Report Number: WUCS-93-34

1993-01-01

The DIM system: WOZ Simulation Results - Phase II

Anne Johnstone, Umesh Berry, and Tina Nguyen

We report an experiment designed to compare human-human spoken dialogues with human-computer spoken dialogue. Our primary purpose was to collect data on the kinds of protocols that were used to control the interaction. Three groups of 12 subjects each were asked to complete tasks over the phone. These tasks involved the use of custom-calling features such as call-forwarding and speed-dialing. The experimental procedure was a new variation on the Wizard of Oz (WOZ) technique that allowed much clearer comparisons to be made between human-human and human-computer interactions. Subjects in the Operator Group were told they were talking to a... Read complete abstract on page 2.

The DIM system: WOZ Simulation Results - Phase II

Anne Johnstone, Umesh Berry, and Tina Nguyen

Complete Abstract:

We report an experiment designed to compare human-human spoken dialogues with human-computer spoken dialogue. Our primary purpose was to collect data on the kinds of protocols that were used to control the interaction. Three groups of 12 subjects each were asked to complete tasks over the phone. These tasks involved the use of custom-calling features such as call-forwarding and speed-dialing. The experimental procedure was a new variation on the Wizard of Oz (WOZ) technique that allowed much clearer comparisons to be made between human-human and human-computer interactions. Subjects in the Operator Group were told they were talking to a human operator. Subjects in the second two groups were told they were talking to a machine when in fact their queries were dealt with by two humans. Of these two groups, one listened to a sample dialogue before attempting the tasks (Computer Group (T)), while the other did not (Computer Group (NT)). Previous studies have shown that there are significant differences between human-human and human-computer interactions. The aim of this study is to examine such differences more closely. While some effects can be attributed to the beliefs about computers the subjects bring to the task, others appear to be connected with the on-going interaction styles of the speakers. Our study focuses on effects created by differences in interaction style. An important feature of the study is the use of two wizard, a technique which results in a realistically degraded communication channel. This created an interaction style that was very distinct from that of the human operator.

The DIM system :

WOZ Simulation Results - Phase II

Anne Johnstone, Umesh Berry, and Tina Nguyen

WUCS-TR-93-34

August 15, 1993

Department of Computer Science
Campus Box 1045
Washington University
One Brookings Drive
St. Louis, MO 63130.

Abstract

We report an experiment designed to compare human-human spoken dialogues with human-computer spoken dialogues. Our primary purpose was to collect data on the kinds of protocols that were used to control the interaction. Three groups of 12 subjects each were asked to complete tasks over the phone. These tasks involved the use of custom-calling features such as call-forwarding and speed-dialling. The experimental procedure was a new variation on the Wizard of Oz (WOZ) technique that allowed much clearer comparisons to be made between human-human and human-computer interactions. Subjects in the Operator Group were told they were talking to a human operator. Subjects in the second two groups were told they were talking to a machine when in fact their queries were dealt with by two humans. Of these two groups, one listened to a sample dialogue before attempting the tasks (Computer Group (T)), while the other did not (Computer Group (NT)).

Previous studies have shown that there are significant differences between human-human and human-computer interactions. The aim of this study is to examine such differences more closely. While some effects can be attributed to the beliefs about computers the subjects bring to the task, others appear to be connected with the on-going interaction styles of the speakers. Our study focuses on effects created by differences in interaction style. An important feature of the study is the use of two wizards, a technique which resulted in a realistically degraded communication channel. This created an interaction style that was very distinct from that of the human operator.

Table of Contents

Table of Contents

Page

List of Tables

List of Figures

1. Introduction

Recent advances in speech processing and natural language processing have raised the possibility of creating systems that allow interactions through dialogues rather like the cooperative dialogues we have with each other. Research has begun to focus on what "rather like" might mean in the near and more distant future. Is speech a natural medium for human-computer interaction? Can system designers apply models developed from human-human interaction to human-computer interaction? Researchers speculate about whether people will use a computer as a tool, a sort of limited listening and talking appliance, or welcome it as a collaborative partner in the performance of some unfamiliar task. It is likely that, as in other areas of its application, the computer will completely change our concept of what a tool can be.

There is some conflict in published accounts that contrast human-human interactions with human-computer interactions. (Amalberti et al, 1993) have suggested that these conflicts are due to differences in the simulated systems' behavior. They cite as evidence the observation that differences in user behavior are reduced when differences in system behavior are reduced. Our study is designed to shed further light on this area by presenting a case where the human and simulated computer's interlocutory styles are very distinct. We find that patterns of user behavior can be related to this difference. We then discuss the implications of these findings for system design. Because there is a positive result, a marked difference between the human-human and the human-computer cases, we are better able to compare and contrast the two types of processing and to offer an improved computational model suited to human-computer interaction.

2. Background and Motivation

Representations used with computers have moved from the zeros and ones of machine code to languages that people use to communicate with each other, such as natural languages, both typed and spoken, and visual languages. There has also been an increasing emphasis on improving the quality of interaction. The Macintosh interface, for example, employed various visual and linguistic symbols to make it easier to control applications, switch between applications, and get relevant help.

It is still unclear how system designers will use spoken natural language dialogue to improve human-computer interaction. It is conceivable that spoken language could be used to provide flexible structured dialogues for help in problem-solving and information gathering. An important benefit of this type of interaction is that the system would no longer require people to learn its command language in advance. All they would need to know is how to get help, and the better the dialogue system the easier this would be since, in theory, they would be able to use natural language and draw upon standard conventions of cooperative dialogue.

There may be important differences in human-computer interactions compared with human-human cooperative dialogues, however. For example, (Amalberti et al, 1993) report that subjects in a spoken interaction experiment "tended to solve the problem on their own and not to use linguistic devices to facilitate understanding on the part of the computer. Subjects in the Computer Group thus tended to use the computer as a tool rather than as a participant in collective problem solving." This result may reflect a fixed constraint on human-computer interaction or may only indicate that we are still in the early stages of integrating speech processing, discourse processing and reasoning into human-computer interfaces.

Two areas of research are contributing to the goal of achieving spoken dialogue processing. The first involves improving speech processing to the point where machines can handle limited spontaneous dialogue robustly. The second involves experimental work using simulated systems to gather data on how people might actually speak to such a system. A common goal is to specify systems that are habitable, that is, they do not feel unnaturally constrained to the user, and yet they are constrained enough for successful system performance. Clearly the two research programs are not unconnected. The results of the experiments may influence the short term goals at least of some system designers; and the work of the system designers may indicate what is currently realistic to the experimenters.

There is some evidence that the behavior of the dialogue system affects the linguistic and problem-solving behavior of the user. An important issue, then, is the specification of the simulated system. Is it clearly specified? Is it within the reach of current systems or is it closer to a human's performance? Experiments may tell us that people talk to a computer exactly as they would to a person when the computer appears to offer the same capabilities as a person, but that does not help the designers of today's systems which fall far short of such a standard. In the following sections we will briefly review, first, the current capabilities of speech processing and dialogue processing; second, the characteristics and results of simulation experiments.

2.1 Speech and Dialogue Processing

Commercially available Automatic Speech Recognition (ASR) systems can recognize isolated words and phrases, but not fluent continuous speech. Typically such systems also allow only limited interaction, although a dialogue of sorts could be conducted in isolated phrases if the context and the goals of the user could be represented and used in selecting responses. For example:

User:	"Microsoft Word"
System:	"Which file?"
User:	"Most recent"
System:	"Please repeat the command."
User:	"Help"
System:	"Which topic?"
User:	"File names"
System:	"Files may be referred to by full identifier or by "last""
User:	"Last"
System:	*<opens application with the correct file>*

Speaker independent continuous speech processing is now achieving high accuracies for careful speech in restricted domains, as in for example the SPHINX system (Lee, 1989). Like the isolated phrase systems, these systems still have a limited capacity for interaction. Typically, they do not construct a representation of the on-going dialogue and so can only respond to single questions or to simple commands. The underlying technology uses the Hidden Markov Model (HMM), which employs local pattern-matching techniques to find the most likely path through a string of word or phoneme models. This technique can make little or no use of more global prosodic, syntactic, or semantic information. One utterance is processed at a time in its entirety. The system outputs a string of word hypotheses sometimes with associated probability scores and some syntactic annotation.

In the area of discourse processing, a great deal of work has been done recently on how problem-solving knowledge, knowledge of the user's plans, and discourse knowledge can be used to represent the structure of a dialogue. Early work in discourse processing focused on story-understanding and so did not address the issue of interaction. More recent work (Litman and Allen, 1987; Lambert and Carberry, 1991) has looked at interactive dialogues, but the work has typically been based on typed text.

There are thus several issues specific to the combination of speech and dialogue that have barely been addressed. The ASR systems do not yet deal adequately with prosodic cues, broken or incomplete utterances, hesitations and repetitions. The models of dialogue do not take into account misunderstandings and the complex patterns of acknowledgement, disagreement and repair by which people try to avoid such misunderstandings. Work on these issues has begun, but it is still unclear what knowledge people use, and how such knowledge could be integrated into a speech processing system (Allen and Schubert, 1991; Traum, 1991; Young et al; 1989).

2.2 Experiments in Human-Computer and Human-Human Interaction

The rationale behind Wizard of Oz (WOZ) studies (Kelley, 1983) is that people adapt their speech according to the identity of their interlocutor. If they are led to believe that they are talking to a computer then they will speak accordingly and data can be collected that will be useful in designing an automated system.

There have been a number of studies to investigate how people talk to computers (Baber and Stammers, 1988; Thompson, 1980; Kennedy et al, 1988; Morel, 1989; Richards and Underwood, 1984; Amalberti et al, 1993). The general consensus seems to be that people do change their style of speaking when talking to a computer. They typically use a smaller vocabulary, shorter sentences, make fewer exchanges, interrupt less, and offer fewer justifications for their requests. Our first study (Balentine, Berry, and Johnstone, 1992) found very similar results. These changes in speaking style should make the recognition task easier. Such results therefore provide support for the claim that useful and powerful communication tools can be built by combining dialogue capability with current speech technology.

It is unclear from previous work, however, whether people speak differently because they believe they are talking to a machine, or because they are mimicking the conversational style of a computer-like interlocutor. Morel, for example reported that subjects were non-polite when talking to a computer system. Richards and Underwood, on the other hand, found that subjects *did* use polite phrases, but only when the computer system also used them, indicating that the users were responding according to the style of the interlocutor rather than to its identity as system or human. Amalberti et al performed an experiment in which the only difference was the supposed identity of the interlocutor: the wizard's dialogue style was no different in either case. They found that, over time, differences between the Operator group and the Computer group decreased, suggesting that the subject's linguistic behavior was due less to the fact that they were interacting with a computer and more to the interaction style of the interlocutor. Clearly more work is needed in order to clarify how people respond to systems with different linguistic and problem-solving capabilities, both real and simulated.

3. General Methodology

Our general methodology covers the following areas of research:

1) WOZ / designers' guess work
2) Conversational Analysis / Discourse Analysis for data analysis and computational modelling

We will discuss each of these in turn.

System development requires an understanding of what users say to the system. This is impossible because of the design paradox: it is essential to know how users will communicate with a system in order to build it; yet user interactions with the system cannot be explored until the system is built (Kelley, 1983). This paradox can be avoided by adopting the Wizard of Oz (WOZ) simulation technique first proposed by Kelley. He describes the WOZ paradigm as one in which experimental participants are given the impression that they are interacting with a computer that understands English as well as another human would. The experimenter, acting as the "Wizard", surreptitiously intercepts communication between the participant and the computer, supplying answers and new inputs as needed. (Fraser and Gilbert, 1991) proposed an iterative WOZ methodology which uses three phases of experiments in the specification and evaluation of a dialogue system. The first phase is pre-experimental. During this phase the experimenters decide on the boundaries of the domain and the basic functionality of the system. In the second phase a simulated system is used to collect data on such things as vocabulary, grammar, and dialogue model. The system is as unconstrained as possible, in order to find out the natural boundaries of the task, the way that people would behave if they were not constrained by the system's performance. The third phase involves as many iterations as is possible. In this phase, more realistic constraints are placed on the system. Wherever possible, actual components are implemented and used. We are in agreement with Fraser and Gilbert that this methodology is preferable by far to the method of relying on the system builders' intuitions and guesswork. However, in our own experimentation we have become more aware of the difficulties involved in i) limiting what work the Wizard is doing, and ii) the complicated ways in which different constraints affect subjects' behavior. For example, in our first WOZ (Balentine, Berry and Johnstone, 1992) it was not at first apparent that the Wizard was using prosodic and other turn-taking cues to ease *grounding*, or mutual understanding (Clark and Schaefer, 1989). In a more realistic system this capability would not be available and this could have far-reaching effects on interlocution style. For these reasons we have moved more towards using the third phase primarily to test specific hypotheses. The current study focused on turn-taking behavior, contrasting human-human dialogues with realistically simulated human-computer dialogues.

Our approach to model building is also strictly empirical, along the lines of the techniques used in Conversation Analysis. We did not form any premature theories, based on our intuitions (as is often done in Discourse Analysis), about what is and what is not true for human-human or human-computer interaction. Instead, we collected a lot of data and analyzed the data extensively for recurring patterns, which were used to test hypotheses and develop theories. A comparison of the two techniques can be found in (Levinson, 1983). An interesting approach is that of (Litman and Hirschberg, 1990) who have recently begun to use groups of subjects to segment and label dialogues. In our study both the human-human and the human-computer dialogues are analyzed so that we can compare two approaches to the same communication task. Wherever possible, the human-human data will be used to build a more natural model. Where this is not feasible or not desired by the users (e.g. politeness), the model based on human-computer interactions can be used. Finally, it is necessary to actually implement components and try to integrate them into a working system. It is only through trying to implement a dialogue management system in a realistic setting that many problems to do with understanding and grounding have come to light. This is partly because, as in many AI systems, the problem of integrating the different knowledge sources is often harder than the problem of building any one of the knowledge sources.

4. Experimental Methodology and Procedure

In the following sections we discuss how various design requirements were met for this particular experiment. Requirements include:

- **Realistic simulation**: the simulated system must be functional,
- **Detailed specification**: wizards must know what they can and cannot do,
- **Plausibility**: subjects must believe they are talking to a computer,
- **Wizard consistency**: wizards should not change their style over time, or according to the identity of the subject,
- **Richness of domain**: the domain should be small enough to be plausible, but rich enough to generate interesting dialogues.

5. System

Subjects called over a telephone line to speak to either the human operator or to the "computer system". In accordance with our goals, we used two wizards as the "computer system" in this experiment. One wizard performed the speech recognition and language understanding (the natural language wizard) and the second performed the dialogue management and the response generation (the dialogue wizard). The natural language wizard passed on a list of *detectables* , words and their synonyms we expected the automated system to be able to recognize, together with indications of silence and words that were not understood ("babble") , to the dialogue wizard. A partial list is given in Figure 1 below, and the complete list is given in Appendix A.1. The dialogue wizard was in a separate room and could not hear the subject. The dialogue wizard had a menu of pre-recorded utterances that she could select in response to the current list of detectables and the current state of the dialogue. A partial list is given in Figure 2, and the complete list is given in Appendix A.2. Both wizards used workstations to control their part of the interaction. Detectables and dialogue responses were quickly selected using menu systems.

HELP

confused, oops, directions, how, what, information

DIAL

call

CALL-FORWARD

call-forwarding, forwarded, transfer, reach, reached, receiving

CALL-WAITING

waiting

INTERRUPTED

interrupt

WANT

like, need

Figure 1. Partial List of Detectables

The detectables are given in boldface and the word lists following them are their synonyms, which were considered to be the same as the detectable.

General

01) Hello, this is Southwestern Bell's voice phone service. Can I help you?

02) Do you need help?

03) Please say what service you want.

04) Please rephrase the command.

05) Is there anything else?

06) Thank you.

07) Thank you for participating..... hang up now.

08) I'm sorry, let me get someone who can help you. Please wait.

09) Yes.

10) No.

Help

11) Please say what feature you want help with.

12) The features available are Call-Forwarding, Speed-Dialing, Call-Waiting and Call-Forward Scheduling.

13) With Call-Forwarding, your phone automatically forwards all calls to a location you choose.

Figure 2. Partial List of Responses

There were two reasons for setting up the system in this way. First, the dialogues would be more realistic in that what was and what was not recognized would be closer to an automated system. Secondly, the communication channel carrying information about turn-taking, agreement or acknowledgement was broken. The dialogue wizard received only lexical items. There were no prosodic cues, few backchannel responses, no indications about when the wizard should take a turn or wait for more information, other than silence. The dialogue wizard had to select an appropriate response to play back on the basis of the information that was provided by the natural language wizard. Once the response was selected, it was played in its entirety and could not be interrupted. However discernible speech was recognized when spoken over a response. Each session was logged and recorded with the subjects' permission.

One concern in designing such systems is that the behavior of the wizard be as consistent as possible for each subject even after repetition of the same task. The use of predefined detectables and responses helped to ensure that this condition was met.

A second concern is that the system be convincing as a computer system. In a previous experiment we tested the use of both synthesized responses and prerecorded responses. Subjects were all convinced that the prerecorded responses were generated by a computer system, and so we continued to use this method of response in the present experiment. Only fifty responses were required to cover all situations.

6. Subjects

Thirty-five subjects were used in this experiment. They were divided into the following three groups:

(i) The Operator Group: There were twelve subjects in this group and were told that they would be interacting with a human operator.

(ii) The Computer (T) Group: There were twelve subjects in this group and were told that they would be interacting with a computer system. Before they began their interactions, they were made to hear a sample dialogue (given in Appendix B), between a user and the computer system.

(iii) The Computer (NT) Group: There were eleven subjects in this group and were told that they would be interacting with a computer system. However, they did not hear a sample dialogue.

Each subject was given four tasks. The tasks were the same across different subjects but they were not necessarily given in the same order. The subjects were compensated for their participation. See Appendix C.1, C.2 and C.3 for demographic information on the subjects from each of the three groups.

7. Materials

7.1 Scenarios

Four scenarios, depicting different custom calling features provided by Southwestern Bell Telephone, were used (See Appendix D). The scenarios were described using as little specific vocabulary and syntax as possible, since we wanted the subjects to use their own vocabulary and speech patterns instead of adopting those found in the scenarios. The scenarios ranged from tasks that were easy to ones which were difficult in regard to the number of things that had to be done.

This domain was chosen for a number of practical reasons. Firstly, it requires very little domain or common sense knowledge and so is feasible given the current state of reasoning systems. Second, it is includes tasks of some complexity that elicit rich dialogues. These include call-forward scheduling and speed-dial directory updating. Simple scheduling and updating tasks are likely to arise in a number of different applications. A third reason, is that this is a domain where users typically understand the end goal, e.g. call-forwarding, but may not remember the steps, commands or codes required to implement it. This elicits help dialogues, which again can be expected to be necessary in a number of domains.

7.2 Questionnaires

Subjects answered three types of questionnaires during the experiment. They completed Questionnaire A after each task. Questionnaire B was given after the final task had been completed. The experimenter presented Questionnaire C after the experiment was completed. See Appendices E.1, E.2 and E.3 for the questionnaires that were given to each of the three groups.

8. Procedure

Subjects were placed in a room with a telephone, the human operator or the wizards were in another room (unseen by the subjects). The experimenter (who was different from the operator or the wizards) read the instructions (See Appendices F.1, F.2 and F.3) to each subject (the instructions for the different groups were different) and answered any questions. The experimenter instructed the subjects to read the first of the four scenarios and call the human operator or the system after they understood the task. When the subjects completed all the scenarios, the experimenter provided the subjects with the questionnaires.

9. Results

9.1 Transcribing the Dialogues

All the dialogues were captured and stored digitally affording much greater control than digital recording. The dialogues were first transcribed by an audio-typist who typed all the words spoken. Next, two of the experimenters, working independently, checked that all words had been correctly transcribed and in addition included pauses, speech phenomena such as *um*, and *uh-huh*., and overlapping speech. Soundtool on the Sparc IPC workstation was used to play back stretches of dialogue.[1] The transcribed dialogues for the three groups are documented in (Johnstone, Berry, and Nguyen, 1993b).

Figures 3 and 4 below show transcribed dialogues from the Computer (NT) group and the Operator group respectively. S is the System and U the User. The start and end points of overlapping speech are shown using square brackets []. {short} and {long} indicate periods of silence. They are subjective measures. However, this is in some ways more accurate than an absolute measure, since the perceived shortness of a pause depends on many factors including, for example, the speaker's normal speaking rate, or the reason for the pause. Different listeners agree on what feels short or long in the context of a particular utterance.

Computer (NT) Group

S: hello {short} this is southwestern bell's voice phone service {short} can I help you
U: I'd like to schedule some call forwarding
S: the call forward schedule is {short} monday through friday from 7 30 am {short} to 6 o'clock pm {short} the calls are not forwarded {short} at all other times the calls are forwarded to 2 2 4 {short} 0 0 9 9
U: um hmm {short} ok I would like to have some calls forwarded {short} at a different time
S: please give the day {short} the time interval {short} and the telephone number where you want your calls forwarded
U: ok {short} on monday {short} I would like my calls forwarded between the time of 2 pm and 4 pm {short} to 7 7 6 {short} 1 2 3 4 {long} on friday {short} I would like my calls
S: on monday {short} your calls are forwarded between 2 o'clock pm and 4 o'clock pm {short} to 7 7 6 {short} 1 2 3 4
U: ok on friday {short} I would like my calls {short} forwarded between 7 30 am and 12 noon {short} to 3 5 6 {short} 4 9 3 0
S: on friday {short} your calls are forwarded {short} from 7 30 am to noon {short} to 3 5 6 {short} 4 9 3 0
U: ok on the same day {short} I would like my calls forwarded to 3 3 9 {short} 2 3 2 3 {short} until 8 pm that evening {long}
S: please give the day
U: that would be on friday
S: please specify the start {short} and the end {short} of the time interval
U: ok the start would be {short} from noon {short} until 8 pm
S: on friday {short} your calls are forwarded from noon {short} to 8 pm {short} to 3 3 9 {short} 2 3 2 3
U: thank you
{hang up}

Figure 3. Transcribed Computer (NT) Dialogue

[1] A third much finer-grained transcription that marked prosodic and other cues was also done later for building the computational model of turn-taking. This is reported in Berry and Johnstone 1993.

Operator Group

S: hello this is southwestern bell's phone service can I help you{short}
U: yes I need to make {short} a few changes in my weekly call forwarding schedule {short}
S: certainly what can I change for you
U: on monday {short} I have a meeting from 2 in the afternoon until 4 {short}
[ ]
S: um hmm
S: ok
U: and those numbers {short} or they should call 7 7 6 {short} 1 2 3 4 during that time
S: ok so that's 7 7 6 {short} 1 2 3 4 between 2 and 4 on monday
U: yes {short}
S: fine {short}
U: then on friday in the morning {short} from 7 30 until noon {short} I will be at 3 5 6 {short} 4 9 3 0
[ ]
S: ok
S: ok {long} that's friday 7 30 to noon {short} and the number is 3 5 6 {short} 4 9 3 0 {short}
U: that's right {short}
S: I have that
U: and then {short} in the afternoon {short} through the whole time period {short} that I have scheduled through 6 o'clock {short} I will be at 3 3 9 {short} 2 3 2 3 {short}
S: ok so is that every afternoon
U: um just friday afternoon
[ ]
S: just friday ok and that was from noon until 6 o'clock {short}
U: right {short}
S: and that's 3 3 9 {short} 2 3 2 3
U: yes {short}
S: ok {short}
U: that's it then
[ ]
S: that's all the changes {short}
U: yes
S: ok so we've updated the schedule {short}
U: thank you {short}
S: byebye
U: um bye
{hang up}

Figure 4. Transcribed Operator Dialogue

9.2 Conciseness of Human-Computer Speech

Our first pass at analyzing the data used indices that have proven useful in other studies:

- mean number of words per dialogue
- mean number of exchanges per dialogue
- average utterance length
- vocabulary size

These indices have been used to measure the verbosity or conciseness of subjects when talking to humans or to machines. Some of the indices are relevant to automated speech processing. For example, a smaller task vocabulary and a constrained grammar favour more robust recognition. Other indices, such as the number of exchanges, may measure the efficiency and ease of using spoken dialogue to carry out the task

In line with other studies, we found that people were more concise when talking to the computer. The results per dialogue are summarized in Table 1 below.

Table 1. Linguistic Data

	Operator	Computer (NT)	Computer (T)
Words per Dialogue (Mean)	102.1	77.0	74.7
Utterances per Dialogue (Mean)	9.0	4.7	4.3
Utterance Length (Mean)	11.3	16.3	17.4
Vocabulary (Total)	420	319	307
% of Subject Initiated Utterances	78.5 %	56.5 %	77.7 %

Subjects in the Operator Group used an average of 102.1 words and 9.0 exchanges per task with a mean utterance length of 11.3, as compared to 77.0 words and 4.7 exchanges per task with a mean utterance length of 16.3 for subjects in the Computer (NT) group. The vocabulary for the Operator group was 391 words (in 45 dialogues) as compared to 319 words (in 39 dialogues) for the Computer (NT) group.

ANOVA tests indicate that the differences were significant for each group. A post hoc analysis indicated that in each case the Operator group was significantly different from both the Computer groups with, on an average, a higher number of words ($p < .05$), a greater number of exchanges ($p = 0$), and a lower utterance length ($p < .05$).

We found no significant difference between the Computer (T) and the Computer (NT) groups on these measures. Thus, the training scenario we used did not appear to affect the verbosity, the number of exchanges, or the length of exchanges that users subsequently had with the computer. However, we did find a significant difference in the percentage of the utterances by the subject that were self-initiated. The percentage is significantly higher for the Operator and the Computer (T) group than for the Computer (NT) group. Users took the initiative more often in the Operator and the Computer (T) group as compared to the Computer (NT) group. This is discussed in more detail below.

9.3 The Effects of Turn-Taking Protocols

A result that we did not anticipate was that people used longer exchanges when talking to the computer than when talking to the operator. Subjects in the Operator group had an average exchange length of 11.3 words as opposed to 16.3 words for subjects in the Computer (NT) group and 17.4 in the Computer (T) group, a significant difference ($p < .05$) This is not in line with previous work (Morel, 1986) which suggested that people use shorter exchanges when talking to a computer. A closer examination showed that the dialogues of the Operator group had a lot of rapid turn-switches in the form of short utterances (from both the subject and the operator) signalling agreement, understanding, uncertainty and confirmations (in the form of *OK*, *uh-huh* and verbatim repetitions of parts of previous utterances). In addition to this, the dialogues of the Operator group had many short utterances (from both the subject and the operator) signalling the end of conversation (in the form of *thanks* and *goodbye*). In the following sections we will use the terms *groundings* and *closings* to refer to these two kinds of utterances respectively.

As mentioned earlier, one of the motivations for conducting the experiment was to gather data on turn-taking, particularly in the human-human case. Analysis of the data showed that although the subjects in the Computer (NT) group used approximately 76% of the number of words and 50% of the number of exchanges that subjects used in the Operator group, subjects in the Computer (NT) group needed approximately *twice* the amount of time to complete the task as compared to the time needed by subjects in the Operator group. Closer examination of the dialogues from the Computer (NT) group has revealed that most of the pauses occurred when the turn switched from the user to the computer (it took some time for the "computer" to realize that the turn had been released), whereas exchange of turns from the computer to the user were quick (the users easily recognized when a turn had been released). Clearly two very different turn-taking protocols were being used. In the next section we discuss the effects of these protocols on language behavior.

9.4 Characteristics of the Dialogues with Groundings and Closings Removed

The differences between human-human speech and human-computer speech appear to be affected less by the user's model of what the computer is capable of, than by the patterns of behavior presented to the users during the dialogues. The closer the pattern to normal human behavior, the more the dialogues resemble human-human dialogues.We believed that the main difference between the Operator and the Computer system in our experiment was the protocol for grounding, and the turn-taking behavior that implemented it.

Subjects in the Computer groups talked to a "computer" which had the tasks of speech recognition and language understanding separated from the tasks of dialogue management and response generation. The communication channel carrying information about turn-taking was thus significantly diminished. In particular, there were no prosodic cues available to the dialogue wizard who was managing the turn-taking. By contrast, subjects in the Operator group talked to a human operator who was able to hear everything that was being said and hence had access to prosodic cues and whatever turn-taking mechanisms operate in natural conversation.

The differences in the turn-taking are quite clear in the two sample dialogues given earlier. Compare, for example, the human-computer and human-human dialogues to set up the call-forwarding schedule for Monday.

Computer (NT) Group

U: ok {short} on monday {short} I would like my calls forwarded between the time of 2 pm and 4 pm {short} to 7 7 6 {short} 1 2 3 4 **{long}** on friday {short} I would like my calls
S: **on monday {short} your calls are forwarded between 2 o'clock pm and 4 o'clock pm {short} to 7 7 6 {short} 1 2 3 4**
U: ok on friday {short} I would like my calls {short} forwarded between 7 30 am and ....

Operator Group

U: on monday {short} I have a meeting from 2 in the afternoon until 4 {short}
[ ]
S: **um hmm**
S: **ok**
U: and those numbers {short} or they should call 7 7 6 {short} 1 2 3 4 during that time
S: **ok so that's 7 7 6 {short} 1 2 3 4 between 2 and 4 on monday**
U: yes {short}
S: **fine** {short}
U: then on friday ...
[ ]
S: **ok** {short}

In the Computer (NT) dialogue, there are several short pauses at significant points in the utterance: after the opening, after the day, after the time, and during the giving of the phone number. However, there is no response from the system. Note the long pause after the subject has given Monday's schedule. It seems very likely that she was waiting for some feedback at this point. Getting none, she starts to give the information for Friday. Then, late and disruptively, comes the computer utterance that checks through repetition the schedule for Monday. The subject confirms that the computer has got it right with an **ok**, and then continues on again with the Friday schedule.

In the Operator group, the operator frequently provides feedback in the pauses left open by the subject: the **um hmm** after the day, the **ok** after the times, and finally the repetition and summary after the phone numbers have been given. The user acknowledges that the operator has correctly understood (**yes**), the operator signals that she is comfortable with what has been done so far (**fine**) and the user starts to give the information for Friday.

We hypothesized that, if the utterances relating to such acknowledgments (groundings) and closings were removed from the dialogues, the differences in our measures of human-human and human-computer speech should be substantially reduced. We defined groundings and closings as follows:

i) Grounding: This refers to the various ways in which information is confirmed and clarified in spoken dialogues. Specifically, it takes on two forms: words like *OK* (or any synonym) that signal understanding, or verbatim repetitions of previous utterances that signal understanding or requests for confirmations. For example, in the following dialogue fragment, the two occurrences of groundings are given in boldface.

U: the number is 7 2 7 {short} 7 5 0 7 {short}
S: **7 2 7** {short} **7 5 0 7**
U: **yes** {short}

ii) Closing: This is mostly uttered at the end of conversations in order to thank the other participant or to bid the other participant good-bye. In the following example, examples of closings are given in boldface:

U: ok very good **thank you** {short}
S: **thank you**
U: uh huh **bye**

Specifically, we were interested in testing the following three hypotheses:

1) We had observed that users used more words when talking to the Operator than to the Computer. We hypothesized that the extra verbosity could be accounted for the by the closings and groundings which are manifestations of the distinct interaction styles in the Operator vs. the Computer groups.

2) We had observed that users used more exchanges when talking to the Operator than to the Computer. We hypothesized that the extra utterances could also be accounted for by the closings and groundings of the communication protocol.

3) We had observed that users interacted via shorter utterances with the Operator than with the Computer group. We hypothesized that this could accounted for by the presence of closings and groundings which are relatively shorter than other utterance and hence lower the average utterance length.

We removed closings and groundings from all dialogues, both the Operator dialogues and the two sets of Computer dialogues, and recalculated the measures of conciseness. Table 2 below shows the results per dialogue. We performed ANOVA tests measuring differences per subject.

Table 2. Linguistic data: without Closings and Groundings

	Operator	Computer (NT)	Computer (T)
Words per Dialogue (Mean)	93.1	76.2	73.6
Utterances per Dialogue (Mean)	5.1	4.5	4.0
Utterance Length (Mean)	17.9	16.8	18.3
Vocabulary (Total)	403	315	302

We found that the significant difference between the Operator group and the Computer group had disappeared. There was now *no* significant difference on average number of words, average number of utterances, and mean utterance length.

10. Discussion

Our primary result is that the extra verbosity in the human-human group can be accounted for solely by the words and exchanges typically used to ground and close dialogues. We believe that this result is caused by a difference in turn-taking protocols. However, the purpose of turn-taking, grounding, and politeness, and the relation of these behaviors to the user's model of the interlocutor are quite subtle and complex.

Let us first be clear about what we have removed from the dialogues. We removed utterances that are typically related to closings and groundings as described before. We believe that the utterances we removed are those that require an easy and rapid method of switching turns, i.e. back-channel responses. They seem to provide a minimum level of politeness (stock phrases like *thank you*) and a minimum assurance of understanding (murmurings of *OK* and *uh-huh*, verbatim repetitions). There are other more sophisticated methods of grounding, such as paraphrase, overt requests for acknowledgment of some action, etc., which we did not remove.

We found that we removed more closings and groundings from the Operator dialogues than from the Computer dialogues. This only tells us how people behaved in the Operator group. However, it is an interesting result in its own right. Previous studies have shown only that there is a difference in conciseness between the human-human and human-computer groups. Our study is more specific on exactly what that difference consists in, the use of simple closings and groundings.

We are also claiming, however, that we can say something about the human-computer interaction, specifically, that there was an effect of the two-wizard set-up on the turn-taking protocol that *prevented* the users' from using the simple human-human grounding protocols. An opposing view would be: the differences we observed had nothing to do with the communication protocols but were to do with the identity of the interlocutor. People are polite and concerned with grounding only when talking to people; they feel no need to do this with computers. The users were not prevented from closing and grounding. They would not have made use of such linguistic cues with the computer, *even if* they had had a communication channel that facilitated such an interaction.

This position requires that there always be a difference between human-human and human-computer dialogues, since the difference is caused by the difference in interlocutor identity, rather than in on-going interlocutor style. However there is evidence contradicting this from (Amalberti et al, 1993). They found that, when the communication protocols were identical there was indeed no significant difference on their linguistic indices between the human-human and the human-computer groups. We argue that, in an experiment where we *did* find significant differences on the same measures that Amalberti et al used, and where the major difference was the turn-taking protocol, the linguistic differences were most likely due to the differences in protocols and not to differences in interlocutor identity.

Further evidence is also provided by the fact that we found, when examining our dialogues, that there were numerous instances, especially in the beginning of dialogues, where the users would say what they desired, pause for a while (perhaps expecting a response) and finding none, continue speaking, giving the remaining information or completing the request. We also found a great deal of hesitancy to close the dialogues, as though the users were not reassured that the interaction had been successfully completed. One can speculate, that if the system had been able to give back-channel responses, the subjects would have made use of this protocol.

Finally, there is evidence that people use *other* methods of grounding when talking to a computer, such as asking for explicit acknowledgments that some action has taken place. In our questionnaires, the lack of acknowledgment, of "knowing what was going on" was frequently mentioned as a concern.

10.1 Implications for System Design

As mentioned before, subjects in the Operator group took part in rapid turn switches designed to signal agreement and understanding. The subjects in the Computer group behaved quite differently. It is our belief that subjects were forced by the limitations of the system to mimic the conversational style of the computer and to make statements in their entirety instead of saying them over more turns involving confirmations and agreement. Apart from increasing the length of the utterances, the degraded turn-taking protocol also resulted in large portions of silence so that the tasks took twice as long as in the Operator group. Occasionally, the user and the computer would start speaking at the same time after a pause from the user (each interpreting the pause in a different way). In general, conversation was not smooth (probably due to the erratic cues being used). This seems to suggest that the absence of a turn-taking mechanism operating on prosodic and lexical patterns leads to the conversation being inefficient and unnatural. The tasks *were* successfully completed, indicating that the system constraints nonetheless resulted in a workable system. Our data suggests that, in addition to receiving information about the kinds of vocabulary and syntactic and discourse styles that are appropriate, users are also looking for cues about how turn-taking and grounding will be accomplished when talking to their interlocutor, be it human or computer. Our experiment clearly demonstrates the need for a turn-taking model that is closer to the human-human protocol. We plan to investigate this further by studying the existing data, and through more constrained experiments.

An important question is: would the subjects have made more use of such closings and groundings in the Computer group if the communication channel had been suitable and relevant responses available? The prerequisite to this is to have an automated system able to provide such capabilities. In a further study reported elsewhere, we analyzed the human-human dialogues and developed a model that allows turns to be automatically predicted based on simple prosodic, syntactic and semantic cues. The model is described in (Berry and Johnstone, 1993). While that model is able to predict accurately *when* a turn can be taken, it has little to say about *why* a turn might be taken by one interlocutor rather than another. This requires integrating the model with a model of cognition, a difficult task. To put it mildly, the relation between turn-taking and the grounding of meaning is highly complex.

10.2 Turn-Taking, Grounding, and Understanding

Reddy (1979) dissected "the conduit metaphor" underlying much of our understanding of language and pointed out the dangers for cognitive science of assuming such a metaphor. Under this assumption thoughts are "put into" words as if these were containers from which the contents could easily be extracted at the other end. In its place, Reddy proposed the toolmaker metaphor. Toolmakers live in a state of isolation, unaware of the different conditions in which other toolmakers live, or even that there is a difference. Each toolmaker sends off to the others beautiful blueprints of useful tools, and is astounded to find that these come back with all kinds of incomprehensible changes and add-ons. Under this metaphor, communication is a risky business because neither party is aware of the beliefs and assumptions that are informing the crafting and the interpreting of each utterance.

Speech understanding brings yet another twist to this story. We need to add a hard-pressed pony express service that is used for transporting the blueprints. The hearer is handed a document that may be crumpled and have vital portions torn off, but the panting pony rider cannot wait long for a reply. The blueprint must be interpreted quickly and a response scribbled out in haste. The toolmaker watches the cloud of dust disappearing over the horizon with a sense of misgiving about the fate of his botched blueprint and about the likely relevance of the one he will receive in return.

Fortunately the toolmakers seem to have come up with a number of strategies fro increasing the likelihood of successful communication. One such strategy is turn-taking. The toolmaker sends off a part of a blueprint at a time and waits for confirmation that it has been received. If the information is particularly important the receiving toolmaker may make a copy and send it back as confirmation that the blueprint has been correctly received. Compare again, the strategies used in the Operator group and the Computer (NT) group below. In the Operator group, the toolmakers make use of 15 pony-express rides, whereas in the Computer group they use only 7.

Computer (NT) Group

U: ok on the same day {short} I would like my calls forwarded to 3 3 9 {short} 2 3 2 3 {short} until 8 pm that evening {long}
S: please give the day
U: that would be on friday
S: please specify the start {short} and the end {short} of the time interval
U: ok the start would be {short} from noon {short} until 8 pm
S: on friday {short} your calls are forwarded from noon {short} to 8 pm {short} to 3 3 9 {short} 2 3 2 3
U: thank you
{hang up}

Operator Group

U: and then {short} in the afternoon {short} through the whole time period {short} that I have scheduled through 6 o'clock {short} I will be at 3 3 9 {short} 2 3 2 3 {short}
S: ok so is that every afternoon
U: um just friday afternoon
[ ]
S: just friday ok and that was from noon until 6 o'clock {short}
U: right {short}
S: and that's 3 3 9 {short} 2 3 2 3
U: yes {short}
S: ok {short}
U: that's it then
[ ]
S: that's all the changes {short}
U: yes
S: ok so we've updated the schedule {short}
U: thank you {short}
S: byebye
U: um bye
{hang up}

The turn-taking is quite different in the two dialogues. In the Operator group, the Operator received most of the message in the first utterance. The question "...so is that. every afternoon" signals that the rest of the message has been understood. Once the day has been cleared up, the next 7 utterances are quick exchanges designed to check that the

Operator did indeed correctly understand the first message. The final 7 exchanges are designed to get agreement that there are no more messages and that the conversation is over. Most of the exchanges are to do with checking that grounding has taken place and that there is mutual agreement between the two speakers.

In the Computer group, by contrast, the system is mainly concerned with getting basic information. The phrase "please give the day" gives no clue as to whether the rest of the preceding message was understood. There are no quick checks and confirmations as in the Operator case. The only confirmation comes when the computer repeats back the entire message in the last utterance but one. Finally, there is no negotiation about the end of the dialogue, the user simply says *thank you* and hangs up.

(Clark and Schaefer, 1989) define the Grounding criterion as follows:

> "The contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for the current purpose."

Other work (Sperber and Wilson, 1986; Traum, 1991), has questioned the need for such a strong concept as mutual belief. It would seem that in both the human-human and the human-computer dialogues, grounding was sufficient for the current purpose: all subjects felt that they had accomplished their tasks. The intent to ground is clearly apparent in the Operator group. In the Computer group, we hypothesize that it is only the fact that the computer is unable to give and receive signals about turn-taking that prevents the speaker from making the frequent and efficient checks on grounding that are so prevalent in human-human conversations.

11. Conclusions

(Amalberti et al, 1993) write:

> "Most studies on operator-computer interaction have focused primarily on linguistic differences between human-computer and human-human exchanges. The present study confirms the presence of linguistic differences, but only in Session 1. Differences tend to reduce as subjects gain familiarity with the system (either human or computer)."

Our work also suggests that the differences between human-human and human-computer speech are less pronounced and less permanent than previously thought and depend, rather, on the kind of interaction the system is capable of. We found that the only significant difference on a variety of linguistic indices was in the area of turn-taking and grounding. In the human-computer case, turn-taking was rapid and about 50% of the utterances were to do with grounding. Many were of the form that Clark and Schaefer classify as acknowledgment, though all forms of grounding were present. In the human-computer case, the computer was only able to ground by *display*, i.e. repetition of what the speaker had said. The subjects appeared to be forced into a simple grounding scheme by the system's limitations. Turn-taking was slow and cumbersome, although eventually successful.

11.1 Relevance to WOZ Research

WOZ experiments have often been viewed simply as a way of collecting data such as vocabulary and syntactic forms with a simulated system. (Fraser and Gilbert, 1991) describes such a program. The roots of this technique go back to the incremental simulation work of (Woods et al, 1976) when building the HWIM system. More recent studies that have a human-human control show the usefulness of WOZ experiments for hypothesis testing. In this experiment, we combined both approaches in a way that is very useful for system design. By using a human operator and a constrained computer simulation for the same task, we were better able to focus on the differences between the way in which people communicate with each other and the way they are forced to communicate using a machine. Our first WOZ (Balentine, Berry and Johnstone, 1992), which conformed to (Fraser and Gilbert's, 1991) unconstrained phase I, did not make apparent the amount of turn-taking behavior and grounding that was going on. The second WOZ, with contrasting protocols made it very clear that two different types of behavior were taking place. If we had not run the Operator group side by side with the simulated system and compared the two sets of data, we might not have been so aware of the need for good turn-taking and grounding, since the tasks were, after all, carried out successfully. It would have been easy to assume that that just was the way people talked to computers. The experimental set-up helped us to see that people talked to computers in that manner when computers were limited in their responses.

This seems to be a problem with WOZ experiments in general. One of the conditions given in Fraser and Gilbert is that the system be fully specifiable. This is particularly difficult in the area of speech recognition where we do not understand well exactly which cues are being used and in what manner. It is very hard, first, to be aware of all the cues that the Wizard is using, and second, to make sure the Wizard is ignoring them. We find that the iterative methodology is most useful when it consists of controlled experiments comparing slightly differing systems.

11.2 Relevance to Speech and Dialogue Research

It is often thought that people will never speak to computers the way they would talk to a human while engaged on a similar task. (Amalberti et al, 1993) have challenged this idea as far as linguistic behavior is concerned finding no difference in subject behavior when interlocutor styles do not differ. However, even they conclude that there will always be a difference in problem-solving behavior. They write:

> "The Computer Group tended to solve the problem on their own and not to use linguistic devices to facilitate understanding on the part of the computer....Overall, these differences in problem-solving suggest that human-computer interaction cannot be successfully analyzed in terms of human-human dialogues since the collective dimension of the task is lacking in the human-computer situation."

While this may turn out to be the case, we feel that spoken interaction with machines is so recent and so primitive that such conclusions are somewhat premature. Our findings on turn-taking and grounding are particularly pertinent to these issues since grounding has often been defined in terms of mutual belief.

Recent work on human-human communication has led to a weakened notion of mutual belief and understanding. Understanding is seen as something that only has to be "good enough" for the task in hand. It may be more a matter of making manifest, through language and action, some correspondence between the participants representations of the world. If this is the case, then it is very likely that people engaged in a task will use the mechanisms of language to reassure themselves that the system has correctly interpreted what they have said and is carrying out the actions they have requested. One such mechanism is the simple turn-taking strategies noted in our study. These ideas suggest two interesting areas of research to explore further:

1) Experiments with a simulated system capable of more sophisticated turn-taking and grounding behavior.

2) Experiments using a task where grounding is more critical to the subject, e.g. involving dependent sub-goals.

Such experiments will be useful in designing automated dialogue systems. They may also provide a controlled way of examining the kinds of mechanisms people employ in limited collaborative tasks, and thus provide more data on the complex process of establishing mutual belief and understanding in spoken interactions.

APPENDIX A.1

Detectables

Note: The detectables are given in bold-face and the word lists following them are their synonyms, which were considered to be the same as the detectable.

01) **HELP**

confused, oops, directions, how, what, information

02) **DIAL**

call

03) **SPEED-DIAL**

speed-call, speed-dialing, speed-calling, directory, name

04) **CALL-FORWARD**

call-forwarding, forwarded, transfer, reach, reached, receiving

05) **CALL-WAITING**

waiting

05a) **CALL-FORWARD-SCHEDULING**

06) **INTERRUPTED**

interrupt

07) **WANT**

like, need

08) **YES**

yeah

09) **NO**

nope

10) **FROM**

11) **TO**

12) **CAN**

could

13) **DON'T**

do not

14) **WHO**

15) **OKAY**

alright, sure, correct, good, right, fine

16) **CANCEL**

cancelled, stop, stopped, off

17) **PLEASE**

18) **THANKS**

thank you

19) **CHANGE**

changed, modify, adjust

20) **NOT**

21) **REPEAT**

again

22) **ALL**

23) **NONE**

24) **ADD**

put, include

25) **DELETE**

remove, erase, drop

26) **SET**

27) **FEATURES**

feature, services, service

28) **ENTRIES**

entry

29) **BYE**

good-bye, bye-bye

30) **MENU**

prompt

31) **REVIEW**

show, hear, list

32) **MIKE**

Michael

33) **POPPLESTEIN**

34) **POOPSIE**

35) **JOE'S-AUTO-BODY**
Joe's, Auto-Body
36) **JANE**
37) **727-7507**
38) **225-1320**
39) **776-1234**
40) **356-4930**
41) **339-2323**
42) **MONDAY**
43) **FRIDAY**
44) **0**
45) **1**
46) **2**
47) **3**
48) **4**
49) **5**
50) **6**
51) **7**
52) **8**
53) **9**
54) **7:30**
half-past-seven
55) **12**
56) **NOON**
57) **A.M.**
morning, day
58) **P.M.**
afternoon, evening, night
59) **babble**
60) **silence**
61) **HANGUP**

APPENDIX A.2

Responses

General

01) Hello, this is Southwestern Bell's voice phone service. Can I help you?

02) Do you need help?

03) Please say what service you want.

04) Please rephrase the command.

05) Is there anything else?

06) Thank you.

07) Thank you for participating..... hang up now.

08) I'm sorry, let me get someone who can help you. Please wait.

09) Yes.

10) No.

Help

11) Please say what feature you want help with.

12) The features available are Call-Forwarding, Speed-Dialing, Call-Waiting and Call-Forward Scheduling.

13) With Call-Forwarding, your phone automatically forwards all calls to a location you choose.

14) With Speed-Dialing, you can reach important numbers more quickly and easily. You store a person's name and phone number under a one or two digit code. Then you can dial using either the person's name or their code instead of their phone number.

15) You can add a new entry to your speed-dial directory by giving the name of the person followed by their phone number.

16) You can delete an entry from your speed-dial directory by giving the name of the person or the phone number.

17) You can change an entry in your speed-dial directory by first specifying which entry you want to change and then saying what the change is.

20) You can dial by speaking a phone number, a speed dial code, or a speed dial name.

18) Cancelling Call-Waiting stops important calls from being interrupted. Before you dial the call in which you don't want to be interrupted, you can ask for Call-Waiting to be cancelled.

19) With Call-Forward Scheduling you can forward your calls to different locations at different times. Call forwarding is activated for the following week only. Give the day, the time-interval, and the telephone number where you want your calls forwarded.

Speed-Dial

20) Please say what operation you want to perform on the Speed-Dial directory.

21) Please select from the following features: Dial by name or number, add, delete, change entry or review Speed-Dial directory.

22) Please give the name and telephone number of the person you want to add.

23) Please give the name for the telephone number you want to add.

24) Please give the telephone number for the name you want to add.

25) Now adding Mike at 727-7507.

26) Now adding Joe's Auto Body at 225-1320.

27) Please give the name or telephone number you want to dial.

28) Now dialling Mike.

29) Please give the name or telephone number of the entry you want to delete.

30) Now deleting the entry for Jane.

31) Now deleting the entry for Popplestein.

32) Please give the name or telephone number you want to change.

33) What do you want to change it to?

34) Now changing Popplestein to Poopsie.

35) Stored entries are: Popplestein at 756-8743, Jane at 675-0786.

36) Are these all the changes you wanted to make?

37) Specify the rest of the changes now.

38) Stored entries are: Poopsie at 456-8743, Joes's Auto Body at 225-1320.

Call-Waiting

39) Which service do you want off?

40) Call-Waiting is now off.

41) Call-Waiting is now on.

Call-Forward Scheduling

42) The Call-Forward Schedule is: Monday to Friday from 7:30 a.m. to 6:00 p.m. the calls are not forwarded. At all other times the calls are forwarded to 224-0099.

43) Please give the day, the time-interval and the telephone number where you want your calls forwarded.

44) Please give the time-interval and the telephone number.

45) Please give the day and the telephone number.

46) Please give the day and the time-interval.

47) Please give the day.

48) Please give the time interval.

49) Please give the telephone number.

50) Please specify a.m. or p.m. when you give the time interval.

51) Please specify the start and the end of the time interval.

52) On Monday your calls are forwarded from 2:00 p.m. to 4:00 p.m. to 776-1234.

53) On Friday your calls are forwarded from 7:30 a.m. to noon to 356-4930.

54) On Friday your calls are forwarded from noon to 8:00 p.m. to 339-2323.

55) Are these all the changes you wanted to make for the next week?

56) Specify the rest of your changes now.

57) The Call-Forward Schedule is: Monday, 7:30 a.m. to 2:00 p.m. calls are not forwarded, 2:00 p.m. to 4:00 p.m. calls are forwarded to 776-1234, 4:00 p.m. to 6:00 p.m. calls are not forwarded. Tuesday through Thursday from 7:30 a.m. to 6:00 p.m. the calls are not forwarded. Friday, calls are forwarded from 7:30 a.m. to noon to 356-4930, from noon to 8:00 p.m. to 339-2323. At all other times the calls are forwarded to 224-0099.

APPENDIX B

Sample Dialogue

S: hello this is Southwestern Bell's voice phone service can I help you {short}

U: yes, I would like to know how to add a number to my speed call directory {long}

S: you can add a new entry to your speed-dial directory by giving the name of the person followed by their phone number {short}

U: oh ok, add mike at 7 2 7 {short} 7 5 0 7 {short}

S: now adding mike at 7 2 7 {short} 7 5 0 7 {short}

U: thank you {long}

S: thanks for using Southwestern Bell's voice phone service

{hang up}

APPENDIX C.1

Demographic Information for subjects from the Operator Group

Subject	Age	Occupation	Education*	Sex	Language**
1	55	Housewife	B	F	E
2	37	Homemaker	HS	F	E, S
3	36	Flight Attendant	B	F	E, F
4	32	Salesman	HS	M	E
5	38	Optometrist	PG	F	E, F
6	64	Pipe Fitter	HS	M	E
7	62	Professor	D	M	E
8	34	Manufacturer's Rep	B	M	E
9	65	Retired	HS	M	E
10	41	Sales Rep	B	M	E, F
11	67	Asphalt Worker	----	M	----
12	29	Waitress	HS	F	E

* HS: High School, B: Bachelor's, PG: Post Graduate, D: Doctorate

** E: English, S: Spanish, F: French

APPENDIX C.2

Demographic Information for subjects from the Computer (T) Group

Subject	Age	Occupation	Education*	Sex	Language**
1	48	Self Employed	HS	M	E
2	49	Quality Control	HS	M	E
3	50	Housewife	HS	F	E
4	68	Retired	B	M	E
5	47	Homemaker	B	F	E
6	42	Secretary	B	F	E
7	53	Building Developer	B	M	E
8	43	Dance Teacher	B	F	E
9	65	Retired	HS	F	E
10	47	Electronics	HS	M	E
11	42	Child Advocate	B	F	E
12	42	Supervisor	HS	M	E

* HS: High School, B: Bachelor's

** E: English

APPENDIX C.3

Demographic Information for subjects from the Computer (NT) Group

Subject	Age	Occupation	Education*	Sex	Language**
1	58	Sales Rep	B	M	E, G
2	47	Social Worker	M	F	E, T
3	65	Receptionist	HS	F	E
4	53	Homemaker	M	F	E
5	57	Retired	B	F	E
6	49	Teacher	B	F	E
7	44	Bookkeeper	B	F	E, H
8	55	Accountant	B	M	E
9	44	Sales	B	M	E
10	48	Business Owner	M	F	E, F, R
11	---	---	---	--	---

* HS: High School, B: Bachelor's, M: Master's

** E: English, G: German, T: Tagalog, H: Hebrew, F: French, R: Russian

APPENDIX D

Scenarios

Scenario 1

On Saturday night, you bumped into Mike, an old friend from school. Now that you know his phone number, you plan on calling him often. You would like to add Mike to your Speed Call Directory; his telephone number is 727-7507. After you have an entry for Mike, you want to dial him by name. Using the telephone service, complete the task.

Scenario 2

You plan to call your mother long distance and do not want to waste time and money by answering other calls while you are talking to her. You are unsure how to do this so you are going to ask for assistance and after you get the desired information you will go ahead and carry out the task. Using the telephone service, complete the task.

Scenario 3

It is Sunday. You have just set up your weekly call forwarding schedule so that your calls will be forwarded from your home to your office. Your calls will be forwarded from your home number to your office from 7:30 a.m. to 6:00 p.m., Monday through Friday. At all other times you will receive your calls at your home number (224-0099).

However, on Monday you have a meeting with your colleague from 2:00 p.m. to 4:00 p.m. You want your calls to reach there (at 776-1234) during that time. On Friday you will not be coming to office because you have to meet someone at the Sheraton Hotel (at 356-4930). You will be there from 7:30 to noon. After you leave the Sheraton, you will be at your grandmother's house (339-2323) till 8:00 p.m. Friday.

Using the telephone service, complete the task.

Scenario 4

It has been a long time since you have updated your speed call directory and you want to know who is in your directory. Make the following changes to your Speed Calling Directory:

--You and Mrs. Popplestein have become better acquainted. You want the speed dial directory to now contain "Poopsie" instead of "Popplestein".

--Your brother works at "Joe's Auto Body" at 225-1320, you want this number to be in your directory

--Your neighbor Jane has moved to Miami so you no longer want the entry "Jane" in your speed call directory.

Using the telephone service, complete the task.

APPENDIX E.1

Questionnaires for the Operator Group

Questionnaire A

1. "Overall, this task was ... "

|-----|-----|-----|-----|-----|-----|-----|-----|

1	2	3	4	5	6	7	8	9
very difficult		difficult		average		easy		very easy

2. What was the hardest part of this task?

3. "I was able to accomplish the task ..."

|-----|-----|-----|-----|-----|-----|-----|-----|

1	2	3	4	5	6	7	8	9
very slowly		slowly		average		quickly		very quickly

4. What took the most time?

Turn the page and read the next scenario. When you understand your task, dial 7695. After you hang up, turn the page and complete the following questionnaire.

Questionnaire B

1. The responses you heard were: (check all that apply)

o enthusiastic	o clear	o too fast
o friendly	o garbled	o too slow
o neutral	o too quiet	o boring
o too brief	o too loud	o pleasant
o too wordy		

2. What did you like about the service?

3. What did you dislike about the service?

Questionnaire C

NOTICE: Any information you offer below will be held confidential. This information will only be used for purposes of statistical analysis.

Name: ______________________________
(Last) (First) (M.I.)

Sex: _______ Birth Date: ___/___/___
(m) (d) (y)

Occupation: ____________________________

Highest level of education attained: ________________________

Place(s) lived for at least six months from ages 0 to 18:

example: 0-5 Austin, Texas, USA
5-13 Sydney, Australia

____

____

____

Place(s) you lived for at least six months during the past three years:

____

____

____

Languages spoken, and the approximate ages during which you spoke them:

example: 0-4 German
4-present English

____

____

Do you presently have any condition such as a cold, sore throat etc., which might make your speech different than usual?

○ yes ○ no

If yes, briefly describe it:

____

APPENDIX E.2

Questionnaires for the Computer (T) Group

Questionnaire A

1. "Overall, this task was ... "

```
|-----|-----|-----|-----|-----|-----|-----|-----|
1     2     3     4     5     6     7     8     9
very       difficult    okay       easy       very
difficult                                     easy
```

2. What was the hardest part of this task?

3. "The voice phone let me accomplish the task ..."

```
|-----|-----|-----|-----|-----|-----|-----|-----|
1     2     3     4     5     6     7     8     9
very       slowly      okay      quickly      very
slowly                                      quickly
```

4. What took the most time?

Turn the page and read the next scenario. When you understand your task, dial 7695. After you hang up, turn the page and complete the following questionnaire.

Questionnaire B

1. "Speaking to the system was ..."

```
|-----|-----|-----|-----|-----|-----|-----|-----|
1     2     3     4     5     6     7     8     9
very                   fairly                very
unnatural              natural               natural
```

2. How helpful was the sample dialogue that you heard at the beginning of the experiment?

```
|-----|-----|-----|-----|-----|-----|-----|-----|
1     2     3     4     5     6     7     8     9
not                    fairly                very
helpful                helpful               helpful
```

3. The responses you heard were: (check all that apply)

o enthusiastic	o clear	o too fast
o friendly	o garbled	o too slow
o neutral	o too quiet	o boring
o too brief	o too loud	o pleasant
o too wordy		

4. What did you like about the system?

5. What did you dislike about the system?

6. Did you feel at any time that the system replied too quickly or slowly? If so, what was your reaction?

7. Do you think that a human operator would be faster?

8. Did you feel unnatural talking to the system? If yes, what was the thing that felt most unnatural?

9. Would it have been more helpful to hear more sample dialogues?

10. Overall, what was your strongest impression of the system?

Questionnaire C

NOTICE: Any information you offer below will be held confidential. This information will only be used for purposes of statistical analysis.

Name: ____________________________________
(Last) (First) (M.I.)

Sex: _______ Birth Date: ___/___/___
(m) (d) (y)

Occupation: ________________________________

Highest level of education attained: __________________________

Place(s) lived for at least six months from ages 0 to 18:

example: 0-5 Austin, Texas, USA
5-13 Sydney, Australia

____

____

____

Place(s) you lived for at least six months during the past three years:

____

____

____

Languages spoken, and the approximate ages during which you spoke them:

example: 0-4 German
4-present English

____

____

Do you presently have any condition such as a cold, sore throat etc., which might make your speech different than usual?

○ yes ○ no

If yes, briefly describe it:

____

APPENDIX E.3

Questionnaires for the Computer (NT) Group

Questionnaire A

1. "Overall, this task was ... "

```
|-----|-----|-----|-----|-----|-----|-----|-----|
1     2     3     4     5     6     7     8     9
very       difficult      okay        easy       very
difficult                                        easy
```

2. What was the hardest part of this task?

3. "The voice phone let me accomplish the task ..."

```
|-----|-----|-----|-----|-----|-----|-----|-----|
1     2     3     4     5     6     7     8     9
very       slowly      okay        quickly       very
slowly                                         quickly
```

4. What took the most time?

Turn the page and read the next scenario. When you understand your task, dial 7695. After you hang up, turn the page and complete the following questionnaire.

Questionnaire B

1. "Speaking to the system was ..."

|-----|-----|-----|-----|-----|-----|-----|-----|
1 2 3 4 5 6 7 8 9

very unnatural — fairly natural — very natural

1	2	3	4	5	6	7	8	9
very unnatural				fairly natural				very natural

2. The responses you heard were: (check all that apply)

o enthusiastic
o friendly
o neutral
o too brief
o too wordy

o clear
o garbled
o too quiet
o too loud

o too fast
o too slow
o boring
o pleasant

3. What did you like about the system?

4. What did you dislike about the system?

5. Did you feel at any time that the system replied too quickly or slowly? If so, what was your reaction?

6. Do you think that a human operator would be faster?

7. Did you feel unnatural talking to the system? If yes, what was the thing that felt most unnatural?

8. Overall, what was your strongest impression of the system?

Questionnaire C

NOTICE: Any information you offer below will be held confidential. This information will only be used for purposes of statistical analysis.

Name: ______________________________________
(Last) (First) (M.I.)

Sex: _______ Birth Date: ___/___/___
(m) (d) (y)

Occupation: _________________________________

Highest level of education attained: ___________________________

Place(s) lived for at least six months from ages 0 to 18:

example: 0-5 Austin, Texas, USA
5-13 Sydney, Australia

____

____

____

Place(s) you lived for at least six months during the past three years:

____

____

____

Languages spoken, and the approximate ages during which you spoke them:

example: 0-4 German
4-present English

____

____

Do you presently have any condition such as a cold, sore throat etc., which might make your speech different than usual?

○ yes ○ no

If yes, briefly describe it:

____

APPENDIX F.1

Instructions for the Operator Group

Thank you for volunteering to use our new Custom Calling Features (CCF) service. CCF are enhancements to your telephone service which make receiving and managing your calls easier. The features that are available are:

- Call Forwarding - this service forwards your calls from your home number to another number.
- Remote Call Forwarding - this allows you to forward your calls from your home number to another number, but you can activate the service from any telephone.
- Call Forward Scheduling - this allows you to forward your calls, but you can specify the day and time of day you want your calls forwarded.
- Call Waiting -- this allows you to receive calls while talking on the phone with someone else. You can turn this service on and off before making a call.
- Speed Calling - this allows you to dial a person by their name. You set up a list of names that the service calls when you say the name. For example, "Mom", or "Office".

You will be talking to an operator in order to accomplish tasks related to the CCF. If you need more specific information about how to use these services the operator can provide help.

The following pages contain four scenarios, each describing a CCF related task. There will be a questionnaire after each scenario which you should complete before you go on to the next scenario.

Turn to the next page for the first scenario. When you have read it and decided what action you will take, dial 7695 to talk to the operator.

APPENDIX F.2

Instructions for the Computer (T) Group

Thank you for volunteering to use our new "voice phone" Custom Calling Features (CCF) service. CCF are enhancements to your telephone service which make receiving and managing your calls easier. The features that are available are:

- Call Forwarding - this service forwards your calls from your home number to another number.
- Remote Call Forwarding - this allows you to forward your calls from your home number to another number, but you can activate the service from any telephone.
- Call Forward Scheduling - this allows you to forward your calls, but you can specify the day and time of day you want your calls forwarded.
- Call Waiting - this allows you to receive calls while talking on the phone with someone else. You can turn this service on and off before making a call.
- Speed Calling - this allows you to dial a person by their name. You set up a list of names that the service calls when you say the name. For example, "Mom", or "Office".

You will be talking to a computer system which will allow you to use selected CCF by speaking naturally into the phone. If you need more specific information about how to use these services the system can provide help. It will provide information, acknowledge, and occasionally ask questions to help you accomplish your goal. You may ask the system for help, or correct a statement at any time.

You will now hear a sample dialogue with the system. Please ask the experimenter to play the recording.

The following pages contain four scenarios, each describing a CCF related task. There will be a questionnaire after each scenario which you should complete before you go on to the next scenario.

Turn to the next page for the first scenario. When you have read it and decided what action you will take, dial 7695 to talk to the system.

APPENDIX F.3

Instructions for the Computer (NT) Group

Thank you for volunteering to use our new "voice phone" Custom Calling Features (CCF) service. CCF are enhancements to your telephone service which make receiving and managing your calls easier. The features that are available are:

- Call Forwarding - this service forwards your calls from your home number to another number.
- Remote Call Forwarding - this allows you to forward your calls from your home number to another number, but you can activate the service from any telephone.
- Call Forward Scheduling - this allows you to forward your calls, but you can specify the day and time of day you want your calls forwarded.
- Call Waiting - this allows you to receive calls while talking on the phone with someone else. You can turn this service on and off before making a call.
- Speed Calling - this allows you to dial a person by their name. You set up a list of names that the service calls when you say the name. For example, "Mom", or "Office".

You will be talking to a computer system which will allow you to use selected CCF by speaking naturally into the phone. If you need more specific information about how to use these services the system can provide help. It will provide information, acknowledge, and occasionally ask questions to help you accomplish your goal. You may ask the system for help, or correct a statement at any time.

The following pages contain four scenarios, each describing a CCF related task. There will be a questionnaire after each scenario which you should complete before you go on to the next scenario.

Turn to the next page for the first scenario. When you have read it and decided what action you will take, dial 7695 to talk to the system.

Acknowledgements

We owe a great deal to all the team members working on the DIM system: George Velius, Jeff Scruggs, John Tadlock, Alan Asper, and Sean Dickman, for all the help in conducting the experiment and in the transcriptions. The Computer Science Department at Washington University and the speech lab of Southwestern Bell Technology Resources provided an ideal setting for all this work.

Bibliography

Allen, J. and L. Schubert, "The TRAINS project", *TRAINS Technical Note 91-1, Computer Science Department, University of Rochester*, (1991).

Amalberti, R., N. Carbonell and P. Falzon, "User representations of computer systems in human-computer speech interaction", *International Journal of Man-Machine Studies, 38*, (1993): 547-566.

Baber, C. and R. Stammers, "Is it Natural to Talk to Computers? An Experiment using the Wizard of Oz Technique", *Contemporary Ergonomics Proceedings of the Ergonomics Society's Annual Conference*, (1988): 234-239.

Balentine, J., U. Berry, and A. Johnstone, "The DIM System: WOZ Simulation Results - Phase I", *Technical Report WUCS-TR-92-20, Department of Computer Science, Washington University, St. Louis*, (1992).

Berry, U., and A. Johnstone, "The DIM System: Turn-Taking in Dyadic Telephone Dialogues", *Technical Report WUCS-TR-93-33, Department of Computer Science, Washington University, St. Louis*, (1993).

Clark, H. and E. Schaefer, "Contributing to Discourse", *Cognitive Science*, 13, (1989): 259-294.

Lee, K., "Automatic Speech Recognition: The Development of the SPHINX System", *Kluwer Academic Publishers* (1989).

Fraser, N. and G. Gilbert, "Simulating Speech Systems". *Computer, Speech, and Language*, 5.1, (1991): 81-99.

Green, G., *Pragmatics and Natural Language Understanding*, Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers, (1989).

Johnstone, A., Berry, U., and Nguyen, T., "The DIM System: WOZ II Transcribed Dialogues", *Technical Memorandum WUCS-TM-93-01, Department of Computer Science, Washington University, St. Louis*, (1993b).

Kelley. J., "An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications", *Proceedings from CHI*, (1983): 193-196.

Kennedy, A., A. Wilkes, L. Elder, and W. Murray, "Dialogues with Machines", *Cognition*, 30, (1988): 37-72.

Lambert, L. and S. Carberry, "A Tripartite Plan-based Model of Dialogue", *Proceedings from the Association for Computational Linguistics*, (1991): 47-54.

Levinson, S., *Pragmatics*, Cambridge University Press, (1983).

Litman, D. and J. Allen, "A Plan Recognition Model for Subdialogues in Conversations", *Cognitive Science*, 11.2, (1987): 163-200.

Litman, D. and J. Hirschberg, "Disambiguating cue phrases in text and speech", *Papers presented to the 13th International Conference on Computational Linguistics, Helsinki,* (1990): 251-256.

Morel, M., "Computer-Human Communication", In Taylor, M., F. Neel, and D. Bouwhuis (Eds.), *The Structure of Multimodal Dialogue*, (1989): 323-330.

Reddy, M. J., "The Conduit Metaphor - a case of the Frame Conflict in our Language about Language", In A. Ortony (Ed.), *Metaphor and Thought*, Cambridge University Press, (1979).

Richards, M. and K. Underwood, "How should People and Computers Speak to Each Other?", *INTERACT* 84, (1984): 33-36.

Sperber, D., and D. Wilson, *Relevance: Communication and Cognition*, Harvard University Press, (1986).

Thompson, B., "Linguistic Analysis of Natural Language Communication with Computers", *Proceedings of COLING-80*, Tokyo, Japan, (1980): 190-201.

Traum, D., "Towards a Computational Theory of Grounding in Natural Language Conversation", *Technical Report 401, Computer Science Department, University of Rochester*, (1991).

Young, S., Hauptmann, A., Ward, W., Smith, E., and Werner, P., "High Level Knowledge Sources in Usable Speech Recognition Systems", *Communications of the ACM 32, 2* (1989).

Woods, W., Bates, M., Brown, G., Bruce, B., Cook, C., Klovstad, J., Makhoul, J., Nash-Webber, B., Schwartz, R., Wolf, J., and Zue, V., "Speech Understanding Systems: Final Technical Progress Report", *Bolt Beranek and Newman, Inc. Report No. 3438, Cambridge, MA (in 5 volumes),* (1976).