

# Amount of Information Presented in a Complex List: Effects on User Performance

Dawn Dutton  
AT&T Labs - Research  
180 Park Avenue  
D103, Bldg. 103  
Florham Park, NJ 07932  
+1.973.236.6522  
dldutton@att.com

Selina Chu  
Information and Computer Science  
444 Computer Sci. Bldg.  
UC Irvine  
Irvine, CA 92697  
+1.949.509.9762  
selina@ics.uci.edu

James Hubbell  
12 Lawrence Avenue  
West Long Branch, NJ 07932  
+1.732.728.9351  
jameshubbell@hotmail.com

Marilyn Walker  
AT&T Labs - Research  
180 Park Avenue  
E103, Bldg. 103  
Florham Park, NJ 07932  
+1.973.360.8956  
walker@research.att.com

Shrikanth Narayanan  
Speech and Image Proc. Institute  
Department of EE-Systems  
EEB 430, USC  
Los Angeles, CA 90089  
+1.213.740.6432  
shri@sipi.usc.edu

## ABSTRACT

AT&T Communicator is a state-of-the-art speech-enabled telephony-based application that allows the end-user to, among other things, select and reserve airline itineraries. This experiment explores how the amount and structure of information presented in complex lists influences the user experience and the ability of subjects to successfully complete a selection task. Presenting all the relevant information needed for a decision at once was the factor that most positively influenced successful task completion and the user experience. Subjects preferred hearing all of the relevant information about each flight, without initiating additional dialog with the system. Additionally, successful task completion rates improved when all of the flights were presented at once, without any intervening questions from the system.

## Keywords

dialogue, user interface, human factors, spoken language, lists, selection

## 1. INTRODUCTION

AT&T Communicator is a state-of-the-art spoken dialogue system that allows the end-user to select and reserve various travel related resources, in particular, airfare, hotel, and rental cars [2]. One of its most challenging dialog modules is that for presenting

information that allows the user to select a desired air travel itinerary. While selecting between multiple possible itineraries in a visual domain is a relatively simple task because most of the selection criteria are listed in a single scrollable page, the same task is likely to have a much higher cognitive load in an audio-only domain [1, 3]. The selection criteria for each candidate flight must be presented to the user serially, leading to higher cognitive demands that could result in errors in itinerary selection.

A sample AT&T Communicator dialog with the prototype used in this experiment, showing the outbound flight only, follows.

- System: Hello, this is AT&T Communicator.  
How may I help with your travel plans?
- User: Yeah, I'd like to fly from Newark to San Francisco.
- System: OK, a flight from Newark, New Jersey to San Francisco, California. And will this be a round trip?
- User: Yes, and I need to arrive in San Francisco by 4 PM on November 22<sup>nd</sup>.
- System: I found five outbound United Airlines flights.  
Option 1 is flight number 635. It has one stop; it departs at 10:15 A.M. and arrives at 2:42 P.M.  
<pause>  
Option 2 is flight number 7535. It has one stop; It departs at 12:20 P.M., and arrives at 4:37 P.M.  
<pause>  
Option 3 is flight number 837. It's a nonstop flight; it departs at 2:00 P.M. and arrives at 5:01 P.M. <pause>  
Would you like to hold any of these flights?

User: I guess I'll take option one.

System: Got it.

Note that the second to the last system utterance, in which a set of itineraries are presented to the user for selection, is by far the longest system utterance in the dialog. This utterance includes a complex list of information. Unlike a more simple list, each list item (option) contains a number of sub-items, for example flight number, number of stops, etc. In early dialog evaluations of the Communicator system, the selection of a single itinerary from such a complex list of flights was a source of substantial usability problems; users frequently became frustrated. In fact, many users dropped out of the dialog before successfully selecting an itinerary.

This paper describes an experiment in which we vary the amount and structure of information presented about available itineraries. We predict that the amount and structure of information presented affects the ability of users to successfully select the optimal itinerary within a set, and influences subjective measures such as user satisfaction.

## 2. METHODS AND PROCEDURES

### 2.1 Subjects

Sixty-four subjects were run at a local shopping mall over a five day period. Subjects were recruited from the shoppers frequenting the mall.

### 2.2 Wizard of Oz

A Wizard of Oz (WOZ) experiment was run to determine the optimal way for the end-user to select a desired itinerary in the Communicator project.

A Wizard of Oz experiment is one in which no real automatic speech recognition (ASR) or natural language understanding (NLU) is used. Instead, the user interface is prototyped and a 'wizard,' or experimenter, acts in place of the ASR and NLU. Consequently, subjects believe that ASR/NLU is being used. The WOZ methodology allows competing user interface strategies to be prototyped and tested with end users in a shorter period of time than would be required to implement multiple fully-functioning systems with competing user interfaces.

### 2.3 Apparatus & Materials

Relevant aspects of the AT&T Communicator user interface were prototyped using the Unisys Natural Language Speech Assistant (NLSA) software. NLSA runs on a PC using the Windows NT operating system. Subjects called into the Communicator prototype using an analog telephone and interacted with the system by voice. The wizard categorized the subject's speech using the NLSA Wizard graphical user interface (GUI). Each subject completed 5 surveys in pen and paper format. During the course of the experiment, subjects also had access to a pad of paper.

### 2.4 Experimental Design

All itineraries presented to the subjects were round-trip.

#### 2.4.1 Independent Variables

This was a factorial experiment with two factors, one factor between subjects and the other within subject (see Table 1).

**Selection Itinerary Content.** There were two levels of this between subjects factor:

--*Terse.* The presented itineraries included: airline, number of stops, and departure time<sup>1</sup>. In order to get additional information, the user could ask the system questions (e.g. "When does that flight arrive?").

--*Verbose.* The presented itineraries included: airline, flight number, number of stops, departure time, and arrival time. All the information relevant to the tasks specified in the experiment are presented about each flight; the user did not need to ask questions to get additional information.

**Number of Flights Before Question.** Each level is actually a combination of two separate, but related, factors.

--*Combined vs. Separate.* Whether outbound and return flights are presented separately or in combination.

--*Number of flights.* The number of flights that are presented before asking the subject to make a decision.

Four levels of this factor were chosen. In all cases (1) the total number of flights 'found' was 5, and, (2) the question was, "Would you like to hold [that flight/any of those flights]?"

--*Separate 1.* The outbound and return flights of the trip are presented separately and after each flight the subject is asked the question.

--*Separate 3.* The outbound and return flights of the trip are presented separately and after the third flight the subject is asked the question.

--*Separate 5.* The outbound and return flights of the trip are presented separately and after the last flight the subject is asked the question.

--*Combined.* The outbound and return flights of the trip are presented at the same time and after each set of two flights the subject is asked the question.

**Table 1: Factors used in this experiment.**

Selection Itinerary Content (Between)			
Terse		Verbose	
Outbound / Return	# of Flights Before Question	Outbound / Return	# of Flights Before Question
(Within)		(Within)	
Separate	1	Separate	1
Separate	3	Separate	3
Separate	5	Separate	5
Combined	2	Combined	2

**Example.** The following example could have been used in the Separate 3 condition. Text that is unformatted is common to both the terse and verbose conditions. Text in *italics is found only in the verbose condition.*

"I found 5 outbound Delta flights. Option 1 is *flight number 323. It's a non-stop leaving at 9:10 and arriving at 2:01.* Option 2 is *flight number 798. It has one stop; it departs at 11:13 and arrives at 5:07.* Option 3 is *flight number 295. It*

<sup>1</sup> All times in this experiment were presented to the minute, with either AM or PM, e.g. "...departs at 6:01 AM."

has two stops; it departs at 1:52 and arrives at 6:57. Would you like to hold any of those flights?"

#### 2.4.2 Mixed Initiative and Flow of Control

The dialog strategy was mixed initiative. The first prompt was open-ended, e.g. "How may I help you with your travel plans?" All subsequent prompts requested specific information from the user (e.g. "What date did you want to depart?") The prototypes were built to allow the user to provide multiple informational elements (e.g. departure city and departure date) to either open-ended or specific requests. Subsequent steps in the flow of control could be skipped if multiple pieces of information were presented at a single dialog point.

#### 2.4.3 Tasks

Each subject was asked to complete four tasks in the course of this experiment. In each task the subject was given a set of criteria that the subject had to meet in selecting both an outbound and a return flight. The tasks used in this experiment exercise selection criteria that are representative of selection criteria typically used by individuals actually purchasing airline tickets. The four tasks given to subjects follow:

**Departure Only.** The task criteria for both the outbound and return flights require the subject to choose flights based on departure time only.

**Arrival Only.** The task criteria for both the outbound and return flights require the subject to choose flights based on arrival time only.

**Departure & Arrival.** The task criteria require the subject to choose the outbound flight based on departure time and the return flight based on arrival time.

**Specific Flight.** The task requires the subject to book a particular flight for both the outbound and return flights.

**Example.** The following example was used for the Departure & Arrival task (it has been edited for presentation here).

*You want a round trip ticket from Boston to Charleston.  
You want to leave Boston about 5 in the evening of  
Friday November 10th. You want to arrive in Boston no  
later than 8 PM on Tuesday November 14th.*

An important selection criterion for many purchasers of airline tickets is price. The price of the ticket was not a selection criterion used in this experiment because it would introduce possible confounds. Many users are willing to trade-off other important selection criteria, e.g. arrival time and departure time, in order to minimize price. Therefore, it was decided, *a priori*, to postpone the use of price as a selection criterion to a later experiment.

#### 2.4.4 Counterbalancing

A Balanced Greco-Latin Square was used to counterbalance the orders of the conditions and tasks.

#### 2.4.5 Dependent Measures

A rich set of dependent measures were gathered in this experiment:

- After each system prompt was played, NLSA recorded what subjects said.
- At the end of each task, the wizard determined whether that task was successfully completed.

- At the end of each task, subjects completed paper and pen surveys rating the overall dialog for that task.
- After experiencing all four tasks, subjects told the experimenter which of the flight selection criteria were important to them.

**Objective measure.** Successful task completion was the one objective measure used in determining the optimal method for presenting complex lists in an audio-only domain. For each task the subject was given a set of required criteria for selecting both the outbound and a return flight. Task completion was binary, successful or unsuccessful, and was determined by the experimenter (wizard) at the time the subject completed each task. In order for a subject to successfully complete a task, the subject had to select both the outbound and return flight that best fit the clear criteria given to subjects in the task description.

**Subjective measures.** Other data gathered in this experiment included a number of subjective measures. After each task, subjects were asked:

*Overall, how satisfied were you with AT&T Communicator while booking this flight?*

- [1] Very Satisfied
- [2] Somewhat Satisfied
- [3] Neither Satisfied, Nor Dissatisfied
- [4] Somewhat Dissatisfied
- [5] Very Dissatisfied

*Overall, how easy was it to use AT&T Communicator while booking this flight?*

- [1] Very Easy
- [2] Somewhat Easy
- [3] Neither Easy, Nor Difficult
- [4] Somewhat Difficult
- [5] Very Difficult

*Overall, how quickly did AT&T Communicator respond to your speech?*

- [1] Much Too Fast
- [2] A Little Too Fast
- [3] Just the Right Speed
- [4] A Little Too Slow
- [5] Much Too Slow

*After you told Communicator the date and time to book your flight, Communicator responded with possible flights to choose from. For EACH of the possible flights, did Communicator present the right amount of information?*

- [1] Too Much Information about Each Flight
- [2] Just the Right Amount of Information About Each Flight
- [3] Too Little Information about Each Flight

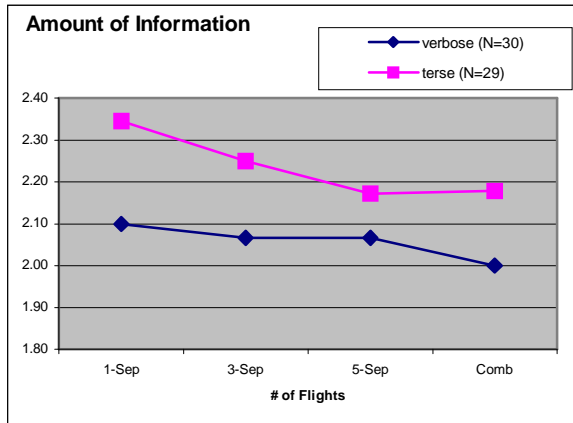
After completing all four tasks, subjects were asked to (1) rank order the criteria they personally use when selecting between multiple itineraries, and (2) specify the information that Communicator should present about every flight for selection purposes in the future.

## 3. RESULTS AND CONCLUSIONS

### 3.1 Terse or Verbose?

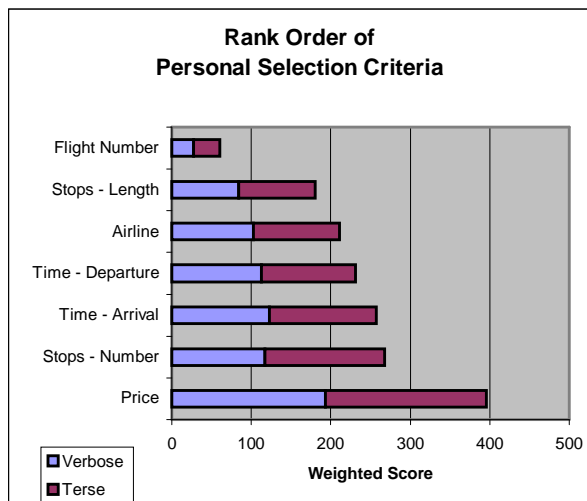
A two-way, 2x4, Analysis of Variance (ANOVA) was run for each of 5 dependent measures: successful task completion, amount of information presented about each flight, satisfaction,

ease of use, and speed of interaction. For each dependent measure, no significant interactions were found<sup>2</sup>. A significant main effect for Terse/Verbose was found for the subjective measure of the amount of information presented about each flight ( $p=.001$ ), see Fig. 1.



**Figure 1: Verbose and Terse subject ratings to the Amount of Information question (2=Just the Right Amount of Information about each flight).**

No other significant main effects were found for any of the dependent measures. The optimum value for the dependent measure amount of information is '2' (Just the right amount of information about each flight). The average value for the Verbose condition (across the 4 levels of # of Flights) was 2.06, while the equivalent average for the Terse condition was 2.24.



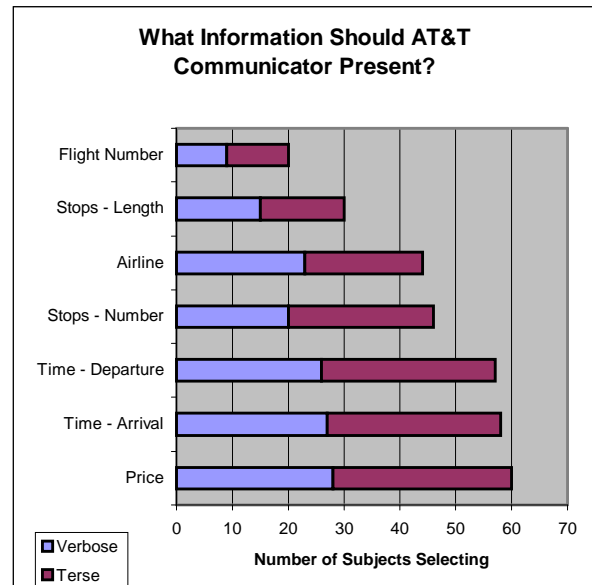
**Figure 2: Summed Weighted Scores of subjects' rank ordering of their personal selection criteria.**

Related to these results is a question that was asked of all subjects at the end of the experiment. Figure 2 shows the weighted scores based on the rank ordering of the selection

<sup>2</sup> Throughout the experiment, the alpha level used to determine significance of an effect was  $p<.05$ .

criteria subjects personally use when selecting among multiple flights. A rank order of 1 was given a score 7 points, a rank order of 7 was given a score of 1 point, etc. The Weighted Score for each selection criteria shown in Figure 2 is the sum of the Weighted Scores for all subjects.

Similarly, a second question was asked of all subjects at the end of the experiment: "In the future, what information should AT&T Communicator present about each flight when you are choosing between multiple flights?" Figure 3 shows the compiled responses to this question.



**Figure 3: Number of subjects indicating that each selection criterion should, by default, be presented by AT&T Communicator.**

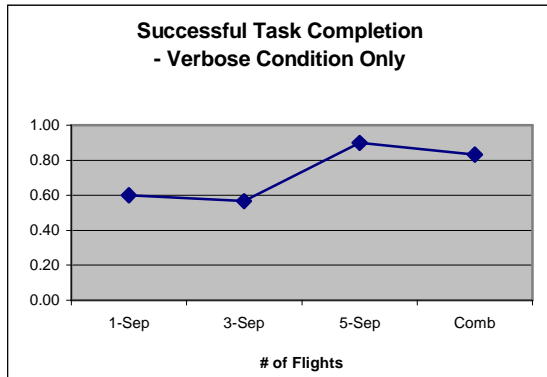
Information that should definitely be presented to subjects when selecting between multiple flights includes: price, arrival time, departure time, number of stops and airline. The value to users of the length of stops is ambiguous. It probably should not be presented by default, although it might be useful to present the length of stops if they will be inordinately long, e.g. greater than 2 hours, or inordinately short, e.g. less than 45 minutes. Flight number was judged to be least valuable and should not be presented.

### 3.2 Number of Flights?

The above analyses indicate that the amount of information presented in the Verbose condition better met the expectations of subjects. The next question then was, *within* the verbose condition, which level of the number of flights before the question factor showed the best performance. A one-way, 1x4, ANOVA was run for the verbose condition for each of five dependent measures: successful task completion, amount of information about each flight, satisfaction, ease of use, and speed of interaction<sup>3</sup>. A significant main effect was found for successful

<sup>3</sup> As noted at the beginning of the Results section, subject responses to the satisfaction, ease of use, and speed of the interaction questions may be attributable to the subject's reactions to the novel user-system

task completion ( $p=.005$ ). Figure 4 shows the percentage of successful task completions in the Verbose condition only. No significant effects were found for the other four dependent measures.



**Figure 4: Successful task completion (in percent) across the four levels of the # of Flights Before Question condition (Verbose only).**

The significant main effect was probed using the Tukey test<sup>4</sup>. Separate 5 was the condition with the highest successful task completion rate. Only one pairwise comparison was significant ( $p<.05$ ). Tasks attempted in the Separate 5 condition were significantly more likely to be completed successfully than tasks attempted in the Separate 3 condition.

**Table 2: Results of Tukey test for successful task completion across the four levels of # of Flights Before Question condition (Verbose only).**

	Sep. 3 =.567	Sep. 1 =.600	Comb. =.833	Sep. 5 =.900
Sep. 3 =.567		$p>=.05$	$p>=.05$	$p<.05^*$
Sep. 1 =.600			$p>=.05$	$p>=.05$
Comb. =.833				$p>=.05$
Sep. 5 =.900				

Among the three Separate conditions (Separate 1, Separate 3, and Separate 5), subjects were much more likely to successfully complete a task in Separate 5. That is, when all the flights for a given flight (outbound or return) were presented at once, without any intervening questions. Also, based on subject comments, it appeared that at least some subjects in the Separate 3 condition were confused about the number of flights they had available to select between. These subjects didn't realize that there were more flights available after the system presented them with the first

interaction style, rather than to the experimentally varied presentation of the flight selection criteria.

<sup>4</sup> The Tukey is a test of significance for pairwise comparisons of treatment means that controls for familywise error.

three in a total set of five flights. This is in spite of the fact that in all tasks, including the Separate 3 condition, the subjects heard a sentence like "I found five outbound Northwest Airlines flights," before the options were presented for selection.

It not possible, on the basis of the experimental data gathered in this study, to unambiguously choose one of the # of flights before question conditions over the others. It may be that a more difficult set of tasks would elicit stronger differences in both the objective and subjective measures for the levels of this factor. However, in absolute terms, the task completion rates with Separate 5 and Combined were both high (90% and 83%, respectively), relative to the Separate 1 and Separate 3 conditions (60% and 57%, respectively).

Anecdotal evidence sheds some additional light on the issue of which condition (Separate 5 or Combined) is preferred by subjects. In the Verbose condition, the last 17 subjects run in the experiment were asked a few questions that provide evidence concerning their subjective impressions of the four levels of the number of flights before question factor. The first question was "Did you notice any difference between the different versions of the system?" Twelve of seventeen subjects stated that they had noticed a difference between the four versions. Those 12 subjects were then asked to choose the version they liked the best, and then the version they considered to be the worst.

**Best?**

- 3-Did not specify a 'best' version
- 2-Combined only
- 2-Separate 1 only
- 1-Separate 3 only
- 1-Separate 5 only
- 2-Any of the Separate versions
- 1-Either Separate 3 or Separate 5

**Worst?**

- 7-Did not specify a 'worst' version
- 5-Combined only
- 0-Separate 1 only
- 0-Separate 3 only
- 0-Separate 5 only

In response to the question of which version of the system was best, the subjects stated no consistent preference for any of the versions of the system. On the other hand, the responses to the question concerning which version of the system was 'worst' resulted in a more consistent set of responses; the Combined version was selected by 5 of 12 of the subjects as the version they considered to be the 'worst.' From subject comments, it appeared that subjects didn't like it when they heard one flight that matched their constraints (e.g. outbound), while the other flight did not match their constraints (e.g. return). Some subjects found this to be frustrating, confusing, and/or tedious.

**4. DISCUSSION**

Presenting all the relevant information about a given flight *at once* seemed to be the single overarching factor that most positively influenced successful task completion and the user experience. Subjects wanted to hear *all of the relevant information* about a flight needed to make the best choice.

Within the Separate conditions (Separate 1, Separate 3 and Separate 5), the task completion rate was highest for the Separate 5 condition. That is, when *all of the flights* were presented at

once, without any intervening system questions. The Separate 5 and Combined conditions had similar task completion rates and were not significantly different. However, the Combined condition was the only condition considered 'worst' by subjects. Thus, the condition that maximized both successful task completion and user experience was the Verbose Separate 5 condition.

A major concern in the design of this experiment was that the audio presentation of lists of complex information, in this case lists of multiple airline flights each containing multiple pieces of information, would result in cognitive overload. These findings argue that, for this task, our concern about the increased cognitive load in an audio-only domain was unfounded. There are a couple of possible reasons that cognitive load did not appear to have the influence that we anticipated. First, users knew their constraints when listening to the flights. Some users may have used a strategy of only remembering the options that potentially matched their constraints. Second, many subjects apparently dealt with the increased cognitive load by taking notes, with flight times, etc., while completing the experimental tasks. Such behaviour is certainly common when making airline reservations with a human travel agent over the telephone. Further, it is anticipated that users would generally book flights in a hands-free environment, and would be able to take notes, when it is desirable. Although less ecologically valid, future experiments primarily concerned with the issue of cognitive load should explicitly preclude subjects from taking notes.

From the above data, it appears that the standard user interface practices for Prompt & Collect (in which the system 'prompts' the user for a specific piece of information and then 'collects' their response) should be different than the standard practice for Selecting from a List. For Prompt & Collect, the standard UI practice is for the system to be as terse as possible when prompting the user. The user generally knows the information the system is asking for, they just need to be prompted so that they know what piece of information the system is currently ready to accept. When Selecting from a List, however, the level of knowledge of the system and user are very different. The system knows the possible items that can be selected. The user begins the interaction only knowing their own constraints (and their relative priority). The system must present all the relevant information so that the user can select the optimal item from the list based on their constraints and priorities. Consequently, when the user must select an item from a list, the system should be as verbose as necessary and present all the information relevant to their decision at once.

For the task of selecting between multiple airline flights, the following information should definitely be presented to users about each flight: price, arrival time, departure time, number of stops and airline. While there was agreement between subjects on the important selection criteria for flights, there are also some individual differences. For example, some users like to fly only a particular airline or have strong preferences concerning the

locations of layovers. This experiment indicates that users like to hear all of the information relevant to their decision at once. Therefore, if a user asks a question concerning information that is not by default presented by the system about each flight, the information presented to that user for each subsequent flight on that call, should include the information they have requested. For example, if after the first flight the user asks "Where does that flight stop?", the information for all subsequent flights in that call should include the location of any stops.

It thus appears likely that subjects in this experiment were using AT&T Communicator as a tool, rather than as a conversational or negotiating partner. That is, their goal was to use Communicator to quickly and efficiently select the single flight that best matched the criteria given to them in each task. Asking the system questions in order to get information relevant to this selection process, would decrease the speed and efficiency with which they were able to accomplish this task. So, subjects preferred the most verbose presentation of information in order to increase the speed and efficiency of the overall task.

Nevertheless, eighty percent of the subjects stated that the most important criterion when personally selecting a flight was price. A number of subjects commented that they were willing to trade off other important criterion, e.g. airline, number of stops, in order to get a better price. In a more complex selection task, where the user is choosing a set of flights based on multiple, *competing* selection criteria, the user may wish to use AT&T Communicator as a negotiating partner in order to get the best flight available. Such a negotiation might lend itself well to exploring machine-user dialog in a natural language telephony-based system.

## 5. ACKNOWLEDGMENTS

We would like to thank the AT&T Communicator team for their assistance on this project.

## 6. REFERENCES

- [1] Blanchard, H.E. & Lewis, S.H. (1999), The Voice messaging user interface, in D. Gardner-Bonneau (ed.), *Human factors and voice interactive systems*, Kluwer Academic Publishers, pp.257-284.
- [2] Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Di Fabbrizio, G., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M., Ruscitti, P., and Walker, M. (2000), *The AT&T-DARPA Communicator mixed-initiative spoken dialog system*, Proc. of the International Conference of Spoken Language Processing, (Beijing, China), pp. 122-125.
- [3] Schneiderman, B. (1992). *Designing the user interface* (2<sup>nd</sup> ed.). Reading MA: Addison Wesley.