Comparing Several Aspects of Human-Computer and Human-Human Dialogues

Christine Doran, John Aberdeen, Laurie Damianos and **Lynette Hirschman**
The MITRE Corporation
202 Burlington Road
Bedford, MA 01730 USA
{cdoran,aberdeen,laurie,lynette}@mitre.org

Abstract

While researchers have many intuitions about the differences between human-computer and human-human interactions, most of these have not previously been subject to empirical scrutiny. This work presents some initial experiments in this direction, with the ultimate goal being to use what we learn to improve computer dialogue systems. Working with data from the air travel domain, we identified a number of striking differences between the human-human and human-computer interactions.

1 Introduction

In our initial experiments comparing human-human (HH) and human-computer (HC) interaction we have annotated dialogues from the air travel domain with several sets of tags: dialogue act, initiative and unsolicited information. Our aim is to begin an empirical exploration of how these aspects of the dialogue shed light on differences between HH and HC interactions. We found striking differences between the human-human and human-computer interactions. With many of the issues we examine here, researchers have voiced strong intuitions about the differences between HH and HC communication, but these intuitions have not previously been subject to empirical scrutiny.

Why do we want to compare HH and HC interactions? We believe that an examination of the differences between HH and HC dialogues can help those working on the HC interactions to improve their systems. This will not necessarily mean making the HC interactions "more like" HH interactions; rather, we believe that such analysis can give us insights about the appropriateness and success of various communicative approaches in different settings. We are also interested in quantifying what it means for a dialogue to be "mixed-initiative". There is liberal use of this term in work on human-computer dialogues, but there does not seem to be a clear sense of what it really means and how to define it.

This work is being done in the context of the Communicator Travel task. Communicator is a DARPA-funded program involving major industry and academic sites, established to provide the next generation of intelligent conversational interfaces to distributed information. The goal of the program is to move beyond the current system initiated voice menu style of dialogue interaction towards a more flexible strategy of shared conversational control between human and system. Work up to the fall of 2000 concentrated on the travel task, but groups are now moving into other domains with their dialogue systems.

2 Our Data

We have tagged 20 HH dialogues and 40 HC dialogues; the HH dialogues consist of 25,208 words and 1742 turns, and the HC dialogues consist of 8313 words and 957 turns. The HH dialogues are from the SRI ATIS collection (available at `www.ai.sri.com/~communic/amex/amex.html`) and the CMU Travel data (`www.speech.cs.cmu.edu`). The HC dialogues are from a subset of the systems participating in the DARPA Communicator data collection which was run in June and July of 2000.[1] Subjects were selected by the U.S. National Institute of Standards and Technology (NIST) and each subject was paid to call all of the 9 participating sites with either a fixed or open travel itinerary. The fixed itineraries were intended to balance flight types (one-way, round-trip, multi-leg) and destinations (large US airport, small for-

[1] Ideally, of course, we would want HH dialogues collected under similar circumstances to the HC dialogues. However, no such data are available and collecting new HH data is not currently a feasible endeavor for us.

eign airport, etc.).[2] We made no attempt to balance the subset of data we analyzed; we merely selected 10 random users. We then identified four sites whose systems we expected to differ quite widely (based on researchers' descriptions of their systems), and extracted the 10 users' dialogues for each site. Each dialogue was tagged with three sets of annotations—for dialogue acts, for initiative and for provision of unsolicited information by users (HC only). Sections 2.1, 2.2 and 3.5 describe these in more detail. The passage below is excerpted from one of the human-computer interactions, using the transcribed versions of the user's turns.[3]

```
User said: i would like to fly from miami
 florida to sydney australia on wednesday
 october eleventh please
System said: okay, from Miami to Sydney on
 Wednesday October 11. Can you provide the
 approximate departure time or airline?
User said: [um] i prefer american airlines
 and i can leave anytime
System said: There are no nonstop flights.
 I have a flight connecting in Los Angeles
 leaving at 7:15 a.m. a flight connecting
 in Los Angeles leaving at 5:20 p.m. and a
 flight connecting in Los Angeles leaving
 at 7:30 p.m. Do you like any of these?
User said: yes how about the morning one
System said: American flight 912 connecting
 to American flight 7362 in Los Angeles has
 been added to your itinerary. you depart
 Miami at 7:15 am and arrive in Sydney at
 9:40 pm the following day. Where would you
 like to go next?
```

2.1 Dialogue Act Tagging

We have done two rounds of tagging. Our starting point was a 20-tag subset of the CSTAR Consortium tags (Consortium, 1999). We found the CSTAR tag set to be the most suitable for our task of the tagging schemes we considered, based on information provided in the survey of dialogue act annotation schemes assembled by the MATE project (Klein et al., 1998). We picked the CSTAR tags because they had been designed for task-oriented dialogues, the tag categories were sufficiently clear and simple that we believed we would be able to tag the data reliably and, finally, the categories captured the sorts of distinctions we believed would be relevant. We rejected the DAMSL tag set (Core and Allen, 1997; Core et al., 1999) on the grounds that is was too sophisticated for our purposes, covering many aspects of dialogue structure that were not necessarily relevant for our task such as intentionality, grounding and context tracking. In addition, the interannotator agreement levels reported for this scheme are quite low. Some of the other tag sets we considered were (Carletta et al., 1995; Nakatani et al., 1995; van Vark et al., 1996; Di Eugenio et al., 1998; Jurafsky et al., 1997).

In collaboration with AT&T, we arrived at a set of changes to our tag set that would make it compatible with their efforts to tag system utterances automatically (Walker and Passonneau, 2001), in the hopes of being able to share results with them more easily. We added a situation/conversation/task distinction to a number of our tags (e.g. GIVE-INFORMATION split into GIVE-TASK-INFO, GIVE-SITUATION-INFO and GIVE-CONVERSATION-INFO). We also added a NOT-UNDERSTAND tag and collapsed some original tags into super-categories. Our revised tag set had 26 tags, and two people (one who had also done the first round of tagging) tagged the same data set. The situation/conversation/task distinction turned out to be extremely difficult for the taggers to make; we believe that revisions to the tagging guidelines could lead to some improvement on this front, but without enumerating the kinds of utterances which fall into each category, this will remain a difficult task.

We tagged each utterance that contained some speech, i.e. was not composed entirely of non-speech annotation like *pause* or [click], and we split turns[4] into utterances using guidelines that had been developed internally for another purpose. Utterances on this definition were roughly clause-sized units, and possibly fragmentary.[5] This meant that there were often multiple dialogue acts (DAs) per turn, and where there were multiple sequential DAs of the same type, we collapsed them under a single tag on the assumption that they were combining to "perform" that DA. We initially split some of the CSTAR tags

[2] It is important to note that the fixed itineraries were assigned, i.e. not real trips, and it was also evident from the transcripts that few if any of the open itineraries represented real trips. This may well have had some impact on the data that were collected.

[3] Both human-human and human-system dialogue transcripts are reproduced with their native formatting, i.e. all caps, no punctuation, etc. The system side contains punctuation resulting from the generation process.

[4] Chunk of text labelled with either *User said* or *Expert said*. It was possible for a single speaker to have more than one sequential turn, i.e. turn $\neq$ speaker change.

[5] In hindsight, it would have been preferable to segment the dialogues in a separate step.

into IMPLICIT and EXPLICIT versions, but found that the IMPLICIT cases were so hard to identify that we were not using those tags, and they were dropped from the tag set.

Tables 1 and 2 show roughly parallel subdialogues from the HH and HC data.[6] Each turn is tagged with its DA, and the first expert turn in Table 2 shows multiple DAs within a turn, a GIVE-INFORMATION followed by an OFFER.

Expert:WHAT TIME DO YOU NEED TO DEPART	[req-task-info]
User:AS SOON AS POSSIBLE AFTER FIVE P.M.	[give-task-info]
Expert:THE FIRST FLIGHT AFTER FIVE P.M. ON THAT DATE IS AT FIVE THIRTY FIVE P.M. ARRIVING IN CHICAGO AT SIX OH SIX P.M. ON U.S. AIR	[give-task-info]
User: IS THAT O'HARE	[req-task-info]

Table 1: DA tagging in an HH Exchange

Expert: i have an American Airlines flight departing Seattle at twelve fifty five p.m., arrives Tokyo at three p.m. the next day.	[give-task-info]
Is that OK?	[offer]
User: yes I'll take it	[accept]
Expert: Will you return to seattle from tokyo?	[req-task-info]
User: what airport	[req-task-info]
Expert: Will you return to seattle from tokyo?	[req-task-info]

Table 2: DA tagging in an HC Exchange

With our first tag set, our Kappa score for interannotator agreement on these dialogues is 0.90 (with two annotators). Not surprisingly, our Kappa score on the second, more complex tag set (cf. Table 10 for a list of the tags) was lower, 0.71 (0.74 on the HC data and 0.66 on the HH data). Both scores are in line with scores reported in similar tagging tasks (Klein et al., 1998): 0.56 for DAMSL (overall average), 0.83 for Maptask (experienced coders), 0.8-0.84 for Switchboard DAMSL and 0.83 for VerbMobil. The drop in score between our two tag sets emphasizes an issue which we continue to wrestle with–the tradeoff between tag set complexity and tagging accuracy. At what point is it more useful to have reliable results from an impoverished tag set than results of questionable value from a sophisticated tag set?

[6]Throughout the paper, we will use *expert* to refer to either the human or the computer travel agent, *system* to refer exclusively to the computer travel agent, and *user* to refer to the travelers.

2.2 Initiative Tagging

There is not a clearly agreed upon definition of *initiative* in the literature on dialogue analysis (but see e.g., (Chu-Carroll and Brown, 1998; Jordan and Di Eugenio, 1997; Flammia and Zue, 1997)), despite the fact the terms *initiative* and *mixed-initiative* are widely used. Intuitively, it seems that control rests with the participant who is moving a conversation ahead at a given point, or selecting new topics for conversation.

After experimenting with several tagging methods, we concluded that the approach presented in Walker and Whittaker (1990) adopted from (Whittaker and Stenton, 1988) best captured the aspects of the dialogue we were interested in and, as with the DAs, could be tagged reliably on our data.

Each turn is tagged with which participant has control at the end of that turn, based on the utterance type. Again, we did not tag turns composed entirely of non-speech annotation, and we also excluded conventional openings and closings, following Walker and Whittaker. Below, we list the rules for tagging each utterance type; a PROMPT is an utterance "which did not express propositional content, such as *Yeah*, *Okay*, *Uh-huh*, ..." (Op cit, p. 3) The classification refers to the illocutionary force of the item, rather than to its particular syntactic form.

Assertion: speaker has initiative unless it is a response to a question *or command*[7]

Question: speaker has initiative unless it is a response to a question or command

Command: speaker has initiative

Prompt: hearer has initiative

Tables 3 and 4 show the same passages used above, but this time tagged for initiative. To give a sense of how the tagging rules are applied, let us step through the HC example (Table 4). Turn (1) is assigned EXPERT-INITIATIVE, because it is an assertion which is not a response to any preceding question or command. Turn (2) is still EXPERT-INITIATIVE, because it is an answer to the question *Is that OK?*. The third turn is a question and EXPERT-INITIATIVE, but turn (4) is USER-INITIATIVE because it is a question that is not a response to the previous question. The system

[7]Italics show our modification to the rule.

does not address the user's question, but rather repeats its own question, so the final turn (5) is EXPERT-INITIATIVE.

Expert:	WHAT TIME DO YOU NEED TO DEPART	[exp-init]
User:	AS SOON AS POSSIBLE AFTER FIVE P.M.	[exp-init]
Expert:	THE FIRST FLIGHT AFTER FIVE P.M. ON THAT DATE IS AT FIVE THIRTY FIVE P.M. ARRIVING IN CHICAGO AT SIX OH SIX P.M. ON U.S. AIR	[exp-init]
User:	IS THAT O'HARE	[user-init]

Table 3: Initiative tagging in an HH Exchange

(1)	Expert: i have an American Airlines flight departing Seattle at twelve fifty five p.m. , arrives Tokyo at three p.m. the next day. Is that OK?	[exp-init]
(2)	User: yes I'll take it	[exp-init]
(3)	Expert: Will you return to seattle from tokyo?	[exp-init]
(4)	User: what airport	[user-init]
(5)	Expert: Will you return to seattle from tokyo?	[exp-init]

Table 4: Initiative tagging in an HC Exchange

Our Kappa scores for interannotator agreement on the initiative tagging were somewhat lower than for DA tagging. Here, κ=0.68. In fact, our agreement was rather high, at 87%, but because there were so few instances of user initiative in the HC dialogues, our agreement would have to be exceptional to be reflected in a higher Kappa score. While we had believed this to be the easier task, with quite clear guidelines and only a binary tagging choice, it in fact proved to be quite difficult. We still believe that this tag set can give us useful insights into our data, but we would be interested in attempting further revisions to the tagging guidelines, particularly as regards the definition of an "answer", i.e. when an answer is responsive and when it is not.

3 Analysis

We found a number of interesting differences between the HH and HC dialogues. While we have not yet been able to test our hypotheses about why these differences appear, we will discuss our ideas about them and what sorts of further work we would like to do to subject those ideas to empirical validation.

3.1 Initiative Distribution

Based on researchers' descriptions of their systems (i.e. for the most part, "highly mixed-initiative"), we had expected to find some variance in the distribution of initiative across systems. As is evident from Table 5, the HC systems do not differ much from each other, but taken as whole, the dialogues differ dramatically from the HH dialogues. In the HH dialogues, users and expert share the initiative relatively equitably, while in the HC data the experts massively dominate in taking the initiative. Here, we are simply counting the number of turns tagged as USER-INITIATIVE or EXPERT-INITIATIVE.[8]

We also show turns to completion and overall user satisfaction scores for each system as a reference point. User satisfaction was calculated from five questions asked of each user after each dialogue. The questions use a 5-point Likert scale. Turns to completion measures the total number of on-task turns. We found no significant correlations here, but cf. Walker et al. (2001) which provides more detailed analyses of the Communicator dialogues using user satisfaction and other metrics, within the PARADISE framework. It is worth noting, however, that the HC D has both the highest percentage of expert initiative and the highest satisfaction scores, so we should not conclude that more initiative will necessarily lead to happier users.

	% Exp Init	% User Init	Turns to Comp	User Sat
HC A	86.8%	13.2%	40.5	60.0%
HC B	89.9%	10.1%	41.4	71.5%
HC C	90.6%	9.4%	36.0	68.5%
HC D	93.7%	6.3%	43.9	82.8%
HH SRI	48.3%	51.7%	N/A	N/A
HH CMU	54.0%	46.0%	N/A	N/A

Table 5: Percentages of User and Expert Initiative in HH and HC Dialogues

In the HC dialogues, we also see a difference in success rate for user-initiative turns. By our definition, the user "succeeds" in taking the initiative in the dialogue if the system responds to the initiative on the first possible turn. The rate of success

[8] A cautionary note is warranted here. We are not suggesting that more user-initiative is intrinsically preferable; it may well turn out to be the case that a completely system-directed dialogue is more pleasant/efficient/etc. Rather, we are seeking to quantify and assess what it means to be "mixed-initiative" so that we can better evaluate the role of initiative in effective (task-oriented) dialogues.

is the ratio of successful user-initiatives attempts to total user-initiatives attempts. There appears to be a negative relationship between number of initiative attempts and their success rate. See Figure 1, below. HC D has a high success rate for a relatively small number of user-initiative attempts. HC A has many more occurrences of user initiative, but does not incorporate them as well.

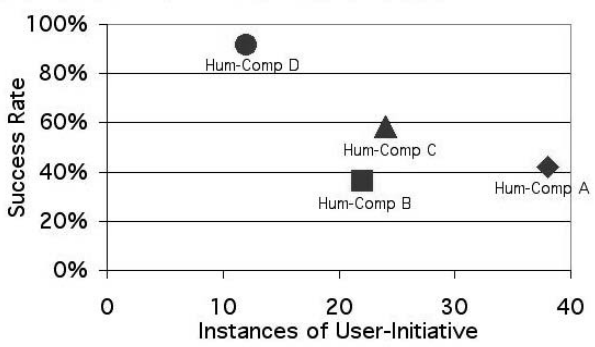

Figure 1: User-Initiative and Success Rate per System

There is no determinable relationship between user experience (i.e., the number of calls per systems) and either the amount of user-initiative or the success rate of user-initiative.

We also looked at user-initiative with respect to dialogue act type. Most user-initiatives are REQUEST-ACTION (26%) and REQUEST-INFORMATION (19%). REQUEST-INFORMATION dialogue acts (e.g., *What cities do you know in Texas?*, *Are there any other flights?*, *Which airport is that?*) are handled well by the systems (83% success rate) while REQUEST-ACTION dialogue acts (e.g., *start over*, *scratch that*, *book that flight*) are not (48%). Most of the user-initiatives that are REQUEST-ACTION dialogue acts are the *start over* command (16% of the total user-initiatives). Corrections to flight information presented by the systems consist of 20% of the total user-initiatives.

3.2 Overall Verbosity

In counting the number of words used, we find that the computer experts are much more verbose than their human users, and are relatively more verbose than their human travel agent counterparts. In the HH dialogues, experts average 10.1 words/turn, while users average 7.2. In the HC dialogues on average, system have from 16.65-33.1 words/turn vs. the users' 2.8-4.8 words/turn. Figure 2 shows these differences for each of the four systems and for the combined HH data.

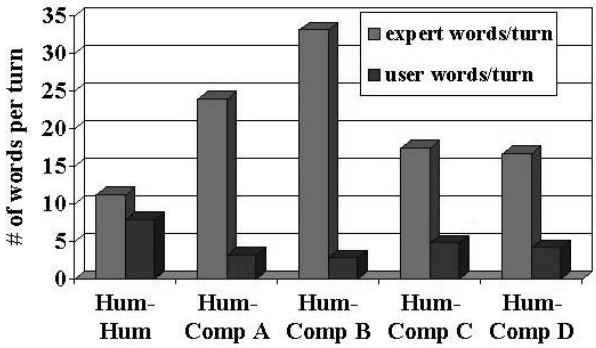

Figure 2: Words per turn for users and experts in the HH and HC dialogues

3.2.1 Short vs. Long Confirmations

One DA which is a basic conversational tool and therefore an interesting candidate for analysis is the use of confirmations. Instances of short confirmation, typically back-channel utterances such as *okay* and *uh huh* were tagged as ACKNOWLEDGE, while instances of long confirmation, as when one participant explicitly repeats something that the other participant has said, were tagged as VERIFY-X, where X=CONVERSATION-ACTION, TASK-INFORMATION and TASK-ACTION, This tagging allows us to easily calculate the distribution of short and long confirmations.

Overall we found in the HC dialogues a rather different confirmation profile from the HH dialogues. In the HC dialogues, the systems use both types of confirmation far more than the users do (246 total system, 8 total user). Moreover, systems use long confirmation about five times more often (210 vs. 36) than they use short confirmation. In contrast, the experts in the HH dialogues use somewhat more confirmations than users (247 vs. 173), but both parties use far more short than long confirmations (340 vs. 80), just the reverse of the HC situation. This difference partially accounts for the total word count differences we saw in the previous section. Tables 6 and 7 show the breakdowns in these numbers for each system and for the two sets of HH data, and begin to quantify the striking contrasts between human and computer confirmation strategies.

3.3 Number of Dialogue Acts

Another observation is that the computer experts appear to be trying to do more. They have significantly more DAs per turn than do their human users, whereas in the HH dialogues, the two participants have nearly the same number of DAs per turn (just over 1.3). In the HC dialogues, sys-

Site	Expert	User	Total
HC A	3 (0.5%)	4 (0.7%)	7 (1.2%)
HC B	13 (1.9%)	0 (0.0%)	13 (1.9%)
HC C	20 (3.1%)	3 (0.5%)	23 (3.6%)
HC D	0 (0.0%)	0 (0.0%)	0 (0.0%)
HH SRI	95 (16.1%)	79 (13.3%)	174 (29.4%)
HH CMU	94 (12.1%)	72 (9.3%)	166 (21.4%)

Table 6: Number of short confirmations, i.e. ACKNOWLEDGE (percentage of total dialogue acts)

Site	Expert	User	Total
HC A	32 (5.7%)	0 (0.0%)	32 (5.7%)
HC B	74 (10.6%)	0 (0.0%)	74 (10.6%)
HC C	59 (9.2%)	1 (0.2%)	60 (9.4%)
HC D	45 (8.6%)	0 (0.0%)	45 (8.6%)
HH SRI	11 (1.9%)	11 (1.9%)	22 (3.7%)
HH CMU	47 (6.1%)	11 (1.4%)	58 (7.5%)

Table 7: Number of long confirmations i.e. VERIFY-X (percentage of total dialogue acts)

tems have, on average 1.6 DAs per turn where users have just 1.0, as Figure 3 shows. If we take a DA as representing a single dialogue "move", then users in the HC dialogues are managing one move per turn, where the systems have at least one and often more. A common sequence for the computer experts is a VERIFY-TASK-INFORMATION followed by a REQUEST-TASK-INFORMATION, such as *A flight to Atlanta. What city are you departing from?*.

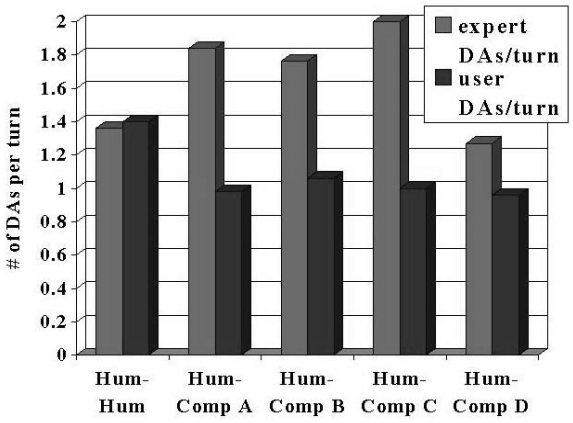

Figure 3: Dialogue acts per turn for users and experts in the HH and HC dialogues

3.4 Types of Dialogue Acts

One of our main questions going into this work was whether there would be interestingly different distributions of DAs in the HH and HC dialogues, and whether different distributions of DAs across systems would be correlated with user satisfaction. Unfortunately, we do not have user satisfaction scores for the HH data, but if new data were to be collected, this would be an essential addition.

Tables 8 and 9 illustrate some of the main differences between the HH and HC dialogues, and as regards our first research question, definitely give an interesting view of the differences between the HH and HC conversations.

DA	Overall	Expert	User
GiveTaskInfo	27.7%	29.7%	25.5%
Acknowledge	24.9%	26.9%	22.7%
RequestTaskInfo	11.0%	10.7%	11.4%
VerifyTaskInfo	5.4%	7.5%	3.2%
Affirm	4.8%	4.3%	5.4%

Table 8: Five most frequent DAs in Human-Human dialogues, by percent of total DAs for column

DA	Overall	Expert	User
GiveTaskInfo	23.7%	12.9%	46.3%
RequestTaskInfo	15.3%	22.1%	1.3%
Offer	7.7%	11.5%	0.0%
VerifyTaskInfo	7.1%	10.5%	0.1%
Apology	4.5%	6.6%	0.1%

Table 9: Five most frequent DAs in Human-Computer dialogues, by percent of total DAs for column

As expected in this domain, all DAs involving exchange of task information (GIVE-TASK-INFO, REQUEST-TASK-INFO, and VERIFY-TASK-INFO are frequent in both sets of dialogues. However, in the HH dialogues, ACKNOWLEDGE (e.g. the tag for back-channel responses and general confirmations such as *right*, *uh huh* and *okay*) is the second most common DA, and does not even appear in the top five for the HC dialogues. The DA for positive responses, AFFIRM, is also in the top ranking for the HH dialogues, but does not appear in the list for the HC dialogues. Finally, OFFER and APOLOGY appear frequently in the HC dialogues and not in the top HH DAs. The appearance of these two is a clear indication that the systems are doing things quite differently from their human counterparts.

Turning to differences between experts and users in these top categories, we can see that human users and experts are about equally likely to ask for or give task-related information (GIVE-TASK-INFO and REQUEST-TASK-INFO). In contrast, in the HC dialogues nearly half of the users' DAs are giving task information and hardly any

are requesting such information, while almost a quarter of expert DAs are requesting information. There is some inequity in the use of VERIFY-TASK-INFO in the HH dialogues, where experts perform about twice as many verifications as users; however, in the HC dialogues, virtually all verification is done by the expert. All of these patterns reinforce our finding about initiative distribution; in the HC dialogues, one disproportionately finds the expert doing the asking and verification of task information, and the user doing the answering, while in the HH dialogues the exchange of information is much more balanced.

DA	HC A	HC B	HC C	HC D
accept	3.9%	3.1%	4.8%	3.4%
acknowledge	1.2%	1.9%	3.6%	**0.0%**
affirm	1.8%	2.4%	0.8%	**9.5%**
apologize	4.6%	3.7%	8.9%	**0.0%**
demand-conv-info	1.1%	0.0%	0.0%	0.0%
demand-sit-info	0.0%	1.6%	1.4%	1.3%
demand-task-info	**3.4%**	0.3%	0.0%	1.3%
give-sit-info	5.7%	6.3%	4.7%	1.9%
give-task-info	**34.8%**	16.0%	24.8%	20.8%
negate	2.1%	1.7%	0.8%	**5.2%**
not-understand	2.5%	3.7%	7.2%	**0.0%**
offer	**3.5%**	8.4%	9.4%	9.4%
open-close	2.3%	3.1%	4.8%	3.4%
please-wait	0.0%	6.2%	1.6%	3.1%
reject	1.1%	4.1%	0.3%	2.5%
req-conv-action	2.7%	4.4%	2.5%	1.0%
req-sit-action	1.1%	1.4%	0.2%	1.9%
req-sit-info	0.0%	3.3%	0.2%	3.2%
req-task-action	1.1%	1.4%	0.3%	0.2%
req-task-info	**17.9%**	12.6%	10.9%	**21.6%**
suggest-conv-action	1.6%	0.1%	2.0%	0.0%
thank	2.1%	3.4%	1.4%	1.7%
verify-conv-action	0.7%	0.7%	0.0%	0.0%
verify-task-action	2.5%	0.4%	1.9%	0.0%
verify-task-info	2.5%	9.4%	7.5%	8.6%
user satisfaction[9]	60.0%	71.5%	68.5%	82.8%

Table 10: Distribution of DAs by System

Table 10 gives an interesting snapshot of each system, in terms of its overall distribution of DAs. These numbers are reflective of the system designers' decisions for their systems, and that means all DAs are not going to be used by all systems (i.e. 0.0% may mean that that DA is not part of the system's repertoire).

We will concentrate here on the best and worst received systems in terms of their overall user satisfaction, HC D and HC A; the relevant numbers are boldfaced. They also have very different dialogue strategies, and that is partially reflected in the table. HC D's dialogue strategy does not make use of the 'social nicety' DAs employed by other systems (ACKNOWLEDGE, APOLOGIZE, NOT-UNDERSTAND), and yet it still had the highest user satisfaction of the four. This system also has the highest proportion of AFFIRM (more than three times as many as the next highest system) and REQ-TASK-INFO DAs, which suggests that quite a lot of information is being solicited and the *users* (because we know from Table 9 that it is primarily the users responding) are more often than average responding affirmatively. The fact that the percentage of GIVE-TASK-INFOs is somewhere in the middle of the range and AFFIRMs is so high may indicate that the HC D uses more yes/no than content questions.

Looking at the lower scoring system, HC A, we see very different patterns. HC A has most of the DEMAND-TASK-INFOs, the second highest percentage of REQ-TASK-INFOs and by far the most GIVE-TASK-INFOs, so its dialogue strategy must involve a large number of attempts to extract information from the user, and yet it has the fewest OFFER DAs, so these don't appear to be resulting in suggestions of particular travel options.

Turning to correlations between DA use by expert and user (combined across systems) and user satisfaction, we see some expected results but also some rather surprising correlations. Not unexpectedly, apologies and signals of non-understanding by the system are highly negatively correlated with satisfaction (-0.7 and -0.9, respectively). While it may seem counter-intuitive that OPEN-CLOSE by the user is negatively correlated (at -0.8), those familiar with this data will undoubtedly have noticed that users often try to say *Goodbye* repeatedly to try to end a dialogue that is going badly. Discussion of situational information (e.g. phone use) by the expert is highly negatively correlated, but by the user, the DA REQ-SITUATION-INFO is perfectly positively correlated. We cannot account for this finding.

3.5 Unsolicited Information

In the HC data we noticed that users often provided more information than was explicitly solicited–we call this 'unsolicited information'. For example, when a system asks for one piece of information, *On what day would you be departing Portland?*, the user might respond with additional information such as, *Thursday, October 5th before six pm from Portland back to Seattle.*

[9]This figure combines the scores on five user satisfaction questions. A perfect score is 100%.

78% of that unsolicited information is offered in response to open-ended questions (e.g., *How can I help you?* or *What are your travel plans?*). While our initiative tagging partially captures this, there are cases where the answer may be considered responsive (i.e. initiative does not shift away from the participant asking the question) and yet unsolicited information has been offered. Thus, this category is somewhat orthogonal to our characterization of initiative, although it is clearly one way of seizing control of the conversation.[10]

To get at this information, we developed a third tagging scheme for annotating unsolicited information. We began examining just the HC documents, because the phenomenon is prevalent in these data; we hope to perform a similar analysis on the HH data as well. We found that the systems we examined in general handle unsolicited information well. 70% of all unsolicited information is handled correctly by the systems, 22% is handled incorrectly, and the rest could not be accurately classified. Information offered in response to open-ended questions is handled correctly more often by the systems than unsolicited information offered at other points in the dialogue (74% versus 56%). The former figure is not surprising, since the systems are designed to handle "unsolicited" information following open-prompts. However, we were surprised the systems did as well as they did on unsolicited information in contexts where it was not expected. Figure 4 shows the relationship between frequency of various types of unsolicited information and how well the system incorporates that information. There appears to be some correlation between the frequency of unsolicited information and the rate of success, but we do not have enough data to make a stronger claim.

Furthermore, systems vary in response delay to pieces of unsolicited information. We define response delay as the number of system turns it takes before the information is acknowledged by the system (either correctly or incorrectly.) If a system responds immediately to the unsolicited information, a count of zero turns is recorded. Figure 5 shows the difference among systems in responding to unsolicited information. We graphed both the average total number of system turns as well as the average number of turns minus repetitions. HC B responds almost immediately to

[10]This issue may also be related to where in the dialogue errors occur. We are pursuing another line of research which looks at automatic error detection, described in (Aberdeen et al., 2001). We believe we may also be able to detect unsolicited information automatically, as well as to see whether it is likely to trigger errors by the system.

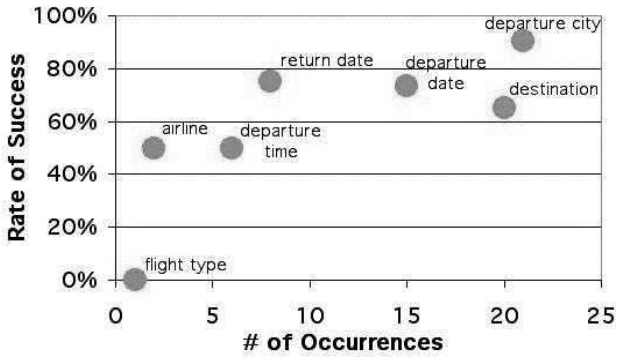

Figure 4: Unsolicited Fields vs. Success Rate of Incorporation

unsolicited information while HCs A and C take more turns to respond. HC D has trouble understanding the unsolicited information, and either keeps asking for clarification or continues to ignore the human and prompts for some other piece of information multiple times.

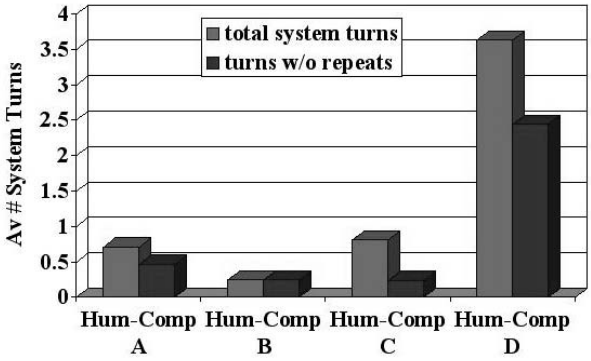

Figure 5: Variation of System Response to Unsolicited Information

Figure 6 shows the different rates at which systems acknowledge unsolicited information for different fields. For example, departure city is recognized and validated almost immediately. Return date and flight type are incorporated fairly quickly when the system understands what is being said.

If we look at the effects of experience on the amount of unsolicited information offered, as shown in Figure 7, we can see that users tend to provide more unsolicited information over time (i.e., as they make more calls to the systems). This effect may be the result of increased user confidence in the systems at handling unsolicited information. It also may be attributed to user boredom; as time goes on, users may be trying to finish the task as quickly as possible. Even if this is true, however, it demonstrates attempts by users to take more control of the interactions as

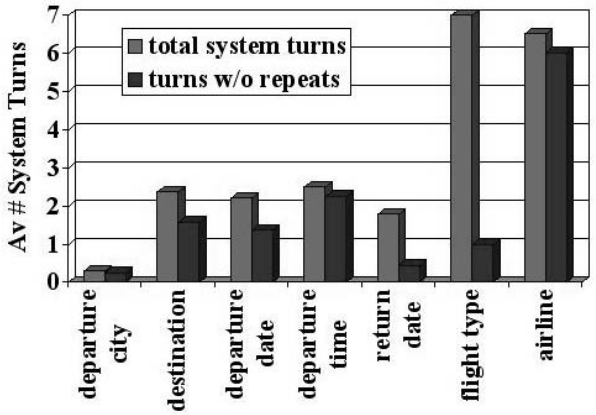

Figure 6: System Response to Different Types of Unsolicited Information

they become more experienced.

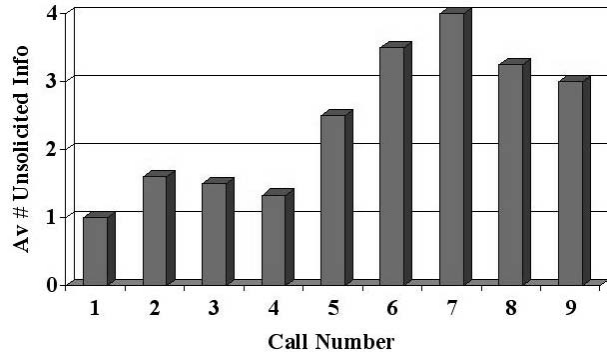

Figure 7: Effect of Experience on Unsolicited Information

Our data also show that the success rate of incorporating unsolicited information improves with user experience. The ratio of successes to failures increases in later calls to the systems (Figure 8).

4 Discussion

This was a relatively small study, but many of the results are sufficiently striking that we expect them to hold over large sets of dialogues. First, it is clear that (for our definition of the term) initiative is skewed towards the computer expert in the human-computer dialogues, despite claims of developers to the contrary. Whether this is desirable or not is a separate issue, but we believe it is a move forward to be able to quantify this difference. Second, there are clear differences in dialogue act patterns between the HH and HC dialogues. When the DAs correspond to basic dialogue moves, like questions or signals of agreement, we can begin to see how the dialogue dynamic is different in the human computer situation. In general, the conversation was much more balanced between traveler and expert in the HH setting, in terms of amount of speech, types of dialogue acts and with respect to initiative. In the HC conversations, the system dominated, in number of words and dialogue acts and in initiative.

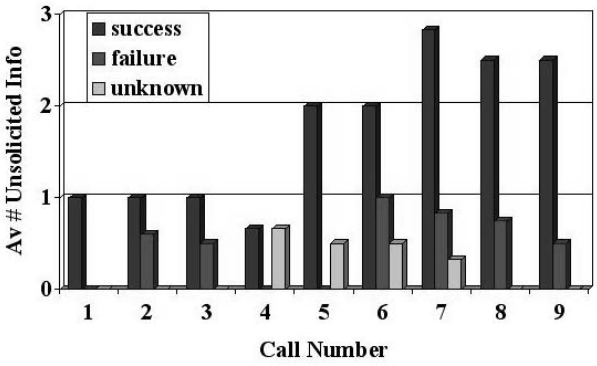

Figure 8: Experience versus Success Rate of Incorporating Unsolicited Information

We are very interested in the selection of the 'right' tag set for a given task. As we noted in our discussion of DA tagging, we had very different outcomes with two closely related tag sets. Clearly the choice of tag set is highly dependent on the use the tagged data will be put to, how easily the task can be characterized in the set of tagging guidelines, and what trade-offs in accuracy vs. richness of representation are acceptable.

A central question we are left with is "Why don't the users talk more in HC dialogues?" Is it that they are happy to just give short, specific answers to very directed questions? Or do they "learn" that longer answers are likely to cause the systems problems? Or perhaps users have preconceived notions (often justified) that the computer will not understand long utterances? We may speculate that poor speech recognition performance is a major factor shaping this behavior, leading system designers to attempt to constrain what users can say, while simultneously attempting to hold onto the initiative. (Walker et al. (2001) found sentence accuracy to be one of the significant predictors of user satisfaction in the Summer 2000 DARPA Communicator data collection.) There are some cases where the experts in the HC dialogues say things their human counterparts need not. One obvious case, which appears in even the small example dialogues we are using here, is that the systems tend to repeat utterances when there is some processing difficulty. In the same vein, errors and misunderstandings are more frequent in the HC data, resulting in (some

fairly verbose) efforts by the systems to identify the problem and get the conversation back on track.

5 Future Work

We are currently working with other Communicator sites who are also looking at dialogue issues. In addition, we are beginning to look at two new aspects of these dialogues: task complexity and conversational failure analysis (at the turn level, (Aberdeen et al., 2001)). We are also interested in examining patterns of initiative tags, i.e. control shift types and length of initiative runs, and at relations between DAs and user satisfaction.

6 Acknowledgments

Thanks to Lori Levin and Alon Lavie at CMU for sharing the CSTAR tagging guidelines and their sample tagged corpus.

References

J. Aberdeen, C. Doran, L. Damianos, S. Bayer, and L. Hirschman. 2001. Finding errors automatically in semantically tagged dialogues. In *Notebook Proceedings of the First International Conference on Human Language Technology Research*, San Diego, CA, March.

J. C. Carletta, A. Isard, S. Isard, J. Kowtko, G. Doherty-Sneddon, and A. Anderson. 1995. The coding of dialogue structure in a corpus. In J. A. Andernach, S. P. van de Burgt, and G. F. van der Hoeven, editors, *Proceedings of the Twente Workshop on Language Technology: Corpus-based approaches to dialogue modelling*, Enschede, The Netherlands. Universiteit Twente.

Jennifer Chu-Carroll and Michael K. Brown. 1998. An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction*, 8(3-4):215–253.

CSTAR Consortium. 1999. Dialogue act annotation. Unpublished Manuscript, October.

Mark Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *Proceedings of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA, November.

Mark Core, Masato Ishizaki, Johanna Moore, Christine Nakatani, Nobert Reithinger, David Traum, and Syun Tutiya, editors. 1999. *The Report of The Third Workshop of the Discourse Resource Initiative*, Chiba Univeristy. Technical Report No.3 CC-TR-99-1.

Barbara Di Eugenio, Pamela W. Jordan, and Liina Pylkknen. 1998. The COCONUT project: dialogue annotation manual. Technical Report ISP Technical Report 98-1, University of Pittsburgh, December.

Giovanni Flammia and Victor Zue. 1997. Learning the structure of mixed initiative dialogues using a corpus of annotated conversations. In *Proc. Eurospeech 97*, pages 1871–1874, Rhodes, Greece, September.

Pamela W. Jordan and Barbara Di Eugenio. 1997. Control and initiative in collaborative problem solving dialogues. In *AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction*, Stanford, CA.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. Technical Report Technical Report 97-02, University of Colorado Institute of Cognitive Science, August.

Marion Klein, Niels Ole Bernsen, Sarah Davies, Laila Dybkjær, Juanma Garrido, Henrik Kasch, Andreas Mengel, Vito Pirelli, Massimo Poesio, Silvia Quazza, and Claudia Soria, 1998. *Supported Coding Schemes, MATE Deliverable D1.1*, July. http://mate.nis.sdu.dk/.

Christine H. Nakatani, Barbara J. Grosz, David D. Ahn, and Julia Hirschberg. 1995. Instructions for annotating discourse. Technical Report TR-21-95, Harvard University.

R.J. van Vark, J.P.M. de Vreught, and L.J.M. Rothkrantz. 1996. Analysing ovr dialogues, coding scheme 1.0. Technical Report 96-137, Delft University of Technology.

Marilyn Walker and Rebecca Passonneau. 2001. Dialogue act tags as qualitative dialogue metrics for spoken dialogue systems. In *Notebook Proceedings of the First International Conference on Human Language Technology Research*, San Diego, CA, March.

Marilyn Walker and Steve Whittaker. 1990. Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of ACL90.*

M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayan, 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. Submitted., April.

Steve Whittaker and Phil Stenton. 1988. Cues and control in expert client dialogues. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics (ACL88)*, pages 123–130.