What would you ask a Conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting

Susan Robinson, David Traum, Midhun Ittycheriah, Joe Henderer
Institute for Creative Technologies/University of Southern California
13274 Fiji Way,
Marina del Rey, CA 90292
{robinson, traum, henderer}@ict.usc.edu, ittycher@usc.edu

Abstract

Embodied Conversational Agents have typically been constructed for use in limited domain applications, and tested in very specialized environments. Only in recent years have there been more cases of moving agents into wider public applications (e.g. Bell et al., 2003; Kopp et al., 2005). Yet little analysis has been done to determine the differing needs, expectations, and behavior of human users in these environments. With an increasing trend for virtual characters to 'go public', we need to expand our understanding of what this entails for the design and capabilities of our characters. This paper explores these issues through an analysis of a corpus that has been collected since December 2006, from interactions with the virtual character Sgt Blackwell at the Cooper Hewitt Museum in New York. The analysis includes 82 hierarchical categories of user utterances, as well as specific observations on user preferences and behaviors drawn from interactions with Blackwell.

1. Introduction

Interactive Agents have come a long way in the past 10 years. Most commonly developed for specific task or limited domain dialogues tested in specialized environments, there has been an increasing move both to get the Agents out of the lab into wider public access, as well as to implement a more ambitious range of conversational abilities. As agents have become more realistic—both in dialogue capabilities as well as visual and sometimes multimodal renderings—there has also been an increasing concern with studying and implementing theories of social dialogue, or 'small-talk'. Several studies have observed that more realistic agents in public user settings are frequently addressed by users with social dialogue (Bell and Gustafson 2000, Bernsen and Dybkjaer, 2004).

To date, however, work in dialogue development has focused primarily on adapting models of dialogue for agent behavior, with little attention paid to a fuller understanding of the users' behavior and input. Rea, for example, an embodied Agent that engages in the task of a real-estate sales interview, was given 'small talk' moves to build rapport with prospective buyers (Bickmore and Cassell, 2000). In this phase, however, she controls the dialogue, and user input is irrelevant to her next conversational move until she enters the task phase of the dialogue. Although more sophisticated agents have been developed in recent years designed specifically to handle a wider dialogue, there is still a tendency to focus more heavily on discourse models and agent initiative than full development of a domain for comprehension of user initiative. Bernsen and Dybkjaer (2004) state this problem with their prototype of the H.C. Andersen system:

> HCA's main problem seems to be that he cannot always pursue in depth a topic launched by his interlocutor because, at this stage of development, at least, his knowledge and conversational skills are still somewhat limited, and we do not have sufficient information about the key interest zones of his target audience. This is where the rhapsodic nature of conversation may come to his rescue to some extent. When, during conversation, and despite his following an agenda in conversation, HCA is lost and repeatedly does not understand what the user is saying, he changes topic or even domain in order to recover conversational control.

The strategy of exerting agent initiative to recover from poor agent performance is a common one, and is an effective method to increase the realism of the dialogue, as well as get the agent back on track. And yet there seems a danger of overuse of this mechanism to boost apparent performance in that it keeps us stuck in techniques that have been learned and work effectively for simpler task-oriented agents, where a narrow delimitation of the dialogue domain and agent control of the dialogue flow are two major assets contributing to system performance. With the more ambitious goal of creating agents with more truly conversational abilities, a very central measure of progress will necessarily be based on both discovering and utilizing sufficient information about the central interests and language of the target audience. But paradoxically, the more the agent utilizes initiative to present its own topics, the less information we may obtain about the users' interests and social dialogue behaviors, and further system development based on such data runs a risk of perpetuating its own dialogue design and domain coverage.

Mixed-initiative dialogue systems attempt to gain the advantages of allowing the user to take initiative and say what they want, while allowing constrained problem solving (Levin et al 2000). These systems usually can only understand information used to fill in a form in a task-oriented domain rather than responding to input that might be on a different topic of the user's choice.

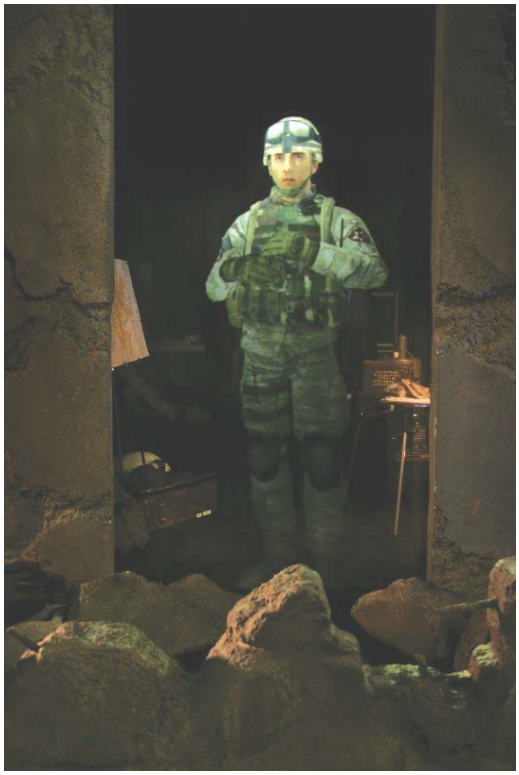

Figure 1: Sgt Blackwell

Gathering a thorough understanding of a domain as unbounded as 'social dialogue' is a daunting prospect, but there are some advantages. A certain amount of data will appear in common across varied situations and characters, thus generalizations drawn from the interactions with one agent may be applied to others, insofar as both the system and the method of data classification are reasonably domain independent. The goal of the present study is to give such a characterization, from data collected at a museum installation of the agent Sgt. Blackwell.

An advantage of using Blackwell's data for an analysis of public user dialogue tendencies is that Blackwell is a QA based character that relies on user initiative and has no clear functional domain. While any character will have unique idiosyncrasies that suggest certain character-specific questions, Blackwell had no clearly defined role, interactive task, or conversational game to play that would steer the users' questions into a functionally domain-specific application. He is simply 'himself', presented as a character to freely interact with.

The paper is organized as follows: Section 2 gives an overview of the system; section 3 describes Sgt Blackwell's installation at the museum and an overview of the corpus. Section 4 discusses the characteristics of user data and our method of categorization in detail. Section 5 gives the main conclusions of our analysis and some future directions relevant to agent design and revision.

2. System

Sgt Blackwell (figure 1) was originally constructed as a technological showpiece at an information kiosk for the 2004 Army Science Conference. Since then, Blackwell and other characters based on his design have been developed for a variety of entertainment, educational and training applications (e,g.,. Traum et al, 2007). Blackwell is based on a QA Character model (Leuski et al, 2006), and has a finite set of pre-recorded responses. ASR input is sent to a classifier based on cross-language information retrieval techniques, which was trained on a set of utterance-answer mappings.

To prepare for the museum installation, we did further user testing, expanded Blackwell's training set of user utterances to over 1,700, and expanded his set of possible responses from 83 to 104. The majority of Blackwell's utterance set consists of content-focused responses covering greetings, closings, and a wide range of topics such as biographical information, his technology and training, and the Institute for Creative Technologies (ICT). In addition, Blackwell has a wide variety of 'off-topic' responses, as question answering in a conversational context presents its own problems. Where a standard QA system is asked factual questions about the world, they are either clearly in or out of its domain of knowledge. If it is out of domain, a reasonable response is 'I don't know the answer to that.' While this may not be the desired response, it is a conversationally and pragmatically coherent one—that is, the agent may have failed a knowledge test, but passed the conversational one. And any query outside of its defined domain may reasonably elicit that response. With a character based QA dialogue, users will invariably ask more personal questions that the agent reasonably should be able to answer about his own experience or background, but are not in its training data. Thus Blackwell's has a range of 18 'off topic' responses that use a variety of evasive tactics, such as "I'd like to know that myself", "What does that mean?", and "That's classified." Finally, Blackwell has the following dialogue tracking capabilities: If several user utterances are classified as 'off topic', he attempts to direct the user to some of his range of topics by using one of 5 prompts. If he detects a repetition of content, he prefaces his response with one of 4 'pre-repeat' lines that indicate he's repeating himself.

3. Museum Data Collection

Sgt Blackwell was installed in the Cooper-Hewitt National Design Museum in New York, from December 2006 until July 2007, as part of the National Design Triennial. The user directions at the museum installation were simple and straightforward: "Meet Sgt. John Blackwell, the Interactive Character created at the University of Southern California's Institute for Creative Technologies (ICT). He can take one query at a time and respond to it." This was followed by simple technical directions for operating the push to talk button, and a short list of sample questions:

When did you join the army?
Where were you born?

What's your name?
Tell me about ICT.
What's your favorite music?

To date, we have logged approximately 150,000 user utterances from first three months of Blackwell's presences in the museum, of which 12,000 were transcribed and examined to determine the range of user behavior. The sample used for this study was a random subset of 1,000 utterances from 251 speakers. Transcribed utterances were coded with speaker number, gender and dialect. Speaker demographics from our sample are shown in table 1.

GENDER	**Total = 251**
Male	131 (52.2%)
Female	118 (47%)
Child	2 (0.8%)
LANGUAGE	**Total = 251**
Standard American English	160 (63.7%)
Other Native English Varieties	29 (11.6%)
Non-Native Speakers	62 (24.7%)

Table 1: Speaker Demographics

The average length of speaker turns per dialogue was 4, though the dialogues ranged in length from one to eighteen. While we have no absolute sense of the conditions in the museum, several factors may be deduced from the data that may have contributed to shorter dialogues. The quantity of data collected and background noise from the audio files suggest a fairly large traffic flow by the installation, which may have encouraged participants to keep their interactions brief. In addition, there are a number of cases where it is clear that groups of museum goers took turns interacting with the agent (where two or three participants alternated and the same speaker reappeared several utterances later). In some cases, poor agent response likely contributed to shorter interactions. Also, some of Blackwell's responses about his technology and ICT were rather lengthy, which may have bored some users, particularly if they were inappropriate responses. In some cases there is evidence a speaker left in the middle of one of these responses, as Blackwell's voice can be heard in the background of the initial utterance of the next speaker.

We have yet to do a full evaluation of Blackwell's performance at the museum, but have coded a sample of 150 utterances from our data set above. WER averaged 0.70, which was considerably higher than in our standard demonstrations. Audio conditions in the museum were poor and noisy due to a console mounted microphone and ambient noise. Another major factor was user demographics. The speech models were trained on male Standard American English speakers, which amounted to only 71 speakers from our data (28.3%).

We also rated response quality, based on a 1-6 scale of coherence and appropriateness (Gandhe et al., 2006). Only the upper end (5-6) is considered a good response, and used to train Blackwell's QA pairs. Of these 150 utterances, Blackwell responded with 41 answers of this quality (27.3%); the median answer rating was a 2. Of his answers, 79 were 'off-topic' responses (52.7%).

Performance also appears lower in a sense because Blackwell exerts no control over the conversation, aside from occasional suggestive prompts. Yet, as discussed in section (4.2.1.) below, a majority of speakers rejected Blackwell's prompts, in favor of pursuing their own topics.

4. Data and Categorization

For our study, a first pass was made through 12,000 user utterances to determine the range of data and to draft an initial categorization of user utterances. This initial categorization was based only on content of the utterances. However, it became clear that a context free categorization of utterances led to occasional misclassification, and missed some important tendencies reflecting user behavior. So we revised the classification to view each utterance in context, to determine whether the utterance was of the user's initiation (initial utterances and mid-dialogue utterances introducing a new topic) or a reactive utterance (response to either an agent prompt or follow-up on a topic in Blackwell's utterance). For the analysis described in this paper, we randomly selected a section of the corpus consisting of the interactions of 251 consecutive speakers, for a total of 1000 user utterances. These utterances were coded in context with the refined contextual categories. An overview of the final categories and their utterance frequencies is shown in table 2.

Since the ultimate aim of this study is to provide a characterization of user dialogue with agents in a public setting that may contribute to agents other than Sgt Blackwell, we found two previous studies that were particularly useful to provide a point of comparison, both for method of categorization and range of user behavior. The first, (Gustafson & Bell, 2000), describes August, an animated spoken dialogue system which was displayed at a museum in Sweden. August could field questions and offer information about its home institute, speech technology, and information about Stockholm, but otherwise was not presented with any clear domain to users. The second study, (Kopp et al, 2005), describes the implementation of Max, an animated agent utilizing a keyboard interface, which was installed as a museum guide at a museum in Germany. Both studies gave some detailed discussion and categorization of user utterances, and we draw comparison with their findings where relevant below.

4.1 User-Initiated Categories

User-initiated utterances are defined as utterances that initiate a topic of the user's choosing, regardless of Blackwell's previous utterance. They far outnumbered

CATEGORY OF USER UTTERANCE	TOTAL
I. USER INITIATED UTTERANCES	**789**
1. DIALOGUE FUNCTIONS	**82**
GREETING	44
POLITE SOCIAL PHRASES	24
CLOSING	14
2. USER-INITIATED INFORMATION REQUESTS	**634**
BIOGRAPHICAL QUESTIONS	293
PERSONAL PREFERENCES	136
MILITARY EXPERIENCE & KNOWLEDGE	114
GENERAL PURPOSE & ABILITIES	53
IMMEDIATE EXPERIENCE (Emotive and Physical)	38
3. OTHER USER-INITIATED	**73**
HAZING/TESTING PERCEPTION	40
FLAMING	24
IMPERATIVES (Do movement, 'shut up')	9
II. REACTIVE UTTERANCES	**182**
1. OVERT RESPONSES TO PROMPTS	**69**
PROMPT ACCEPTANCE (Reformulated Requests)	53
PROMPT REJECTION (Overt)	16
2. RESPONSIVE UTTERANCES	**113**
RESPONSIVE Qs	50
EVALUATION/META COMMENTARY	43
MISC RESPONSIVE STATEMENTS	20
III. OTHER	**29**

Table 2: Frequency of user utterances by Category

reactive utterances in our data, not surprising as the system was based on a QA model with little initiative. What is more striking is the persistence of users to humanize the agent, both in the content and form of the questions they asked. While this is noted with other agents, the trend in our data is much stronger. Gustafson & Bell (2000) classify 33% of user data as 'social', which covers greetings and remarks of a personal nature. Greetings, social 'commonplace phrases' and 'anthropomorphic questions' account for only 13% of utterances coded by Kopp et al (2005), but as their agent had more initiative and the majority of user utterances were answers, a comparison with user questions is more appropriate. Of these, 38% could be considered social or 'humanizing the agent'. Of the user initiated questions in our data, the vast majority treated Blackwell as a human interlocutor; when combined with greetings and polite phrases these accounted for 70% of our total data.

4.1.1. Common Dialogue Functions

Common dialogue functions included greetings, closings, and common polite socializing utterances such as "it's very nice to meet you." Common dialogue functions accounted for 10.4% of user-initiated utterances.

4.1.2. User-Initiated Information Requests

Of the user-initiated information requests, the vast majority (96.7%) were questions about Blackwell's biographical information, personal experiences, opinions and preferences, or his 'personal' military experience, questions which in either content or phrasing cast Blackwell as a human participant. Only 21 of these questions (3.3%) were direct questions about his technology ('how many questions can you answer') or biographical questions which cast Blackwell as machine rather than human (e.g. 'when were you created', rather than 'when were you born'). As these questions are at the heart of socializing dialogue with the agent, providing the core answer to the question of what users choose to ask an agent, we will discuss them in some detail.

The largest group (46.2%), biographical questions, consisted of general questions about Blackwell, mostly name, origin, and age, followed by a surprising number of questions about his marital status (7.8% of all biographical questions). Other questions included where he was

stationed, his height, questions about his family and whether he had parents or children, his country of origin, birth-date, hobbies, shoe size and whether he wore glasses.

The second group, covering 21.5% of user questions, could also be in a sense considered biographical, but for distinction of phrasing, as well as a sense of attributing tastes and desires to the agent, were categorized separately as 'personal preferences'. The most frequent, due most likely to the suggested questions list was 'What is your favorite music?', followed by queries about his favorite food, color, whether he enjoyed the army, what movies and television shows he liked, and his sexual preference. Finally we grouped together a miscellaneous category of questions only asked once, two of which showed relation to his context (do you like art and design? Do you like New York City?), others harder to anticipate (what is your favorite flower? Do you like vegetarians?) While perfect coverage for such questions will never be possible, it is striking how many questions from many different users fell into the same categories—the first six categories above accounted for 96% of all questions on personal preference.

The next most frequent category of questions (18%) were those that, of all, are most 'domain specific', suggested by Blackwell's role as a solider. These included both questions about Blackwell's presumed personal military experience, as well as more general questions about the military and the current politics of war. Following the trend of users focusing on the 'personal', 86% of the military related questions focused on Blackwell's presumed 'personal experience' as a soldier; they covered when he joined or intended to leave the army, why he joined the army, details about his combat training, guns, whether he has killed anyone, his role in the military, and miscellaneous questions about experience of war, such as whether he has been wounded and how hot it is in Baghdad. The two final categories covered general military questions and questions about current politics. Though by content classed as impersonal, they were still not encyclopedic, but tended to refer to his opinion or knowledge. The former included questions on the philosophy of war and germ warfare. The latter covered questions about current events and politics relevant to Blackwell's role as a soldier, such as whether we should leave Iraq and his opinion of George Bush.

The next category, 8.4% of user-initiated questions, covered a variety of classes of what might be termed general questions about Blackwell's abilities, purpose and functions. While a number of these still maintained the personification prevalent throughout, the largest number of meta questions—acknowledging the nature of 'machine'--occurred here. Nearly half of these questions were general prompts about Blackwell's purpose (why are you here, what do you do). Others covered requests to know about our institute, or who made him, and questions about his technology and abilities ("how many questions can you answer?", "do you speak any foreign languages?")

The category of 'immediate experience', though less frequent (6%) seems an important one for a sense of realism of the agent. Though it covers disparate question types, the common factor is the questions address the agent's awareness of his presumed spatial and temporal environment and experience. In some sense, many of these questions could be classed as 'test' as they have a sense of probing the agent's awareness of the world around him and his emotional capacities. Yet the distinction we draw is they seem reasonably good faith probes into his local context, thus are distinguished from 'tests' proper, which are further discussed in the next section. Immediate experience queries covered items from his local animated context, particularly details of his military costume ("are you comfortable in your uniform?", "what are your gloves made of?"), as well as his emotive state ("are you having a good day?", "are you happy?"). Also covered are questions about his experience at the museum ("have you met any interesting people?", "what have you been doing since I last talked to you?") and questions that presume an understanding of the broader real world context in which he is situated, both spatial awareness of the museum setting ("listen soldier where is the toilet?") and temporal awareness ("do you know what day it is today?").

4.1.3. Other User-Initiated Utterances

Of other types of user-initiated utterances in our data, the category we've designated 'hazing/testing' is the most frequent (4% of total utterances). Utterances classed as 'testing the system' have been widely remarked on in the literature, as have other utterance types commonly found in public human-agent dialogues, such as other usually out of domain factual questions and flaming. These categories however, seem poorly defined, particularly since the boundaries between them are rather fuzzy in many cases. As these types of utterances are very specific to human-agent dialogue, they may be especially useful as a point of comparison across different systems with regard to how users accept an agent. Yet if we are to compare such behavior across corpora, clearer definitions are necessary.

Gustafson and Bell (2000), for example, classify user utterances into six broad categories, four of which are relevant here. Insult is defined as expletives and swear words ("you are stupid"); test contained utterances apparently designed to deceive the system ("what is my name"); meta contained both questions about the system as well as all comments about the dialogue ("what can I ask you", "yes that was a smart thing to say"), and facts were factual out of domain questions of an encyclopedic nature, or questions people might expect a computer to handle well ("what is the capital of Finland", "what is two times two"). Kopp et.al. (2005), on the other hand, characterize their data with considerably more categories; though explicit definitions are not given, some examples suggest they cut across some boundaries of those above, as

well as ours. While some types of flaming by their examples (abuse, name calling) overlap with Gustafson and Bell's 'insult' category, they distinguish a category of 'negative feedback to agent', which would cross both insults and 'meta'. In examining our own data it became clear the categories above cover too broad a range of user behavior, so we refined ours as discussed below. Additionally, the distinction between initiative and reactive behavior also was particularly useful for disambiguation.

One factor that makes these categories difficult to delineate is they are perhaps the most predicated on gauging user intent. The difference between a test or teasing / flaming question relies on the presumed hostility of the user—whether the question is asked sincerely in good faith, or as a means for entertaining one's friends. For example, "raise your right hand" seems a sincere test of the agent's ability to respond to movement requests or commands; "pull down your pants" does not. Context can go a long way to disambiguate these. The former speaker in our example asked several general biographical questions of Blackwell, but the latter speaker initiated the dialogue with the above phrase, then continued for a number of turns to swear and call Blackwell names for the duration of his dialogue. Most events of what we call flaming seem to have this characteristic—an ongoing rant of numerous hostile or pornographic utterances. Only context can really disambiguate a phrase like "you're stupid" between flaming and a negative evaluation of the agent's poor response. And while it is possible frustration with an agent's poor performance can devolve into an episode of name-calling or flaming, most cases seem to be an attitude flamers have brought to the interaction from the start. Lying between these categories is something we've called 'hazing'. These are questions or statements that are testing in the sense they seek a response from the system, but are not particularly reasonable; they are like flaming in the sense they are toying with the system, but lack the apparent hostility of flaming.

To summarize these categories, hazing/testing is grouped together, with the following subtypes. A 'test' proper is a reasonably good faith question to test the system's boundaries. This includes movement tests ("can you turn around for us?"), perceptual tests ("how many fingers am I holding up") and factual tests. These are defined as beyond the agent's reasonable local context (as discussed in 4.1.2.), but within general cultural knowledge, or more encyclopedic in nature, such as "who is the president?" and "what is the theory of relativity?"

Hazing is a form of testing, but distinguished by toying with the system, and the content also occasionally pushes its own agenda; users don't seem to be honestly expecting a reasonable response, and this category includes a wider range of utterances, including questions ("do you sympathize with Rambo"), offers ("would you like a big mac and fries?") and somewhat random statements ("we come in peace").

Flaming is more overtly hostile and includes direct insults, swearing and offensive utterances, as described above, and accounts for 2.4% of our total data. The remainder of utterances are miscellaneous imperatives ('shut up', 'go to it').

4.2 Reactive Categories

Reactive categories cover any topics that are not explicitly user-initiated. These include answers, meta comments on the dialogue, and questions that follow up on a topic in Blackwell's previous utterance. Reactive utterances accounted for 18.2% of the total data.

4.2.1. Prompts

Response to Blackwell's attempts to steer users in a particular direction were interesting in light of user interests in socializing. Blackwell has five prompts overall, one rather open-ended, and four covering specific topics. Response to the general prompt ("Why don't you ask me something I know about?") was reasonably favorable. We defined an acceptance as a positive question on something the character might reasonably be expected to know (though not necessarily in his actual range of knowledge). In this case there were 31 acceptances (67.4%) plus 5 clarification requests on his range of knowledge, which could also be viewed as a form of acceptance (10.9%). This suggests users were generally trying to be agreeable and work within Blackwell's domain of knowledge. This is fairly close to a response rate cited by Gustafson & Bell(2000), where 63% of prompted users immediately followed up on the suggested topic. Yet response to Blackwell's topic specific prompts was considerably lower. The four prompts are:

"Want to hear something about my training? Just ask me."
"Wouldn't you like to know something about ICT?"
"Ask me why I can understand what you're saying right now"
"You should ask me instead about my technology"

Of these prompts, only 35% were accepted, 55% implicitly rejected by the user pursuing their own question, and 10% overtly rejected ("no", "not right now"). The much lower rate of positive acceptance to Blackwell's four topical prompts could be a result of the users simply wanting to pursue their own topics. Yet given the line of topics the vast majority of users chose to pursue, it seems likely that the prompts were rejected out of disinterest: all four prompts are on topics characterizing Blackwell 'as machine', which conflicts with the strategy of interacting with Blackwell as a human character.

4.2.2. Responsive Categories

Although responsive categories covered the smallest number of utterances, they cover the widest range of functional categories in the dialogue. The most frequent type in our classification, covering 44% of responsive utterances, is questions that follow up in some manner on

Blackwell's previous utterance. These questions followed up by direct questions on a specific topic that had been peripherally mentioned by Blackwell, or by requesting clarification or elaboration on Blackwell's topic, as shown below:

User: "What is your favorite color?"
Blackwell: "I like red, white and blue"
User: "Why do you like red?

Such questions are not uncommon in human-human dialogue, and we might reasonably expect them to become more frequent as the accuracy of an agent's responses increases.

The next most common responsive category consisted of meta commentary. This covered 38% of responsive utterances and were largely critiques, both positive and negative, of Blackwell's responses. Even though a number of cases were in direct response to poor (machine-like) answers from Blackwell, the comments were still, paradoxically, largely 'humanizing' in their format ("Oh I see you have an attitude", and "Private you're not listening…")

The remaining 18% of responsive utterances consisted of answers, contradictions of corrections ("don't call me 'sir'"), canceling a topic, and reformulations.

5. Conclusions

Aside from the observations on user behavior preferences that suggest specific modifications for future versions of our system, the categorization of content of user questions will be useful for expanding the basic social domain of many agents. Future work will include expanding and refining these categories into a coded database, from which the relevant data can be utilized to give wider initial coverage for a variety of new agents.

6. Acknowledgements

We would like to thank members of our team who helped design and implement Sgt Blackwell, as well as the project members who were responsible for his installation at the Cooper-Hewitt museum, especially Diane Piepol, Jarell Pair, Anton Leuski, David Hendrie, Dick Lindheim, Jillian Gerten, and Bill Swartout. We would also like to thank Shanus Adams and Michael O'Shea at the Cooper-Hewitt museum. The project described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

7. References

Bell, L. & Gustafson, J. (2003). Child and Adult Speaker Adaptation during Error Resolution in a Publicly Available Spoken Dialogue System. In *Proceedings of Eurospeech 2003*

Bernsen, N.O. and Dybkjaer, L. (2004). Domain-Oriented Conversations with H.C. Andersen

Bickmore, T. & Cassell, J. (2000). "How about this weather?" Social Dialogue with Embodied Conversational Agents

Gandhe, S., Gordon, A.S. & Traum, D. (2006). Improving Question-Answering with Linking Dialogues. In *Proceedings of the 11th international conference on intelligent user interfaces (IUI'06)*

Gustafson, J. & Bell, L. (2000). Speech technology on trial: Experiences from the August system. *Natural Language Engineering,* 1 (1), pp. 1-15.

Kopp, S., Gesellensetter, L., Kramer, N., & Wachsmuth, I. (2005). A Conversational Agent as Museum Guide--Design and Evaluation of a Real- World Application. In Panayiotopoulos et al., (Eds.), *Intelligent Virtual Agents*, LNAI 3661, pp. 329-343.

Levin, E., Narayanan, S., Pieraccini, R., Biatov, K., Bocchieri, E., Di Fabbrizio, G., Eckert, W., Lee, S., Pokrovsky, A., Rahim, M., Ruscitti, P., & Walker, M. (2000). The AT&T-Darpa Communicator Mixed-initiative Spoken Dialog System. In *Proceedings of the Intnl Conf. Spoken Lang. Processing.* Beijing, China, pp. 122–125.

Leuski, A., Patel, R., & Traum, D. (2006). Building Effective Question Answering Characters, In *Proceedings of SIGDial 2006*

Traum, D., Roque, A., Leuski, A., Georgiou, P., Gerten, J., Martinovski, B., Narayanan, S., Robinson, S., & Vaswani, A. (2007). Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, Antwerp, Belgium, September 2007*, pp. 71-74.