

A Research Platform for Multi-Agent Dialogue Dynamics *

Thomas K. Harris, Satanjeev Banerjee, Alexander I. Rudnicky,
June Sison, Kerry Bodine, and Alan W. Black
Carnegie Mellon University, Language Technologies Institute
Pittsburgh, Pennsylvania, USA

E-mail {tkharris, banerjee+, air, sison, kbodine, awb} @cs.cmu.edu

Abstract

Dialogue agents are often designed with the tacit assumption that at any one time, there is only one agent and one human, and that their communication channel is exclusive. We are interested in evaluating complications that arise when multiple heterogeneous dialogue agents interact with one or more human interlocutors, and their communication channel is necessarily shared. In this paper we describe a multi-agent dialogue test-bed that we have implemented to study such dialogue coordination issues. We also describe an initial pilot study that uses this test-bed to experiment with a very simple addressee disambiguation algorithm.

1 Introduction and Motivation

1.1 Spoken Language Interfaces

Spoken language interfaces are the modality of choice for fictionalized accounts of future robots, but have not been a modality of practical consideration until the rather recent computer speed advances and statistical speech recognition breakthroughs of the past few years. Today, given relatively small domains of understanding, automatic speech recognition is at least able to offer itself as an alternative modality, and dialogue systems that take into account the loss of accuracy are able to mitigate its effects.

Thus spoken language interfaces are in the process of becoming technologically feasible, but feasibility aside, what may the best modalities be? The preferred modality of course depends on the tasks at hand, but one ethnographic study (Brumitt & Cadiz, 2001) shows that for some simple and very common tasks, people prefer speech interfaces to any other modality. Although much more is still left to be learned about the preferences of communication modalities, spoken

language's high status in human-human communication is likely to be a continuing driving force behind the preference for spoken language in human-machine communication.

1.2 A Future of Heterogeneous Spoken Dialogue Agents

The recent feasibility and preference for spoken language interfaces has led to an explosion of dialogue agents. The number of users of spoken dialogue agents would be difficult to estimate; but, a single service, AT&T's toll-free directory assistance, alone services around 200,000,000 calls per year. It is conceivable that the number of spoken dialogue agent users exceeds even the number of home computer users worldwide.

Spoken dialogue agents, which were only recently specialty products and have their mainstay in computer applications and telephone services, are now found embedded in mobile phones, information kiosks, audio/video equipment, automobiles, toys, personal digital assistants, video games, and robots (Hoge et al., 1999). They are accessible, but the interfaces are almost always integrated into their applications. As the number of appliances with embedded spoken dialogue agents increases, we will be living in environments of multiple heterogeneous dialogue agents.

1.3 Problems with the Communication Channel

Multiple spoken dialogue agents will face some of the same main communication issues that have been active areas of research in the domain of computer networks, namely the issues of message identification, message addressing, channel contention, and session identification.

1.3.1 Message Identification

With multiple spoken dialogue agents in an environment, there is the unintended potential (and some-

*This work is supported by Boeing. Thanks to Dan Bohus and Antoine Raux for help with the Ravenclaw dialogue system and to Kishore Prahallad for integrating the multi-voice text-to-speech component

times the need) for dialogue agents to speak to each other. Dialogue agents will undoubtedly misbehave unless they can identify who is speaking. Speaker identification has had some success recently and may be employed to address this issue, but serious issues remain to be resolved, such as its scalability.

1.3.2 Message Addressing

When an environment contains more than one spoken dialogue agent, each agent must resolve who a particular utterance is addressed to. Evidence from acoustic, linguistic, and pragmatic sources of knowledge, combined with additional information from other communication modalities such as gesture, gaze, and touch have been used to perform address resolution. Systems that perform such resolution however often make use of deep domain knowledge, or make the simplifying assumption that other dialogue agents in the environment have domains that are sufficiently different from its own.

1.3.3 Channel Contention

With multiple spoken dialogue agents in an environment, there is the potential that they will speak simultaneously, or that they will interrupt each other. Methods currently employed to resolve these contentions often consist simply of waiting until nobody else is speaking for a second or so, and then to begin speaking until someone interrupts. In environments with one dialogue agent and one human interlocutor, this algorithm is usually sufficient, but studies of larger human group dynamics show that many important subtleties are missing. Agents who do not indicate their desire to speak, or who do not introduce themselves before they speak are likely to not be understood by their listeners, or, at the very minimum, such agents are likely to add to the cognitive load of their listeners.

1.3.4 Sessions Identification

Although communication between humans and agents is improved when those agents introduce themselves before speaking, they cannot introduce themselves before every utterance. A concept of a communication session must be developed.

2 An Experimental Multiple Agent Dialogue System

We have engaged in a systematic approach to finding efficient solutions through empirical experiments to the four communication issues identified above. In

particular, we have developed a multi-agent dialogue (MAD) system, which can accommodate multiple dialogue agents in a single experimental framework (see Figure 1). The system works both with real robots adapted for the Carmen robot platform (Montemerlo, Roy, & Thrun, 2002), and in a simulated Carmen environment.

The front-end architecture is an instance of the Galaxy-II spoken dialogue system reference architecture (Seneff et al., 1998). We use Sphinx-II (Huang et al., 1993) for automatic speech recognition, Phoenix (Ward, 1994) for context-free grammar parsing, Helios (Bohus & Rudnicky, 2002) for confidence annotation, Ravenclaw (Bohus & Rudnicky, 2003) for dialogue management, ROSETTA (Oh & Rudnicky, 2000) for natural language generation, and Festival (Black, Taylor, & Caley, 1998) for text-to-speech rendering. The robots, named Bashful and Clyde (ghosts from Namco Ltd.'s Pac-Man[®]), each have their own Ravenclaw dialogue system. Ravenclaw is a generalized tree-based dialogue management framework that provides the designer of a dialogue management system with mechanisms through which to specify dialogue tasks. Essentially a designer specifies the various actions that must take place in the system (e.g. the action to be taken when the user asks a robot where it is) and the flow of the dialogue.

The back-end consists of programs that use the Carmen set of libraries to communicate with the robots. The libraries currently utilized in our project include those that allow the user to send messages to the robots to get them to move a specified distance in a certain direction, and those that allow the user to set a goal position and then allow the robot to plan a route to that position.

3 User Studies

3.1 Goals

We have used the system described above to conduct a preliminary set of pilot studies. We have two goals in performing these studies. The first goal is to establish the *usability* of our system, that is, we ask the question: Can this system be used by a human being to successfully interact with the robots? A large system like the one above can fail to be usable for a variety of reasons: the speech recognition may be too error prone, the speech synthesis may be unintelligible, the pace of the interaction too slow, the robot navigation libraries too unreliable, etc. Our experimentation is designed to show that our system can indeed be used to interact naturally with the robots. The second goal of our studies is to experiment with a very simple mechanism for dealing with multi-agent

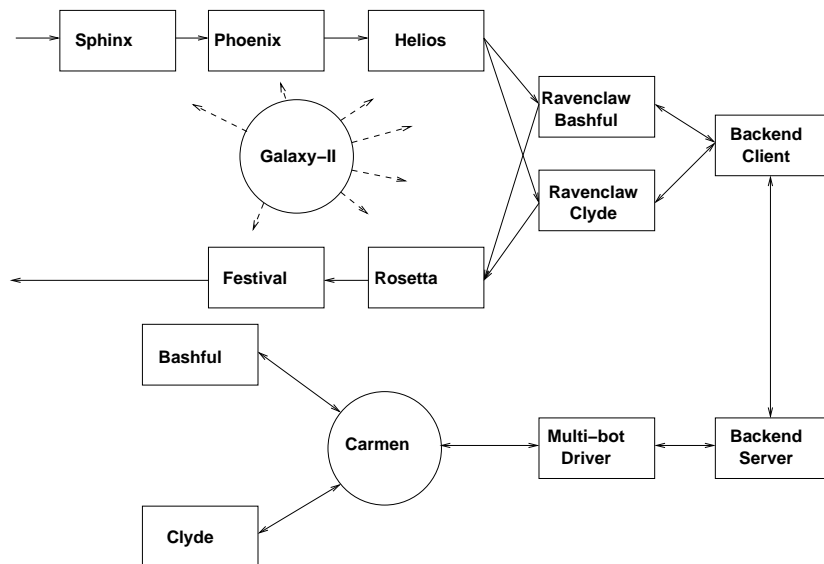


Figure 1: Multiple Agent Dialogue System

communication issues. Specifically, we tested a simple strategy for disambiguating the intended addressee of each user utterance.

3.2 Addressee Disambiguation Algorithm

We have implemented a simple, first-cut algorithm for disambiguating the intended addressee of each user utterance as follows:

If an utterance starts with the name of a robot, then that is the robot this utterance is addressed to. We call this form of addressing *explicit addressing*, and the robot being addressed the *explicit addressee*. For example, in the utterance *Bashful, where are you?*, the form of addressing is explicit, and Bashful is the explicit addressee.

If an utterance does not start with the name of a robot, then the last explicitly addressed robot is being addressed in this utterance. We call this form of addressing *implicit addressing*, and the robot being addressed the *implicit addressee*. For example, if the utterance above is followed by the utterance *Go ten meters north*, the form of addressing is implicit, and Bashful is the implicit addressee.

3.3 Task Description

In our experiment, users were required to navigate the two robots (Bashful and Clyde) through a maze of corridors using only the speech channel to communicate with the robots. The users were not allowed to see the robots and therefore had to rely on spoken dialogue to query the robots regarding their positions in the maze at all points of time.

Specifically, the task involved first finding out the initial positions of the two robots in the maze, and then navigating them to the specified destination marked on the map. Figure 2 shows a map of the maze used. Participants were provided with a hard-copy of this map (without the initial locations of the robots) and were asked to mark on it the positions they believed the robots were at initially. Users were given a maximum of 30 minutes to finish the task, and were allowed to give up earlier if they wished. Note that users were not informed of the addressing mechanism described above since one of the aims of this experiment is to determine if the addressing mechanism can be intuited by the users and, in general, if it makes for naturalistic dialogue.

3.4 Grammar for User Utterances

We used the following simple grammar to parse the user’s utterances:

HumanReportCommand \rightarrow
 ([RobotName]? report) | ([RobotName])

RobotName could be either Bashful or Clyde. The user could either utter “Bashful” or “Bashful report” to address Bashful explicitly, or just utter “report” to implicitly address the last explicitly addressed robot.

HumanLocationQuery \rightarrow
 ([RobotName]? where are you)

This command could be used to query (either

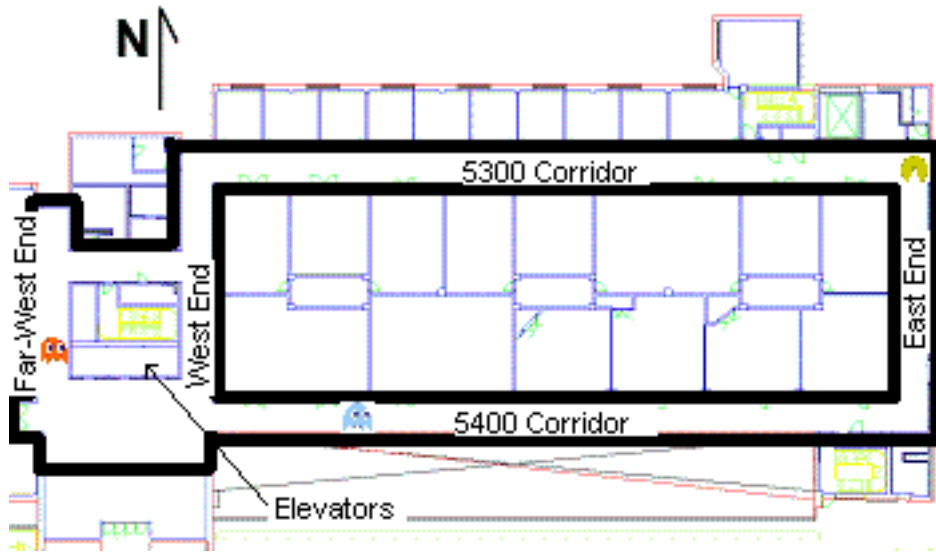


Figure 2: The map of the maze. Initial robot positions are marked with the icons in the far-west end and the 5400 corridor; the destination is marked with the icon in the corner of the 5300 corridor and the east-end.

explicitly or implicitly) a robot’s location.

```
MoveVector →
([RobotName]? MOVE [Direction]?
[Distance]?)
```

This command could be used to direct a robot to move a certain distance along a certain direction. We restricted **distance** to be any integer distance from 1 to 20 meters, and **direction** to be either north, south, east or west. Although utterances that did not contain both a direction and a distance were parsed by the grammar, both pieces of information were needed to perform the move. Hence in a situation where the user provided only one or neither of the pieces of information, the dialogue manager would ask the user to supply the missing information.

Although users were not presented with the above grammar, they were informed that they could ask “Where are you?” and that they could instruct the robot to move between 1 and 20 meters in one of the four directions.

3.5 Robot Responses to User Utterances

The Ravenclaw dialogue system used in this platform requires the designer to specify the response of the system for every parse-able user utterance. Responses may include speech output, back-end actions taken by the system, or a combination of the two. Following is a description of the system responses for each of the three families of user utterances described

above.

Response to HumanReportCommand: Robots responded to this family of utterances by saying *Bashful here*, or *Clyde reporting*, etc. This dialogue helped the user initiate a communication channel with a robot.

Response to HumanLocationQuery: Robots responded to this family of utterances by specifying where in the maze they were. Each part of the map a robot could be in was pre-assigned a name as shown in figure 2. The system’s back-end mapped the robot’s absolute (x, y) coordinates obtained from the CARMEN robot API to the corresponding area name. The system also computed the approximate distance from the closest end of the area. A typical reply to a HumanLocationQuery would be *I am now in the fifty three hundred corridor, about five meters from the east end*.

Response to MoveVector: The addressed robot responded to this family of utterances by first making sure it had a value for both the distance and the direction components. If one or both values were missing, the system engaged the user in a follow-up dialogue by asking, for example, *How far do you want me to go east?* or *In which direction do you want me to go five meters?* Once both values were provided, the system used the robot’s current (x, y) coordinates to compute the destination position, and then used CARMEN’s autonomous navigation API to move the robot to the new position. At the same time, the robot would inform the user that it was following the command by

uttering, for example, *Going five meters toward the north*.

3.6 Other Details

For this pilot study, we used simulated robots in a simulated environment. These robots were initially placed at the positions shown in Figure 2 for each participant in this user study. Participants used a single head-mounted close-talking microphone to speak to both robots, and the speech from both the robots was routed through a single set of speakers. To help the user to distinguish between the speech from the two robots, we used a male voice to synthesize the speech from Bashful, and a female voice to synthesize the voice from Clyde.

3.7 Results

Table 1: *Pilot-study Results*

Part. #	Task Success	Time Taken (mins)	Addressing Mechanisms Used
1	Both	28	Only explicit
2	One	21	Both forms
3	Both	28	Both forms
4	None	18	Both forms
5	One	20	Only explicit
6	One	12	Only explicit

We ran the experiment with 6 different participants. Every participant could correctly identify the approximate initial positions of the robots on the map. We defined task success as follows: A participant was completely successful (denoted as “both”), partially successful (“one”) and completely unsuccessful (“none”) if he or she managed to navigate both, one or none of the robots to their destinations respectively. We also measured the time taken until the end of the experiment. Table 1 shows task success and time taken for each of the 6 participants.

During the experiment we noted what addressing mechanisms, explicit or implicit, the participant was using in his or her utterances. Three participants used only the explicit form of addressing; that is, each of their utterances was prefaced with the name of the robot. When asked at the end of the experiment whether they understood that they could engage in implicit addressing and simply chose not to, all three replied that they did not realize that implicit addressing was possible. The remaining three participants used both forms of addressing.

We also noted the strategy used by each participant in moving the robots to their final destinations.

Four participants elected to simultaneously move both robots, conversing with one robot after giving the other robot a move command, for example. The two remaining participants chose to complete moving one robot to its final destination before commencing to maneuver the other robot.

3.8 Analysis and Lessons Learned

Table 2: *Utterance Analysis*

Part. #	# utts w/o recognition errors	# utts with recognition errors	# utts outside grammar
1	–	–	–
2	52 (77.6%)	9 (13.4%)	6 (9.0%)
3	88 (91.7%)	6 (6.2%)	2 (2.1%)
4	52 (70.3%)	18 (24.3%)	4 (5.4%)
5	103 (76.3%)	31 (23.0%)	1 (0.7%)
6	37 (43.0%)	47 (54.7%)	2 (2.3%)

We performed a post-experiment analysis of the participants’ utterances, and, for each participant, counted the number of utterances that the system recognized and processed without any errors, the number of utterances that had at least one error in them, and the number of utterances that were outside the system’s grammar described in section 3.4. Utterance errors included either speech recognition errors (such as misrecognizing the word *east* as *west*) or dialogue processing errors (such as ignoring a *HumanLocationQuery* command while the robot was in the process of engaging the user in confirmation dialogue regarding a previous *MoveVector* command). Table 2 shows the results of this analysis. Corresponding participant numbers in tables 1 and 2 refer to the same participants. Due to a technical glitch we lost the utterances of participant #1.

Note from table 2 that more than 70% of the utterances for 4 participants had no errors at all, and that less than 10% of the utterances for each participant were outside the system’s grammar. These numbers result from the fact that the system used a very constrained grammar, making the speech recognizer’s task relatively easy. The number of recognition errors seems to correspond somewhat with task success. Thus, the participant who managed to move both robots to their destination had the largest percentage of correctly recognized utterances, while the participant who had the lowest success had the second lowest percentage. Due to the small scale of this pilot experiment however, these results are only suggestive of trends; as future work we plan to perform larger

experiments to better test such hypotheses.

When asked to rate the interaction after the experiment, every participant replied that he or she found both the dialogue and the pace of the interaction naturalistic. These reports established that our implemented system can be used successfully to interact with robots. Furthermore, the fact that every participant understood the explicit addressing mechanism and half the participants understood the implicit mechanism implies that our simple addressee disambiguation algorithm is easy to understand and makes for natural dialogue.

Participants were also asked to provide feedback on any aspect of the experiment. Every participant felt that the robots needed to provide more feedback. For example in the current design when a robot is asked to go further than it can, it does not report this inability. Participants expressed satisfaction at having two robots to work with instead of one; they felt that the pace of the interaction would have been too slow if there was only one robot, since robots take some time to move from one point to another. Participants also found the set of user commands somewhat limiting. For example, when the robot was stuck against an unknown obstruction, the participants felt that more exploratory commands such as *What can you see?* would have been very useful. Some participants felt that descriptions of the locations as spoken by the robots were sometimes unintuitive. For example, when the robot said *I am in front of the elevators about 3 meters from the east end*, the robot was referring to the east end of “in front of the elevators”. Since “in front of the elevators” is not normally the type of area with clear boundaries, the participant naturally thought that “east end” referred to the east end of the map. Further research is necessary to determine how to describe the current location of a robot such that the description is maximally intuitive from a human’s point of view.

4 Conclusions

Heterogeneous interface agents cannot act in concert, achieving a globally optimal interface strategy by understanding or predicting each others’ behavior. Some research groups have taken the approach that constructs an aggregated spoken dialogue front-end for a community of under-specified agents (e.g. the Speech Graffiti Personal Universal Controller (Harris, 2004)). Such systems often severely limit the expressive power of natural language however, and any aggregating spoken dialogue front-end will potentially sacrifice the integration of domain knowledge into the dialogue.

In order to directly address heterogeneous multi-agent communication problems, we have established an understanding of the issues and built a platform for experimentation in that domain. The platform, with a few simple strategies, has yielded interesting lessons and some relatively positive results in a small pilot study.

References

- Black, A., Taylor, P., & Caley, R. (1998). *The Festival speech synthesis system*. (<http://festvox.org/festival>)
- Bohus, D., & Rudnicky, A. (2002, November). *Integrating multiple knowledge sources for utterance-level confidence annotation in the CMU Communicator spoken dialog system* (Tech. Rep. No. CMU-CS-02-190). Pittsburgh, Pennsylvania: School of Computer Science, Carnegie Mellon University.
- Bohus, D., & Rudnicky, A. I. (2003). Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Eurospeech*. Geneva, Switzerland.
- Brumitt, B., & Cadiz, J. J. (2001). “let there be light” examining interfaces for homes of the future. In *Proceedings of interact 2001* (pp. 375–382).
- Harris, T. (2004). *The speech graffiti personal universal controller*. Unpublished master’s thesis, Carnegie Mellon University, Pittsburgh.
- Hoge, H., Burchard, B., Comeyne, R., Diehl, F., Fischer, V., Hakkinen, J., & Marasek, K. (1999). *Market analysis*.
- Huang, D., Alleva, F., Hon, H. W., Hwang, M. Y., Lee, K. F., & Rosenfeld, R. (1993). The Sphinx-II speech recognition system: An overview. *Computer, Speech, and Language*, 7(2), 137–148.
- Montemerlo, M., Roy, N., & Thrun, S. (2002). *Carnegie mellon robot navigation toolkit*. <http://www-2.cs.cmu.edu/~carmen/>.
- Oh, A. H., & Rudnicky, A. (2000, May). Stochastic language generation for spoken dialogue systems. In *Anlp/naacl workshop on conversational systems* (pp. 27–32).
- Seneff, S., Hurley, E., Lau, R., Pao, C., Schmid, P., & Zue, V. (1998). Galaxy-II: A reference architecture for conversational system development. In *Proceedings of the international conference on spoken language processing*.
- Ward, W. (1994, September). Extracting information from spontaneous speech. In *Proceedings of the international conference on spoken language processing*.