Crowdsourcing HRI Through Online Multiplayer Games

Sonia Chernova and **Jeff Orkin** and **Cynthia Breazeal**
MIT Media Laboratory
20 Ames St.
Cambridge, MA 02139

Abstract

The development of hand-crafted action and dialog generation models for a social robot is a time consuming process that yields a solution only for the relatively narrow range of interactions envisioned by the programmers. In this paper, we propose a data-driven solution for interactive behavior generation that leverages online games as a means of collecting large-scale data corpora for human-robot interaction research. We present a system in which action and dialog models for a collaborative human-robot task are learned based on a reproduction of the task in a two-player online game called *Mars Escape*.

Introduction

Robots require a broad range of interaction skills in order to work effectively alongside humans. They must have the ability to detect and recognize the actions and intentions of a person (Kelley et al. 2008; Gray et al. 2005), to produce functionally valid and situationally appropriate actions (Breazeal 1998; Mutlu et al. 2009), and to engage in social interactions through physical cues (Sidner and Lee 2007) and dialog (Kulyukin 2004).

A number of robotic platforms capable of these types of interactions have been developed for different applications, including museum guidance (Burgard et al. 1998), reception desk assistance (Lee and Makatchev 2009) and elder care (Graf, Hans, and Schraft 2004). Research for action and dialog generation has also been conducted in the gaming community in the context of character development for role-playing games (McNaughton et al. 2004; Kacmarcik 2005). All of the above approaches present successful solutions for their respective applications based on carefully hand-crafted models for action and dialog generation. While the time required for the development of these models is not typically reported, in the case of one recent interactive online game called *Façade* (which takes about 15 minutes to complete), development by two researchers took approximately 5 years (Mateas and Stern 2005). The typical result for this type of development process is a system that is capable of natural and engaging interaction for some range of topics, but only for those that were predetermined by the programmers.

Copyright © 2010, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Data-driven techniques present an alternate solution to hand-crafted models. These approaches utilize datasets of hundreds of example behaviors, often from a multitude of different users, to generate appropriate responses to input. Successful data-driven techniques have been demonstrated in a number of interactive applications, such as dialog management systems trained to produce appropriate responses based on recorded caller data (Gorin, Riccardi, and Wright 1997; Singh et al. 2002).

The question we explore in this paper is whether similar data-driven approaches can be developed for interactive robotic systems. Can robot behaviors be *crowdsourced* to produce natural, engaging and functionally appropriate actions and dialog based on data from hundreds of people? Ideally, such an approach would benefit from the "power of the masses", requiring less total development time while also producing a more general result by incorporating examples from many users.

The challenge for this approach is to develop a method for gathering robot data on a large scale. One solution is to utilize the Wizard-of-Oz technique in which data is recorded as the robot is puppeteered through the task by a human subject. The significant drawback of this approach is that it requires extensive amounts of time to recruit and train subjects, with the result that such studies are typically limited to a few dozen participants.

In this paper, we propose the use of online games as a means of generating large-scale data corpora for human-robot interaction research. We present a system in which action and dialog models for a collaborative task involving a person and a robot are learned based on a reproduction of the task in an online multiplayer game. Similar to projects such as Games with a Purpose (von Ahn and Dabbish 2008) and the ESP Game (von Ahn and Dabbish 2004), our goal is to make work fun in order to harness the computational power of internet users. Our work is inspired by the Restaurant Game project (Orkin and Roy 2007; 2009), in which data collected from thousands of players in an online game is used to acquire contextualized models of language and behavior for automated agents engaged in collaborative activities.

Robot Task Domain and Data Collection

The goal of our research is to enable a robot to perform a collaborative task with a human by leveraging a corpus of example interactions collected in an online game. For evaluation we have selected a general search and retrieval task in which the robot and human must work together to collect multiple objects. The task has no strictly assigned social roles, however, the domain is developed to encourage collaborative behaviors such as action synchronization, sequencing and dialog.

Our project is organized into three phases. During the first two phases we collect data using a custom-made online two-player game called *Mars Escape*. The game records the actions and dialog of two players as they take on the roles of a robot and an astronaut on Mars. We then utilize the resulting interaction corpus, consisting of data from hundreds of online players, to learn action and dialog models for the collaborative retrieval task. In the final phase of the study, scheduled to take place in September, 2010 at the Boston Museum of Science, we will evaluate the learned models by using them to generate behaviors for an autonomous robot in a real-world retrieval task. Each of the research phases is described in detail in the following sections.

Online *Mars Escape* Game with Text-Based Dialog

Mars Escape is a two-player online game in which two randomly-paired players take on the roles of a human astronaut and a robot on Mars. The object retrieval task is incorporated into the back-story of the game, in which the players are told that they are located on a remote research station on which the oxygen generator has failed. In order to successfully complete the mission, the pair must locate and salvage their five most valuable research items and return to the spaceship before oxygen supplies run out. The list of required items is presented in a side-bar inventory screen. The players have 15 minutes to complete the mission.

During the game, the players are able to navigate in the environment, manipulate objects using a number of predetermined actions (e.g., look at and pick up) and communicate with each other through in-game text-based chat (see Figure 1). All player actions and dialog is recorded by the system. At the completion of the game, the players are given individual scores based on the number of items collected and the time required to complete the mission. Players are also asked to complete a survey evaluating their gaming experience and the performance of their partner. Players are asked to rate how much they enjoyed working with the other person, as well as to speculate on whether the other character was controlled by a person or an AI.

Although the game is designed to represent and provide data for a real-world task, the virtual environment does not represent a precise 3D model of our target domain because, in this application, knowledge of abstract actions (e.g. astronaut picked up the alien) is sufficient for modeling high-level collaborative behavior. However, character attributes must be chosen carefully in order for the in-game interactions to accurately reflect real-world preferences. Specifically, since timing and the relative durations of actions plays

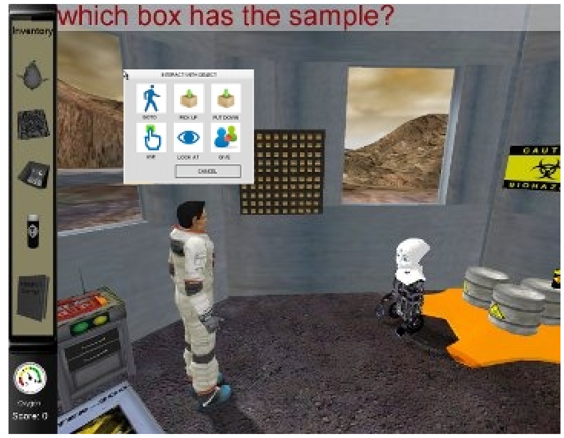

Figure 1: A screenshot of the *Mars Escape* game showing the action menu and dialog between the players.

a significant role in the selection of collaborative behaviors, the movement speeds of the players and the durations of certain actions are modelled from real-world data. Similarly, the physical characteristics of the avatars in terms of reach distance, lifting strength and climbing ability also reflect those of an average human and our robotic platform (the MDS robot, Figure 2). Finally, in *Mars Escape* the robot is equipped with a special sensor, an "organic spectral scanner", that enables it to detect things not visible to the human eye. Together, these different characteristics of the avatars influence the way in which players select who performs different elements of the task. Table 1 presents a list of the five inventory items in the game, how they can be accessed by each player, and how each object can be generalized to a broader class of problems.

During the first three months of the release of the game[1] we captured data from 558 two-player games. Of these, approximately 700 player logs are retained for analysis after excluding logs in which a player exited the game prematurely. Below is an example transcript showing an interaction in which the astronaut (A) and the robot (R) first retrieve the alien and then attempt to find the sample box.

<robot *go to* boxes>
R: "hey, look at that checkerboard thing on the wall..."
R: "it seems to light up if you scroll over it"
R: "perhaps there is something there"
<astronaut *go to* boxes>
A: "oh no. do you think that gadget is in one of these?!"
A: "there are a lot to check..."
R: "maybe, perhaps we will have to check all of them"
A: "booo! i start from the bottom and you the top?"
<astronaut *look at* boxes>
<robot *analyze* boxes>
R: "woot"
A: "you find it!?"
R: "ok, I have a special analyzer that helps me see"
A: "lol. have you been holding out on me???"
R:"the object is in the box that is 3rd column from the right..."
R: "and 4th row from the bottom"

[1] http://www.robotic.media.mit.edu/MarsEscapeGame

Item	Game Context	Generalization
Research Journal	Located on a stack of boxes. Reachable only by the astronaut by climbing the nearby objects.	a task that can be performed only by one of the players
Captured Alien	Located in a cage on a raised platform. Reachable by either player after lowering the platform using wall mounted controls.	a task that can be performed by either player
Canister	Located near a spill of toxic biohazardous chemicals. Reachable by either player, but the astronaut loses 10 points for coming in contact with the chemicals.	a task for which one player is better suited than the other
Memory Chip	The appearance of this item is triggered by both players standing on a weight sensor at the same time. Once active, this item can be retrieved by either player.	a task that requires action synchronization.
Sample Box	One of 100 identical boxes located on a high shelf. Astronaut must pick up and look at each box until the correct one is found. Robot can identify the exact box using its organic spectral scanner, but can not reach the box due to its hight. Optimal solution is for the robot to scan the boxes and then tell the astronaut the sample's exact location.	a task that requires coupled actions and dialog

Table 1: Description of the five objects players must obtain to successfully complete the game.

As can be seen from this example, players can engage in problem solving dialog as they attempt to perform the task. One challenge presented by the typed speech data, however, is the frequent use of colloquialisms, slang, acronyms and abbreviations that are rarely used in spoken dialog, such as the use of "*lol*" and "*woot*" to convey amusement and celebration. In fact, we find that crowdsourced typed dialog is better suited for identifying likely patterns and topics of conversation and for building an initial language model than for generating verbal speech. In the following section we discuss how we leverage the strengths of the typed corpus to develop a spoken dialog version of the game.

Online Mars Escape Game with Spoken Dialog

The second phase of our project aims to address the disparity between typed and spoken dialog by introducing a modified version of the game in which players are able to communicate by speaking through a headset microphone. However, instead of allowing one player to hear the other's speech, as in a phone conversation, we use speech recognition software to transcribe the spoken language into text, which is then displayed at the top of the player's screen as in the original version of the game.

The technique of using speech recognition was chosen for several important reasons. First, it introduces minimal changes to the game, modifying only the way in which players produce dialog but not the way in which it is received. This allows us to use direct comparison between the two game versions to identify changes in communication patterns. Second, textual presentation preserves the anonymity of the players and prevents characteristics such as age, gender and nationality from impacting the interaction. Third, this approach provides the listening player with exactly the same information as what can be perceived by the robot. In particular, this prevents users from utilizing social cues such as tone of voice, which are typically too subtle for computer systems to interpret correctly. Finally, users will be forced to deal with the same speech recognition errors as will be encountered by the robot during real-world experiments (e.g., incorrectly transcribing "scale" as "stale"). The user may respond in different ways to such errors, possibly by asking for clarification, or by making an assumption as to the intended meaning of the phrase given the objects in the environment. Such information will form an invaluable part of the data corpus that will aid the physical robot in dealing with similar errors.

To perform speech recognition we utilize the WAMI web-based toolkit (Gruenstein, McGraw, and Badr 2008) from which recognition results are communicated directly to the *Mars Escape* game server. WAMI utilizes a constrained language model based on the Java Speech Grammar Format (JSGF)[2] standard, which enables the system to perform incremental recognition in real time. We initialize the grammar by seeding it with phrases that make up the typed dialog corpus, in effect limiting the range of possible recognition results to previously observed utterances. We then incrementally improve the grammar model by incorporating manual transcriptions of games played with spoken speech to grow the grammar corpus. In preliminary evaluation we have found the post-game survey to be helpful in selecting games for transcription based on ratings of communication difficulty between players.

For data collection using the speech-based game, we have deployed two *Mars Escape* gaming stations at the Boston Museum of Science. We are currently in the process of building the dialog corpus and not enough data has yet been gathered to provide a full comparison to the text-based version. Preliminary results indicate that the model works well for common topics of conversation; critical key words are often picked up that accurately convey the gist of the phrase. However, recognition errors are frequent, often resulting in semantically different output. Table 2 presents four example recognition results based on the grammar derived from the

[2]http://java.sun.com/products/javamedia/speech/forDevelopers/JSGF

Spoken Phrase	Recognition Result
pick up the yellow canister	could you get the yellow canister
use the button	are you human
give me the alien	and get the alien
please put the alien in the yellow bucket	pick the alien and put in box

Table 2: Example WAMI speech recognition results based on grammar derived from text-based game corpus.

text-based game corpus.

Physical Object Collection Task

In the final phase of this project, action and dialog models learned from the acquired interaction data corpus will be evaluated in a real-world variant of the collaborative retrieval task. Evaluation will be performed at the Boston Museum of Science, where museum visitors will be recruited to perform the task in collaboration with our autonomous MDS robot *Nexi* (Figure 2). The MDS robot platform combines a mobile base with a socially expressive face and two dexterous hands that provide the capability to grasp and lift objects. The robot is equipped with a biologically-inspired vision system that supports animate vision for shared attention to visually communicate the robot's intentions to human observers. Auditory inputs support a microphone array for sound localization, as well as a dedicated channel for speech recognition via a wearable microphone.

Due to the complexity of the search and retrieval task, a high precision offboard Vicon MX camera system will be used to supplement the robot's onboard sensors. The Vicon system will be used to track the position of the robot, human, and objects in the environment in real time using lightweight reflective markers attached to object surfaces. This tracking system will enable the robot to have a greater degree of environmental awareness that is comparable to that of a human. The human teammate will be fitted with uniquely marked hat and gloves to enable the system to accurately identify the direction of the person's gaze and gestures. This information will be critical for inferring the contextual meaning of the user's utterances.

The physical space in which the evaluation will be conducted will resemble *Mars Escape* game in terms of the general size and layout of the environment. It will contain five objects that the players must collect, in similar placements to their in-game counterparts.

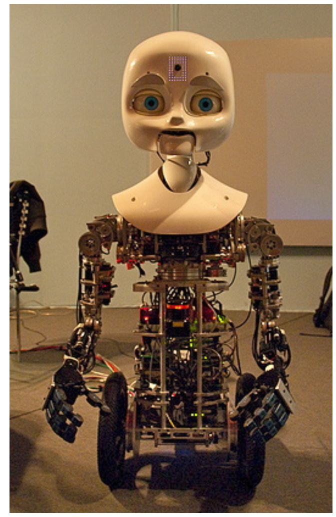

Figure 2: The MDS robot platform.

Data Processing

Although the entire span of the project is not yet complete, we have performed preliminary analysis on logs from approximately 700 players from the text-driven version of the game. In the following sections we present separate models and analysis for the action and dialog components. Our long-term aim is to combine both datatypes into a single model for robot behavior control.

Action Processing

In our analysis of physical actions, we are interested in identifying statistically significant action transitions that represent typical player behavior. To model what a "typical" interaction might look like, we utilize the Plan Network representation developed by Orkin and Roy for the Restaurant Game (Orkin and Roy 2007). A Plan Network is a statistical model that encodes context-sensitive expected patterns of behavior and language. Given a large corpus of data, a Plan Network provides a mechanism for analyzing action ordering and for visualizing the graphical structure of action sequences. Within our work, we have additionally extended the representation to include temporal action information, resulting in a Temporal Plan Network that extracts information on action duration, providing an estimate of expected time for each activity.

Using the current data corpus, we are able to detect significant patterns in the behavior of both player characters using the Plan Network representation. Figure 3 presents the Temporal Plan Network for the astronaut role in which all action transitions with a likelihood of less then 0.08 have been eliminated, resulting in a graph that shows only transitions that were taken by more than 8% of the players. Numerical values along graph edges represent the average number of seconds players take to transition between two actions. While many players choose to deviate from the norm at some point in their gaming experience, aberrant interactions wash away statistically when compared to the larger number of examples of typical behavior.

In addition to providing a model of typical behavior, the Plan Network can be used during task execution to detect

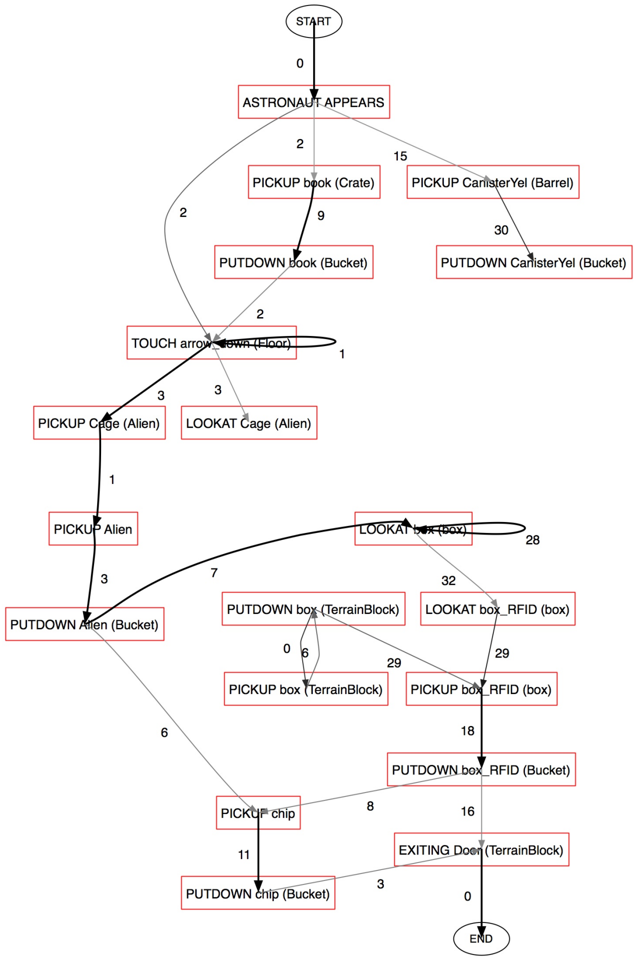

Figure 3: A Temporal Plan Network visualizing typical behavior for the astronaut role across 350 games.

unusual behaviors performed by a user that are outside the expected norms. Using likelihood estimates, the model is able to score the typicality of a given behavior by evaluating how likely this behavior is to be generated by an agent controlled by the learned model. Atypical behaviors not directly related to the task are highly likely to occur in many instances of human-robot interaction. Providing the robot with an ability to respond to such events appropriately may have powerful impact on how the person perceives the robot and on the overall success of the interaction. An interesting direction for future work is to examine how the robot should behave when its teammate goes "off-script"; politely reminding the person to get back to work, or simply continuing to perform the task alone are two possible options.

Dialog Processing

The goal of our dialog system is to automatically generate contextually appropriate responses to natural language input by drawing from a corpus of hundreds of dialogues observed between humans playing the online game. An agent may respond to natural language input with a single utterance or a sequence of utterances.

Game logs are preprocessed to facilitate efficiently finding responses to natural language input at runtime. Our approach is inspired by Gorin et al.'s system for extracting salient phrases from phone conversations correlated with call routing decisions (Gorin, Riccardi, and Wright 1997). In our game logs, we may find a dialogue composed of an arbitrary number of utterances from one or both players between any pair of physical actions. We cluster these dialogues by their preceding physical action, which provides context for dialogues. For each cluster, we extract all sequences, between one and five words long, observed in more than some specified minimum number of game logs, and store these in a phrase dictionary. The final preprocessing step employs the phrase dictionary to abstract every utterance from every game log into an unordered set of the salient (recurring) phrases found within the line of text. This set may contain phrases which overlap, or may be empty if no salient phrases are found in the utterance. Abstracted dialogues are stored with their original associated text, and with the physical action that concludes the dialogue.

At runtime, natural language input is abstracted into an unordered set of recurring phrases through the same process described above. This abstraction is then used to retrieve dialogues from the corpus containing the most similar utterances, based on the number of shared recurring phrases. These candidates are further filtered down to the list of dialogues with histories that best match the interaction observed at runtime, based on previously observed recurring phrases. One dialogue is arbitrarily selected from the remaining candidates, and the utterance or action that follows the matched input is selected for execution by the agent. If the response is a physical action attributed to the other player, this is treated as an expectation for the agent to observe this action in the near future. In cases where the utterance is completely unrecognized, the system can make an assumption based on a previous recognized utterance, and issue a response based on the assumed input. The assumption may not be correct, but a contextually appropriate response often allows the interaction to recover (and we sometimes observe the Eliza-effect, where the human explains away the communication failure and continues, possibly rephrasing his/her input).

Below is an example conversation between a researcher (regular text) and the dialog generation system (italicised text) based on data acquired from 350 game logs:

R: hello
A: hi
R: it is my first time
A: ok great. can you get the canister?
R: Sure
R: what would you like me to do with it?
A: put it in the yellow box by the door
R: ok
A: thanks
R: No problem
A: now can you come stand on the scale here
R: roger dodger

The system successfully selects semantically meaningful responses, even in the case of the rather unusual colloquial phrase at the end. In the additional example below, dialog is automatically generated for both sides of the conversation:

R: what do you want me to do?
A: find the items on the left
A: place them in the yellow box by the door
R: ok
A: need an elektronic circuit
A: and some box with a thing in it
R: ah ha... my organic scanner thingy says that there is something glowing
R: in one of the boxes
R: third row from the right
R: fourth box down
A: got it
R: thanks
A: lets get the alien now

While our approach completely lacks any semantic interpretation of the dialogue, the intuition is that if many players said the same thing in the same context, it must be meaningful at that moment. In previous work, we have shown quantitatively that this dialogue system preserves the texture of human-human dialogue, but unsurprisingly does occasionally lead the agent to say things out of context or repeat itself, due to the absence of semantics (Orkin and Roy 2009). There are a number of ways we might address these problems in future work, including filtering out responses that lead to actions that have already been observed in the interaction, including some representation of current state in the criteria for clustering dialogues, or incorporating a small amount of human annotation to associate utterances with the goals they help achieve.

Conclusions

In this paper, we propose a novel data-driven approach to behavior generation for interactive robots based on a data collection method that utilizes online multiplayer games. While the full scope of this work has not yet been completed, preliminary results in action sequence analysis and dialog generation show promising ability to identify typical behaviors and produce contextually meaningful output. We believe that this approach of crowdsourcing behavior data has potential to provide solutions to a broad range of computationally challenging problems in human-robot interaction research.

Acknowledgments

This research is partially supported by ONR grant N00014-09-1-0112 and by the Singapore-MIT GAMBIT Game Lab. Additionaly, the authors wish to thank Elisabeth Morant for her assistance with the project.

References

Breazeal, C. 1998. A motivational system for regulating human-robot interaction. In *AAAI/IAAI*, 54–62. Menlo Park, CA, USA: American Association for Artificial Intelligence.

Burgard, W.; Cremers, A. B.; Fox, D.; Hähnel, D.; Lakemeyer, G.; Schulz, D.; Steiner, W.; and Thrun, S. 1998. The interactive museum tour-guide robot. In *AAAI/IAAI*, 11–18.

Gorin, A. L.; Riccardi, G.; and Wright, J. H. 1997. How may I help you? *Speech Commun.* 23(1-2):113–127.

Graf, B.; Hans, M.; and Schraft, R. D. 2004. Care-o-bot II—development of a next generation robotic home assistant. *Auton. Robots* 16(2):193–205.

Gray, J.; Breazeal, C.; Berlin, M.; Brooks, A.; and Lieberman, J. 2005. Action parsing and goal inference using self as simulator. In *Workshop on Robot and Human Interactive Communication (Ro-Man05)*, 202–209. IEEE.

Gruenstein, A.; McGraw, I.; and Badr, I. 2008. The wami toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *ICMI*, 141–148. New York, NY, USA: ACM.

Kacmarcik, G. 2005. Question-answering in role-playing games. In *Workshop on Question Answering in Restricted Domains, AAAI-05*.

Kelley, R.; Tavakkoli, A.; King, C.; Nicolescu, M.; Nicolescu, M.; and Bebis, G. 2008. Understanding human intentions via hidden markov models in autonomous mobile robots. In *HRI*, 367–374. New York, NY, USA: ACM.

Kulyukin, V. 2004. Human-robot interaction through gesture-free spoken dialogue. *Auton. Robots* 16(3):239–257.

Lee, M. K., and Makatchev, M. 2009. How do people talk with a robot?: an analysis of human-robot dialogues in the real world. In *CHI*, 3769–3774. New York, NY, USA: ACM.

Mateas, M., and Stern, A. 2005. Procedural authorship: A case-study of the interactive drama façade. In *Digital Arts and Culture*.

McNaughton, M.; Schaeffer, J.; Szafron, D.; Parker, D.; and Redford, J. 2004. Code generation for ai scripting in computer role-playing games. In *Challenges in Game AI Workshop at AAAI-04*.

Mutlu, B.; Shiwa, T.; Kanda, T.; Ishiguro, H.; and Hagita, N. 2009. Footing in human-robot conversations: how robots might shape participant roles using gaze cues. In *HRI*, 61–68. New York, NY, USA: ACM.

Orkin, J., and Roy, D. 2007. The restaurant game: Learning social behavior and language from thousands of players online. *Journal of Game Development*.

Orkin, J., and Roy, D. 2009. Automatic learning and generation of social behavior from collective human gameplay. In *AAMAS*, 385–392.

Sidner, C., and Lee, C. 2007. *Engineering Approaches to Conversational Informatics; Attentional Gestures in Dialogues between People and Robots*. Wiley and Sons.

Singh, S.; Litman, D.; Kearns, M.; and Walker, M. 2002. Optimizing dialogue management with reinforcement learning: experiments with the NJFun system. *J. Artif. Int. Res.* 16(1):105–133.

von Ahn, L., and Dabbish, L. 2004. Labeling images with a computer game. In *CHI*, 319–326. New York, NY, USA: ACM.

von Ahn, L., and Dabbish, L. 2008. Designing games with a purpose. *Communications of the ACM* 51(8):58–67.