Department of Computer Science and Engineering

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

Donna K. Byron
email:dbyron@cse.ohio-state.edu

The Ohio State University
Department of Computer Science and Engineering
Columbus, Ohio 43210

Technical Report: OSU-CISRC-805-TR57

September 6, 2005

Abstract

This report describes the OSU Quake2004 corpus of two-party exploration dialog in a simulated 3D world. This report describes the experimental setting and recording conditions. The corpus is freely available for research and educational use, and is downloadable from the website http://slate.cse.ohio-state.edu/quakeref.

This material is based on work supported by the OSU Computer Science and Engineering department and the General Electric Faculty for the Future grant.

Contents

1 Motivation and introduction to the domain

The last few years have seen a growing interest in creating embodied conversational agents that act as partners to humans in a variety of problem-solving domains, such as robotic partners in search and rescue or in-home applications, and 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 kinds of information navigation tasks.

The corpus described in this report was collected as the first step in our attempt to understand how dialog is used by partners performing tasks in 3D space. Partners performing such situated tasks use dialog to coordinate not only their actions relative to the ongoing task, but also their beliefs about the world around them. Our goal in collecting this corpus is to provide researchers with a set of human-human dialogs to study, that help us understand the linguistic behaviors needed by automated agents that are built to collaborate with human partners in similar 3D problem-solving domains.

This report describes the details of the experimental stimuli and technical recording specifications of the corpus. The corpus itself is available for research and educational purposes, and is currently distributed at http://slate.cse.ohio-state.edu/quakeref.

1.1 Why collect a new corpus?

The existing data collections of human-human task-oriented dialog were collected under conditions that limit their utility to provide information about dialog behavior on situated tasks. Existing data collections such as TRAINS (HA95), Maptask (ABB$^+$91), or ATIS (cor) have provided the research community with an invaluable data resource to investigate dialog phenomena such as grounding, speech act patterns, disfluencies and spoken language parsing, etc., but the partners in those data collections were not speaking face-to-face, and were not able to move about in a spatially-extended task world. When dialog partners work on their task within a world that they can jointly perceive, manipulate, and discuss, this provides additional constraints on the language they use and the order in which the task is accomplished. For an automated agent to be able to cooperate in these domains, it must be able to calibrate linguistic acts against the spatial configuration of the task world, the current state of the task and of the discourse, and the partners' beliefs about all of this. This creates new challenges for information integration in a conversational agent, that have not been systematically studied in the past. Rather than imagining a solution to a problem, as in the TRAINS93 or Monroe corpus dialogs (Ste00), in our experiments, the partners are carrying out the task at the same time they are discussing it. This also changes the way they discuss results and re-plan in the face of unexpected outcomes. To create automated agents that can navigate this coordination between language and the world, we first need to record how humans perform this coordination with each other.

Our corpus, the OSU Quake2004 corpus, includes spontaneous 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 was chosen to exhibit the following characteristics:

- 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.

- The two partners are instructed to work together to solve the task.
- 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 partners speak spontaneously to each other.
- 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.
- Because the two partners are each mobile in the world, their experience and knowledge of the world can diverge from each other. This is different than other corpora, in which the participants' knowledge of the task is kept synchronized through the discourse itself. In this domain, the partners are sometimes together and sometimes apart, therefore they sometimes have joint perceptual access to events in the world. This shared experience modifies their attentional state and their mutual knowledge, both cognitive effects that might manifest in their linguistic behavior.
- The objects of interest to the task are not always co-present in space with the participants at the moment when they want to talk about the objects. Therefore their exophoric reference (reference to objects in the task world) are sometimes accompanied by gestures and sometimes gesture is not possible. This provides a different condition than other studies of exophoric reference.

The participants performed their task in our experiments in a virtual world rather than in a real world, which eases the difficulty of recording (e.g. we did not need wireless microphones or a large multi-room interior physical space to conduct the experiment) while still simulating moving about in a realistically-large space. Each partner in the virtual world is represented by an animated avatar character, which the participant moves by pressing the arrow keys on his computer keypad. The participant never sees the body of his own avatar, but he can see the avatar representing his partner. This embodiment condition limited the degree to which each speaker could use his body to convey conversational information - they could not use fine-grained gestures other than moving their entire body closer to an object of interest, and it is somewhat difficult for the partners to track each other's gaze direction. So some fine-grained gesture and gaze clues, which partners working in the real world would have utilized, are artificially absent from this data. Other than these differences from the actual world, the partners working in the virtual world seem to be sufficiently engrossed in the task, and seem to buy into the pretense that they are actually moving about in the virtual task world.

1.2 Uses for the corpus

Spontaneous spoken dialog corpora collected in the past have provided an invaluable resource for studying human dialog behavior. The dialogs collected in this experiment demonstrate human competence for spontaneous dialog as an element of collaborative situated problem solving. The dialogs can be used to examine interactional phenomena such as coordination and grounding, spatially-sensitive expressions such as "on your left" or "here", coordination of beliefs in a 3D world, prosodic characteristics of speech that has been produced while moving and observing a changing world, and the interaction between task structure, dialog structure, and spatial topology, among other things.

In the first few 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:

1. 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 (BMSX05), 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-modal attentional state estimate.

2. 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 (BDG$^+$05) 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.

3. 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 (BS05), 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.

2 Experimental method

This section provides detailed information about the experiment, including the subjects, the stimulus, the recordings, etc. Please email the author if any additional information not covered here is needed.

2.1 Subject recruitment and consent forms

2.1.1 Demographic profile

Subjects were recruited from student populations at the Ohio State University. They were required to be native speakers of North American English and to have normal or corrected-to-normal vision and no hearing impairments. Subjects were asked to sign up in pairs, so as a result, the pair of partners in each recorded session knew each other. We recruited the participants in pairs to mitigate against the risk that they would feel inhibited or intimidated by any disparity in their experience with computers or computer games 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.

2.1.2 Consent forms

We designed the consent process to happen in two phases, which allowed the participants to determine whether they consented to our distributing the audio recording of their session only after they knew what was on the recording. The text of the consent forms is shown in Appendix A-1. When the subjects arrived, they gave their consent to participate in the study before the experiment began. At the conclusion of the session, subjects were told that they would receive compensation regardless of their consent over the data, and then they were asked to decide what level of data usage consent they were willing to convey to us, and asked to sign the form choosing one of four data distribution levels:

- No consent: instructing the researchers to destroy the data
- Limited use: use the data only in the SLATE lab at OSU
- Unlimited use: the data will be distributed to the worldwide research community
- Defer consent: the subject could ask to listen to the audio recording or review the final transcript before consenting to distribution of the data.

2.2 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 a graphically-rendered world shown to them in first-person perspective on a computer monitor placed on the desk in front of them. First person graphics means that each person's view of the world is created from the perspective of someone standing at a particular position in 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. 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 you turned your head to the right while standing in the scene.

QuakeII renders the world as a true first-person view, and each player in the world cannot see any part of the body of his own avatar. Also, there are no controls for fine-grained movements such as arm movements that could signal linguistically-relevant 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.

2.3 Virtual World used in the Experiment

The map we used in the experiments is an interior space consisting of around 15 rooms and 2 staircases. Most of the rooms in the space contain only a few objects. Figure 1 shows an example screen shot of a room in the collaboration task world, and the instructions in Appendix A-2.2 show several more views, which depict the type of interior spaces the subjects were moving about in during the experiment. The little blue/black boxes (shown on the left-hand wall in the Figure 1) are buttons that depress when they are approached, and some buttons trigger various state-changes of items in the world (some buttons have no effect). For example, in the sample scene, 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 the picture is a sliding door, which will open if a participant 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. The world was managed in run-time by the QuakeII game engine, which imposes certain limitations on what kinds of manipulations the subjects, as players, can perform on objects in the virtual world. Some subjects had prior experience with QuakeII embodiment, and so they were better able to use (or modify in some cases) the command keys to control their avatar in the task world.

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 used in the collaboration phase of the experiment is included as Appendix A-3.

2.4 Subject Instructions

The subjects were seated at two computer workstations in our lab which were separated by a partition so that the partners could not make eye-contact during the session. The computer workstations were typical personal computer workstations with 17" monitors and keyboard and mouse input devices. The two subjects were allowed to select which seat they would take, and the seat

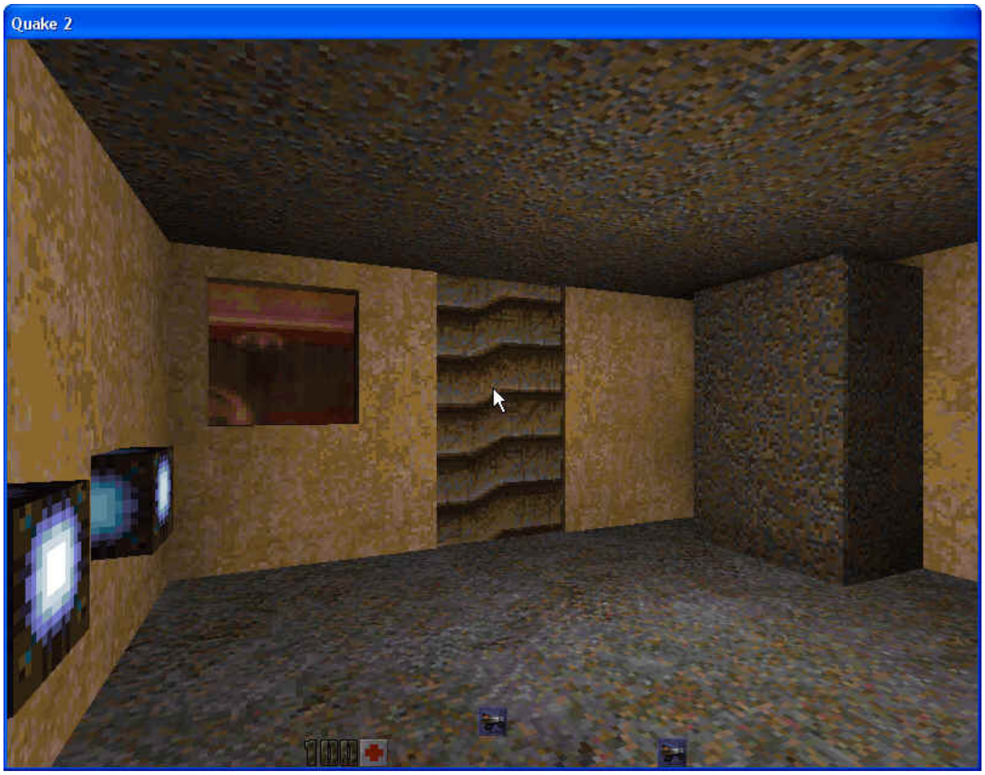


Figure 1: Sample interior view of a room in the treasure-hunt world

they chose determined who would be the leader, but this fact was not known to them at the time they selected which workstation to sit at. After they sat down, the experimenter informed them both that “Person X is sitting in the leader seat, so he/she will get the list of objectives for your task.” The leader was assigned in this manner so that the subjects couldn’t self-select as leader, and it was also clear that the experimenter had not chosen one of them to be the leader based on some assessment of leadership-worthiness, it was just the luck of the draw. Subjects were then fitted with headsets, which have fully-enclosed earcups to block out external noise, but allowed the subjects to hear each other’s voices through the headphones.

After being fitted with headsets, both participants were given instructions to complete a short training exercise on their own in a small training world.

Training Phase The instructions for the training phase are shown in Appendix A-2.1. The subjects were placed into a small world with only three rooms, and were instructed to push a button which opens a cabinet, pick up an object found inside the cabinet, and then drop the object. This simple task allowed them to get practice with all of the user controls that they would need in the actual experiment for movement and manipulating objects in the world. In the training session, each subject was in his own version of the world, so the two partners did not interact with each other. During the training phase, the subjects were also allowed to pick the character that would be their avatar for the session from four available characters.

The subjects’ time spent in the training phase was not recorded.

Collaborative Task Phase After completing the training phase, both subjects joined the same world map, so that they could collaborate on the set of tasks that constitute the experiment proper. The leader's instructions, shown in Appendix A-2.2, included screen-shots of the tasks to be completed, and both the leader and follower were given general instructions about the task. These instructions are also in Appendix A-2.2. The task included changing the position of seven objects within the virtual world. The subjects 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. Each pair of subjects was free to do their own negotiation to decide what to do next and which partner should do which task.

A blueprint (2D) map of the collaboration phase world is shown in Appendix A-3. Subjects took anywhere from 8 minutes to 35 minutes to complete the task. Some subject pairs did not complete the task correctly, since one of the objectives was particularly difficult.

2.5 QuakeII level and configuration

The experiment used the QuakeII game engine as a virtual world simulator. QuakeII was chosen because the software is open-source and runs on many platforms. It does not require special-purpose high-end graphical simulators running on special-purpose hardware, and it is relatively simple to build virtual worlds in QuakeII using free tools. Because QuakeII is an older version of QuakeII and therefore the source code is available, it is easy to modify, so we chose it in our research program so that we can build an automated agent that inhabits the same task world that the humans inhabited during this experiment. Other game engines could be used in a similar manner, for example HalfLife, Halo, and Neverwinter Nights have been used to build intelligent agents because they are designed to accept 'mods': new add-on software agents and world entities created by programmers externally and added to the game's standard repertoire.

Our experiment ran QuakeII on windows XP workstations. Although the QuakeII software is free source, some datafiles and textures needed to run the game are not included in the GPL files distributed on idsoftware.com, so each workstation running QuakeII needs to have QuakeII installed from the retail disk. We had no difficulty finding copies of the commercial QuakeII games for sale on ebay.

This section contains some details on how the experimental world was built. We used a variety of free tools to construct the experimental stim, including QEradiant (www.qeradiant.com) and Quake Army Knife (http://dynamic.gamespy.com/q̃uark/) to build the map, and a tool called wally to create wall textures (http://www.telefragged.com/wally/). The avatar models were found on planetquake (http://www.planetquake.com/polycount), and we were careful to select character models that did not violate intellectual property such as 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.

QuakeII source code is available on idsoftware.com
(http://www.idsoftware.com/business/techdownloads/). When the QuakeII software is installed, it includes a directory named baseq2 that includes a configuration file with many parameter settings used during gameplay. The settings that we found important for the experiment were:

- "rate" is the speed at which the avatars move, which we set on the low side to prevent them from zipping eratically through the level

- “skin” is the default avatar.
- set hand “2” will prevent showing a weapon in the frame of view, if the map allows the players to pick up weapons (but see item 2 below)

We made two other changes that were designed to create a more IRB-friendly gameplay experience for our subjects:

1. Disabling Damage: Even in what seem like completely harmless maps, there might be ways a player could get hurt - for instance, by getting pinched in a closing door or falling off a ledge. To prevent this, we completely disabled the function T_Damage in file g_combat.c that deals with damage.
2. Disabling the default weapon: The default configuration in QuakeII starts each player with a gun that is shown in view in the frame as though the player is holding the gun. To disable this feature required commenting out all of the code in the Think_Weapon function in file p_weapon.c, which removes graphics commands for the weapon, and also changing the default weapon's name, since otherwise it would still appear in each subject's inventory. This was done by changing the code in two places: in g_items.c, look for the following comment:

   ```
   /* weapon_blaster (.3 .3 1) (-16 -16 -16) (16 16 16)
   always owned, never in the world
   */
   ```

   (it's found in the declaration of a *very* large array called gitem_t itemlist[]). now go down a few lines, to the comment /* pickup */. Right after it, there should be the word ”Blaster”, inside quotation marks. Change it to whatever you like; I chose ”No tea” . Make sure to leave the quotation marks and comma as they are, of course. Now, go to p_client.c and find the function InitClientPersistant. Near its beginning there should be a call to FindItem with a parameter of ”Blaster”. Change that parameter to whatever you renamed the blaster, such as ”No tea”. Make sure it's *exactly* the same or QuakeII won't run because it won't be able to initiate the player properly.

Another feature of QuakeII that we used extensively was the ability to record a logfile of one session of a player's gameplay, which can be re-played by QuakeII. At replay time, QuakeII renders the video to the computer monitor exactly as it looked when the subject originally played the game, including the position and orientation of other players. This replay capability came in handy for recording both the leader and follower's visual record to tape without having to purchase two cameras. Recording a file is initiated by pressing tilde while the game is running, and typing the following at the prompt:

```
record <name>
```

This command stores a file named baseq2/demos/name.dm2, and to replay a log, execute quakeII with the command:

```
quake2 +map name.dm2
```

3 Technical specs of the distributed corpus

This section describes the technical details of the corpus recordings.

3.1 Recording Process and Equipment

Our experiment involves two people, each on their own computer and each seeing a first-person view of the task from their own position in the virtual world. This means that we needed to create two different audio-video recordings, one rendered from each player's perspective, to capture a time-aligned version of each partner's experience in the study. We used a different process to record the leader and follower experience, as described below. Only the leader's experience was recorded simultaneously to audio/video while the experiment was happening. The audio track of the follower's movie was dubbed onto the video record of his experience in the experiment.

To record the video stream that the subjects were seeing on their computer monitors as the experiment played out, we tried a variety of video-frame grabber and machinimation software (see Lessons learned below) to try to write the video frames directly to the computer's hard disk, but none of these techniques worked for us. So in the end our solution was to mirror the subject's computer monitor to the second video-output port on the computer, and use the SVideo connection on the computer's video card to feed the video stream to a digital video camera. We used a Canon Optura 200 camera with stereo input.

During the experiment, the camera was fed video from the leader's computer monitor and stereo audio was fed to the camera from our mixer board at the same time. This allowed us to record the video stream as viewed by the leader, along with the dialog from both partners, directly into the camera as it happened with no latency. To make the movie of the follower's experience, we replayed the QuakeII log of the follower's gameplay, and piped the rendered video frames to the video camera. This video track was captured in Adobe Premiere and the stereo audio track from the leader's movie was pasted onto the audio track in Premiere by hand. This dubbing process introduces the possibility for alignment errors between the audio and video track. The hand-alignment was performed by choosing an alignment point in the leader's movie when the follower and leader jointly perceived a visual event in the virtual world that happened simultaneously with an audio event, such as the onset of a word. The audio track was aligned in the follower's video in as close as possible to the same relationship to the visual alignment event, using human perception to judge the alignment.

In addition to recording the audio signal to the Canon digital movie camera, we also made a redundant stereo audio recording directly to hard disk using the Wavesurfer software and an M-audio Audiophile 2496 stereo card connected to a Linux workstation. However, these audio recordings were unusable because some audio frames were dropped (about a 1-second chunk of audio was dropped every 10 minutes). We have since experimented with using the same stereo audio card on an Apple G4 and it worked with no frame drop.

3.2 Video Properties

The movie files that are available on the website were converted to MPEG1 format using Adobe Premiere Pro software. The video properties of the files are: NTSC standard, 29.97 frames per second, 652 x 480 frame size.

3.3 Audio Properties

To make the audio recordings of the dialog, we combined the headset microphone output from both subjects into a mixer and recorded in stereo to the video camera. The microphones used are Sennheiser HMD280-Pro noise-cancelling supercardiod mics. The audio from the two subjects was combined with a Mackie Eurorack UB1202 stereo mixer and recorded using the stereo inputs of the digital video camera. Using the multi-track mixer, we separated the audio by panning the inputs full-left and full-right: the leader's microphone was recorded panned full-left and the follower's microphone 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 channels can be separated using the Sox audio tool. The audio is MPEG-1, Layer II audio, recorded at 16bit sample size and 16Khz.

Since the experiment took place in an office environment rather than in a noise-free environment, occasionally the audio track records the voice of other people in the lab, a telephone ringing, the voice of the experimenter, etc.

The separate audio WAV recording of each session, which is also available on the website, was extracted from the leader movie, so it has identical audio event timing as the leader's movie.

3.4 Alignment Issues

The movies for the leader/follower of a particular session start in approximately the same moment in relation to the beginning of the experiment, but there is no automatic clock synchronization between the two recordings. Therefore, a discourse event that occurs at offset X of the leader's movie does not necessarily occur at exactly offset X of the follower's movie, but it will be close.

Because of the offline process we used for making the follower's movie, very precise measurements cannot be made of the synchronization between the dialog and the visual stimuli presented to the follower during the experiment.

It is possible to align events from the movie with actual world events in QuakeII, such as the exact position of the subjects in the world at any particular moment in the movie using the dm2 files. These QuakeII logs can be made available on request. We hope to make them available along with synchronization software to convert log-time-stamps into movie time-stamps at a future date.

3.5 Preparation of Dialog Transcripts

The transcripts of the dialogs were created using a plain-text editor. A basic word transcription was created first, and the turns were then further segmented into utterances.

The transcription policy we follwed was adapted from the ICSI meeting corpus transcription rules (http://www.icsi.berkeley.edu/Speech/mr/icsimc_doc/index.html).

- Word Transcription
  - Partial words: Are indicated with a terminal hyphen
  - Non-standard words: We used a standard spelling convention for so-called communicative grunts that occur frequently in conversation, such as UH, UM, OH, AH, UHHUH, MMHM, OOPS, OP (an ope sound they make when they see something new). The same spelling was used no matter how the word was intoned or lengthened.

  - Non-word noises: Sounds such as door slams or a telephone ringing are enclosed in curly braces with the tag NVC: and a description of the sound. Vocalized sounds such as heavy breath, laughing, or sound effects like shooting are enclosed in curly braces with the tag VOC: and a description of the noise.
  - Words that are intoned with a Singing intonation are indicated with <singing> <singing>
  - Non-standard pronunciation: If the speaker used an informal version of a word such as "gonna", we typed it as it was said.

- Segmentation

  The dialogs contain many inter-clause and inter-turn pauses, so silence was not used as a criteria for breaking contributions into turns or utterances.

  - Each turn continued until a speaker-transition occured
  - Turns were segmented into utterances by choosing the smallest unit that maintained a predicate and its argument in the same utterance
  - Overlapping speech is surrounded by + +

3.6 Lessons Learned

Recording high frame-rate data to hard-disk in real time is still difficult with 2004 Wintel technology, and we used work-arounds for both audio and video recording (these problems could have been alleviated by using a Mac in the first place, but we didn't have a QuakeII port to OSX at that time). We initially tried to capture the video frames from each player's monitor in real time to disk while the experiment was running. This turned out not to work. Although there are many tools that record screenshots into a file (robodemo, camtasia, hypercam), we were not able to get any of them to work at the same rate that QuakeII delivers video frames (30 frames/sec), resulting in files that were mangled in one way or another.

Our initial attempt to record audio straight to hard disk also failed, although we did not detect this problem at first. Luckily, we had made a redundant recording to the video camera. I would recommend always recording redundantly to tape even if you don't anticipate hard-disk trouble, because a software recording device can crash during recording or when you press the SAVE button to write the file to disk, and an entire session can be lost.

Multimedia files are giant. Each leader or follower video coming from the video camera in AVI format for a 30 minute session is around 8Gigs, so there are 16Gigs total just for the original file, and your multimedia workstation needs that much working directory space to edit the file and render it into MPEG. Plan to buy alot of disk space before attempting to collect a movie-related corpus.

4 Obtaining the corpus

The corpus files are available on the web at http://slate.cse.ohio-state.edu/quakeref. Additional files will be added to the website as the corpus grows. The movie files are very large even though they have been compressed to MPEG-1 format. Each problem-solving session requires from 200Megabytes to as much as 1 Gigabyte disk space for storage. Using a high-speed internet connection, it should be possible for you to download these files in a couple of hours. If your installation cannot provide storage space for these files, contact slate@cse.ohio-state.edu to request a DVD of the corpus. The files are too large even for CD-based distribution. When the corpus reaches a more mature phase, we may contemplate DVD or CD distribution through the Linguistic Data Consortium, as long as it can be done at minimal cost to the user.

We will also be happy to share the QuakeII map files if you should want to run additional subjects using this same virtual world, or to use our map as a starting point to create your own.

5 Pointers to Resources

5.1 Resources used for corpus preparation and analysis

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

2. Transcription: The Soundscriber tool only runs on Windows, but it is useful for transcribing because it loops each small segment of audio several times, so the person transcribing can keep typing instead of hitting the rewind key. It is downloadable from http://www.lsa.umich.edu/eli/micase/soundscriber.html

3. The transcriber tool from LDC can be used to align the word transcripts with the audio signal http://www.ldc.upenn.edu/mirror/Transcriber/

4. Several tools have appeared recently for multi-media annotation, including MMAX and Nite and Anvil. Anvil did not work for us because it has a very small maximum file size. The Nite toolkit looks promising. We have used MMAX primarily, but it there is a licensing fee for the new version.

5.2 Resources used to create the Quake world

1. Level editors to make a quake map: QEradiant (www.qeradiant.com) and Quake Army Knife (http://dynamic.gamespy.com/ quark/)

2. Textures created with wally (http://www.telefragged.com/wally/)

3. Prefabs and avatar models were found on planetquake (http://www.planetquake.com/polycount),

6 Summary

The quake game engine has turned out to be a useful tool for collecting face-to-face dialog in a simple virtual world for collaborative problem solving. 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.

7 Appendices

A-1 Consent Form

A-1.1 First-stage consent

CONSENT FOR PARTICIPATION IN SOCIAL AND BEHAVIORAL RESEARCH

CONSENT TO INVESTIGATIONAL TREATMENT OR PROCEDURE

Description of the research

1. Description of the procedure or treatment. Reason for performing this procedure or treatment:

   The procedure involves two partners who discuss simple tasks. The conversational data will be used for language research and to support building spoken language technology.

2. Discomforts and risks that might reasonably be expected from participation in this study:

   Dizziness or queasiness are possible. Please stop the experiment if you experience any physical discomfort.

3. Possible benefits for participants or for society:

   Development of intelligent, conversational software agents.

4. Estimated amount of time it will take (number of sessions; length of each

   session, period of time).

   1 or 2 sessions of approximately 75 minutes total

5. Use of audiotapes, videotapes or photographs to collect information for this study.

   An audio recording will be made of your conversation during the study. If you consent, this data will be retained and a transcript of the conversation will be made. Both the audio data and transcripts will be used for spoken language technology research in the SLATE lab, and made available to the world-wide research community to be used for various other speech and language research projects. This data will be distributed through a scientific data clearinghouse such as the Linguistic Data Consortium (¡http://www.ldc.upenn.edu¿), which offers a repository of speech data that is kept indefinitely and made available to the research community. This data will not include any identifying information such as your name, but will include an anonymous numeric id (allowing multiple problems solved by the same speaker to be associated), and high-level demographic information such as your sex and university affiliation. You will be given an opportunity to allow or prohibit us from using your data in this way at the end of the session on a separate consent form.

CONSENT: I consent to my participation in research being conducted by Donna K. Byron and Eric Fosler-Lussier of The Ohio State University and his/her assistants and associates.

The investigator(s) has explained the purpose of the study, the procedures that will be followed, and the amount of time it will take. I understand the possible benefits, if any, of my participation.

The investigator(s) has explained the risks, if any, and I understand what they are.

I know that I can choose not to participate without penalty to me. If I give my consent to participate, I can withdraw from the study at any time, and there will be no penalty.

I have had a chance to ask questions and to obtain answers to my questions. I can contact the investigators at (614) 292-6350 or (614) 292-4890. If I have questions about my rights as a research participant I can call the Office of Research Risks Protection at (614) 688-4792.

I understand in signing this form that, beyond giving consent, I am not waiving any legal rights that I might otherwise have. My signature on this form does not release the investigator, the sponsor, the institution, or its agents from any legal liability for damages that they might otherwise have.

I have read this form or I have had it read to me. I sign it freely and voluntarily. A copy has been given to me.

(Signatures follow)

A-1.2 Second-stage consent

CONSENT FOR PARTICIPATION IN SOCIAL AND BEHAVIORAL RESEARCH

CONSENT FOR USE OR DISTRIBUTION OF DATA

Please check the appropriate box below:

[ ] I withhold consent for any use of audiotapes or transcripts.

[ ] I consent to use of audiotapes and transcripts by OSU researchers only. These materials will be kept indefinitely. I understand how the materials will be used for this study.

[ ] I consent to audiotapes and transcripts being made available to members of the world-wide research community. The materials will be made available indefinitely. I understand how the tapes will be used for this study.

[ ] I would like an opportunity to review transcripts of my conversation in the study before deciding whether to grant or withhold consent for their use by OSU researchers or distribution. If you checked box 4 above, please provide the following information so that we may contact you when the transcripts are ready for your review: I can be reached at: Home address:
Email:
Phone:

I have had a chance to ask questions and to obtain answers to my questions. I can contact the investigators at (614) 292-6350 or (614) 292-4890. If I have questions about my rights as a research participant I can call the Office of Research Risks Protection at (614) 688-4792.

I have read this form or I have had it read to me. I sign it freely and voluntarily. A copy has been given to me. Print the name of the participant: (Signatures follow)

A-2 Participant Instructions during the experiment

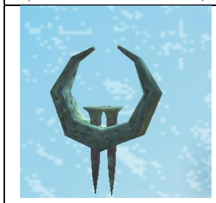
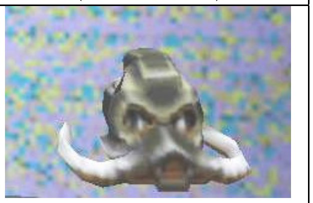
A-2.1 Instructions for the Training phase

Learning Phase

First, you will be put into a small level with no partner. This is a small level that you can use to get accustomed to the quake controls. The keyboard controls are posted on the tip sheet in front of your monitor.

You can ask to stop the experiment at any time if you feel queasy or dizzy. Please let the experimenter know as soon as you experience any discomfort.

1. **Moving around**: Use the arrow keys on the keyboard to move around.
2. **Manipulating the world**: There are only 2 items in the training level that you can pick up and move:

Quake logo (quad damage)	**Helmet** (rebreather)
	

Practice these actions:

(a) Pick up the helmet and quake logo

(b) Push the blue button to open the cabinet.

(c) Drop the logo and helmet near the cabinet.

Take your time getting accustomed to the interface, and let the experimenter know when you are ready to proceed.

3. **Pick an avatar**:

(a) Press escape

(b) Use the arrow key to move down to multiplayer and press enter

(c) Use the arrow key to move down to player setup and press enter

(d) Use the arrow keys to move between characters. You have 4 possible avatar characters to pick between.

(e) Press escape 3 times when you are done.

4. **Audio setup**: Make sure your headphones are comfortable on your head, and that you can hear yourself and your partner.

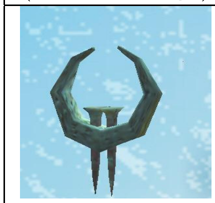
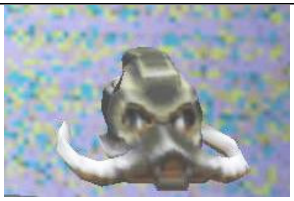
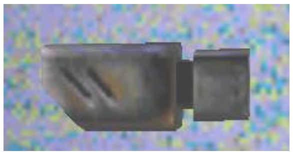
A-2.2 Instructions for the Collaboration phase

(This instruction page was given to both participants)

Treasure Hunt Task

In this level, you will work as a team to go on a modified treasure hunt. The team leader has pictures of how the world should look when you complete the game. The goal of the task is to change the world so that it matches the pictures. There are 4 things to change by pressing buttons plus 3 items to pick up and move to a new location. Try to complete all of the tasks, but it is ok if you cannot complete a task.

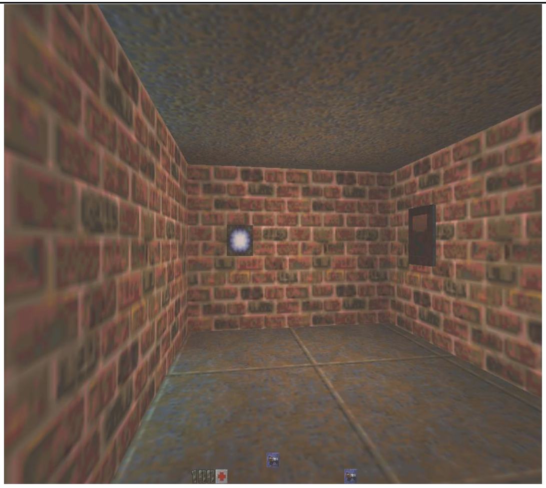
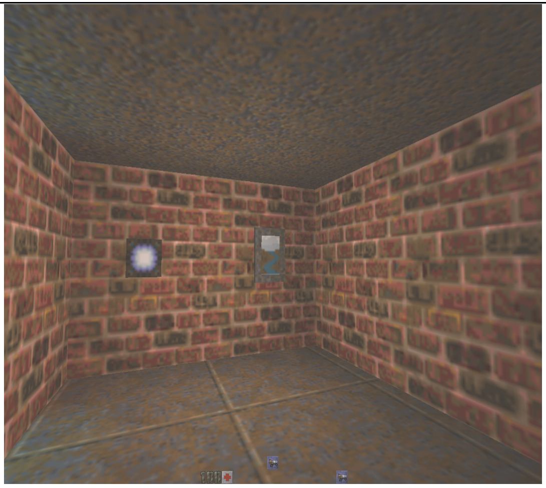
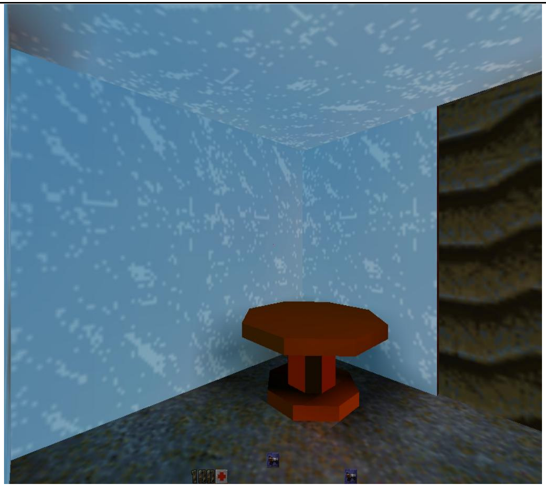
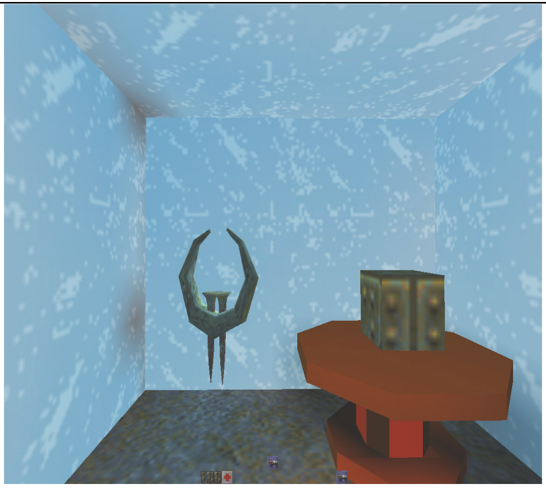
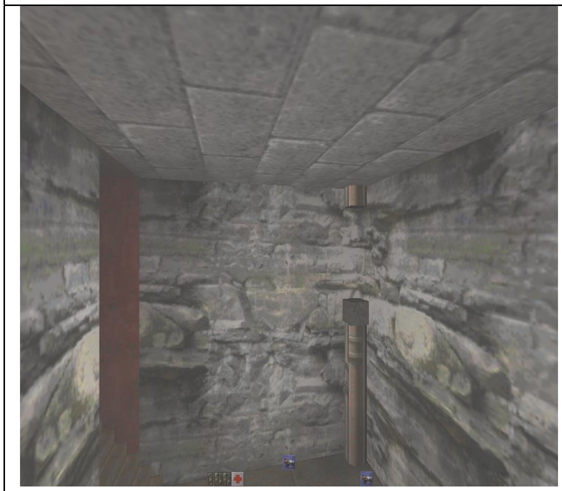
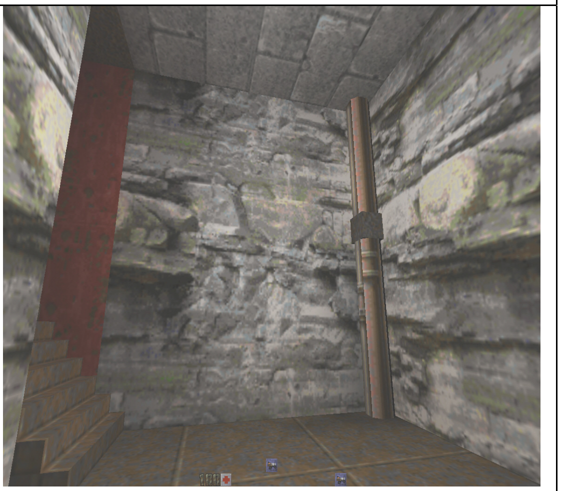
The three items you can pick up and move are:

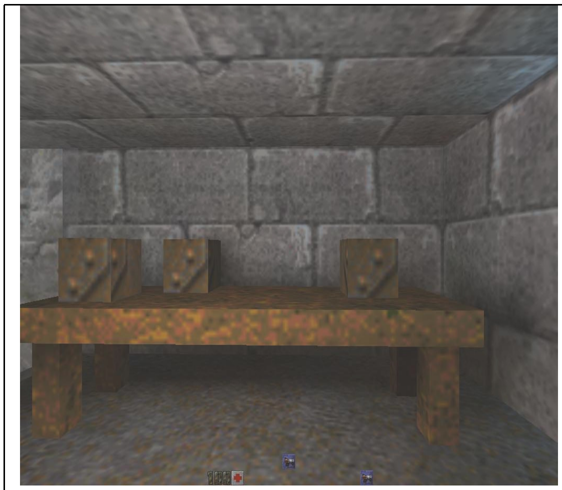
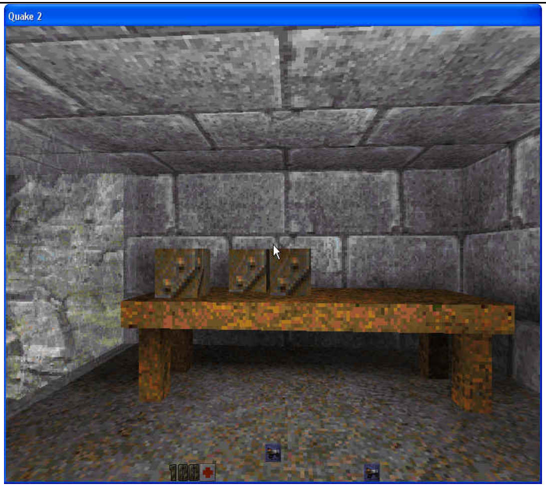
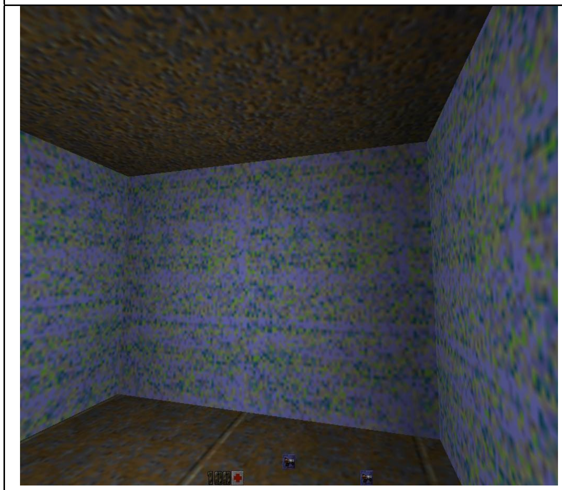
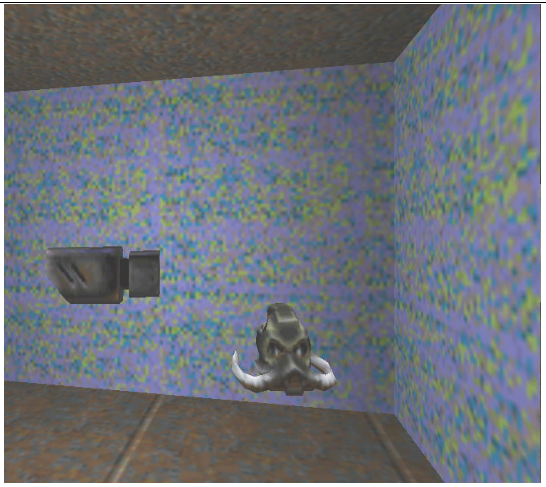
Quake logo (quad damage)	**Helmet** (rebreather)	**Silencer** (ammo clip/silencer)
		

The two pages that follow show the instructions presented only to the leader. The instructions in Figure A-2.2 show only the before/after pictures because I wanted the subjects themselves to choose how to describe the objects and their spatial arrangement, rather than putting words in their mouths. If they asked for clarification about what they were supposed to do, they were told that the pictures showed what the world would look like initially, and what they wanted to make it look like when they were finished.

Your Objectives

These pictures represent how the world should look when you are finished.

Before	After
	
	
	

Before	After
	
	

A-3 Map of the task world

A layout of the two-story virtual world used in the experiment is shown in Figure 2 and Figure 3. A walk-through of the world in 3D is also available on the corpus distribution web page.

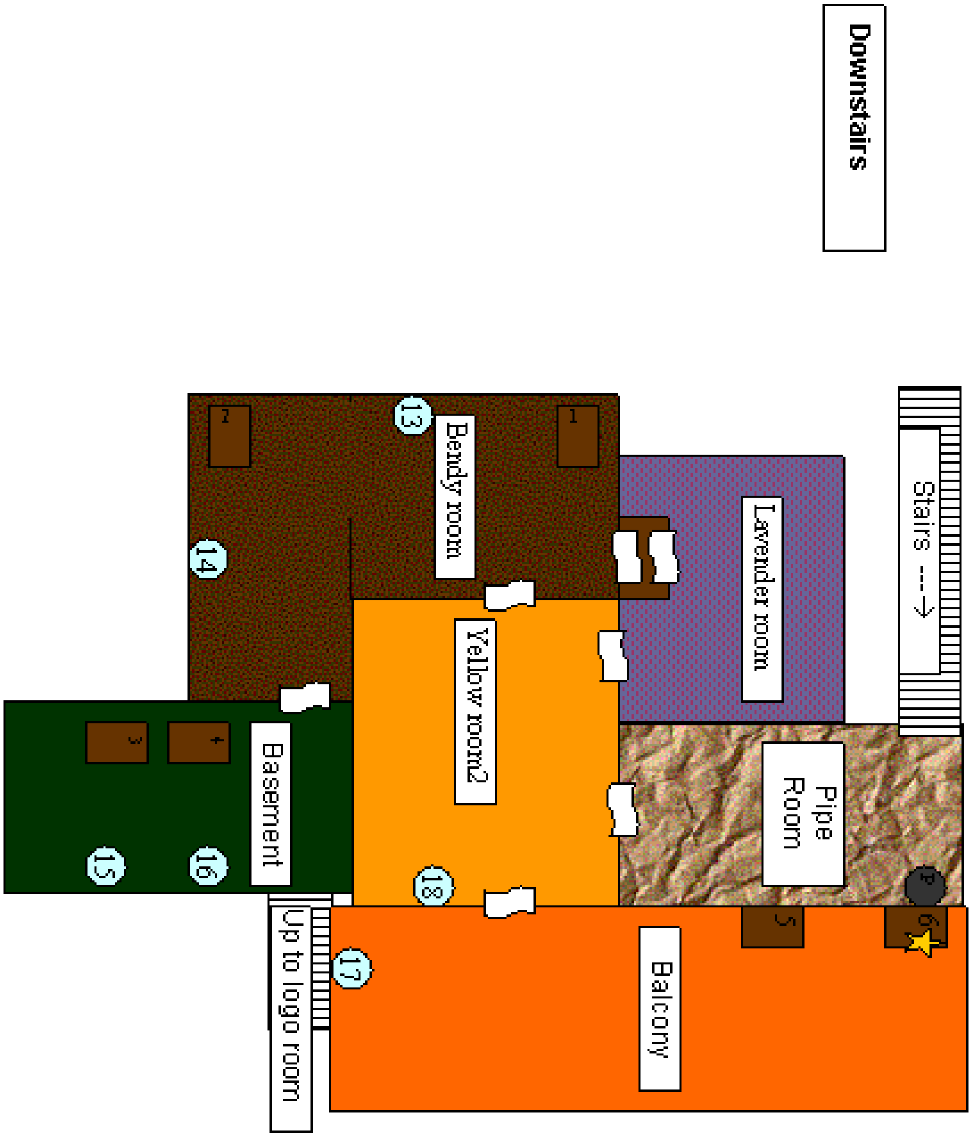


Figure 2: Map of the task world: downstairs

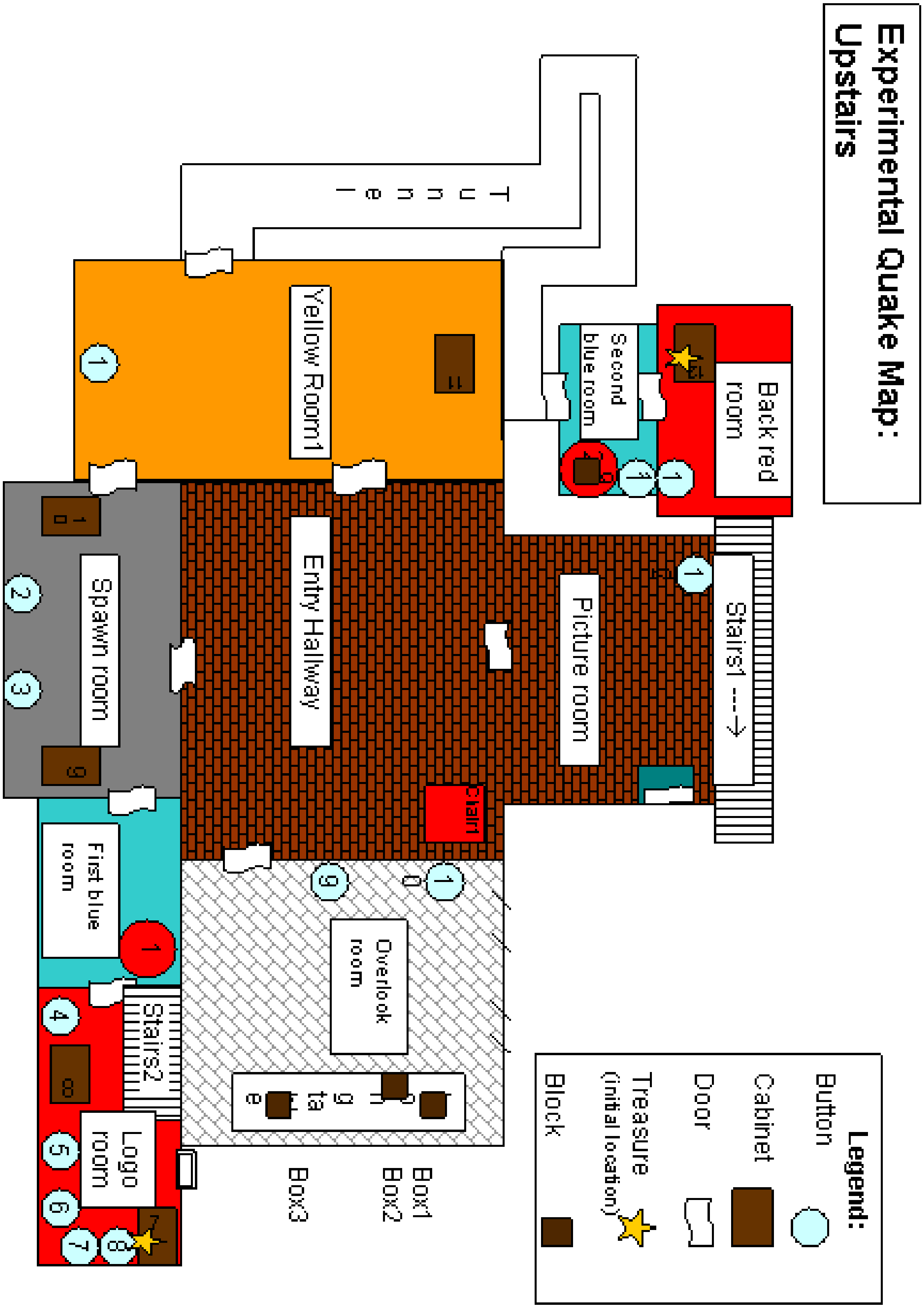


Figure 3: Map of the task world: upstairs

References

[ABB$^+$91] Anne H. Anderson, Miles Bader, Ellen Curman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry Thompson, and Regina Weinert. The HCRC map task corpus. *Language and Speech*, 34(4):351–366, 1991.

[BDG$^+$05] Donna K. Byron, Aakash Dalwani, Ryan Gerritsen, Mark Keck, Thomas Mampilly, Vinay Sharma, Laura Stoia, Timothy Weale, and Tianfang Xu. 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 2005. AAAI.

[BMSX05] Donna K. Byron, Thomas Mampilly, Vinay Sharma, and Tianfang Xu. Utilizing visual attention for cross-modal coreference interpretation. volume 3554/2005, pages 83–96, 2005. Springer Lecture Notes in Computer Science: Proceedings of Context-05.

[BS05] Donna K. Byron and Laura Stoia. An analysis of proximity markers in collaborative dialog. In *Proceedings of the 41st annual meeting of the Chicago Linguistics Society.* Chicago Linguistic Society, 2005.

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

[HA95] P. Heeman and J. Allen. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, 1995.

[Ste00] Amanda Stent. The monroe corpus. Technical Report 728, University of Rochester Computer Science Department, 2000.