient in the TRIPS domain, and hypothesizing the correct words based on phonetic similarities and pragmatic constraint.

Positive feedback

The following list itemizes the types of positive feedback that we encountered in our corpus. The list is given roughly in order of feedback providing the least evidence on which to detect misunderstanding to the feedback providing the most evidence. Note that feedback that provides less explicit feedback of the wizard's understanding demonstrates greater confidence on the part of the wizard in the communication channel.

1. Using an acknowledgment term, such as "okay" or "right"

   ```
   u:  PICK UP THE PEOPLE AT . EXODUS AND TAKE
       THEM . TO BARNACLE
       (pick up the people at Exodus and take
       them to Barnacle)
   s:  Okay.
   ```

2. Providing a simple response to a question, that does not repeat any part of the question

   ```
   u:  . HOW_LONG WOULD IT TAKE . TO DRIVE A
       TRUCK . FROM DELTA . TO  EXODUS .
       (how long would it take to drive a
       truck from Delta to Exodus)
   s:  5 hours 27 minutes
   ```

3. Providing a conventional response to a highly conventionalized utterance, such as a greeting or thanks.

4. Continuing the dialogue with unsolicited information that is relevant to the user's utterance

   ```
   u:   . . FLY A TO BARNACLE .
        (fly to Barnacle)
   s:   that takes an additional 2 hours
   ```

5. Clarifying or correcting the user's previous utterance

```
u:   HOW_LONG IS IT ARRIVE FROM DELTA TO
     ABYSS . . .
     (how long is the drive from Delta
     to Abyss)
s:   on a truck?
```

6. Repeating part or all of the content of the user's utterance in the response to it

```
u:  . HOW_MANY . PEOPLE . ARE AT . HI .
    JUMPING . .
    (how many people are at high
    junction)
s:  there are no people at high junction
```

There were a total of 312 wizard utterances in the collected corpus. Eighty of these utterances provide negative feedback, and 232 provide positive feedback. Of the wizard utterances that provide positive feedback, 107 provide the most explicit form of feedback in our taxonomy. In 45 cases, this most explicit form was used in response to utterances that were correctly recognized. On the other hand, the use of the two least explicit forms of positive feedback were used almost exclusively in response to user utterances that were correctly recognized or had very minor errors.

Conclusion

The observations that we made about the use of positive and negative feedback in our corpus are consistent with the psychological models involving grounding, developed by Herbert Clark and his colleagues. For example, the wizards used negative feedback (repair requests and requests for verification) only when they could not extract enough of the users' utterance for their current purposes.

Furthermore, unless the speech recognition was perfect or very close to perfect, the wizard would give some explicit indication of their understanding of the content of the users previous utterance, thus giving the user the opportunity to detect and repair any misunderstanding that may have occurred. Many times the wizard provided explicit positive feedback even when the speech recognition was perfect. There are two possible explanations for this behavior. Wizards may have been considering the user's perception of the quality of the communication channel when forming their response. Since the wizards typed their responses, they may have been concerned that the delay would lower the user's confidence that they had been correctly understood. To compensate, the wizard may have chosen to provide more explicit feedback than they would have if they had spoken their response. Another possibility is that some wizards misunderstood their instructions. During the training process, wizards were instructed to avoid saying anything that would give them away as being human. Some of the wizards may have misunderstood this directive as meaning that they should give responses that sounded like machine responses, and that could have caused them to give the more verbose (and "machine-like") responses.

Future Work

The next step in our research is to develop a conversational agent capable of employing the same type of feedback and repair strategies that we observed in the collected dialogues, and that is capable of compensating for many of the misrecognition errors in ways similar to those used by our human wizards. The testbed for our new and (hopefully) improved conversational agent will be a baselined version of the TRIPS system. Our evaluation of the conversational agent will be based on overall task efficiency of the end-to-end system as was done in the TRAINS-95 evaluation (Sikorski & Allen 1996), along with user satisfaction metrics.

Acknowledgments

This work was supported by DARPA grant 5-28838. Thanks to Len Schubert, Lyn Walker and Mark Core for their advice and to George Ferguson and Alex Rau for developing the WOZ data collection environment.

References

Allen, J.; Miller, B.; Ringger, E.; and Sikorski, T. 1996. Robust understanding in a dialogue system. In *Proceedings of the 34th Annual Conference of the Association for Computational Linguistics.*

Brennan, S., and Hulteen, E. 1995. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems* 8.

Clark, H., and Brennan, S. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition.* American Psychological Association.

Clark, H., and Schaefer, E. 1987. Collaborating on contributions to conversation. *Language and Cognitive Processes* 2.

Clark, H., and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition* 22.

Ferguson, G., and Allen, J. 1998. Trips: An intelligent integrated problem-solving assistant. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98).*

Heeman, P., and Allen, J. 1995. The trains 93 dialogues. Technical report, University of Rochester.

Oviatt, S.; Levow, G.; MacEarchern, M.; and Kuhn, K. 1996. Modeling hyperarticulate speech during human-computer error resolution. In *Proceedings of the International Conference on Spoken Language Processing*.

Ringger, E., and Allen, J. 1996. Error correction via a post-processor for continuous speech. In *Proceedings of ICASSP-96*.

Sadek, M. D.; Ferrieux, A.; Cozannet, A.; Bretier, P.; Panaget, F.; and Simonin, J. 1996. Effective human-computer cooperative spoken dialogue: The ags demonstrator. In *Proceedings of the International Conference on Spoken Language Processing*.

Schegloff, E.; Jefferson, G.; and Sacks, H. 1977. The preference for self-correction in the organization of repair in conversation. *Language* 53.

Schegloff, E. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology* 97.

Sikorski, T., and Allen, J. 1996. Trains-95 system evaluation. Technical report, University of Rochester.