Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialogue Systems

Marilyn A. Walker
AT&T Labs – Research
180 Park Ave, E103
Florham Park, NJ. 07932
walker@research.att.com

Rebecca Passonneau
AT&T Labs –Research
180 Park Ave, D191
Florham Park, NJ. 07932
becky@research.att.com

Julie E. Boland
Institute of Cognitive Science
University of Louisiana at Lafayette
Lafayette, LA 70504
boland@louisiana.edu

Abstract

This paper describes the application of the PARADISE evaluation framework to the corpus of 662 human-computer dialogues collected in the June 2000 Darpa Communicator data collection. We describe results based on the standard logfile metrics as well as results based on additional qualitative metrics derived using the DATE dialogue act tagging scheme. We show that performance models derived via using the standard metrics can account for 37% of the variance in user satisfaction, and that the addition of DATE metrics improved the models by an absolute 5%.

1 Introduction

The objective of the DARPA COMMUNICATOR program is to support research on multi-modal speech-enabled dialogue systems with advanced conversational capabilities. In order to make this a reality, it is important to understand the contribution of various techniques to users' willingness and ability to use a spoken dialogue system. In June of 2000, we conducted an exploratory data collection experiment with nine participating communicator systems. All systems supported travel planning and utilized some form of mixed-initiative interaction. However the systems varied in several critical dimensions: (1) They targeted different back-end databases for travel information; (2) System modules such as ASR, NLU, TTS and dialogue management were typically different across systems.

The Evaluation Committee chaired by Walker (Walker, 2000), with representatives from the nine COMMUNICATOR sites and from NIST, developed the experimental design. A logfile standard was developed by MITRE along with a set of tools for processing the logfiles (Aberdeen, 2000); the standard and tools were used by all sites to collect a set of core metrics for making cross system comparisons. 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 experiment was designed to make it possible to apply the PARADISE evaluation framework (Walker et al., 2000), which integrates and unifies previous approaches to evaluation (Price et al., 1992; Hirschman, 2000). 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. Our results from applying PARADISE include that user satisfaction differed considerably across the nine systems. Subsequent modeling of user satisfaction gave us some insight into why each system was more or less satisfactory; four variables accounted for 37% of the variance in user-satisfaction: task completion, task duration, recognition accuracy, and mean system turn duration.

However, when doing our analysis we were struck by the extent to which different aspects of the systems' dialogue behavior weren't captured by the core metrics. For example, the core metrics logged the number and duration of system turns, but didn't distinguish between turns used to request or present information, to give instruc-

tions, or to indicate errors. Recent research on dialogue has been based on the assumption that dialogue acts provide a useful way of characterizing dialogue behaviors (Reithinger and Maier, 1995; Isard and Carletta, 1995; Shriberg et al., 2000; Di Eugenio et al., 1998). Several research efforts have explored the use of dialogue act tagging schemes for tasks such as improving recognition performance (Reithinger and Maier, 1995; Shriberg et al., 2000), identifying important parts of a dialogue (Finke et al., 1998), and as a constraint on nominal expression generation (Jordan, 2000). Thus we decided to explore the application of a dialogue act tagging scheme to the task of evaluating and comparing dialogue systems.

Section 2 describes the corpus. Section 3 describes the dialogue act tagging scheme we developed and applied to the evaluation of COMMUNICATOR dialogues. Section 4 first describes our results utilizing the standard logged metrics, and then describes results using the DATE metrics. Section 5 discusses future plans.

2 The Communicator 2000 Corpus

The corpus consists of 662 dialogues from nine different travel planning systems with the number of dialogues per system ranging between 60 and 79. The experimental design is described in (Walker et al., 2001). Each dialogue consists of a recording, a logfile consistent with the standard, transcriptions and recordings of all user utterances, and the output of a web-based user survey. Metrics collected per call included:

- **Dialogue Efficiency**: Task Duration, System turns, User turns, Total Turns
- **Dialogue Quality**: Word Accuracy, Response latency, Response latency variance
- **Task Success**: Exact Scenario Completion
- **User Satisfaction**: Sum of TTS performance, Task ease, User expertise, Expected behavior, Future use.

The objective metrics focus on measures that can be automatically logged or computed and a web survey was used to calculate User Satisfaction (Walker et al., 2001). A ternary definition of task completion, Exact Scenario Completion (ESC) was annotated by hand for each call by annotators at AT&T. The ESC metric distinguishes between exact scenario completion (ESC), any scenario completion (ANY) and no scenario completion (NOCOMP). This metric arose because some callers completed an itinerary other than the one assigned. This could have been due to users' inattentiveness, e.g. users didn't correct the system when it had misunderstood them. In this case, the system could be viewed as having done the best that it could with the information that it was given. This would argue that task completion would be the sum of ESC and ANY. However, examination of the dialogue transcripts suggested that the ANY category sometimes arose as a rational reaction by the caller to repeated recognition error. Thus we decided to distinguish the cases where the user completed the assigned task, versus completing some other task, versus the cases where they hung up the phone without completing any itinerary.

3 Dialogue Act Tagging for Evaluation

The hypothesis underlying the application of dialogue act tagging to system evaluation is that a system's dialogue behaviors have a strong effect on the usability of a spoken dialogue system. However, each COMMUNICATOR system has a unique dialogue strategy and a unique way of achieving particular communicative goals. Thus, in order to explore this hypothesis, we needed a way of characterizing system dialogue behaviors that could be applied uniformly across the nine different communicator travel planning systems. We developed a dialogue act tagging scheme for this purpose which we call DATE (Dialogue Act Tagging for Evaluation).

In developing DATE, we believed that it was important to allow for multiple views of each dialogue act. This would allow us, for example, to investigate what part of the task an utterance contributes to separately from what speech act function it serves. Thus, a central aspect of DATE is that it makes distinctions within three orthogonal dimensions of utterance classification: (1) a SPEECH-ACT dimension; (2) a TASK-SUBTASK dimension; and (3) a CONVERSATIONAL-DOMAIN dimension. We believe that these distinctions are important for using such a scheme for evaluation. Figure 1 shows a COMMUNICATOR dialogue with each system ut-

terance classified on these three dimensions. The tagset for each dimension are briefly described in the remainder of this section. See (Walker and Passonneau, 2001) for more detail.

3.1 Speech Acts

In DATE, the SPEECH-ACT dimension has ten categories. We use familiar speech-act labels, such as OFFER, REQUEST-INFO, PRESENT-INFO, ACKNOWLEDGE, and introduce new ones designed to help us capture generalizations about communicative behavior in this domain, on this task, given the range of system and human behavior we see in the data. One new one, for example, is STATUS-REPORT. Examples of each speech-act type are in Figure 2.

Speech-Act	Example
REQUEST-INFO	*And, what city are you flying to?*
PRESENT-INFO	*The airfare for this trip is 390 dollars.*
OFFER	*Would you like me to hold this option?*
ACKNOWLEDGE	*I will book this leg.*
STATUS-REPORT	*Accessing the database; this might take a few seconds.*
EXPLICIT-CONFIRM	*You will depart on September 1st. Is that correct?*
IMPLICIT-CONFIRM	*Leaving from Dallas.*
INSTRUCTION	*Try saying a short sentence.*
APOLOGY	*Sorry, I didn't understand that.*
OPENING/CLOSING	*Hello. Welcome to the C M U Communicator.*

Figure 2: Example Speech Acts

3.2 Conversational Domains

The CONVERSATIONAL-DOMAIN dimension involves the domain of discourse that an utterance is about. Each speech act can occur in any of three domains of discourse described below.

The ABOUT-TASK domain is necessary for evaluating a dialogue system's ability to collaborate with a speaker on achieving the task goal of making reservations for a specific trip. It supports metrics such as the amount of time/effort the system takes to complete a particular phase of making an airline reservation, and any ancillary hotel/car reservations.

The ABOUT-COMMUNICATION domain reflects the system goal of managing the verbal channel and providing evidence of what has been understood (Walker, 1992; Clark and Schaefer, 1989). Utterances of this type are frequent in human-computer dialogue, where they are motivated by the need to avoid potentially costly errors arising from imperfect speech recognition. All implicit and explicit confirmations are about communication; See Figure 1 for examples.

The SITUATION-FRAME domain pertains to the goal of managing the culturally relevant framing expectations (Goffman, 1974). The utterances in this domain are particularly relevant in human-computer dialogues because the users' expectations need to be defined during the course of the conversation. About frame utterances by the system attempt to help the user understand how to interact with the system, what it knows about, and what it can do. Some examples are in Figure 1.

3.3 Task Model

The TASK-SUBTASK dimension refers to a task model of the domain task that the system supports and captures distinctions among dialogue acts that reflect the task structure.[1] The motivation for this dimension is to derive metrics that quantify the effort expended on particular subtasks.

This dimension distinguishes among 14 subtasks, some of which can also be grouped at a level below the top level task.[2], as described in Figure 3. The TOP-LEVEL-TRIP task describes the task which contains as its subtasks the ORIGIN, DESTINATION, DATE, TIME, AIRLINE, TRIP-TYPE, RETRIEVAL and ITINERARY tasks. The GROUND task includes both the HOTEL and CAR subtasks.

Note that any subtask can involve multiple speech acts. For example, the DATE subtask can consist of acts requesting, or implicitly or explicitly confirming the date. A similar example is provided by the subtasks of CAR (rental) and HOTEL, which include dialogue acts requesting, confirming or acknowledging arrangements to rent a car or book a hotel room on the same trip.

[1] This dimension elaborates of each speech-act type in other tagging schemes (Reithinger and Maier, 1995).

[2] In (Walker and Passonneau, 2001) we didn't distinguish the price subtask from the itinerary presentation subtask.

Task	Example
TOP-LEVEL-TRIP	*What are your travel plans?*
ORIGIN	*And, what city are you leaving from?*
DESTINATION	*And, where are you flying to?*
DATE	*What day would you like to leave?*
TIME	*Departing at what time?.*
AIRLINE	*Did you have an airline preference?*
TRIP-TYPE	*Will you return to Boston from San Jose?*
RETRIEVAL	*Accessing the database; this might take a few seconds.*
ITINERARY	*I found 3 flights from Miami to Minneapolis.*
PRICE	*The airfare for this trip is 390 dollars.*
GROUND	*Did you need to make any ground arrangements?.*
HOTEL	*Would you like a hotel near downtown or near the airport?.*
CAR	*Do you need a car in San Jose?*

Figure 3: Example Utterances for each Subtask

3.4 Implementation and Metrics Derivation

We implemented a dialogue act parser that classifies each of the system utterances in each dialogue in the COMMUNICATOR corpus. Because the systems used template-based generation and had only a limited number of ways of saying the same content, it was possible to achieve 100% accuracy with a parser that tags utterances automatically from a database of patterns and the corresponding relevant tags from each dimension.

A summarizer program then examined each dialogue's labels and summed the total effort expended on each type of dialogue act over the dialogue or the percentage of a dialogue given over to a particular type of dialogue behavior. These sums and percentages of effort were calculated along the different dimensions of the tagging scheme as we explain in more detail below.

We believed that the top level distinction between different domains of action might be relevant so we calculated percentages of the total dialogue expended in each conversational domain, resulting in metrics of TaskP, FrameP and CommP (the percentage of the dialogue devoted to the task, the frame or the communication domains respectively).

We were also interested in identifying differences in effort expended on different subtasks. The effort expended on each subtask is represented by the sum of the length of the utterances contributing to that subtask. These are the metrics: TripC, OrigC, DestC, DateC, TimeC, AirlineC, RetrievalC, FlightinfoC, PriceC, GroundC, BookingC. See Figure 3.

We were particularly interested developing metrics related to differences in the system's dialogue strategies. One difference that the DATE scheme can partially capture is differences in confirmation strategy by summing the explicit and implicit confirms. This introduces two metrics ECon and ICon, which represent the total effort spent on these two types of confirmation.

Another strategy difference is in the types of about frame information that the systems provide. The metric CINSTRUCT counts instances of instructions, CREQAMB counts descriptions provided of what the system knows about in the context of an ambiguity, and CNOINFO counts the system's descriptions of what it doesn't know about. SITINFO counts dialogue initial descriptions of the system's capabilities and instructions for how to interact with the system

A final type of dialogue behavior that the scheme captures are apologies for misunderstanding (CREJECT), acknowledgements of user requests to start over (SOVER) and acknowledgments of user corrections of the system's understanding (ACOR).

We believe that it should be possible to use DATE to capture differences in initiative strategies, but currently only capture differences at the task level using the task metrics above. The TripC metric counts open ended questions about the user's travel plans, whereas other subtasks typically include very direct requests for information needed to complete a subtask.

We also counted triples identifying dialogue acts used in specific situations, e.g. the utterance *Great! I am adding this flight to your itinerary* is the speech act of acknowledge, in the about-task domain, contributing to the booking subtask. This combination is the ACKBOOKING metric. We also keep track of metrics for dialogue acts of acknowledging a rental car booking or a hotel booking, and requesting, presenting or confirming particular items of task information. Below we describe dialogue act triples that are significant predictors of user satisfaction.

Metric	Coefficient	P value
ESC	0.45	0.000
TaskDur	-0.15	0.000
Sys Turn Dur	0.12	0.000
Wrd Acc	0.17	0.000

Table 1: Predictive power and significance of Core Metrics

4 Results

We initially examined differences in cumulative user satisfaction across the nine systems. An ANOVA for user satisfaction by Site ID using the modified Bonferroni statistic for multiple comparisons showed that there were statistically significant differences across sites, and that there were four groups of performers with sites 3,2,1,4 in the top group (listed by average user satisfaction), sites 4,5,9,6 in a second group, and sites 8 and 7 defining a third and a fourth group. See (Walker et al., 2001) for more detail on cross-system comparisons.

However, our primary goal was to achieve a better understanding of the role of qualitative aspects of each system's dialogue behavior. We quantify the extent to which the dialogue act metrics improve our understanding by applying the PARADISE framework to develop a model of user satisfaction and then examining the extent to which the dialogue act metrics improve the model (Walker et al., 2000). Section 4.1 describes the PARADISE models developed using the core metrics and section 4.2 describes the models derived from adding in the DATE metrics.

4.1 Results using Logfile Standard Metrics

We applied PARADISE to develop models of user satisfaction using the core metrics; the best model fit accounts for 37% of the variance in user satisfaction. The learned model is that User Satisfaction is the sum of Exact Scenario Completion, Task Duration, System Turn Duration and Word Accuracy. Table 1 gives the details of the model, where the coefficient indicates both the magnitude and whether the metric is a positive or negative predictor of user satisfaction, and the P value indicates the significance of the metric in the model.

The finding that metrics of task completion and recognition performance are significant predictors duplicates results from other experiments applying PARADISE (Walker et al., 2000). The fact that task duration is also a significant predictor may indicate larger differences in task duration in this corpus than in previous studies.

Metric	Coefficient	P value
ESC (Completion)	0.40	0.00
Task Dur	-0.31	0.00
Sys Turn Dur	0.14	0.00
Word Accuracy	0.15	0.00
TripC	0.09	0.01
BookingC	0.08	0.03
PriceC	0.11	0.00
AckRent	0.07	0.05
EconTime	0.05	0.13
ReqDate	0.10	0.01
ReqTripType	0.09	0.00
Econ	0.11	0.01

Table 2: Predictive power and significance of Dialogue Act Metrics

Note that the PARADISE model indicates that system turn duration is *positively* correlated with user satisfaction. We believed it plausible that this was due to the fact that flight presentation utterances are longer than other system turns. Thus this metric simply captures whether or not the system got enough information to present some potential flight itineraries to the user. We investigate this hypothesis further below.

4.2 Utilizing Dialogue Parser Metrics

Next, we add in the dialogue act metrics extracted by our dialogue parser, and retrain our models of user satisfaction. We find that many of the dialogue act metrics are significant predictors of user satisfaction, and that the model fit for user satisfaction increases from 37% to 42%. The dialogue act metrics which are significant predictors of user satisfaction are detailed in Table 2.

When we examine this model, we note that several of the significant dialogue act metrics are calculated along the task-subtask dimension, namely TripC, BookingC and PriceC. One interpretation of these metrics are that they are acting as landmarks in the dialogue for having achieved a particular set of subtasks. The TripC metric can be interpreted this way because it includes open ended questions about the user's travel plans both at the beginning of the dialogue and also after

one itinerary has been planned. Other significant metrics can also be interpreted this way; for example the ReqDate metric counts utterances such as *Could you tell me what date you wanna travel?* which are typically only produced after the origin and the destination have been understood. The ReqTripType metric counts utterances such as *From Boston, are you returning to Dallas?* which are only asked after all the first information for the first leg of the trip have been acquired, and in some cases, after this information has been confirmed. The AckRental metric has a similar potential interpretation; the car rental task isn't attempted until after the flight itinerary has been accepted by the caller. However, the predictors for the models already include a ternary exact scenario completion metric (ESC) which specifies whether any task was achieved or not, and whether the exact task that the user was attempting to accomplish was achieved. The fact that the addition of these dialogue metrics improves the fit of the user satisfaction model suggests that perhaps a finer grained distinction on how many of the subtasks of a dialogue were completed is related to user satisfaction. This makes sense; a user who the system hung up on immediately should be less satisfied than one who never could get the system to understand his destination, and both of these should be less satisfied than a user who was able to communicate a complete travel plan but still did not complete the task.

Other support for the task completion related nature of some of the significant metrics is that the coefficient for ESC is smaller in the model in Table 2 than in the model in Table 1. Note also that the coefficient for Task Duration is much larger. If some of the dialogue act metrics that are significant predictors are mainly so because they indicate the successful accomplishment of particular subtasks, then both of these changes would make sense. Task Duration can be a greater negative predictor of user satisfaction, only when it is counteracted by the positive coefficients for subtask completion.

The TripC and the PriceC metrics also have other interpretations. The positive contribution of the TripC metric to user satisfaction could arise from a user's positive response to systems with open-ended initial greetings which give the user the initiative. The positive contribution of the PriceC metric might indicate the users' positive response to getting price information, since not all systems provided price information.

As mentioned above, our goal was to develop metrics that captured differences in dialogue strategies. The positive coefficient of the Econ metric appears to indicate that an explicit confirmation strategy overall leads to greater user satisfaction than an implicit confirmation strategy. This result is interesting, although it is unclear how general it is. The systems that used an explicit confirmation strategy did not use it to confirm each item of information; rather the strategy seemed to be to acquire enough information to go to the database and then confirm all of the parameters before accessing the database. The other use of explicit confirms was when a system believed that it had repeatedly misunderstood the user.

We also explored the hypothesis that the reason that system turn duration was a predictor of user satisfaction is that longer turns were used to present flight information. We removed system turn duration from the model, to determine whether FlightInfoC would become a significant predictor. However the model fit decreased and FlightInfoC was not a significant predictor. Thus it is unclear to us why longer system turn durations are a significant positive predictor of user satisfaction.

5 Discussion and Future Work

We showed above that the addition of dialogue act metrics improves the fit of models of user satisfaction from 37% to 42%. Many of the significant dialogue act metrics can be viewed as landmarks in the dialogue for having achieved particular subtasks. These results suggest that a careful definition of transaction success, based on automatic analysis of events in a dialogue, such as acknowledging a booking, might serve as a substitute for the hand-labelling of task completion.

In current work we are exploring the use of tree models and boosting for modeling user satisfaction. Tree models using dialogue act metrics can achieve model fits as high as 48% reduction in error. However, we need to test both these models and the linear PARADISE models on unseen data. Furthermore, we intend to explore methods

for deriving additional metrics from dialogue act tags. In particular, it is possible that sequential or structural metrics based on particular sequences or configurations of dialogue acts might capture differences in dialogue strategies.

We began a second data collection of dialogues with COMMUNICATOR travel systems in April 2001. In this data collection, the subject pool will use the systems to plan real trips that they intend to take. As part of this data collection, we hope to develop additional metrics related to the quality of the dialogue, how much initiative the user can take, and the quality of the solution that the system presents to the user.

6 Acknowledgements

This work was supported under DARPA GRANT MDA 972 99 3 0003 to AT&T Labs Research. Thanks to the evaluation committee members: J. Aberdeen, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Narayanan, K. Papineni, B. Pellom, A. Potamianos, A. Rudnicky, G. Sanders, S. Seneff, and D. Stallard who contributed to 2000 COMMUNICATOR data collection.

References

John Aberdeen. 2000. Darpa communicator logfile standard. http://fofoca.mitre.org/logstandard.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Barbara Di Eugenio, Pamela W. Jordan, Johanna D. Moore, and Richmond H. Thomason. 1998. An empirical investigation of collaborative dialogues. In *ACL-COLING98, Proc. of the 36th Conference of the Association for Computational Linguistics*.

M. Finke, M. Lapata, A. Lavie, L. Levin, L. Mayfield Tomokiyo, T. Polzin, K. Ries, A. Waibel, and K. Zechner. 1998. Clarity: Inferring discourse structure from speech. In AAAI Symposium on Applying Machine Learning to Discourse Processing.

Erving Goffman. 1974. *Frame Analysis: An Essay on the Organization of Experience*. Harper and Row, New York.

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

Amy Isard and Jean C. Carletta. 1995. Replicability of transaction and action coding in the map task corpus. In *AAAI Spring Symposium: Empirical Methods in Discourse Interpretation and Generation*, pages 60–67.

Pamela W. Jordan. 2000. *Intentional Influences on Object Redescriptions in Dialogue: Evidence from an Empirical Study*. Ph.D. thesis, Intelligent Systems Program, University of Pittsburgh.

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

Norbert Reithinger and Elisabeth Maier. 1995. Utilizing statistical speech act processing in verbmobil. In *ACL 95*.

E. Shriberg, P. Taylor, R. Bates, A. Stolcke, K. Ries, D. Jurafsky, N. Coccaro, R. Martin, M. Meteer, and C. Van Ess-Dykema. 2000. Can prosody aid the automatic classification of dialog acts in conversational speech. *Language and Speech: Special Issue on Prosody and Conversation*.

M. Walker and R. Passonneau. 2001. Date: A dialogue act tagging scheme for evaluation. In *Human Language Technology Conference*.

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

M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. 2001. Darpa communicator dialog travel planning systems: The june 2000 data collection. In *Submitted to EUROSPEECH 2001*.

Marilyn A. Walker. 1992. Redundancy in collaborative dialogue. In *Fourteenth International Conference on Computational Linguistics*, pages 345–351.

Marilyn Walker. 2000. Communicator evaluation committee webpage. http://www.research.att.com/ walker/eval/eval.html.

Speaker	Utterance	Speech Act	Task	C-Domain
SYS	Welcome.	opening/closing	nil	task
	You are logged in as a guest user.	instruct	nil	frame
	You may say repeat, help me out, start over, or, that's wrong,	instruct	nil	frame
	you can also correct and interrupt the system at any time.	instruct	nil	frame
	What airport woodja like to fly out of?	request-info	origin	task
USER	MIAMI FLORIDA			
SYS	Leaving from Miami,	implicit-conf	origin	communication
	And, what city are you flying to?	request-info	dest	task
USER	MINNEAPOLIS MINNESOTA			
SYS	Flying from Miami to Minneapolis,	implicit-conf	origin, dest	communication
	What date would you like to fly?	request-info	date	task
USER	UH MONDAY OCTOBER TWENTY THIRD			
SYS	Leaving on the 23rd of October,	implicit-conf	date	communication
	And what time didja wanna leave?	request-info	time	task
USER	UH LATE MORNING			
SYS	From Minneapolis.	implicit-conf	origin	communication
	Are you returning to Miami?	request-info	trip-type	task
USER	UM			
SYS	Are you going back to Miami?	request-info	trip-type	task
	Please say yes or no.	instruct	nil	frame
USER	NO			
SYS	Okay, it will just take a moment while I look.	status-report	retrieval	task
SYS	I found 3 flights from Miami to Minneapolis on October 23rd .	present-info	itinerary	task
	For option one I have a flight with United departing at 10 A M and arriving at 3 P M with one stopover and			
	Ticket price is 390 dollars.	present-info	price	task
	Please say next option, or, flight details, or, I'll take it.	offer	itinerary	task
USER	NEXT OPTION			
SYS	For option two I have a flight with Delta departing at 12 and arriving at 4 P M non-stop.	present-info	itinerary	task
	Ticket price is 450 dollars.	present-info	price	task
	Would you like me to hold this option?	offer	itinerary	task
USER	NEXT OPTION			
SYS	For the last option I have a flight with Northwest departing at 2 P M and arriving at 7 P M with 1 stopover and	present-info	itinerary	task
	Ticket price is 470 dollars.	present-info	price	task
	say next option, or, flight details, or, I'll take it.	offer	itinerary	task
USER	NEXT OPTION			
SYS	Here is your third option again. a flight with Northwest departing at 2 P M and arriving at 7 P M with 1 stopover and	present-info	itinerary	task
	Ticket price is 470 dollars.	present-info	price	task
	say next option, or, flight details, or, I'll take it.	offer	itinerary	task
USER	I'LL TAKE IT			
SYS	Great! I am adding this flight to your itinerary.	acknowledge	booking	task
	Okay, Is Minneapolis your final destination?	request-info	trip-type	task
USER	YES IT IS			
SYS	Would you like to make any ground arrangements?	offer	ground	task
USER	NO			
SYS	Now, would you like to hear a summary of your itinerary?	offer	itinerary	task
USER	YES PLEASE			
SYS	Okay, Here's your itinerary. On October 23rd Northwest Flight 123 departs Miami at 2 P M, arrives Pittsburgh at 4 30 P M connecting to Northwest Flight 146 that departs Pittsburgh at 5 P M and arrives Minneapolis at 7 P M.	present-info	itinerary	task

Figure 1: Dialogue Illustrating the Speech Act, Task-Subtask and Conversational Domain Dimensions of DATE