The OSU Quake 2004 corpus of two-party situated problem-solving dialogs

Donna K. Byron, Eric Fosler-Lussier

Computer Science and Engineering, The Ohio State University
2015 Neil Ave., Columbus Ohio 43221
dbyron|fosler@cse.ohio-state.edu

Abstract

This report describes the Ohio State University Quake 2004 corpus of English spontaneous task-oriented two-person situated dialog. The corpus was collected using a first-person display of an interior space (rooms, corridors, stairs) in which the partners collaborate on a treasure hunt task. The corpus contains exciting new features such as deictic and exophoric reference, language that is calibrated against the spatial arrangement of objects in the world, and partial-observability of the task world imposed by the perceptual limitations inherent in the physical arrangement of the world. The corpus differs from prior dialog collections which intentionally restricted the interacting subjects from sharing any perceptual context, and which allowed one subject (the *direction-giver* or *system*) to have total knowledge of the state of the task world. The corpus consists of audio/video recordings of each person's experience in the virtual world and orthographic transcriptions. The virtual world can also be used by other researchers who want to conduct additional studies using this stimulus.

1. Introduction

The last few years have seen a growing interest in creating embodied conversational agents that act as partners to humans in a variety of task domains, such as robotic partners for search and rescue or domestic applications, and animated computer graphics characters in training or entertainment simulations. The conversational skills of these ECAs (Embodied Conversational Agents) are quite different from those required of spoken dialog agents working in traditional domains, such as the travel agent domain or other information navigation tasks, even when multimodal information exchange is possible. For an automated agent to be able to act as an ECA, it must be able to calibrate linguistic acts against the spatial configuration of the world in which it is embedded, keep track of the current state of the task and of the discourse, and also calculate an internal model of its partners' beliefs about all of this. This creates new challenges for information integration in ECA language processing modules that have not been studied systematically in the past.

Partners discussing a task carried out within a 3D world that they jointly perceive, manipulate, and discuss generate linguistic utterances that are sensitive to situational context. This context sensitivity impacts some obvious expressions such as demonstrative reference, indexical reference such as *here*, and spatial predicates such as *on the right*. The presence of a jointly-perceived situation also impacts the linguistic behavior in other ways, because the dialog participants' attentional state is effected not only by their discourse but by events and objects perceived in the situation.

For certain kinds of collaborative tasks, such as search and rescue, the physical configuration of 3D space provides a constraint on the order in which the task can be executed. This is unlike other planning or repair domains in which the task structure can be represented as a stack or plan, yielding a widely-accepted process for determining which items are salient due to the task state (Grosz and Sidner, 1986). In an exploration task that uses an extended 3D space, the arrangement of objects that the collaborative partners encounter will impact the flow of their task; interrupting planned activity with discovery and opportunistic re-planning. To study these phenomena as they relate to linguistic behavior will require new corpora.

Existing dialog corpora such as TRAINS (Heeman and Allen, 1995), Maptask (Hemphill et al., 1993), ATIS (ATIS, 1993), or COMMUNICATOR (Walker et al., 2002) have provided the research community with a valuable resource to investigate dialog phenomena such as grounding, speech act sequencing, disfluencies and spoken language parsing, etc. However, these corpora were collected in experimental conditions that prevented the partners from sharing extra-linguistic context. The partners in those data collections did not speak from a location within the world they conceived of during the task. These (often purposeful) limitations on the dialog activity of the subjects enable the data to be used for many interesting analyses, but make it impossible to use the data as evidence for modeling other linguistic behaviors of current interest, specifically in relating collaborative dialog to 3D space.

Our corpus, the OSU Quake2004 corpus (Byron, 2005), is meant to provide a new source of attested language examples to be used by researchers who wish to study situated dialog. The corpus includes spontaneous English dialogs that record two partners performing a treasure-hunt task in a graphically-represented world, rendered on their computer monitors from a first-person view. The problem domain exhibits the following characteristics:

1. The partners have asymmetric knowledge of the goals of the treasure hunt task, resulting in one partner assuming the role of the 'leader', but both partners have equal capabilities within the task world to move about and manipulate the world. Therefore, the task initiative is equally shared between the partners.

2. The partners can move about in a graphically-rendered 3D world, and their perceptual access of the world is limited by their position at any one time. The task world is therefore not fully observable.

3. Because the two partners are each mobile in the world, their experience and knowledge of the world can di-

verge from each other. In other words, they have both shared experiences and private experiences over the course of a problem-solving session. This is different than other corpora, in which the participants' knowledge of the task is kept synchronized through the discourse itself.

4. The partners observe sudden state changes to objects in the world, and they also observe task-relevant objects undergoing state changes in their accomplishment of the task. This knowledge is received through perception rather than through the linguistic channel, but nevertheless it modifies the focus of their attention. Therefor, in this corpus, the attentional state of the dialog participants cannot be modeled from information in the dialog alone.

5. The partners do not know what is in the task world when they begin the task. Both partners must explore the world to gain a mental model of the spatial arrangement and what they are able to manipulate in the world. This is unlike the ATIS or TRAINS93 dialogs, in which one participant plays the 'system' who has perfect knowledge of the world.

6. Objects mentioned by the speakers are not always co-present in space with the participants at the moment when they want to mention them. Therefore, exophoric references are sometimes accompanied by gestures and sometimes gesture is not possible. This differs from other studies of exophoric reference, which concentrate on exophors that are accompanied by gesture or gaze (Bolt, 1980; Kaiser et al., 2003).

The corpus is distributed through the OSU webpage: http://slate.cse.ohio-state.edu/quakeref. The corpus is distributed for noncommercial research use only.
This report contains a summary of the data collection conditions and examples of the resulting data and analysis we have conducted so far. Additional details on the experimental setup, including exact copies of the subject instructions, are available in the OSU Technical Report (Byron, 2005).

2. Data Collection Conditions

2.1. Description of Subjects

The dialogs in this corpus contain human-human spontaneous conversational data recording the collaboration of two partners performing a treasure-hunt task within a virtual-reality world. The language of the study is English, and all subjects were native speakers of North American English who were university students or administrators. Subjects were enrolled in pairs, so as a result, the pair of partners in each recorded session knew each other. We recruited sets of partners together to mitigate against the risk that they would feel inhibited or intimidated by any disparity in their experience with computers or computer games, which they might experience if they were partnered with a random stranger. Each participant was compensated $5 cash at the end of the session, which took approximately one hour including equipment setup time and debriefing. The recorded sessions are from 9 to 35 minutes in length.

2.2. Virtual World used in the Experiment

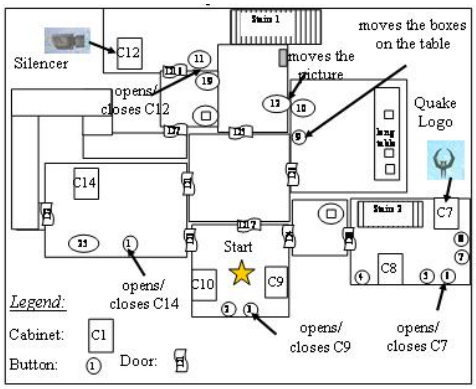

Figure 1: Line-drawing of the upper portion of the virtual world

Our experiment used a multi-player computer game engine called QuakeII to produce a 3D rendering of the small world in which the subjects performed their task. We chose to use a graphically-simulated world in this study rather than using the actual world because the simulator keeps precise track of the position of objects and the state of the world, we can exactly calculate what is in view for each subject at each moment in time, and an intelligent software agent inserted into this world can make sense of the physical world without requiring us to build working computer vision or robotics platforms for the embodied agent.
Subjects in the study started up the software and launched themselves as *players* in the game world. The world that we constructed for this study consists of the interior of a building containing around a dozen rooms, two staircases, an outdoor balcony, and a long hidden passage (Figure 1 shows a map of the upstairs portion of the world). Most of the rooms in the space contain only a few objects. Figures 2 and 3 show two example screen shots of rooms in the collaboration task world.
QuakeII is software that was released in the early 1990's and was the graphical rendering/physics engine for many of the first widely available multi-player first-person shooter computer games. The company that sold QuakeII, Id Software, has produced more recent versions of the game, and has therefore made the source code of the older version available for free under the Gnu Public License. The benefits of using this freely available game are that it runs on consumer hardware platforms and can be built to run under a variety of operating systems (Windows, Linux, MAC OSX), which will allow other researchers to replicate our study without purchasing expensive computers that specialize in producing high-quality graphics. Second, because the source code is freely available, it can be modified to fit particular experimental purposes, for example to add instrumentation/logging for an experiment, to change the behaviors of the world, or to add software bots as partners in the game.
In our experiment, QuakeII ran on Dell workstations running Windows XP. Although the QuakeII software is free source, some datafiles and textures needed to run the game

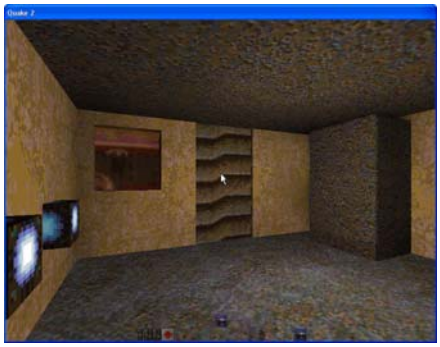

Figure 2: An example room showing (from L to R), buttons a window, a door, and a cabinet

Figure 3: Another example room with different textures

are not included in the generally-available downloaded files, so some files from the retail disk need to be installed on each workstation running QuakeII. We had no difficulty finding licensed copies of the retail QuakeII game for sale on the resale market. QuakeII source code is available at http://www.idsoftware.com/business/techdownloads/.

A limitation of QuakeII is that it can only render virtual worlds built with particular map-building software. Section 5. lists the software and websites we used to construct the virtual world. We were careful to select character models that did not violate intellectual property, for example characters from cartoons or computer games. The available avatars were a penguin, a UFO, a floating/flying girl that was holding a fireball, and a rabbit on a pogo stick.

2.3. Subject's embodiment in the virtual world

The experiment included two phases: first a training exercise, then a collaborative task phase. For both the training and collaboration phase, the subjects performed tasks in the graphically-rendered world shown to them in first-person perspective on a computer monitor placed on the desk in front of them. The term "first person graphics" refers to the fact that each person's view of the world is rendered from the perceptual locus of someone standing at a particular position and orientation within the world. In the first-person rendering we used, the subject doesn't see any portion of his own body, although an avatar (a graphical character) representing each subject is rendered by the graphics engine at his current position in the virtual world, and the avatar of one subject is visible to the other subject when appropriate. The subjects used the arrow keys on their keyboards to change their position or orientation in the virtual world. For example, "Turning to the right" is accomplished by pressing the right-arrow key, which results in the scene depicted on the computer monitor moving as it would if he turned his head to the right while standing in the scene.

Subjects did not have control over fine-grained movements such as using arm movements for communicative gestures, so the body postures that the subjects were able to make use of for communicative purposes were very limited. Each partner could ascertain the approximate gaze direction of the other partner's avatar, and could generate large-scale pointing gestures such as moving closer to an object of interest. Many of our subjects were not experienced with first-person computer games, and had less control over such signals. Some subjects seemed to move as little as possible, for example, rather than panning their field-of-view to place an object of interest at the center of their view, subjects would sometimes pan or move forward just until the object was in view, and then stop. Subjects who had prior experience playing QuakeII were better able to use (or modify in some cases) the command keys to control their avatar in the task world.

2.4. The task

After familiarizing themselves with the QuakeII game controls, both subjects joined the experiment world map, so that they could collaborate on the set of tasks that constitute the experiment proper. The task the subjects were asked to perform was a basic treasure hunt, finding objects in an unknown world. One subject received printed instructions depicting the specific tasks to be completed; this subject took on the role of *leader*. The subjects were not provided with a map, but were expected to find and reposition seven objects within the virtual world (by either moving the object to a different room, or activating a wall switch which caused the object to change position); the task descriptions were purposefully non-linguistic (with pictures demonstrating "before" and "after" conditions). The participants were not given any direction on how they should collaborate, e.g. which partner should do what sorts of tasks in the world, or what order to complete the tasks in.

The small blue/black boxes (shown on the left-hand wall in Figure 2) are buttons that depress when they are approached, which are the wall switches that trigger various state-changes of items in the world. In the sample scene shown in Figure 2, pressing one of the buttons opens the doors of the large cabinet/armoire shown in the room at the right of the frame. The wavy brown square on the wall in roughly the center of Figure 2 is a sliding door, which will open to reveal another room if a player gets close to it. In addition to pushing buttons and opening doors, the participants can pick up and drop objects they find in the world. Figure 4 shows one task from the instructions, in which a box is supposed to be moved from one end of a table to another.

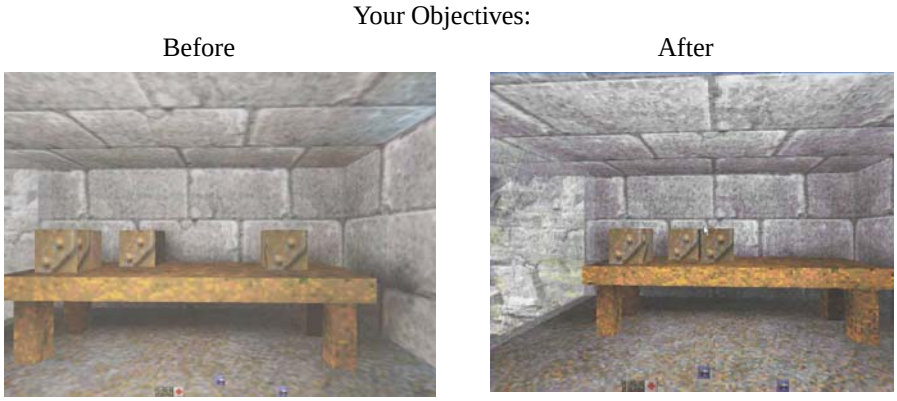

Figure 4: A portion of the instructions provided to the subject playing the Leader role

For more clarification on the stimulus and possible actions of the subjects during the tasks, the reader is encouraged to watch the Guided Tour video available on the corpus website, a 15-minute MPEG movie which guides the viewer through the entire map and explains the user controls. A blueprint view of the map showing the layout of the two-story virtual world used in the experiment is included in the technical report (Byron, 2005).

2.5. Recording procedure

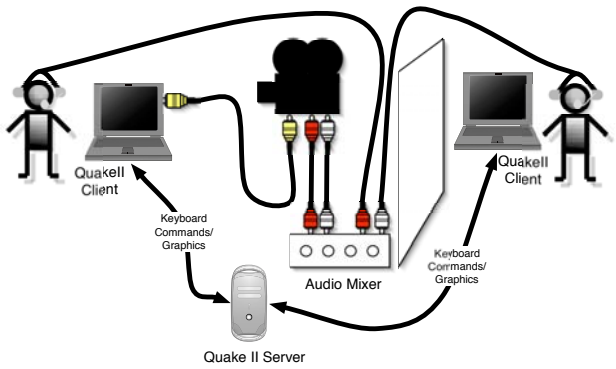

Figure 5: Sketch of recording hardware

For each session, the corpus contains two movies, one recording the virtual-world experience of each partner, a separate audio recording in WAV format, and orthographic transcriptions of the audio. Figure 5 sketches the hardware used in our recording process. Partners spoke to each other through headset-mounted microphones with enclosed-earcup headphones (Sennheiser HMD280-Pro noise-cancelling supercardiod microphones). The audio from the two subjects was combined with a Mackie Eurorack UB1202 stereo mixer and recorded using the stereo inputs of a Canon digital video camera. Using the multi-track mixer, we separated the audio by panning the inputs full-left and full-right: the leader's audio was recorded panned full-left and the follower's was recorded panned full-right. Because the two subjects were sitting approximately 20 feet apart in an office environment, there is a slight amount of bleed-through of the other speaker's voice into the wrong channel.

The video-stream going to the leader's computer monitor was also sent to the video input of the digital video camera to be recorded.[1] Therefore, the audio signal and video experience of the person playing the leader role is aligned by virtue of simultaneous recording to the video camera. For the person playing the follower's role, the video track of their experience in the QuakeII world was recorded after the session was completed, using the replay capability available in QuakeII, and once again feeding the video stream from the computer monitor to the video camera. The audio track containing both audio streams was added onto the video record of the follower's experience, and manually aligned. In order to confirm that the re-recording of the playback of the follower's experience was accurate, we also replayed the leader's viewpoint and verified that it was identical to that which was captured on the camera.

2.6. Annotation

The dialog recordings have been orthographically transcribed. The transcripts do not show timing information, such as overlapping speech or word alignment with the audio file, but plans are in place to complete an alignment using the Anvil toolkit (Kipp, 2004). Transcription practices for non-words and abandoned utterances used the ICSI meeting corpus guidelines (Janin et al., 2003).

3. Sample Data from the corpus

Figure 6 shows a portion of the dialog in session 10. The partners are in a room together, and the leader (dialog lines marked L) is describing the task that must be accomplished to the Follower (marked F). Events external to the dialog are marked with symbols at approximately the point at which they occur. Once the Follower finds the correct trigger button, at line 138-1, the object moves into place. This fragment demonstrates the fact that both partners observe changes to the world together, and the events or situations they perceive become part of their mutual knowledge of the world. Both partners react when the plaque moves, demonstrating their mutual knowledge of the event.
Figure 7 shows another dialog segment from the corpus. In this segment, the Follower presses a button after 137-3 (at

[1] We attempted to do this recording with a variety of screen-capture tools, but found that none of them could keep pace with the frame rate produced by QuakeII.

27-1: L: that brick room straight ahead we need to go in there and move some stuff around ‡
128-1: F: oh okay
129-1: L: so you see that thing on the wall on the right
130-1: F: the plaque-looking thing
131-1: L: you need to pick that up and move it ∗ to the wall um the left to the adjacent wall
132-1: F: okay this right ∗
133-1: L: yeah
134-1: F: um ∗ and control is to pick up an object right ∗
135-1: L: right
136-1: F: yeah it doesn't want me to pick it up
137-1: L: um
138-1: F: yeah oh well maybe I gotta touch this †
139-1: L: oh there you go that's what you had to do
140-1: F: okay
141-1: L: okay we're finished with that room
‡: Both partners move into the brick room
∗: The follower presses the *pickup* key, the CTRL key
†: The follower presses a button on the wall, which causes the plaque to move

Figure 6: Sample dialog fragment from session 10

137-2: f: so we have to find the quake logo
137-3: f: oh well wait ∗
138-1: l: oh the box disappeared
139-1: f: oops ∗
140-1: l: oh
140-2: l: it's back again
141-1: f: okay
141-2: f: so I won't bump **that** again
∗: The follower presses a button on the wall, which causes a box to disappear

Figure 7: Sample dialog fragment from session 5

∗), which causes a box to disappear, and presses it again after 139-1, which makes the box reappear. Although the button has not been mentioned in the discourse, it is salient enough to be the referent of the pronoun *that* in utterance 141-2. This demonstrates the fact that the attention of subjects in the QuakeII world is not completely controled by their discourse, but also by the events they perceive in the world.

3.1. Example results produced from the corpus

In the initial months of analysis, we have used this data in three interesting experiments, which we mention here as a way of potentially stimulating the reader's imagination for using this data:

Gaze is a record of visually-guided attention, just as the discourse is a record of linguistically-modulated attention. Therefore, it should be possible to track a speaker's gaze fixations to objects in the virtual world as a prediction of what item he will speak about in upcoming discourse. However, this task is complicated by the fact that gaze can change quickly and happens on an independent time-scale to language, and also much of the time the speaker's gaze may fall on un-salient objects such as the ceiling or walls of the room he is in. In (Byron et al., 2005b), we created an equation for calculating the visual attentional history for each speaker, taking into account the frequency and recency of looks to an object as well as how visually-distinct each object is. We were able to predict which item would be mentioned in upcoming discourse using only visual salience almost as well as we did using a standard metric of discourse salience. The technique could be combined with linguistic salience to produce a multi-source attentional state estimate.

Noun phrases come in a small but closed set of forms, such as pronouns, indefinite descriptions headed by a common noun, definite descriptions with demonstrative determiners, etc. At any one moment in time when a speaker wishes to refer to a particular item, he could potentially phrase that expression in a variety of ways, but the choices that speakers do make tend to correlate with a set of properties of the extra-linguistic context. For example, a speaker's assumption of what his addressee is attending to may cause him to use a pronoun rather than a descriptive noun phrase. The set of factors that account for the distribution of noun phrase forms is still only partially understood, especially in complex discourse contexts such as our experimental QuakeII world. In (Byron et al., 2005a) we collected a number of factors such as topicality in the discourse, mutual knowledge of the speakers, and the spatial configuration of the world, to predict which form would be used for a referring expression. Using a decision tree created with the WEKA toolkit, we induced a decision procedure whose prediction matched the form actually used by speakers in our corpus on 51% of the training instances. The decision procedure could be put to use in natural language generation systems for situated dialog.

The discourse recorded in this corpus contains reference to spatially-extended objects that have a distance relation to the speaker. Therefore, we see many uses of the proximity-marked expressions *here/there, this N/that N, these Ns/those Ns*. Their distribution, however, is only partially explained by spatial distance. In (Byron and Stoia, 2005), we found that they are also used to convey the expected agency of an act in the task world, similar to an indirect speech act. For example, a speaker might say "What's in there?", even though he is physically close to the reference object, if he expects his partner to answer the question, but if he says "What's in here?", the phrasing implies that he intends to answer the question himself.

These dialogs are an exceptionally rich source of attested evidence to investigate the effects of extra-linguistic context on the linguistic behavior of partners collaborating on a task. As the partners move through the virtual world, multiple factors influence their attention and therefore their phrasing of linguistic constituents, such as their position in the world, the task they are focused on, and the history of the discourse they are engaged in. The dialogs contain many spatial predicates that include an inferred point of view and frame of reference. The dialogs include an extraordinarily high number of reduced references, such as pronouns. Interpretation of each utterance must include complex contextual enrichment that utilizes aspects of the situation as well as the discourse history.

4. Conclusions and Future Work

The QuakeII game engine has turned out to be a useful tool for collecting spontaneous, collaborative dialog in a simple virtual world. The subjects in our experiment gave every indication that they were fully immersed in the virtual world and speaking to each other through their avatars as though they were actually located within the task space. The quake world allowed us to explore many properties of task-oriented dialog that have not been available in pencil-and-paper information navigation tasks like those used in the ATIS or TRAINS dialog collections. The spatial extent of the task world allows researchers to explore spatial constraints on language and task structure, mutual knowledge issues in a world where the interlocutors gain knowledge about the world independently of the dialog, and grounding behavior when subjects can use their physical position in the world to control the attention of the interlocutor. These are just a few specific properties of situated dialog that can be explored using the OSU Quake2004 corpus. We hope this corpus stimulates a wide range of research on these and related issues, to help the spoken dialog systems community make progress on the challenging task of dialog for situated problem-solving agents. We look forward to sharing it with other researchers.

Our future work includes varying the embodiment conditions of subjects in the task world and collecting additional dialogs. For example, subjects might complete the task without being allowed to be in the same room, or one subject will complete the task in a wizard-of-oz configuration in which he is lead to believe that his partner is a software agent.

5. Resources used in corpus preparation

1. Quake map builder: QERadiant (www.qeradiant.com) and Quake Army Knife (http://dynamic.gamespy.com/˜quark/)

2. Quake wall textures: Wally (http://www.telefragged.com/wally/).

3. Avatars and furniture models: planetquake (http://www.planetquake.com/polycount),

4. Video editing: Adobe Premier Pro (academic pricing makes it affordable)

5. Transcription: To create this corpus, we used soundscriber, which runs on Windows only. It is helpful for transcribing because it loops each small segment of audio several times, allowing the transcriber to keep typing instead of hitting the rewind key. It is downloadable from http://www.lsa.umich.edu/eli/micase/soundscriber.html

6. References

ATIS. 1993. The ATIS corpus http://www.ldc.upenn.edu/catalog/catalogentry.jsp?catalogid=ldc93s4a.

R. A. Bolt. 1980. Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270.

Donna K. Byron and Laura Stoia. 2005. An analysis of proximity markers in collaborative dialog. In *Proceedings of the 41st annual meeting of the Chicago Linguistics Society*. Chicago Linguistic Society.

Donna K. Byron, Aakash Dalwani, Ryan Gerritsen, Mark Keck, Thomas Mampilly, Vinay Sharma, Laura Stoia, Timothy Weale, and Tianfang Xu. 2005a. Natural noun phrase variation for interactive characters. In *Proceedings of the First Annual Artificial Intelligence and Interactive Digital Entertainment Conference*, pages 15–20, Marina del Rey, California, June. AAAI.

Donna K. Byron, Thomas Mampilly, Vinay Sharma, and Tianfang Xu. 2005b. Utilizing visual attention for cross-modal coreference interpretation. volume 3554/2005, pages 83–96. Springer Lecture Notes in Computer Science: Proceedings of Context-05.

Donna K. Byron. 2005. The OSU Quake 2004 corpus of two-party situated problem-solving dialogs. Technical Report OSU-CISRC-805-TR57, The Ohio State University Computer Science and Engineering Department, September.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

P. Heeman and J. Allen. 1995. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium.

Charles T. Hemphill, John J. Godfrey, George R. Doddington, John Garofolo, Jonathan Fiscus, Nancy Dahlgren William Fisher, Brett Tjaden, and David Pallett. 1993. Hcrc map task corpus: Linguistics data consortium catalog no. ldc93s12.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI meeting corpus.

Ed Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Phil Cohen, and Steven Feiner. 2003. Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In *Proceedings of the 5th International Conference on Multimodal interfaces (ICMI 2003)*, Vancouver, B.C., Canada.

Michael Kipp. 2004. *Gesture Generation by Imitation - From Human Behavior to Computer Character Animation*. Dissertation.com.

M.A. Walker, A. Rudnicky, R. Prasad, J. Aberdeen, E. Owen Bratt, J. Garofolo, H. Hastie, A. Le, B. Pellom, A. Potamianos, R. Passonneau, S. Roukos, G. Sanders, S. Seneff, and D. Stallard. 2002. DARPA communicator: Cross-system results for the 2001 evaluation. In *ICSLP-2002:Inter. Conf. on Spoken Language Processing*, volume 1, pages 273–276, Denver, CO USA, Sept.