The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus

David Herrera*, David Novick*, Dusan Jan†, David Traum†

*The University of Texas at El Paso
†Institute for Creative Technologies, University of Southern California
herrera78@gmail.com. novick@utep.edu, {jan,traum}@ict.usc.edu

Abstract

To help answer questions about conversational control behaviors across cultures, a collaborative team from the University of Texas at El Paso and the Institute for Creative Technologies collected and partially coded approximately ten hours of audiovisual multiparty interactions in three different cultures and languages. Groups of four native speakers of Arabic, American English and Mexican Spanish completed five tasks and were recorded from six angles. Excerpts of four of the tasks were coded for proxemics, gaze, and turn-taking; interrater reliability had a Kappa score of about 0.8. Lessons learned from the multiparty corpus are being applied to the recording and annotation of a complementary dyadic corpus.

1. Introduction

Conversational corpora are important for a variety of purposes, including analysis of conversational behaviors, evaluation of theories about behavior, and training data for machine learning algorithms. We are particularly interested in comparing and contrasting conversational control behaviors across cultures. This effort includes each of the above corpus requirements. We need basic data for analysis of the differences in these behaviors across cultures. We will use the data to provide parameters for culture-specific models of virtual human conversation (Jan et al., 2007). Finally we will use the data to attempt to validate the models of virtual human behavior as well as theories from the literature and our own analysis.

Our current focus is on three different kinds of behavior that show cultural variation: proxemics, gaze, and turn-taking. In order to study these behaviors, a collaborative team from the University of Texas at El Paso (UTEP) and the Institute for Creative Technologies (ICT) at University of Southern California have collected and partially coded approximately ten hours of audiovisual multiparty interactions in three different cultures and languages.

In the next section, we elaborate on our goals and related existing corpora, none of which quite meet our needs. In sections 3 and 4, we discuss the design of our corpus and our annotations in the three areas of interest. In section 5, we describe issues that arose in the corpus recording and annotation. Finally, in Section 6, we describe dissemination plans and future directions.

2. Corpus Data Requirements

Part of this project involves understanding differences among language cultures with respect to proxemics, which is the pattern of physical distances people maintain from each other. There is some evidence in the literatures of sociolinguistics and cultural anthropology that these distances differ based on culture and context.

A second part of this project involves understanding differences among language cultures with respect to turn-taking, which is the amount of pause or silence that is typical between people's speech when a speaking turn changes. Thus, we sought to collect data of conversational interactions that are tuned for culture and context. For example, we are working on representing groups of people in the background of a scene, where these people are having small-group conversations. Their proxemics and turn-taking should be appropriate to their culture.

In a third part of the project, we seek to provide data for related conversational behaviors such as gaze, which is a factor in turn-taking. The relation of gaze to turn-taking, particularly considered across cultures, is the subject of open research questions (Rossano et al., 2009)

There are many conversational corpora, but it is still difficult to use these to study general cross-cultural conversational behavior. Many corpora record the speech only, which misses important information in face-to-face interaction. Dyadic conversation has been most studied, but this kind of dialogue has a simplified turn-taking scheme, in which actions such as releasing a turn and assigning a turn are not distinguishable. Moreover, addressee identification is trivial, and there is no distinction between individual or group addressing.

These factors affect the usefulness of existing corpora, including the AMI corpus (Carletta et al., 2005), the CUBE-G corpus (Rehm et al., 2008), and the UTEP CIFA corpus (Flecha-Garcia et al., 2008). The AMI corpus contains dialogues with four participants and audio-visual information, however it does not allow the study of proxemics, because the participants were given fixed locations in which to sit. Moreover, the participants were not balanced for cultural background, so it does not allow cross-cultural studies. Finally, the set of tasks is more rigid, with each participant assigned a specific unique role. Likewise, the CUBE-G corpus, while focused specifically on cultural differences for non-verbal conversational behaviors such as gaze and proxemics, has only dyadic dialogues. Moreover, one member of each of dyad was an actor trained by the researchers, so the corpus could be said to reflect individual rather than multiparty responses. The UTEP CIFA corpus also has limitations. While the participants were from different culture/language groups (American and Arab) and the recordings were made with multiple cameras to help with tracking

Figure 1a American Group 1 in Task 3.

Figure 1b Arab Group 3 in Task 4.

Figure 1: Comparison of American and Arab groups in Toy-related tasks. Note the difference in proxemics between the groups as shown by the dots on the carpet.

gaze, the participants were seated so that proxemics would be an independent variable, and all the conversations were dyadic.

3. Corpus Collection

To address research questions for which uni-cultural, dyadic and seated corpora were ill-suited, we designed the UTEP-ICT corpus with four-person groups, with the participants standing and free to move. The participants are selected from three different cultures: Arabs, Americans, and Mexicans, with each group consisting of members of the same culture. The participants were recruited from local churches, restaurants, on campus, and through networks of known members of each cultural group in the El Paso area, which borders Mexico and has, in part because of the university, many representatives of other nations and cultures. We have completed the recording and partial coding of twelve four-person groups. Four of the groups were composed of native speakers of Arabic, four of native speakers of American English, and four of native speakers of northern Mexican Spanish. In recruiting participants, we sought to obtain a mix of people, some of whom were strangers and some of whom knew each other. To facilitate analysis of culture as independent variable, most of the groups were male-only, but we had one group in each language condition with at least two female participants. In Arab group 1, there were two brothers, one friend (for three years), and one stranger. Arab group 2 comprised two brothers and two cousins. Participants from Arab group 3 belonged to the same English-as-a-second-language program (two friends for 15 years, the rest for few months), and Arab group 4 had two sisters and two strangers. In American group 1, there were two sisters and two strangers. Group 2 comprised a female and her friend (seven years). She was acquainted with a second male (two years), who in turn was acquainted with the third male (two years). Group 3 comprised three soldiers taking the same leadership course and a stranger. Group 4 was comprised of two males and two females. One male knew one of the females for 20 years and the other male for 16 years. The females knew each other for five years. In Mexican group 1, all four males were acquainted (three months). In Mexican group 2, two married couples were also friends (ten years). Mexican group 3 had all females (two were friends for two years) and group 4 had three females and one male (two females were friends for 18 years, the rest a few months).

Task 1	Describe your pet peeves
Task 2	Figure out which movies you've all seen and what were the best and worst parts
Task 3	Come up with a good name for a toy
Task 4	Tell a story about the toy
Task 5	Describe an inter-cultural experience

Table 1: Conversation tasks.

The experimenter asked each of the groups to complete five conversational tasks, which were designed to elicit a range of dialog behaviors. The tasks are listed in Table 1. Tasks 1, 4, and 5 are mainly narrative tasks, where the participants can take turns relating stories or reacting to the narratives of others. Task 1 was meant to "break the ice" and get people comfortable talking with each other. Tasks 2 and 3 are constructive tasks, in which the participants must pool their knowledge and work together to reach a group consensus. Tasks 3 and 4 were designed to have possible task-related gaze focus other than the participants themselves, so a plush toy was provided and tasks related to the toy (see Figure 1). This allows gaze patterns with a copresent referent to be contrasted with gaze patterns without this referent. Task 5 is meant to elicit subjective experiences of intercultural interaction, as a possible starting point for future investigation of higher-level cross-cultural differences.

Each task lasted 10 minutes; the recording sessions lasted about 50 minutes total per group. We wanted to make the interactions as natural as possible, so the participants were not given any special tracking equipment (other than wireless microphones), and the camera were made as unobtrusive as possible. The interactions were recorded with six Apple iMac computers, placed around the periphery of a large open room that serves as a computer lab for UTEP's College of Liberal Arts. We thus recorded six simultaneous views of the participants as they conversed, making it possible, with only rare exceptions, to code the participants' proxemics, gaze and turn-state.

The participants were free to move about the room and

Alternate view 2 of group in Figure 1a.

Alternate view 3 of group in Figure 1a.

Figure 2: Alternate views of Figure 1a.

Alternate view 2 of group in Figure 1b.

Alternate view 3 of group in Figure 1b.

Figure 3: Alternate views of Figure 1b.

stand where they liked. The floor of the computer lab was covered with carpet that had dots evenly spaced at one-foot intervals, which facilitated coding the participants' positions in the room. Audio was recorded at high quality with wireless microphones worn by the participants. Figure 1 shows frames from corpus recordings of an American group and an Arab group. In Figure 1a, the subject on the right holds the plush toy involved in tasks 3 and 4. In Figure 1b, the subject on the left holds the toy. Some of the differences in proxemics among groups can be seen by comparing interpersonal distances between Figure 1a and 1b. Figure 2 shows alternate camera angles for the same the American group, illustrating the ranges views provided in the parallel recordings. Figures 3 shows alternate angles for the same Arab group.

4. Data Coding

From the recordings, we produced time-aligned partial codings of each of the twelve conversations. Specifically, we coded two 30-second excerpts of each of the conversations for tasks 1 through 4 for proxemics, turn-taking, and gaze. For proxemics, a matrix was composed of the area created by the four members as points on a polygon. For turn-taking, the data consisted of a subject's state at each tenth of a second, where state could take a value of talk, pause, or laugh. Table 2 summarizes the range of the collected data and annotations that we have completed, and serves as a guide for identifying specific annotation.

Annotation was done using the ANVIL coding tool (Kipp, 2008). Figures 4, 5, and 6 present examples of annotations of the corpus. Figure 4 shows a segment of proxemics annotations for video file `mx1_t3_1.mov`, which represents the first Mexican group performing task 3 and of the first 30-second excerpt. The figure shows participants' proxemics positions from 17 seconds to 29 seconds. Although only a single numeric value is shown in each element, it consists of x,y-coordinate data in feet and inches. Figure 5 shows the gaze of each participant over the stretch of time from 0 to 6 seconds in Arab group 3 task 4, excerpt 2. D is the speaker during this segment and most of the other participants are looking at him. Figure 6 shows information on when each participant in American group 1, task 2 excerpt 2 was speaking, not speaking (indicated as "pause"), or laughing. In this segment we can see one section between 20 and 21 seconds where three participants are speaking simultaneously, as well as a small segment at 23 where no one is speaking, in between utterances by participant C. Figures 7, 8, and 9 show parallel annotations

Culture	American, Arab, or Mexican
Group	1–4 for each culture
Task	Tasks 1–5 from Table 1
Excerpt	One or two (near beginning or end of task)
Time	Range of time within the excerpt
Behavior	Proxemics, Turn-taking, Gaze

Table 2: Corpus Dimensions.

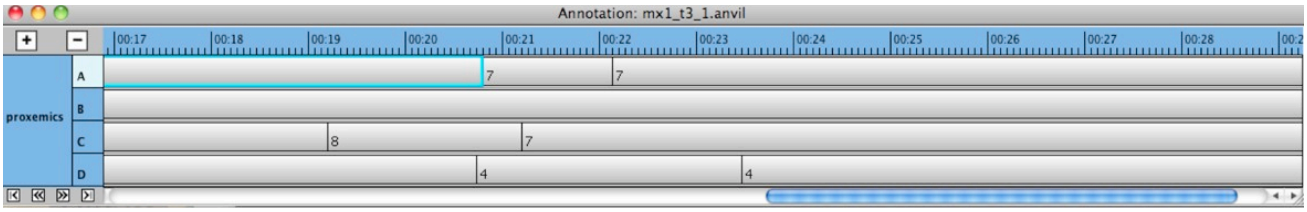

Figure 4: Proxemics coding of Mexican Group 1, Task 3, Excerpt 1.

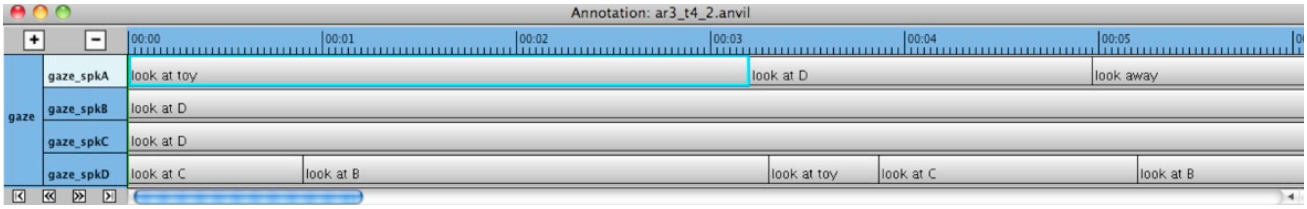

Figure 5: Gaze coding of Arab Group 3 Task 4, Excerpt 2.

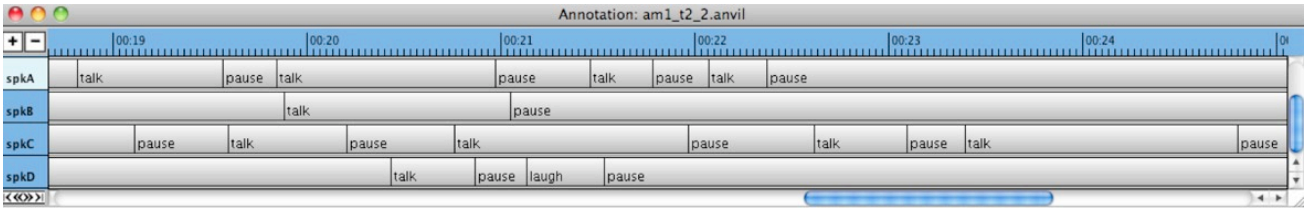

Figure 6: Turn-taking coding of American Group 1, Task 2, Excerpt 2.

for part of a conversation by American group 2 on task 3. In this figure we can see relation of proxemics, gaze and turn-taking behavior of the participants for this period.

5. Corpus Collection & Annotation Issues

In the course of the project, we have had to address a number of issues that arose in annotating the video and audio recordings. At the most basic level, we had to deal with equipment malfunction. Particularly, the software used to record four separate audio channels was sometimes unstable and crashed occasionally. When this occurred, the recovered file would only retain three of the four channels. Coders had to annotate turns for the fourth speaker using audio from the other three channels in conjunction with the video.

Another, more frustrating, problem was the difficulty of placing the cameras to catch the feet and gaze of speakers simultaneously. As speakers were allowed to stand anywhere in an area approximately 20 feet by 10 feet, camera angles could not always capture all gaze and proxemics simultaneously. It was important to capture where the conversants stood, at times sacrificing where the conversant gazed. A camera angle that captures the body of the conversant will not have such a detailed picture of the face making it sometimes hard to see the gaze direction. Additionally, with four conversants, bodies frequently occluded views of others' faces and, even with six camera angles, conversants' eye gaze was not always visible. Originally, we had selected a 30-second excerpts beginning two minutes and six minutes into the conversational task. However, after reviewing some of the videos, there were too many gaze occlusions to obtain useful data, and instead we relied on finding a 30-second excerpt free of occlusions in the first and last five minutes of each conversational task.

A final minor source of error occurred when raters annotated a subject's standing position. As no angle showed all floor marks at once but all marks were relative to the room's top left corner, raters had to determine mark number from one angle to the next by counting the marks. Occasionally, a rater would count incorrectly. Fortunately, these errors were easily spotted when comparing raters' annotations. Counting errors in this case were easily spotted and corrected.

Another difficulty is in recruiting the appropriate subjects. Ideally we would have subjects who had only mono-cultural experience in their native culture, and culture groups would be completely parallel as to their constituent participants, balancing such factors as gender, age, status, how well the participants are known to each other. For this study we were unable to provide such a balance, so it will be difficult to determine which findings are specific to the culture group and which to the particular social relationships of the participants. Broad tendencies across multiple groups (such as proxemic distance in Figure 1) can be attributed to culture group, but many factors will be more subtle, and thus further investigation with additional groups is required.

6. Dissemination and Future Work

We plan to disseminate the corpus to other researchers, subject to privacy-protection restrictions associated with the projects' IRB requirements, beginning January, 2011. Each

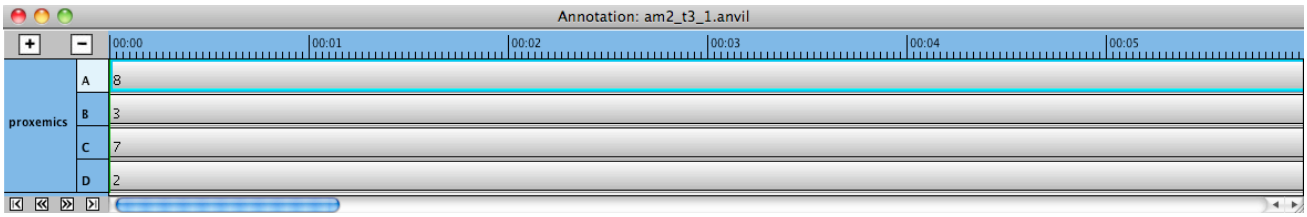

Figure 7: Proxemics coding of American Group 2, Task 3, Excerpt 1.

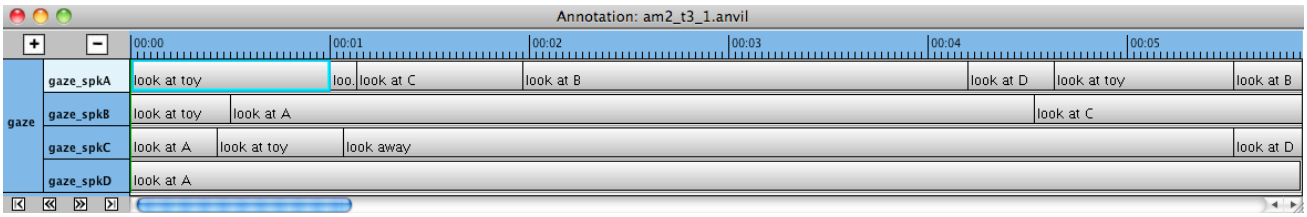

Figure 8: Gaze coding of American Group 2, Task 3, Excerpt 1.

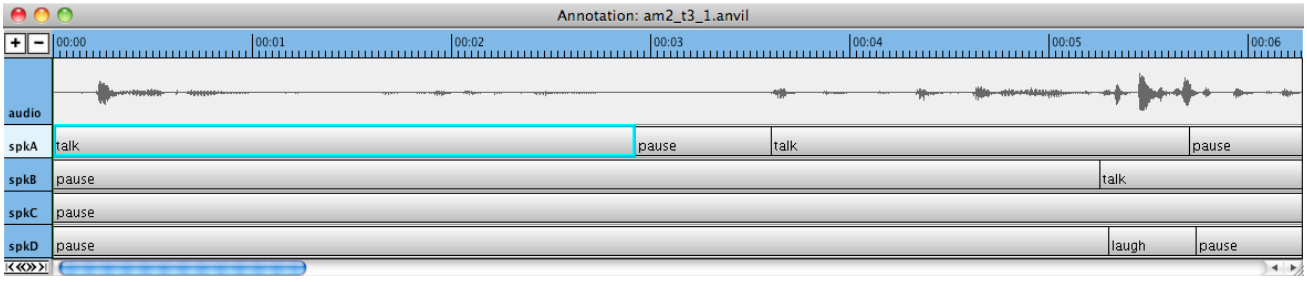

Figure 9: Turn-taking coding of American Group 2, Task 3, Excerpt 1.

file is a combined audio-visual recording. Because there are six cameras per session, the combined corpus files are large; we anticipate that distribution will be via hard-drive or, for subsets, flash drive.

In addition to analyzing patterns of proxemics, gaze, and turn-taking in the multiparty cross-cultural corpus, we are currently collecting a complementary corpus of dyadic conversations with the same tasks across the same cultures. The dyadic corpus should enable us to distinguish differences related to group size from those related to culture.

While six angles were sufficient to determine conversants' proxemic positions, they were not adequate for consistently reliable determination of participants' direction of gaze. For this reason, the corpus collection for dyadic conversations will rely on eight camera angles, although we do not expect as much occlusion as we encountered in the multiparty case. Additionally, the audio software instability problem is no longer anticipated as software for two-channel recording is more stable than that for four-channel recording.

We are also willing to share our partial annotations, which consist of time-aligned notations in ANVIL; we have not yet transcribed the participants' speech, since it was not a major factor in the analyses motivating collection of the corpus, and annotation budget for the initial project was limited. The cross-cultural multimodal phenomena on which our research focuses – proxemics, gaze and turn-taking – appear to be reasonably consistent within groups; our initial analysis suggests that the differences in behaviors between excerpts within groups is much smaller than the differences across groups.

Beyond adding realism to conversational agents in immersive environments, the analysis of the corpus may also help instructors of people who will be conducting conversations with people of different cultures. Because non-verbal behaviors often have different meanings within different culture groups, training in these conversational behaviors may enable conversants to avoid misunderstanding.

Acknowledgments

The work described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position of the United States Government, and no official endorsement should be inferred.

7. References

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain A. McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2005. The ami meeting corpus: a pre-announcement. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI'2005*.

M Flecha-Garcia, D Novick, and N Ward. 2008. Differences between Americans and Arabs in the production

and interpretation of verbal and non-verbal dialogue behaviour. In *Speech and Face-to-Face Communication Workshop, Grenoble France*, pages 47–48.

Dusan Jan, David Herrera, Bilyana Martinovski, David G. Novick, and David R. Traum. 2007. A computational model of culture-specific conversational behavior. In *Proceedings of Intelligent Virtual Agents, 7th International Conference, IVA 2007*, pages 45–56.

Michael Kipp. 2008. Spatiotemporal coding in anvil. In *Language Resources and Evaluation Conference (LREC)*, May.

Matthias Rehm, Yukiko Nakano, Hung-Hsuan Huang, Afia Akhter Lipi, Yuji Yamaoka, and Franziska Grüneberg. 2008. Creating a standardized corpus of multimodal interactions for enculturating conversational interfaces. In *Proceedings of the IUI-Workshop on Enculturating Interfaces (ECI)*.

F. Rossano, P. Brown, and S. C. Levinson. 2009. Gaze, questioning and culture. In J. Sidnell, editor, *Conversation analysis: Comparative perspectives*, pages 187–249. Cambridge University Press.