Article Submitted to Computer Speech and Language

Cross-Site Evaluation in DARPA Communicator: The June 2000 Data Collection

Marilyn Walker[1], Rebecca Passonneau[1], John Aberdeen[3], Julie Boland[2], Elizabeth Bratt[4], John Garofolo[5], Lynette Hirschman[3], Audrey Le[5], Sungbok Lee[1], Shrikanth Narayanan[1], Kishore Papineni[6], Bryan Pellom[7], Joe Polifroni[8], Alex Potamianos[9], Payal Prabhu[1], Alex Rudnicky[10], Greg Sanders[5], Stephanie Seneff[8], Dave Stallard[11] and Steve Whittaker[1]

[1] *AT&T Labs*
[2] *University of Michigan*
[3] *MITRE*
[4] *SRI International*
[5] *NIST*
[6] *IBM*
[7] *University of Colorado*
[8] *MIT*
[9] *Lucent Bell Labs*
[10] *Carnegie Mellon University*
[11] *BBN Technologies*

Abstract

The objective of the DARPA Communicator project is to support rapid, cost-effective development of multi-modal speech-enabled dialogue systems with advanced conversational capabilities. This paper describes the methodology and results of the June 2000 evaluation. The evaluation resulted in a corpus of 662 dialogues, and involved nine different travel planning systems, which varied along three critical dimensions: (1) They targeted different back-end databases for travel information which contained different content; (2) They used different modules for ASR, TTS, NLU; and (3) they applied distinct dialogue strategies for managing mixed-initiative interaction. These sources of system variation provide a challenging context for cross-site evaluation and comparison. We describe the experimental design, the approach to data collection, the metrics collected, and the results.

1. Introduction

The objective of the DARPA Communicator project is to support rapid, cost-effective development of multi-modal speech-enabled dialogue systems with advanced conversational capabilities. Spoken dialogue systems already in use today provide spoken interfaces to relatively simple transaction tasks. For example, Telebanker, an application fielded by SpeechWorks for the Credit Union Australia provides access to account balances and lets users perform certain financial transactions, such as account transfers. Typically, such systems take a slot-filler approach to task execution, soliciting the items of information from the human user that are a prerequisite to assembling a complete database query (e.g., determining whether lost baggage has been located), or to performing a database transaction (e.g., making a ticket purchase). These systems sacrifice the flexibility inherent in human-human dialogues in which users can choose what to say next.

The DARPA Communicator project is designed to encourage research on mixed-initiative systems that give the user more control over the order and manner in which information is provided. The scenario in Figure 1 illustrates the Communicator challenge problem. The assumption of this problem is that a system must be able to support complex conversational interaction in order to complete such a task in 10 minutes or less.

> You are in Denver, Friday night at 8pm on the road to the airport after a great meeting. As a result of the meeting, you need to attend a group meeting in San Diego on Point Loma on Monday at 8:30, a meeting Tuesday morning at Miramar at 7:30, then one from 3-5 pm in Monterey; you need reservations (car, hotel, air).
> You pull over to the side of the road and whip out your Communicator. Through spoken dialogue (augmented with a display and pointing), you make the appropriate reservations, discover a conflict, and send an e-mail message (dictated) to inform the group of the changed schedule. Do this in 10 minutes.

Figure 1: Darpa Communicator Challenge Problem

Despite major advances in component technologies such as speech recognition and natural language generation, appropriate techniques and algorithms for providing such intelligent interaction are still an open research problem [18, 16, 6, 31, 21, 46, 32, 11]. In order to understand which strategies are effective, it is important to be able to evaluate the contribution of various techniques to users' willingness and ability to use a spoken dialogue system [50, 8, 36, 19, 9].

In June of 2000, we conducted an exploratory data collection experiment with nine participating Communicator systems. The experiment was designed by the Communicator Evaluation Committee, chaired by Walker and consisting of representatives from the nine Communicator sites and from NIST. The primary goal of the experiment was to provide a baseline for various metrics that could be used for making comparisons with future versions of the Communicator systems. A secondary goal was to support the application of PARADISE evaluation framework in order to push forward research on evaluation itself [41, 47, 45]. This experiment provides a unique resource for cross-system evaluation because of the complexity and heterogeneity of the systems and the size of the collected dialogue corpus. It is also a challenging data set because cross-system evaluation requires agreement on a common set of metrics which need to be implemented consistently across sites. Although a goal of Communicator is to encourage research on advanced conversational techniques, the requirement for common metrics logging makes it difficult to introduce novel ways of measuring unique innovations in the different systems.

In Section 2 we first briefly describe the dialogue capabilities of the Communicator systems that took part in the data collection. Section 3 describes the experimental design used for the evaluation. Section 4 describes the data processing and handlabelling required prior to analysis of the data. Section 5 presents the subjective results. Section 6 describes the application of PARADISE to the corpus and describes differences across sites for the measures identified by PARADISE as important predictors of user satisfaction. Section 7 discusses what we learned from the data collection and describes future plans.

2. Dialogue Capabilities of the Communicator Systems

The June 2000 data collection experiment involved nine participating Communicator systems from AT&T Labs, BBN Technologies, Carnegie Mellon University, University of Colorado, IBM, Lucent Bell Labs, MIT, MITRE, and SRI International. All of the systems were implemented with a common architecture based on MIT's Galaxy II architecture [33, 26]. This architecture uses a scriptable hub to provide routing and program control, in conjunction with servers that do the actual processing, such as speech recognition (ASR), natural language understanding (NLU), natural language generation (NLG), text-to-speech (TTS), and dialogue management strategies. All of the systems supported travel planning and utilized some strategy for mixed-initiative interaction. However, they varied across three critical dimensions: (1) They targeted several different back-end databases for travel information which contained different content; (2) System modules based on more robust or off-the-shelf technology, such as ASR, TTS, or NLU were typically but not necessarily different across systems; (3) Dialogue capabilities were designed to support mixed initiative dialogue, but these capabilities varied a great deal across systems. Here we briefly describe the range of different dialogue capabilities and dialogue management strategies across the systems. These systems are described in detail elsewhere [49, 22, 30, 29, 27, 35, 34, 40, 38, 37]. Sample dialogues collected as part of the data collection that illustrate some of the differences between systems are provided in Figure 20 and Figure 21.

Systems varied in terms of their basic capabilities and their methods for enrolling the user with the system and acquainting him or her with the system capabilities. Five systems provided an enrollment page, which often included a description of the system's capabilities, a short set of instructions on commands for correcting the system and getting help, and example dialogues. A sample page of instructions included as part of the enrollment for the AT&T Communicator system is in Figure 2. A sample dialogue that was also included as part of the AT&T enrollment page is in Figure 3. Notice that the sample dialogue illustrates the user taking the initiative, using meta-dialogue commands like *That's wrong* and correcting the system when it misunderstands, as indicated by cases where the system implicitly confirms the wrong information. The hints in Figure 2 instruct the user to take note of how corrections are made in the sample dialogue.

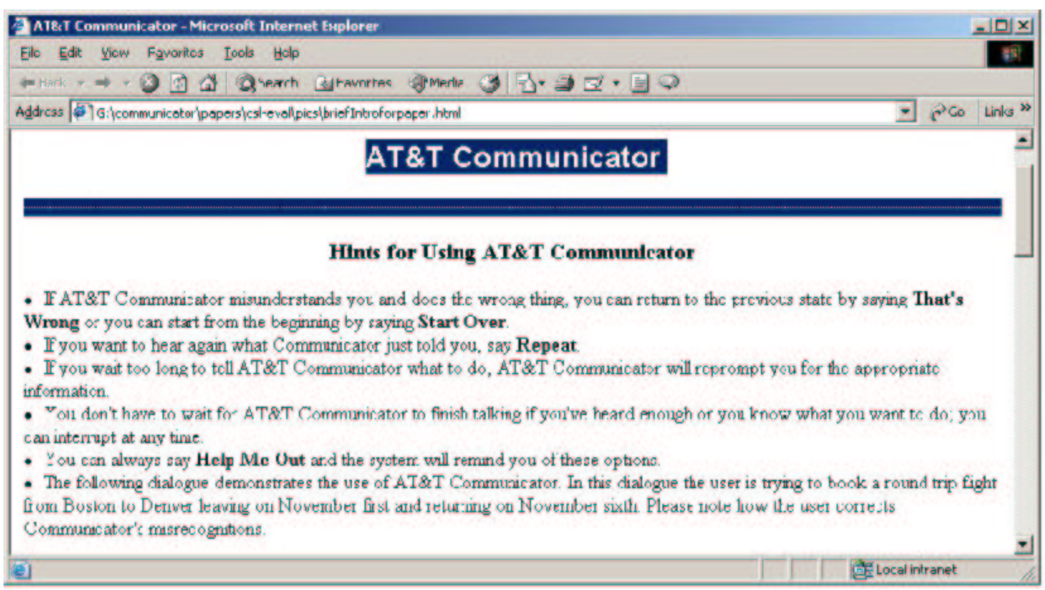

Figure 2: Tutorial on the AT&T Communicator Enrollment Page

A typical dialogue strategy for all of the Communicator systems involved five dialogue phases: OPENING, INFORMATION GATHERING, INFORMATION PRESENTATION, FLIGHT NEGOTIATION AND BOOKING, and CLOSING. The OPENING and CLOSING phases were not counted as part of the on-task portion of the dialogue.

The OPENING phase included the user logging in to the system, either by using a PIN previously assigned or with a user name provided during enrollment with

System: Welcome. You are logged in as a guest user of AT&T Communicator. You may say **repeat**, **help me out**, **start over**, or, **that's wrong**, you can also correct and interrupt the system at any time. What airport would you like to fly out of?
USER: I WOULD LIKE TO BOOK A ROUND TRIP FLIGHT FROM BOSTON TO DENVER
System: Flying from Boston to Denver, Roundtrip flight, ok, and what date did you want to leave?
USER: NOVEMBER FIRST *(System heard: MAY FIRST)*
System: Leaving on the 1st of May, And what time did you want to leave?
USER: NOVEMBER FIRST *(User corrects the date)*
System: Leaving on the 1st of November, What time would you like to depart?
USER: IN THE MORNING
System: And what date did you want to return?
USER: NOVEMBER SIXTH *(System heard: NOVEMBER EIGHT)*
System: Returning on the 8th of November, And what time did you want to return?
USER: THAT'S WRONG *(User cancels his previous input)*
System: Ok, this is where we were before: And what date did you want to return?
USER: NOVEMBER SIXTH
System: Returning on the 6th of November, And what time did you want to return?
USER: ANYTIME
System: Ok, let me see what's available. This will take a minute.
System: Please hold, I'm still waiting to hear back.
System: I found 5 round trips from Boston to Denver on November 1st and returning Boston on November 6th . For option 1 I have an outbound flight with United Airlines departing at 6 AM and arriving at 8 24 AM non-stop. A return flight is with United Airlines departing at 12 35 PM and arriving at 6 26 PM non-stop. Ticket price is 480 dollars. Would you like me to hold this option?
USER: NO
System: For option 2 I have an outbound flight with United Airlines departing at 6 AM and arriving at 8 24 AM non-stop. A return flight is with United Airlines departing at 6 40 PM and arriving at 12 20 AM non-stop. Ticket price is 480 dollars. Would you like me to hold this option?
USER: NO
System: For option 3 I have an outbound flight with American Trans Air departing at 6 15 AM and arriving at 10 oh 5 AM with 1 stopover. A return flight is with American Trans Air departing at 4 PM and arriving at 11 35 PM non-stop. Ticket price is 392 dollars. Please say flight details, i'll take it, or next option.
USER: OPTION ONE
System: For option 1 I have an outbound flight with United Airlines departing at 6 AM and arriving at 8 24 AM non-stop. A return flight is with United Airlines departing at 12 35 PM and arriving at 6 26 PM non-stop. Ticket price is 480 dollars. Would you like me to hold this option?
USER: OPTION THREE
System: For option 3 I have An outbound flight with American Trans Air departing at 6 15 AM and arriving at 10 oh 5 AM with 1 stopover. A return flight is with American Trans Air departing at 4 PM and arriving at 11 35 PM non-stop. Ticket price is 392 dollars. Please say flight details, i'll take it, or next option.
USER: I'LL TAKE IT
System: Great! I am adding these flights to your itinerary. Would you like to make any ground arrangements?
USER: NO
System: Now, would you like to hear a summary of your itinerary?
USER: NO
System: All right you're all set, now please take a minute to answer the following question, Were you able to complete your travel task successfully?
USER: YES
System: Great! Thanks for calling AT&T Communicator. Bye bye. Hope to hear from you again.

Figure 3: Sample Dialogue on AT&T Communicator Enrollment Page

the system. After the user was logged in, the system sometimes provided (additional) instruction for interacting with the system. Some systems provided both

Site Name	Opening Greeting
AT&T	Welcome. You are logged in as a guest user of Ay T and T Communicator. *You may say repeat, help me out, start over, or, that's wrong, you can also correct and interrupt the system at any time.* What airport woodja like to fly out of?
BBN	Welcome to Talk and Travel. *Before we begin, let's go over a few simple instructions First, always wait to hear the beep before you say anything. If I make a mistake, you can correct me, or just say scratch that, or, back up. You can always start over again completely just by saying start over. OK, now we'll begin!* What trip would you like to take?
CMU	Hello. Welcome to the C M U Communicator. (User enters PIN). *You may interrupt these instructions at any time by saying, good enough. The Communicator is a travel planning system with up to the minute flight information. It knows about major U.S. cities, and some international destinations. Here are some tips for a smooth interaction. Please speak clearly and naturally. Do not speak too quickly or too slowly. You can interrupt the system at any time by saying anything you wish. If you need to make a correction, just restate the new information. For example, if you'd like to depart from Cleveland instead of Pittsburgh, you can simply say, i'm departing from Cleveland. Now, here are some keywords you can use. To erase everything so far and start from the beginning, say "start over" To hear the last system response again, say "repeat" To end the call, say "good bye" This is the end of the instructions. If you need help at any time, please say, "help".* This call is being recorded for development purposes, and may be shared with other researchers. Where are you leaving from?
COLORADO	Welcome to the CU communicator. Please enter your personal identification number followed by the pound key. (User enters PIN). *Please remember to speak after the tone. If you get confused at any point you can say start over to cancel your current itinerary.* What are your travel plans?
IBM	Hello, Welcome to our Communicator flight information system. Please tell me about the first leg of your trip. *For example, you can say, I want to fly from Los Angeles to Chicago leaving next Tuesday.*
LUCENT	Hello, welcome to the communicator, an experimental travel reservation system. Hello, this is the Bell Labs travel reservation system.
MIT	Welcome to the initial development version of the MIT air travel planning system, this call is being recorded for system development. *you may hang up or ask for help at any time.* Please say your pin number. (User enters PIN). I'll enroll you temporarily as a guest user. How can I help you?
MITRE	Welcome to the mitre hybrid communicator system. *If you get stuck, you may say start over. When appropriate, to hear the next possible flight you may say next option.* Please enter your pin followed by the pound sign, or just the pound sign to be our guest. (User enters PIN). What are your travel plans?
SRI	Hi! Welcome to S R I's Communicator demonstration. What are your travel plans?

Figure 4: Sample Dialogue Openings: Instructions in Openings are italicized.

an enrollment page and instructions at the beginning of the dialogue, and other systems provided neither an enrollment page nor instructions at the beginning of the dialogue. When instructions were provided, there was a large variation in what and how much information was provided. Figure 4 provides the initial opening prompts for all of the systems. Note that in this prompt, systems varied a great deal in how much information they tried to give the user up front about how to interact with the system and what the system's capabilities were. Systems also varied in terms of how much help was provided during the dialogue when the system was having trouble understanding the user, and the degree to which the information provided in the help prompt was context-specific. For example, in Figure 21 the system makes two different types of suggestions in SYS5 and SYS6 for recovering from the misunderstanding, however some systems would have indicated misunderstanding without making such suggestions.

The INFORMATION GATHERING phase is where differences in initiative and confirmation strategies are primarily in evidence. In terms of initiative, systems started this phase with either: (1) an open-ended question such as *What are your travel plans*; or (2) a directive prompt such as *What city are you traveling to?*. Figure 4 shows for each system the system's opening prompt at the end of the greeting. Subsequent system utterances in the remainder of the information gathering phase were typically direct questions about a single parameter of the flight. See Figures 20 and 21. However all systems utilizing directive prompts were designed to be able to understand open-ended responses to those directive prompts, except possibly in contexts where the system was having trouble understanding the user. Because of this, it is useful to distinguish system initiative using the initiative categorization in Figure 5. As shown in Figure 5, one way to define the type of initiative supported by a system is as as a combination of two factors. The first factor is the type of prompt: the system may use directive prompts that lead the user to respond directly to a specific question [17] or it may encourages the user to take the initiative by using open prompts that don't query for a specific information slot. The second factor is whether, in that dialogue state, the system can understand responses in which the user does take the initiative [22]. Obviously it doesn't make sense for the system to ask an open question, like *What are your travel plans?* and expect the response to a restrictive question like *Where are you traveling to?*. Typically system initiative was only employed by the systems after several misunderstandings had occurred.

Prompt Type	Grammar Type	
	Restricted	Unrestricted
Open	Doesn't make sense	User Initiative
Directive	System Initiative	Mixed initiative

Figure 5: Definition of Types of Initiative in Dialogue Systems

In addition to different types of initiative, different types of confirmation strategies are used by the systems during the INFORMATION GATHERING phase. These include IMPLICIT CONFIRMATION, EXPLICIT CONFIRMATION and VARI-

ABLE CONFIRMATION strategies. The strategy of implicitly confirming information as it is provided by the user is illustrated in the dialogues in Figures 20 and 21. The variable confirmation strategy confirms some parameters during information gathering, but confirms all critical parameters before querying the database. See utterances SYS3 and SYS4 in Figure 6. None of the sites explicitly confirm each item of information as it is provided. However some sites switch to an EXPLICIT CONFIRMATION strategy after several ASR rejections.

Speaker	Utterance
SYS1	....What are your travel plans? ?
USER1	I WISH TO FLY FROM LEXICON KENTUCKY TO SPRINGFIELD MISSOURI THE MORNING OF OCTOBER FIFTH
SYS2	do you have a preferred airline?
USER2	NO
SYS3	I have you going from Kentucky to Gainesville Missouri on October fifth . Is that correct?
USER3	IT'S LEXINGTON KENTUCKY TO SPRINGFIELD MISSOURI
SYS4	I have you going from Lexington to Springfield Missouri on October fifth, Is that correct?
USER4	YES
SYS5	Will you return to Lexington from Springfield?
USER5	NO
SYS6	Is Springfield your final destination?
USER6	YES
SYS7	Hold on while I check availability..... Three options were returned. Option one has a fare of four hundred and thirty three dollars.....

Figure 6: Dialogue Illustrating Variable Confirmation Strategy

In the INFORMATION PRESENTATION phase of the dialogue, systems varied in how many flight options were presented and how much information was presented about each option. For example, contrast the sample dialogues in Figure 20 and Figure 21. In Figure 20, at one point the system presents 12 options and at another point the system presents 23 options. The information provided about each option includes airline, departure time, arrival time and connecting city if such exists. In Figure 21, the system only presents two options at a time and the arrival time is not provided initially; that information is added after the user selects a flight.

Within the FLIGHT NEGOTIATION AND BOOKING phase, systems varied in terms of the type of vocabulary that the user could employ to navigate the set of options returned from the database. Systems also differed in terms of the range of capabilities that were offered to refine the flight selection criterion after the system had presented some options to the user. For example, in Figure 21, the user refines the flight selection criterion after some flights have been returned from the database by asking *Do you have any direct flights?*. In Figure 20 the user responds to the system's query *Is that okay?* by saying *[UH] NO TOO EARLY*. The system then presents some later flights which don't meet the user's constraint of traveling on a day flight, and eventually the user agrees to accept the flight that was first presented. In contrast, some of the systems offered less flexibility in terms of the types of criteria that user's could specify to refine the system's flight selection or gave the user explicit instructions about what to say even if the system did support some types of refinement. For example in the instructional dialogue in Figure 3 the user is asked to *Please say next option, flight details or I'll take it.*

Systems also varied in how the complete booking was confirmed and whether the system offered to send the user an email with all the flight information or not. The system illustrated in Figure 20 offers in SYS12 to read back the itinerary to the user after the booking is completed, whereas the system in Figure 21 emails the itinerary to users who have enrolled with the system.

There were also small differences in the task model employed by the systems. One primary difference was whether the system first got all the information for all the legs of the trip before going to the database, or whether the system booked the trip one leg at a time. A second difference was that some systems asked explicitly about airline preference, and some systems implemented the optional subtasks of car and hotel arrangements.

The dialogue interaction was also necessarily affected by whether the system supported barge-in. Some systems supported voice barge-in, some only DTMF barge-in and some had no barge-in. As would be expected, systems that supported voice barge-in supported more natural dialogue interaction, but typically had lower ASR performance.

In the remainder of the paper, the quantitative performance of the systems will be discussed with the systems anonymously identified by randomly assigned SiteIDs from 1 to 9.

3. Experimental Design

3.1. Overall Setup

The primary goal of the experiment was to provide a baseline for various metrics that could be used for making comparisons with future versions of the Communicator systems. A secondary goals was to support the application of PARADISE evaluation framework in order to push forward research on evaluation itself [41, 47, 45]. The PARADISE evaluation framework integrates and unifies previous approaches to evaluation [28, 12, 13] and has been broadly applied in other work [47, 23, 4, 5, 10]. The framework posits that user satisfaction is the overall objective to be maximized and that task success and various interaction costs can be used as predictors of user satisfaction. Previous work suggests that such predictions can be quite accurate on unseen data, opening up the possibility that dialogue systems could be evaluated automatically with a relatively small amount of training data [47].

Application of PARADISE thus requires the collection of a user satisfaction rating from users and metrics used to predict user satisfaction. As described in more detail below, we collected user satisfaction metrics via a web-based survey that each caller filled out immediately after completing the call. We defined a set of core metrics to use as predictors of user satisfaction [42], and facilitated their collection through the use of a shared logfile standard that was developed by MITRE and used by all the systems [2]. The core metrics were developed during a workshop of the Evaluation Committee and included all metrics that anyone in the committee suggested that could be implemented consistently across systems. NIST's contribution was to recruit the human subjects and to implement the experimental design specified by the Evaluation Committee.

The experimental subjects were 72 native English speakers from the target population of frequent travelers from all over the United States. Sample dialogues were collected for each system by having each subject call each of the nine systems. We expected that the within-subjects design of the experiment would allow us to make statistical comparisons across systems with a smaller dialogue corpus. This was important because we had limited resources for collecting dialogues and because this is not a controlled experiment in the standard sense. Here, there are a tremendous number of different sources of variation across systems. While differences among individuals are typically a large source of variation in most experiments, most other sources of variation would be controlled, and a single experiment parameter (e.g. initiative [44]) would be varied. Control of sources of variation makes it is easy to identify the source of differences in dialogue metrics. We thus had the subjects call each system so that we had a call per subject per system, and perform nine scenarios, which were also controlled to some extent.

There are also three potential sources of bias that our experimental design attempts to minimize:

- Are users influenced by the first system they try? (system-dependent training bias)
- Are users influenced by the first task they try? (task skill bias)
- Does the instruction format for specifying the scenario information to the user influence the user's linguistic behavior? (instruction bias).

In an attempt to eliminate potential bias due to the possibility that subjects would "train" to the first system they call, subjects started the scenarios with different systems. A latin-square design was used to sample from the set of all possible ways to sequence through the 9 systems. Some systems required the users to visit a web page at their site before calling the system. This was implemented as part of the experimental design because it was believed that this approximated a realistic use for those systems that expected users to enroll with the system before making any calls. Figures 2 and 3 for the AT&T system illustrate a typical set of enrollment pages. Subjects carried out the scenarios in a fixed order, with scenarios becoming progressively harder, thus precluding the introduction of bias based on task ease, or differential opportunity among the users to master the tasks.

Recent work has argued that dialogue data collected with fixed scenarios is not realistic [20]. However, to our knowledge, no quantitative or qualitative assessment of the differences between these modes of data collection has ever been published. Furthermore, fixed scenarios serve several purposes in an experiment such as this. First, they can be used to guarantee that tasks of a certain type or level of complexity will occur in the experimental data. Second, they make it possible to control for one source of variation to enable direct comparisons across systems for similar tasks. We collected a combination of fixed and open scenario dialogues from the same user. In an attempt to eliminate instruction bias, scenarios were communicated to the user in a tabular format. Figure 7 illustrates two sample scenarios for a domestic round trip (Task3) and an international round trip (Task6) as they were presented to the users in the data collection experiment.

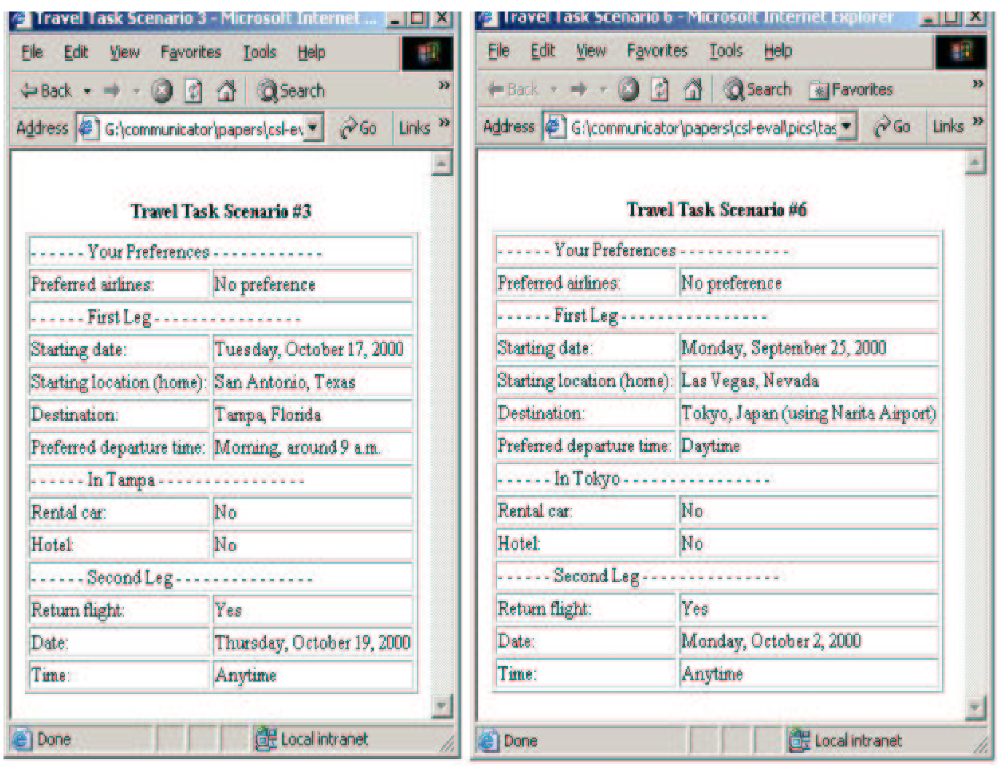

Figure 7: Example Task 3 (Domestic Round Trip) and Task 6 Scenarios (International Round Trip)

The dialogues were recorded in full at NIST by connecting each call through a central call router running on a NIST server. Each site provided a standard logfile as well as transcriptions and recordings of user utterances. At the end of each call, users gave subjective feedback via a web survey. Metrics collected per call consisted of objective metrics extracted from the logging, subjective metrics collected via a survey, and evaluative metrics on task completion, reasons for task failure, and user behavior that were handlabelled at AT&T. The goal was to have 8 dialogues per task per system, but since not all subjects called all systems, the resulting corpus consists of 662 dialogues. The remainder of this section characterizes the data collection setup and experimental design in more detail.

3.2. Task Scenarios

The task scenarios were intended to vary task complexity in order to provide baselines for task completion and other metrics for a range of task types. Task complexity for this purpose was defined simply as the number of constraints that the user had to communicate to the system. The scenarios consisted of 7 *fixed* and 2 *open* scenarios. The *fixed* scenarios consisted of 3 domestic one way (DOW) trips, 2 domestic roundtrip itineraries (DRT), and 2 international round trip (INT) flights. These were presented to the user in tabular format. Example domestic and international round trip tasks are given in Figure 7. Since the data collection occurred over three three-day periods over three weeks, the cities, airport and airline names were changed for each task on a daily basis to ensure that sites could not "game" the system to perform well on a small set of specific tasks.

The *open* scenarios were defined by the user. After completing 7 pre-defined tasks with 7 of the systems, the users were asked to use the remaining two systems to "plan a recent or intended business trip" and "plan a vacation ". The *open* scenarios were intended to approximate the conditions under which these systems would be used in the field [3].

3.3. Subjective Metrics

At the end of each call, each user provided a subjective evaluation of the system's performance via a web survey. The web survey was used to calculate Perceived Completion and User Satisfaction measures. Users report their perceptions as to whether they have completed the task via the yes/no survey (**Perceived Completion**) question in Figure 8.

The User Satisfaction questions on the survey probe different aspects of the users' perceptions of their interaction with the system [36, 24, 14]. They are all stated in terms of positive dimensions of the system; the user specifies the degree to which they agree with these statements in terms of a 5 point multiple choice Likert scale. Each survey response is then mapped into the range of 1 to 5 and the values for all the responses are summed, resulting in a **User Satisfaction** measure for each dialogue ranging from 5 to 25.

- Were you able to successfully complete your task? (**Perceived Completion**)
- In this conversation, it was easy to get the information that I wanted. (**Task Ease**)
- I found the system easy to understand in this conversation. (**TTS Performance**)
- In this conversation, I knew what I could say or do at each point of the dialogue. (**User Expertise**)
- The system worked the way I expected it to in this conversation. (**Expected Behavior**)
- Based on my experience in this conversation using this system to get travel information, I would like to use this system regularly. (**Future Use**)

Figure 8: User Survey assessing Perceived Task Completion and User Satisfaction

3.4. Logfile Metrics

The objective metrics focus on measures that can be automatically logged or computed. They include diagnostic metrics that are comparable across systems for evaluation of component modules, as well as dialogue management and whole dialogue metrics. In addition, we use certain evaluative metrics to categorize system behavior, such as Exact Scenario Completion, which cannot be automated, but which are sufficiently objective that expert judgements are deemed reliable. In general, system metrics are automatically logged unless otherwise indicated. The full set of metrics characterizing system behavior are of three types: efficiency, quality and task success, and are summarized in Figure 9.*

The collection of these metrics in a consistent way across systems is facilitated by the shared logfile standard. The standard specifies that a logfile (document) consists of a number of sessions (typically 1) with the system where each session is composed of a number of system turns and user turns . Each system and user turn contains some number of operations (commands executed by the system within a turn), messages (items sent by the various servers in a system, as well as their replies), and events (such as errors, locks and alarms). Operations, messages, and events may contain data in the form of key/value pairs. All elements are

*Some experimental metrics included in the original specification of the logfile such as the type of prompts, errors, and help messages were not logged by all the sites or were logged inconsistently, and are therefore not discussed here.

- **Dialogue Efficiency Metrics**: Total elapsed time, Time on task, System turns, User turns, Turns on task, Time per turn for each system module
- **Dialogue Quality Metrics**: Word error rate, Response latency, Mean word error rate, Mean response latency, Variance reponse latency
- **Task Success Metrics**: Perceived task completion, Exact Scenario Completion, Any Scenario Completion

Figure 9: Metrics collected via Logfile Standard.

time stamped, to facilitate the calculation of durations. Below we describe the elements of the logfile standard that relate to the objective metrics we wish to calculate. Further details about the structure (as well as the XML format) of the logfile standard are available online [2].

As an example of the logging, consider the sample dialogue in Figure 20. This dialogue can be broadly divided into three sections. The dialogue OPENING in SYS1 and the dialogue CLOSING in SYS14 are not considered part of the on-task portion of the dialogue. The logfile standard encodes the on-task portion of the dialogue with attributes that mark the start and end of the task. By logging system and user turns we can easily calculate the total number of turns in the session (27), as well as the number of system turns (14), number of user turns (13), and the number of turns on task (25). We also log what the system says at each turn of the dialogue and we have human transcriptions of each user turn (human transcriptions can be kept separately or integrated into the logfile after transcription is completed). From these two sources of information the number of user words in a turn and the number of system words in a turn can be calculated, as well as the mean number of user words per turn and system words per turn over the whole dialogue. Because the start and end of the task are marked in the logfile, these metrics can also be calculated for just the on-task portion of the dialogue. The logfile standard also encodes the selected automatic speech recognition hypothesis for each user turn. This, coupled with the human transcription, supports the calculation of word error rate and other ASR metrics such as sentence error, and the number of insertions, deletions and substitutions. As mentioned above, all elements in a logfile are time stamped. This, along with the logfile characteristics described above, enables the calculation of several dialogue efficiency metrics, such as total elapsed time, time on task, mean length of system turn, and response latency. Response latency is calculated by subtracting the value of the end-time attribute of a user utterance tag from the start-time attribute of the following system utterance tag.

4. Data Processing and Hand Labelling

The experiment resulted in 662 dialogues with the number of dialogues per system ranging between 60 and 79. Each site was responsible for collecting the logs, transcribing the user utterances and submitting the logfiles and the user audio to NIST. Variation in the number of dialogues per system and per task resulted from human subjects dropping out of the experiment, problems with system stability and problems with the stability and load on the central call router running at NIST. Thus, although the design was intended to be a within-subjects design, only 49 of the subjects actually called all 9 systems. Below, we will report results from analyzing all of the data; we also separately analyzed the subset of within-subjects data and found no strong differences in the results.

User Commitment: One of our primary concerns was the extent to which the users were actually attempting to achieve the goals that they were given in the tasks. This concern arose because of a widespread conception that users would not "try" to do the tasks that they were given because these tasks were not their own true goals. We developed a labelling scheme and handlabelled each dialogue in order to assess the degree of commitment that the user's exhibited in the dialogues.

Each dialogue was examined and then labelled with one of 5 types of user behavior: Goal-Directed, False-Acceptance, Scenario-Switch, Wrong-Information, Initially-Inattentive. Two labelers labeled the user behavior in concert, then a third labeler confirmed the results; all sites had an opportunity to check the results and request changes. (Note: all hand annotations of the dialogue corpus, including Task Completion, followed this procedure.) A sixth category of user behavior was Unknown, for the dialogues in which no logfile was generated due to system crashes.

Goal-Directed users were completely focused on the task and never exhibited any behaviors of the remaining four types. This category thus represents our ideal user. There were four categories of problematic behavior: (1) False-Acceptance users failed to correct a system misunderstanding; (2) Scenario Switch users described those who changed plans during the dialogue (often in response to repeated recognition error); (3) the Wrong-Information users provided information inconsistent with a fixed scenario; (4) a user was classified as Initially-Inattentive if more than a second elapsed before the user responded to the system, or never responded, or responded incorrectly.

Task Completion: To capture the potentially significant gray area regarding how much of a task the user completed, a ternary definition of Task Completion was annotated by hand at AT&T for each call. We distinguish between exact scenario completion (ESC), other scenario completion (OTHER) and no scenario completion (NOCOMP). It is necessary to distinguish at least these three cases because some callers completed an itinerary other than the one assigned in a fixed scenario. This may have resulted from a caller's inattentiveness, e.g. she didn't correct the system when it misunderstood. In this case, the system could be viewed as having done the best it could with the information provided and we might want to define Task Completion as ESC + OTHER. However, examination of the dialogues suggests that sometimes the OTHER category arose as a rational reaction to repeated recognition error. If the user was being cooperative in trying to complete some task in the face of repeated system error, then it seemed important to maintain a distinction between ESC and OTHER. Furthermore, our initial analysis showed that users were particularly willing to complete an alternative task when performing the *open* scenarios. Therefore, in the analysis below, we present results for both exact scenario completion (ESC only) and ANY scenario completion (ESC + OTHER).

When a dialogue was labelled NOCOMP, we also labelled the reason for task failure. This consisted of assigning the blame to either a system module or to a user behavior. The NOCOMP categories are detailed in Figure 10.

Label	Description
ASR	Task failure due to (repeated) ASR error.
NLU	Task failure due to (repeated) NLU error.
DialogueManager	Task failure due to misbehavior by the Dialogue Manager.
NoFlights	The database returns no flights (sometimes this was a web access error) and the user completes the call.
NoWait	The system asked the user if s/he wanted to continue to wait for the database to come back, and the user said *No* (usually after having already waited for some time).
SystemGaveUp	The system informs the user that it perceives a problem with the dialogue and hangs up (sometimes asking the user to call back later).
CallInterrupted	The dialogue ends in the middle of the dialogue. Some dialogue platforms cannot distinguish the user hanging up from a system crash. If the user commented that s/he hung up due to repeated system error, no completion was blamed on the module causing the error.
NoTaskStart	The system greeted the user but no dialogue ensued. Many of these indicated problems with handshaking with the NIST call router.
Unknown	No Logfile was generated.

Figure 10: Reasons for No Completion and Definitions

The task of hand labeling the Task Completion metric was extremely time consuming. The procedure used was identical to the one described above for annotating user behavior: two labelers performed an initial consensus labeling; their output was checked by a third labeler. These annotations were then distributed to each site for confirmation and adjudication. If a site questioned the labelling, a labeler would re-examine the dialogue and discuss with a site representative the reason for the assigned label. This always resulted in an agreement between the labeler and the site as to the appropriate completion label.

Metrics Derivation: After human transcription and human assessment of task completion and user behavior, the other metrics were calculated automatically using the logfiles generated in logfile standard format with scripts written by MITRE. NIST computed sentence accuracy and word accuracy by comparing the hand transcriptions of each utterance with the recognizer output [25]. Figure 11 summarizes the complete set of metrics available for the analysis.

- **Dialogue Efficiency Metrics**: Total elapsed time, Time on task, System turns, User turns, Turns on task, Time per turn for each system module
- **Dialogue Quality Metrics**: Word error rate, Response latency, Mean word error rate, Mean response latency, Variance reponse latency
- **Task Success Metrics**: Perceived task completion, Exact Scenario Completion, Any Scenario Completion, Reasons for No Completion
- **User Satisfaction**: Sum of TTS performance, Task ease, User expertise, Expected behavior, Future use.
- **User Behavior Metrics**: Classification of User Behavior per dialogue into: Goal-Directed, False-Acceptance, Scenario-Switch, Wrong-Information, Initially-Inattentive.

Figure 11: Complete set of metrics.

5. Subjective Results

There are three different types of information collected from the experiment that we consider subjective. First, as described above, we hand-annotated user behavior in order to assess how cooperative the users were. Second, we collected reports of the users' satisfaction through a web-based user satisfaction survey. Third, the user satisfaction survey included an open text field for free-form user comments.

5.1. User Behavior

We described above in section 4 the labelling scheme for user behavior. Most of the users fell into the category of Goal-Directed users (70.7%); they were completely focused on the task and never exhibited any behaviors of the remaining five types. Of the remaining 29.3%, 10.7% covers those dialogues where no logfile was generated (i.e., the system either crashed or prematurely ended the call). This leaves 18.3% of the users distributed across the three categories of problematic responses. The False-Acceptance users accounted for 8.8%. There were 4.2% Scenario Switch users who changed their plans during a dialogue with an open task (sometimes in response to repeated recognition error). The Wrong-Information users (4.2%) provided information inconsistent with a fixed scenario.

Only 1.9% of the users were Initially-Inattentive. In sum, the hand annotation of 6 types of users indicates that by far the majority (70.7%) were sufficiently goal-directed to follow the instructions as expected, and to attempt to remediate any problems that might arise. However, 10.7% of cases involved system failure such that no logfile of the interaction could be collected. This suggests that in the future there are ways to take better advantage of the high proportion of goal-directed users, or perhaps to increase this proportion; further, in a future data collection effort, it would be useful to forestall failure to generate logfiles.

5.2. User Satisfaction

The mean for User Satisfaction across all sites and all dialogues was 16.2. The box plot in Figure 12 shows the distribution in user satisfaction across the nine systems. The box plot indicates the full range of values for user satisfaction, and the interquartile range as a box within that. The median of the distribution is shown by a horizontal line within the box. A one-way ANOVA for user satisfaction by site (df=8, F=20.0, p=.0001) using the modified Bonferroni statistic for multiple comparisons [51] shows that the user satisfaction metric distinguishes four groups of performers with sites 4,2,1,3 in the top group, sites 3,5,9,6 in a second group, and sites 8 and 7 defining a third and a fourth group. Site 4 has the highest mean user satisfaction, and a relatively low spread, hence is the best in the top group. Note that the mean user satisfaction for site 3, the lowest performer in the top group, is roughly as similar to sites 1 and 2 as it is to sites 5, 9 and 6 in the next group.

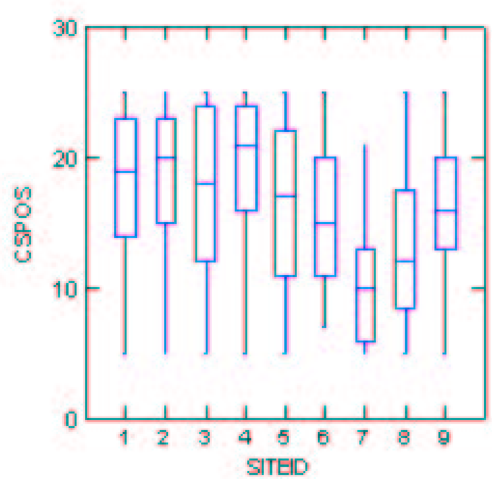

Figure 12: A Box Plot showing the Distribution of User Satisfaction across Sites

We also examined the relationship between the individual components of user satisfaction, namely Task Ease, TTS Performance, User Expertise, Expected Behavior and Future Use and the user satisfaction measure. In contrast to previous work, we found that all of the components contributed similarly to the overall measure [43]. The correlation between User Satisfaction and Task Ease was 0.9, TTS Performance was .72, User Expertise was .83, Expected Behavior was .91 and Future Use was .91. This suggests that if only one question could be asked that the Future Use or Task Ease questions could stand in for the rest. However, we also examined whether there were significant differences across systems in any of these components. As one might expect there were significant differences in all of these components, however the pattern for each component tended in the main to mirror the overall pattern shown in Figure 12.

5.3. User Comments

As described above, users were asked to complete surveys indicating their satisfaction with system performance and were given the option of writing any additional comments in a free text field. Interestingly, 85% of the users were sufficiently motivated to add free-form textual comments about their experience in the dialogue. Figure 13 details some of the comments that users provided. These include a wide range of user observations: the user failing to understand how to execute a desired action; the system not understanding the user; the user not understanding the TTS component; the user's qualitative appreciation of the system voice, vocabulary, or speed of database access, and so on. Our belief is that these aspects of system performance are probably not quantifiable. It would be difficult to design a user survey that anticipates the range of issues that users give feedback on. In addition, in particular cases, these comments were useful in handlabeling the reasons for no completion. For example, when system logfiles didn't distinguish between the user hanging up and the system crashing, the user comments typically resolved this ambiguity. See for example Comments (9) and (10) in Figure 13.

6. Objective Metrics Results

We applied PARADISE to develop models of user satisfaction with the objective metrics as predictors and then examined differences across sites only for the metrics that were strong predictors of user satisfaction. Applying multivariate linear regression to the dataset yielded a linear model with four terms. The learned model is that User Satisfaction is the sum of:

$$.43 * ESC1 - .15 * TaskDur + .21 * SACC + .14 * SysTurnDur$$

This model accounts for 38% of the variance in user satisfaction. However, three of the variables, ESC1, Sacc and TaskDur account for most of this variance (35%; significant at the p=.0001 level using a 2-tailed t-test). As noted below,

Comment ID	User's Comment
1	I was given the wrong option for the flight but it was my fault because I asked for an option by number and was given the first option that was read to me. I should have started over again to obtained the preferred departure time. Also, I thought the names of the cities were a little hard to understand.
2	The system could not understand the city Bismarck, ND so I had to fly into Fargo and take a bus to Bismarck! The voice was very clear, however.
3	The system was somewhat fooled by my faux southern drawl, however I was able to complete the task with very few problems. One thing that I noticed was that the system could not understand the word noon as a replacement for 12:00.
4	How refreshing!!!! It worked perfectly.
5	System gave choice of two flights but didn't indicate how I should express my preference.
6	Pronunciation a little strange at times.
7	Very easy and asked for car or hotel very nice.
8	It was easy to use and even asked if I was really going to take the trip.
9	The system mistook my saying 'fifth' as 'sixth.' It spoke so poorly that I didn't notice the mistake until it was giving me data for the wrong day. Then I had to repeat the whole thing. This time the database was not available and 'he' told me that the internet connection might be broken so he'd try again. I let 'him' try twice, then hung up. There was no way to be nice since he didn't understand any of my words.
10	Everything went well til the database had no record satisfying my constraints. Then neither the system nor I knew what to do next. So I hung up.
11	The computer seemed relaxed, e.g., used words like 'okay' However, when it went through the connecting flights, the voice gained speed and was hard to follow. This system also had quick access to the data base. The volume on this system was extremely faint. I had to boost the gain on my system to hear him. So far this one is numer 2 on my scale of being user friendly.

Figure 13: User Comments (Free Text Field of User Survey), illustrating the range of types of feedback provided.

the fourth variable of SysTurnDur is somewhat counterintuitive; related work suggests that this result reflects the significance of certain types of turns, and therefore a more complete interpretation requires a more subtle analysis of what happens in different types of turns, for example, based on a classification of dialogue acts [39, 48].

The finding that measures of task completion and recognition performance are significant predictors duplicates previous results [47, 15]. The fact that a measure of task duration is also a significant predictor may simply indicate larger differences in task duration in this corpus. In addition, the variables that quantify these aspects of performance, namely ESC1, Sacc and TaskDur, could be expected to be relevant for evaluating system usability across human-computer task dialogues.

TripType	NoComp%	Other%	Exact %	N Calls
DOW	35.111	15.556	49.333	225
DRT	39.103	11.538	49.359	156
INT	43.056	13.889	43.056	144
OPEN	35.766	8.029	56.204	137
Overall%	37.915	12.689	49.396	
N	251	84	327	662

Figure 14: Percentages of Exact Scenario Completion by Trip Type: DOW = Domestic One Way; DRT = Domestic Round Trip; INT = International Round Trip; Open = User's own vacation or business trip

SiteID	NoComp%	Other%	Exact %	N calls
1	29.630	16.049	54.321	81
2	17.949	11.538	70.513	78
3	19.737	14.474	65.789	76
4	22.222	11.111	66.667	72
5	30.137	15.068	54.795	73
6	43.056	13.889	43.056	72
7	72.308	6.154	21.538	65
8	65.278	11.111	23.611	72
9	47.945	13.699	38.356	73
Overall%	37.915	12.689	49.396	
N	251	84	327	662

Figure 15: Percentages of Exact Scenario Completion by Site ID

We turn now to a discussion of the four components in the model in order to interpret their significance and present further qualitative results. As noted in the introduction, there will necessarily be significant variation across sites in many if not all of the variables, due to divergent system modules, external airline databases, and dialogue management strategies. Thus for each variable in the model, we report the results of an ANOVA by site.

6.1. Task Completion (ESC1)

We examined Task Completion by scenario, by task type, and by site. We examined Task Completion by scenario in order to determine whether completion rates increased as users acquired more expertise with the systems. We examined Task Completion by trip type to see whether the experimental manipulation of task complexity had indeed made some tasks more difficult, and whether there were differences in completion between the open tasks (scenarios 8,9) and the fixed tasks (scenarios 1 to 7).

A one-way ANOVA for Exact Scenario Completion by scenario indicated no significant differences between sessions (df =8, F= 1.49, p = .16). The fact that there were no differences suggests that experience with the systems did not improve the users' ability to complete the task. This may have been because users called each system only once. However the experimental design may also obscure any increases in expertise that the users may have had because task complexity increased as expertise increased, at least for the fixed scenarios. Scenarios 1 to 3 were domestic one-way trips; scenarios 4 and 5 were domestic round-trips; scenarios 6 and 7 were international round trips; and scenarios 8 and 9 were the user's intended vacation or business trips.

We then examined Exact Scenario Completion by trip type. A one-way ANOVA for Exact Scenario Completion by trip type also indicated no significant differences (df =3, F= 1.24, p = .30), although the completion rates for the open tasks were higher, and the completion rates for the international round trips were lower. See Figure 14. One reason for the higher completion rates for the open tasks was that, according to our user behavior labelling, users more readily modified their travel plans for the open tasks, i.e. if the system couldn't understand Denpasar airport in Bali, and thought the user wanted to fly to St. Petersburg in Russia, the users changed their vacation plans in order to complete some task. This reflects a high degree of cooperativity in this user population, but such changes in travel plans are unlikely to happen in a real world application.

We then examined Exact Scenario Completion by site. The mean completion rate (ANY) for all sites was 62% (summing OTHER and EXACT) but there was a large variation in completion rate across sites. See Figure 15. A one-way ANOVA for ESC by site using the modified Bonferroni statistic for multiple comparisons indicates significant differences (df = 8, F = 13.9, p = .0001), and defines three groups of performers, with sites 2,3,4,1,5 in the top group, sites 5,6,9 in a second group and sites 8,7 in the lowest group. A one-way ANOVA for ANY Scenario Completion by site using the modified Bonferroni statistic defines the same three groups.

Figure 16 provides the reasons for No Completion for all the calls, independent of site. Problems with ASR accuracy and system stability each account for almost one third of the failures. Problems causes by accessing a web-based source of travel data also accounted for many failures (NoFlights and NoWait). However sites were differentially affected by the causes for no completion. For example, only two sites had task failures due to SystemGaveUp; one site had thirteen NoTaskStart failures due to difficulty completing the handshake with the NIST call router whereas no other site had more than 1 of these; the number of CallInterrupted failures ranged from one to eighteen.

Label	Count	Percentage
AnyComp	411	62.1%
ASR	73	11.0%
NLU	18	2.7%
DialogueManager	3	0.5%
NoFlights	25	3.8%
NoWait	16	2.4%
SystemGaveUp	11	1.7%
CallInterrupted	79	11.9%
NoTaskStart	18	2.7%
Unknown	8	1.2%

Figure 16: Completion and No Completion Counts and Percentages by Type

6.2. Task Duration (TaskDur)

The average task duration across all sites was 294 seconds; for completed tasks the average duration was 300 seconds. Remember that the challenge problem described in Figure 1 is a complex task that must be achieved within 10 minutes. None of the scenarios in the experiment were as complex as this task, but it appears that it should be possible to achieve the task duration goal. A one-way ANOVA for Task Duration by site using the modified Bonferroni statistic for multiple comparisons indicates significant differences in Task Duration across site (df=8, F=10.8, p=.0001). There are three groups of performers with site 3 in the top group (shortest durations), sites 1, 2, 4, 7 in a second group and sites 5, 6, 8, 9 in a third group. However, Task Duration is more relevant for calls in which an itinerary is completed since some failed tasks were due to system crashes early in the dialogue. The box plot in Figure 17 indicates the performance of each site for Task Duration for the ANY task completion subset. A one-way ANOVA for Task Duration by site for this subset also indicates significant differences (df = 8, F= 11.4, p < .0001).

6.3. Sentence Accuracy (Sacc)

A one-way ANOVA for Sentence Accuracy by site using the modified Bonferroni statistic showed significant differences between sites (df = 8, F=40.5, p < .0001) and two groups of performers (1, 2, 4, 8, 9 and 3, 5, 6, 7). Some systems did not support voice barge-in, and this correlated with higher accuracy. However, there was also a strong interaction between gender and sentence accuracy by site; recognition performance at some sites was much better for female speakers, at others better for males, and for some there was no difference. Mean Sentence Accuracy results are in Figure 18. Furthermore, although the experimental design attempted to balance for gender, additional subjects were added as users failed to call. These additional subjects tended to be female, so in the

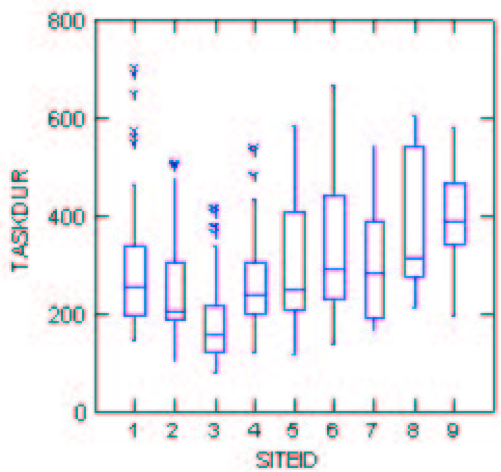

Figure 17: A Box Plot showing the Distribution of Task Duration for Completed Tasks across Sites

SiteID	Female SAcc	Male Sacc
1	66.475	67.873
2	69.259	59.616
3	25.431	43.167
4	78.430	72.963
5	38.317	36.245
6	41.795	49.096
7	45.473	48.982
8	59.414	85.750
9	68.431	67.593

Figure 18: Mean Sentence Accuracy by Site ID by Gender

end, the user population was 64% female and 36% male, causing problems for sites with poor recognition performance for female speakers.

6.4. System Turn Duration (SysTurnDur)

System Turn Duration is positively correlated with user satisfaction. Because flight presentation utterances tend to be longer than other system turns, and because task completion (ESC1) is very highly correlated with User Satisfaction, this probably reflects the presentation of itineraries in dialogues where the task is completed. A one-way ANOVA of SysTurnDur by site using the modified Bonferroni statistic indicates significant cross-site variation (df=8, F=11.2, p=.0001) and distinguishes three groups of performers with sites 1,2,5,9 in the top group, sites 3,4,6,8 in the middle group, and site 7 in the lowest group. Figure 19 shows a box plot for the distribution of system turn duration across sites.

Because the inclusion of SysTurnDur in the model accounts for a relatively small increase in goodness of fit, and because there is a less obvious connection between how long system utterances are on average, and how well an arbitrary utterance functions in facilitating a dialogue, this variable is likely to make a different kind of contribution in different types of dialogues. Two types of system turns that tend to be relatively long, and which may contribute significantly to the presence of this variable in the model, are turns in which a complete flight itinerary is presented, or turns in which the system presents some instructions (see examples in Figure 4). As mentioned above, the initial instructions are a strategy employed by some systems and not others. Consequently, interpreting the role of SysTurnDur is highly dependent on a range of other factors.

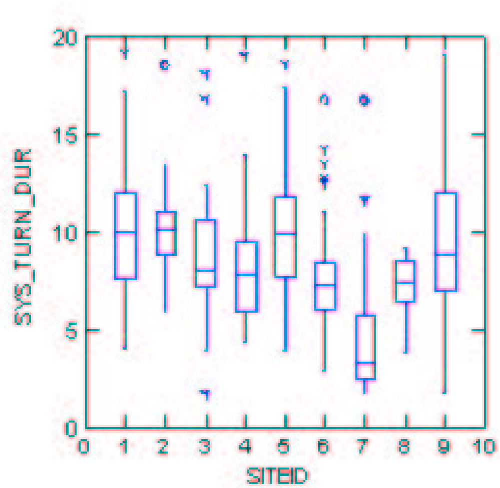

Figure 19: Box Plot of System Turn Duration across Sites

6.5. Summary of Quantitative results

Table 1 summarizes the quantitative results presented above. The dependent variable, User Satisfaction, appears in the first row; the four independent variables in the linear model appear in the remaining rows, along with their coefficients. The last two columns of the table pertain to the groups emerging from the more qualitative analysis of variance yielded by the Bonferroni analysis, and which are discussed above. Two facts, that the model accounts for less than half the overall variance in user satisfaction, and that all the model components show highly significant ANOVAs, indicate that much work remains to be done to arrive at a thorough understanding of the users' subjective judgments. Much of the variability that remains to be accounted for might depend on external factors (such as performance of the airline databases), or on more subtle analyses of functional components of the dialogues, such as how misunderstandings are handled, or how users are instructed in the use of the systems. In the following summary of the quantitative results, we suggest specific targets for meeting or surpassing the sites that seem to have achieved greater overall success, or greater success on individual components.

Variable Name	Coeff.	ANOVA F/p values	Groups	Group Membership, Precedence
User Satisfaction		20.0/.0001	4	4,2,1,3 > 3,5,9,6 > 8> 7
ESC1	.43	13.9/.001	3	2,3,4,1,5 > 5,6,9 > 8,7
Sacc	.21	40.5/.0001	2	1,2,4,8,9 > 3,5,6,7
SysTurnDur	.14	11.47/.0001	3	1,2,5,9>3,4,6,8>7,8
Task Dur	-.15	10.9/.0001	3	3 > 1,2,4,7 > 5,6,8,9

Table 1: User Satisfaction Model and ANOVA by Site results

User Satisfaction: User Satisfaction in the 2000 corpus depends most consistently on four of the metrics collected here: whether the user completes the task (ESC1), how accurately the spoken language understanding components represent what the user says (Sacc), the mean length of system turns (SysTurnDur), and how quickly the task can be completed (TaskDur). The highly significant ANOVAs for all variables in the model demonstrate the enormous variability in this data set. Table 1 illustrates that clustering of sites into distinct groups based on analysis of the individual variables in turn is somewhat stable.

Task Completion: Clearly, User Satisfaction is directly correlated with task completion. The ternary measure of exact scenario completion (ESC1) and other measures of completion such as ANY show the same distribution across sites. One of the things a mixed-initiative dialogue system should offer that a more rigid procedure cannot is flexibility in performing more complex scenarios. In this data, there are no significant differences across sites regarding better or worse performance on specific scenarios, or types of scenarios (e.g., domestic one way versus domestic round trip). We believe that this is likely due to the fact that there was not a wide enough range of task complexities represented in the scenarios.

Sentence Accuracy: The fact that there are two groups (columns 4 and 5) of performers for Sentence Accuracy suggests that improvements in recognition accuracy are still required to achieve high performance in tasks such as this, and that it is important to achieve similar levels of accuracy for both male and female speakers. Sites 1 and 4, for example, have a similarly high sentence accuracy range for both males and females (mean for site 1 is about 67% and for site 4 is 75% irrespective of gender; see Figure 18). Site two also performed better for females than males and thus benefited from the skewed distribution of gender in the experiment. Site 4's high performance is probably attributable to the fact that their recognizer used gender-specific models; they ran two recognizers in parallel in the initial part of the dialogue until the confidence scores disambiguated the subject's gender.

Task Duration: Sites fall into three groups for Task Duration. It is difficult to draw generalizations apart from the fact that shorter durations are preferable, and the observation that the mean is typically closer to the minimum value than to the maximum (see Figure 17), possibly as a result of many long duration outliers, and the fact that there is a more rigid lower bound. For example, site 3, the best performer, with the lowest Min, Max, and Avg values of Task Duration, suggests that a simple flight booking dialogue can be accomplished on the order of 2 minutes at a minimum. Site 5 seems to fall in the lowest performing group rather than the middle group because of the number of outliers falling far above the mean: in Figure 17 the mean for site 5 a respectable 250 seconds while the maximum is about 400. For the future, it is thus important to diagnose more precisely what range of factors leads to increased Task Duration and how to minimize Task Duration. In this evaluation, it is likely that higher Task Durations often result from system misunderstandings of user input, and the length of dialogue time devoted to repairs and clarifications. However Task Duration is also affected by system response time in general and the response time of the external airline database in particular. Accessing different databases, or even accessing the same database at different times of day, may have a more or less significant effect on overall Task Duration. Even though the logfile provides for timestamps for each module, the core metrics did not include measures for database access alone and the tools for extracting metrics from the logfiles did not therefore provide such a metric. In future work, it could be useful to break apart the different components of Task Duration.

7. Discussion and Future Work

We designed a method to collect uniform data for large-scale evaluation of spoken language systems that accommodates systems with different architectures, different dialogue strategies, or different modules, such as TTS. A primary motivation for the experimental design was the desire to evaluate feasibility of migration to real world applications: this led to the use of real subjects, and the use of real-world databases.

The experiment resulted in 662 dialogues, with a logfile, user satisfaction survey, complete recording, and user utterance recordings and transcriptions for each dialogue. We established a performance baseline for one-way and round trip domestic and international flights. The ANY completion rates across all sites for domestic one way trips is 65%, for domestic round trips is 62%, for international round trips is 57% and for open trips is 64%. However as shown in Figure 15 some sites achieved overall completion rates as high as 81%. We established a performance baseline for User Satisfaction of 16.2 across all sites, but User Satisfaction also varied across sites as shown in Figure 12.

One of our major concerns was whether users would behave relatively naturally in an experimental setup such as this. To provide data for assessment of this issue, users were given both fixed scenarios and open scenarios. Using fixed scenarios permits comparison across sites of task complexity, and allows uniform metrics of task completion. In the open scenarios, meaning that users defined their own tasks, the goal was to approximate the conditions under which these systems would be used in the field [3]. Subsequent to the data collection, we hand-annotated user behavior and examined differences across task types.

The user behavior annotation indicates that users were clearly attempting to complete the experimental tasks, perhaps even to a greater extent in many cases than a paying customer could be expected to. Many users (71%) were identified as Goal-Directed, and 85% of users were sufficiently cooperative to add free form comments to the user surveys. Our analysis also indicated no significant differences in task completion between the fixed and open scenarios.

A challenging aspect of this experiment is that the many sources of variation across systems and dialogues make it difficult to draw strong conclusions about which system designs are most effective. There is clearly a tradeoff between large scale evaluation considerations involving portability to the real world (users, databases) and rapid prototyping (use of off-the-shelf modules such as ASR), versus successful exploration of less well understood issues, such as dialogue management strategies that encourage mixed-initiative. Even though a program goal is to develope techniques for mixed initiative interaction, none of the core metrics directly measure the extent of mixed-initiative interaction supported by a system. We examined User Words per Turn as a proxy for user initiative, but found that it is not a significant predictor of user satisfaction. Furthermore, it appears that the level of user initiative was low, as evidenced by a mean length of 3 words per user turn. In completed tasks, the mean was 2.8 words per turn. One explanation for this was that the tabular presentation of tasks (See Figure 7) gave users the impression that their role in the dialogue was to provide values for slots that define the task. Another possibility is that directive prompts (such as those in Figure 20 in SYS2, SYS3, SYS9, SYS10) didn't cue user to take the initiative even when the system supported it. A final possibility is that novice users are more comfortable giving simple responses to system queries. However, a one-way ANOVA for user words per turn by site revealed that there were significant differences among sites (df =8, F = 13.383, p = 0.0001). In particular site 5 was the only site in which at least half the dialogues had an average user words per turn greater than 4. One hypothesis is that this may be due to the fact that site 5 uses more open prompts, both at the beginning of the dialogue, e.g. *Tell me about your travel plans* and at other phases of the dialogue. For example, when system 5 was having trouble understanding the user, it would make open-ended suggestions such as *Try asking for flights between two major cities* rather than using directive prompts such as *Please tell me your destination.* However, this hypothesis remains to be tested.

Our analysis identified several issues with the 2000 data collection. The first issue was the within-subjects design. We thought this would allow us to make comparisons across systems, but we believe this design may result in using behavior reflecting the least common denominator; as users called one system after another, they accommodated their behavior to the least flexible system. Another effect of the within-subjects design is that users did not learn the interaction paradigm of any of the systems since each dialogue of the nine dialogues the user participated in was with a different system. A second issue was the tabular presentation of the fixed scenarios; users took very little initiative and it is possible that the tabular presentation format lead them to believe a conversation is simply filling in the slots in the table. A third issue was that users doing the open scenarios were more likely to change their task midstream (20% vs. 5%); thus these scenarios did not approximate users planning real trips. A fourth issue is that the variation in task complexity in the fixed scenarios was insufficient to investigate how performance might degrade as task complexity increases. We expect to address these problems in several ways.

The 2001 data collection is a longitudinal experiment (6 months) where users repeatedly use the same system. This should more closely approximate the real conditions of use and users should be able to learn how to use the systems as well as providing system designers an opportunity to explore algorithms for system adaptation to users. Second, all users are frequent travelers who call their system to plan real trips. There will be both SHORT and LONG users. The LONG users will perform 4 fixed learning scenarios in the beginning of the data collection; this will provide data for adaptation algorithms and will create an expert population. Third, we plan to use audio presentation of the learning tasks to address the problems of tabular presentation while avoiding the problem of putting words into the user's mouth. The experimental design is described in more detail on the Evaluation Committee web page [42].

A final issue is that methods clearly need to be developed to further automate some aspects of evaluation in order to reduce the cost of doing evaluation. One large cost is in the hand transcription of user utterances. Another cost is labeling and site adjudication required for evaluating user behavior and task completion. In current work, we are applying DATE (Dialogue Act Tagging for Evaluation) to automate the extraction of measures related to task completion and dialogue behaviors [39]. DATE includes a comprehensive set of dialogue act labels providing complete coverage of the system utterances occurring in the 2000 evaluation. The way in which we envision DATE supplementing the current set of metrics can be briefly illustrated for the Task Completion and Task Duration metrics. For Task Completion, alternatives to the hand labeling and site adjudication mentioned above could greatly minimize evaluation costs. When we add metrics derived from DATE labelling to the core metrics for the 2000 corpus, the dialogue act with the single highest coefficient in the PARADISE linear model for user satisfaction is the quantitative measure of the system's utterances that present price information (PriceC). Since price information is presented when a full flight itinerary has been selected, it serves as a landmark for having completed a certain proportion of the task dialogue for some systems. Thus DATE can potentially supplement hand labeled measures of task completion reducing the cost of hand labeling. Regarding Task Duration, the current measure does not differentiate between proportion of a dialogue devoted directly to completing a travel plan with the user, versus ancillary subdialogues consisting of instructions, attempts to recover from errors, or handling failure to retrieve relevant database information. Because DATE includes labels for system instructions, and differentiates between tutorials provided up front versus instructions provided at need, DATE could potentially be used to provide a more informative set of duration measures, such as relative proportion of dialogues on such subdialogues. Our initial results show enhanced fit of User Satisfaction using metrics derived from DATE [48]. In related work, manual semantic tagging of dialogues has been used to find error subdialogues automatically [1]. Given the semantic tagging, concept accuracy measures can be calculated for a dialogue on a per utterance basis. This can also be used to automatically identify the site of a system understanding as well as the location of the correction of the error. Manual dialogue act tagging has also been used to compare human-human dialogues in the travel domain with a subset of the human-computer dialogues analyzed here [11] and to evaluate system performance [7]. This work is closely related to the goals of our automatic DATE tagging. In future work, we hope to be able to further demonstrate that metrics derived from dialogue act tagging can facilitate a deeper understanding of the differences between the systems discussed here.

8. Acknowledments

This work was partially funded by DARPA under contract MDA972-99-3-0003.

References

[1] J. Aberdeen, C. Doran, L. Damianos, S. Bayer, and L. Hirschman. Finding errors automatically in semantically tagged dialogues. In *Human Language Technology Conference*, 2001.

[2] John Aberdeen. Darpa communicator logfile standard, 2000. http://fofoca.mitre.org/logstandard.

[3] Paolo Baggia, Giuseppe Castagneri, and Morena Danieli. Field Trials of the Italian ARISE Train Timetable System. In *Interactive Voice Technology for Telecommunications Applications, IVTTA*, pages 97–102, 1998.

[4] H. Bonneau-Maynard, L. Devillers, and S. Rosset. Predictive performance of dialog systems. In *Language Resources and Evaluation Conference*, 2000.

[5] G. Bouwman and J. Hulstijn. Dialogue strategy redesign with reliability measures. In *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain, 1998, pages 191-198.*, 1998.

[6] S. Boyce and A. L. Gorin. User interface issues for natural spoken dialogue systems. In *Proceedings of International Symposium on Spoken Dialogue*, pages 65–68, 1996.

[7] Roldana Cattoni, Morena Danieli, Andrea Panizza, V. Sandrini, and C. Soria. Building a corpus of annotated dialogues: the ADAM experience. In *Proc. of the Conference Corpus-Linguistics-2001, Lancaster, U.K.*, 2001.

[8] M. Danieli, W. Eckert, N. Fraser, N. Gilbert, M. Guyomard, P. Heisterkamp, M. Kharoune, J. Magadur, S. McGlashan, D. Sadek, J. Siroux, and N. Youd. Dialogue manager design evaluation. Technical Report Project Esprit 2218 SUNDIAL, WP6000-D3, 1992.

[9] M. Danieli and E. Gerbino. Metrics for evaluating dialogue strategies in a spoken language system. In *Proceedings of the 1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, pages 34–39, 1995.

[10] M. Denecke. An integrated development environment for spoken dialogue systems. In *Workshop on Toolsets in NLP, Coling, Saarbrucken*, 2000.

[11] Christine Doran, John Aberdeen, Laurie Damianos, and Lynette Hirschman. Comparing several aspects of human-computer and human-human dialogues. In *SIGDIAL Workshop in conjuction with Eurospeech 2001*, 2001.

[12] L. Hirschman, M. Bates, D. Dahl, W. Fisher, J. Garofolo, D. Pallett, K. Hunicke-Smith, P. Price, A. Rudnicky, and E. Tzoukermann. Multi-site data collection and evaluation in spoken language understanding. In *Proceedings of the Human Language Technology Workshop*, pages 19–24, 1993.

[13] Lynette Hirschman. Evaluating spoken language interaction: Experiences from the darpa spoken language program 1990–1995. In S. Luperfoy, editor, *Spoken Language Discourse*. MIT Press, Cambridge, Mass., 2000.

[14] M.A. Jack, J. C. Foster, and F. W. Stentiford. Intelligent dialogues in automated telephone services. In *International Conference on Spoken Language Processing, ICSLP*, pages 715 – 718, 1992.

[15] C. A. Kamm and M. A. Walker. Design and evaluation of spoken dialog systems. In *Proceedings of the ASRU Workshop*, 1997.

[16] C. A. Kamm, M. A. Walker, and L. R. Rabiner. The role of speech processing in human-computer intelligent communication. *Speech Communication*, 1997.

[17] Candace Kamm. User interfaces for voice applications. In David Roe and Jay Wilpon, editors, *Voice Communication between Humans and Machines*, pages 422–442. National Academy Press, 1995.

[18] L. Lamel. Spoken language dialog system development and evaluation at limsi. In *Proceedings of the International Symposium on Spoken Dialogue*, pages 9–17, 1998.

[19] Lori Lamel and Sophie Rosset. Considerations in the design and evaluation of spoken language dialog systems. In *ICSLP*, 2000.

[20] Lars Bo Larsen. Combining objective and subjective data in evaluation of spoken dialogues. In *ESCA Workshop on Interactive Dialogue in Multi-Modal Systems*, pages 89–92, 1999.

[21] E. Levin, R. Pieraccini, and W. Eckert. A stochastic model of human machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 1:11–23, 2000.

[22] Esther Levin, S Narayanan, R. Pieraccini, K. Biatov, E. Bocchieri, G. Di-Fabbrizio, W. Eckert, S. Lee, A. Pokrovsky, M. Rahim, P. Ruscitti, and M. Walker. The at&t darpa communicator mixed-initiative spoken dialog system. In *Proceedings of the International Conference on Spoken Language Processing,ICSLP00*, 2000.

[23] Diane J. Litman and Shimei Pan. Empirically Evaluating an Adaptable Spoken Dialogue System. In *Proceedings of the 7th International Conference on User Modeling*, 1999.

[24] Stephen Love, R. T Dutton, J. C. Foster, M. A. Jack, and F. W. M. Stentiford. Identifying salient usability attributes for automated telephone services. In *International Conference on Spoken Language Processing, ICSLP*, pages 1307–1310, 1994.

[25] David S. Pallett. Performance assessment of automatic speech recognizers. *J. Res. Natl. Bureau of Standards*, 90:371–387, 1985.

[26] J. Polifroni and S. Seneff. Galaxy-II as an architecture for spoken dialogue evaluation. In *Second International Conference on Language Resources and Evaluation (LREC)*, 2000.

[27] A. Potamianos, E. Ammicht, and J. Kuo. Dialogue management in the bell labs communicator system. In *Proc. ICSLP (Beijing, China), Oct. 2000.*, 2000.

[28] Patti Price, Lynette Hirschman, Elizabeth Shriberg, and Elizabeth Wade. Subject-based evaluation measures for interactive spoken language systems. In *Proceedings of the DARPA Speech and NL Workshop*, pages 34–39, 1992.

[29] A. Rudnicky, C. Bennett, A. Black, A. Chotimongkol, K. Lenzo, A. Oh, and R. Singh. Task and domain specific modelling in the carnegie mellon communicator system. In *ICSLP2000 (Beijing, China., 2000)*, 2000.

[30] A. Rudnicky, E. Thayer, P. Constantinides, C. Tchou, R. Shern, K. Lenzo, W. Xu, and A. Oh. Creating natural dialogs in the carnegie mellon communicator system. In *Eurospeech*, pages 1531–1534, 1999.

[31] M. D. Sadek, A. Ferrieux, A. Cosannet, P. Bretier, F. Panaget, and J. Simonin. Effective human-computer cooperative spoken dialogue: The ags demonstrator. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, pages 169–173, 1996.

[32] R. San-Segundo, B. Pellom, W. Ward, and J.M. Pardo. Confidence measures for dialogue management in the CU communicator system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-2000)*, 2000.

[33] S. Seneff, R. Lau, and J. Polifroni. Organization, communication, and control in the galaxy-ii conversational system. In *Proc. Eurospeech 99*, 1999.

[34] S. Seneff and J. Polifroni. Dialogue management in mercury flight reservation system. In *Satellite Dialogue Workshop, ANLP-NAACL*, 2000.

[35] S. Seneff and J. Polifroni. Formal and natural language generation in the mercury conversationalsystem, 2000.

[36] Elizabeth Shriberg, Elizabeth Wade, and Patti Price. Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In *Proceedings of the DARPA Speech and NL Workshop*, pages 49–54, 1992.

[37] D. Stallard. Dialog management in the talk'n'travel system. In *Automatic Speech Recognition and Understanding Workshop*, Madonna Di Campiglio, Trento, Italy, 2001.

[38] David Stallard. Talk'n'travel: A conversational system for air travel planning. In *In Proceedings of the 6th Applied Natural Language Processing Conference and the 1st Meeting of the North American Chapter of the Association of Computational Linguistics (ANLP-NAACL 2000)*, pages 68–75, 1998.

[39] M. Walker and R. Passonneau. DATE: A dialogue act tagging scheme for evaluation. In *Human Language Technology Conference*, 2001.

[40] M. Walker, O. Rambow, and M. Rogati. Spot: A trainable sentence planner. In *Proceedings of the North American Meeting of the Association for Computational Linguistics*, 2001.

[41] M. A. Walker, D. Litman, C. A. Kamm, and A. Abella. PARADISE: A general framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association of Computational Linguistics, ACL/EACL 97*, pages 271–280, 1997.

[42] Marilyn Walker. Communicator evaluation committee webpage, 2000. http://www.research.att.com/info/walker/eval/eval.html.

[43] Marilyn Walker, Julie Boland, and Candace Kamm. The utility of elapsed time as a usability metric for spoken dialogue systems. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, ASRUU99*, 1999.

[44] Marilyn Walker, Donald Hindle, Jeanne Fromer, Giuseppe Di Fabbrizio, and Craig Mestel. Evaluating competing agent strategies for a voice email agent. In *Proceedings of the European Conference on Speech Communication and Technology, EUROSPEECH97*, 1997.

[45] Marilyn Walker, Candace Kamm, and Julie Boland. Developing and testing general models of spoken dialogue system performance. In *Proc. Language Resources and Evaluation Conference, LREC-2000*, 2000.

[46] Marilyn A. Walker. An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research*, 12:387–416, 2000.

[47] Marilyn A. Walker, Candace A. Kamm, and Diane J. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering: Special Issue on Best Practice in Spoken Dialogue Systems*, 2000.

[48] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. Quantitative and qualitative evaluation of DARPA communicator spoken dialogue systems. In *Proceedings of the Meeting of the Association of Computational Lingustics, ACL 2001*, 2001.

[49] W. Ward and B. Pellom. The CU communicator system. In *Proc. Workshop on Automatic Speech Recognition and Understanding*, 1999.

[50] Steve Whittaker and Phil Stenton. User studies and the design of natural language systems. In *Proc. 4th Conference of the European Chapter of the ACL, Association of Computational Linguistics*, pages 116–123, 1989.

[51] S. Wilks. *Mathematical Statistics*. Wiley, 1962.

Speaker	Utterance
SYS1	Welcome to Talk and Travel. Before we begin, let's go over a few simple instructions. First, always wait to hear the beep before you say anything. If I make a mistake, you can correct me, or just say: scratch that, or, back up. You can always start over again completely just by saying: start over. OK, now we'll begin! What trip would you like to take?
USR1	STARTING FROM ATLANTA GEORGIA TO LONDON ENGLAND
SYS2	flying from atlanta to london. What day are you leaving Atlanta?
USR2	MONDAY SEPTEMBER TWENTY FIVE
SYS3	on monday, september twenty fifth. About what time do you want to leave?
USR3	[UH] DAYTIME
SYS4	Leaving in the daytime. Please stand by while I look up the flight schedules on the web. OK, got them. I have 13 flights. The first flight is on American at six fifty nine eh M, arriving at ten forty five PM, with a connection in Chicago. Is that OK?
USR4	[UH] NO TOO EARLY
SYS5	leaving after six fifty nine eh m. I have 12 flights. The first flight is on Delta at seven ten eh M, arriving at ten forty five PM, with a connection in Chicago. Is that OK?
USR5	[UH] NO
SYS6	The second flight is on Delta at one forty PM, arriving on Tuesday, September twenty sixth at six twenty eh M, with a connection in New York. Is that OK?
USR6	[UH] NO I WANT TO DEPART [UH] DAYTIME
SYS7	leaving in the daytime. I have 13 flights. The first flight is on American at six fifty nine eh M, arriving at ten forty five PM, with a connection in Chicago. Is that OK?
USR7	[UH] YEAH THAT'LL BE OKAY
SYS8	OK, I've got you on American flight six fifty three departing William B. Hartsfield International on Monday, September twenty fifth at six fifty nine eh M. You connect with American ninety in Chicago, departing at nine eh M, arriving Heathrow at ten forty five PM. Do you want a return flight from London to Atlanta?
USR8	[UH] YES
SYS9	Returning from london to atlanta. What day are you leaving London?
USR9	[UH] LEAVING [UM] MONDAY OCTOBER THE SECOND
SYS10	on monday, october second. About what time do you want to leave?
USR10	ANY TIME'S OKAY
SYS11	Any time. Please stand by while I look up the flight schedules on the web... OK, got them. I have 23 flights. The first flight is on United at four fifty PM, arriving at eleven thirty one PM. . Is that OK?
USR11	YES
SYS12	OK, I've got you on United flight nine twenty five departing Heathrow on Monday, October second at four fifty PM, arriving William B. Hartsfield International at eleven thirty one PM. Alright, I think that completes your itinerary. Do you want me to read it back to you?
USR12	[UH] YES ONE TIME
SYS13	I've got you on American flight six fifty three departing William B. Hartsfield International on Monday, September twenty fifth at six fifty nine eh M. You connect with American ninety in Chicago, departing at nine eh M, arriving Heathrow at ten forty five PM. You continue on United flight nine twenty five departing Heathrow on Monday, October second at four fifty PM, arriving William B. Hartsfield International at eleven thirty one PM. Is there anything more I can help you with on this trip?
USR13	[UH] NO THAT'S FINE
SYS14	Goodbye, and thank you for using Talk and Travel

Figure 20: Sample BBN Dialogue for Atlanta,London Round Trip

Speaker	Utterance
SYS1	Welcome to the initial development version of the MIT air travel planning system this call is being recorded for system development you may hang up or ask for help at any time. Please say your pin number.
USR1	ONE THREE NINE ZERO
SYS2	I'll enroll you temporarily as a guest user. How can I help you?
USR2	I'D LIKE TO TRAVEL FROM ATLANTA GEORGIA TO LONDON ON SEPTEMBER TWENTY FIFTH
SYS2	okay, from Atlanta to London on Monday September 25. Can you provide the approximate departure time or airline?
USR3	[uh] [uh] DURING THE DAYTIME
SYS4	Sorry, I'm not sure what you said.
USR4	DURING THE DAY
SYS5	I'm sorry. I'm having trouble understanding you. Try saying a short sentence.
USR5	I'D LIKE TO DEPART DURING THE DAY
SYS6	My mistake again. I'm having trouble understanding you. Try asking for flights between two major cities.
USR6	FLIGHTS BETWEEN ATLANTA GEORGIA AND LONDON
SYS7	okay, from Atlanta to London on Monday September 25. Can you provide the approximate departure time or airline?
USR7	[um] TEN A M
SYS8	What about a Delta flight connecting in New York's La Guardia leaving at 9:30 a.m.? Would that work? I also have a Continental flight connecting in Newark leaving at 8:30 a.m.
USR8	DO YOU HAVE ANY DIRECT FLIGHTS
SYS9	What about a British Airways flight leaving at 4:10 pm? Would that work? I also have a Delta flight leaving at 5:40 p.m.
USR9	[UH] BRITISH I+D PREFER THE BRITISH AIRWAYS FLIGHT
SYS10	Sorry, I'm not sure what you said.
USR10	I+D LIKE THE BRITISH AIRWAYS FLIGHT
SYS11	I have 2 nonstop flights arriving in London: a flight leaving at 4:10 p.m. and a flight leaving at 6: 20 p.m. Are any of these okay?
USR11	YES THE FOUR TEN P M
SYS12	British Airways flight 2228 has been added to your itinerary. you depart Atlanta at 4:10 pm and arrive in London at 5:35 am the following day. Where would you like to go next?
USR12	[um] I'D LIKE TO RETURN FROM LONDON TO ATLANTA
SYS13	What date will you be returning on?
USR13	OCTOBER SECOND
SYS14	I have 2 nonstop flights departing from London: a flight leaving at 9:45 a.m. and a flight leaving at 12:10 p.m. Would one of these work?
USR14	[um] THE TWELVE TEN P M
SYS15	I heard you say you want to depart from Detroit. Is that correct?
USR15	NO I WANT TO DEPART FROM LONDON TO ATLANTA ON THE TWELVE O+CLOCK FLIGHT
SYS16	I have a nonstop British Airways flight leaving at 12:10 p.m. and arriving at 4:20 pm? Would that work?
USR16	YES
SYS17	British Airways flight 2227 has been added to your itinerary. you depart London at 12:10 pm and arrive in Atlanta at 4:20 pm. Would you like me to find out how much your trip will cost?
USR17	[uh] YES PLEASE
SYS18	the cheapest price for your trip is $901.70. Is there something else I can do for you?
USR18	NO THANK YOU
SYS19	Thanks for using Mercury. Have a nice day

Figure 21: MIT Sample Dialogue for Atlanta,London Round Trip