Grounding and Turn-Taking in Multimodal Multiparty Conversation

David Novick, Iván Gris

Department of Computer Science, The University of Texas at El Paso,
500 West University Avenue, El Paso, TX 79968-0518 USA
novick@utep.edu, ivangris4@gmail.com

Abstract. This study explores the empirical basis for multimodal conversation control acts. Applying conversation analysis as an exploratory approach, we attempt to illuminate the control functions of paralinguistic behaviors in managing multiparty conversation. We contrast our multiparty analysis with an earlier dyadic analysis and, to the extent permitted by our small samples of the corpus, contrast (a) conversations where the conversants did or did not have an artifact, and (b) conversations in English among Americans with conversations in Spanish among Mexicans. Our analysis suggests that speakers tend not to use gaze shifts to cue nodding for grounding and that the presence of an artifact reduced listeners' gaze at the speaker. These observations remained relatively consistent across the two languages.

Keywords: Dialog, proxemics, gaze, turn-taking, multicultural, multiparty

1 Introduction

Most studies of multimodal grounding and turn-taking have been based on analysis of dyadic conversation (e.g., [1, 2]). Others have tackled grounding and turn-taking in multiparty conversation, but this was typically either not multimodal (e.g., [3]) or was approached from a theoretical rather than an empirical perspective (e.g., [4]). In the present study, we apply a conversation-analytic approach to begin understanding the mechanisms of grounding and turn-taking in multimodal multiparty conversation.

In this study, our principal objective was to explore the empirical basis for multimodal conversation control acts in multiparty conversation, such as those discussed in ([4]). We were interested in questions such as

- Do grounding behaviors such as nodding get cued in ways similar to those observed (e.g., by [1] and [5]) in dyadic conversation?
- How do the mechanisms of turn-transitions function?
- Does the presence of an artifact lead to changes in grounding behaviors?
- How, if at all, do these behaviors differ across cultures?

We contrasted our multiparty analysis with an earlier dyadic analysis [5] and, to the extent permitted by our small samples of the corpus, contrasted (a) conversations where the conversants had an artifact (a plush toy that they were tasked with naming) and or did not have an artifact, and (b) conversations in English among Americans with conversations in Spanish among Mexicans.

• 3 Background

Research on multimodal multiparty conversation is conducted from multiple perspectives, including observing interaction, describing the conversational functions necessary for effective interaction by conversational agents, and implementing these behaviors in agents.

The observational perspective, through a discourse-analytic approach, can provide a systematic account of grounding and paralinguistic behaviors in conversation. For example, gaze patterns in multiparty conversation have been analyzed statistically, providing a detailed account of the frequency of different gaze patterns associated with turn-taking [6]. This has led to a probablistic model of interaction that has been been validated empirically, but the model's realism might be considered validated at a descriptive level rather than at a causal level. Such a probabilistic model could lead to relatively reliable functioning of, for example, an automatic gaze-dependent video editor for recordings of conversations [7]. Observational studies of multiparty conversation have also described the role of gesture, beyond gaze, in the process of interaction. For example, conversants' gestures, both head and hand, appear to be a function of the conversant's conversational role and the dialog state; conversants clearly coordinate their utterances and gestures, and this may relate to task structure [8]. However valuable, models produced by discourse-analytic studies do not necessarily provide a deep explanation of how the gaze and turn-taking functions actually work. Going beyond surface simulation—even if highly plausible and effective—requires understanding the specific functional mechansisms for, and the context-specific purposes associated with, conversants' use of paralinguistic behaviors.

While the mechanisms of human-human multiparty conversation management may remain only partially understood, the need for them is clear. The kinds of conversational roles and the broad functions of conversational management needed for effective interaction by embodied conversational agents have been comprehensively catalogued [4]. The functions of interaction management include turn management, channel management, thread/conversation management, initiative management, and attention management.

Models of some paralinguistic behaviors have been validated through implementation in conversational agents. For example, a multiparty gaze model based on the findings of Argyle and Cook [9] was validated through simulation in an embodied conversational agent [10].

Whether modeled based on observation or validated through simulation, some aspects of paralinguistic behaviors and dialog management in multiparty conversation have conversational functions that are relatively clear. Other aspects remain confirmed but unexplained from the standpoint of conversational function. For example, conversants in multiparty interaction use a great deal of overlap of utterances [11], but the functional reasons for the overlap are not yet clear. Indeed, a multiparty conversation may actually involve multiple simultaneous conversations, and the conversants must accordingly manage multiple simultaneous conversational floors [12]. The most plausible account of this management involves different types of conversational moves of splitting the conversation ("schisiming") or bringing separated threads back together ("affiliating"). These moves can be categorized as schism-inducing turns, schisming by aside, affiliating by turn-taking, or affiliating by coordinated action [12].

Finally, we note that discourse-analytic comparison of multiparty and dyadic conversation has indicated that differences as a function of the number of conversants vary across cultures [13]. Thus while the use of gaze to coordinate turn-transition differs between speakers of American English and of Mexican Spanish, for Americans this process is a function of group size: gaze plays a relatively smaller role in Mexican multiparty conversation than it does in American [13].

To explore the specific functional mechansisms for conversants' use of paralinguistic behaviors in multiparty conversation, in this study we oriented our study around four principle issues: whether grounding behaviors such as nodding get cued in ways similar to those observed in dyadic conversation, how the mechanisms of turn-transitions actually function, whether the presence of an artifact leads to changes in grounding behaviors, and how, if at all, these behaviors differ between speakers of American English and of Mexican Spanish.

3 Methodology

To address these questions, we conducted conversation analyses of four 20-second excerpts of conversations from the UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus [14]. The corpus comprises approximately 20 hours of audiovisual multiparty interactions among dyads and quads of native speakers of Arabic, American English and Mexican Spanish. The subjects 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. In the present research, we focused on interaction in quads of Spanish and English speakers. And because we were particularly interested in grounding and turn-taking, we based our analysis on conversations that had multiple turns over a short period of time.

Tasks 1, 4, and 5 were mainly narrative tasks, where the participants can take turns relating stories or reacting to the narratives of others. Tasks 2 and 3 were 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 a toy provide a possible gaze focus other than the subjects themselves, so that gaze patterns with a copresent referent could be contrasted with gaze patterns without this referent. Task 5 was meant to elicit subjective experiences of intercultural interaction.

For each of the four excerpts that formed the basis of the study reported here, we transcribed the speech and annotated the gaze, nods and upper-body gestures of the four conversants in the conversation. Timings were noted with the Elan Linguistic Annotator [15]. From the observed behaviors we then attempted to produce a plausible explanation of how these actions served the conversants in grounding (or not) each other's contributions to the conversation and in taking conversational turns.

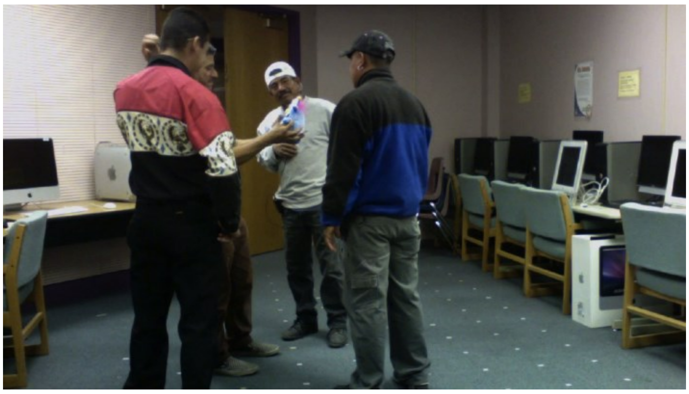

Figure 1. Speakers of Mexican Spanish conversing, with artifact

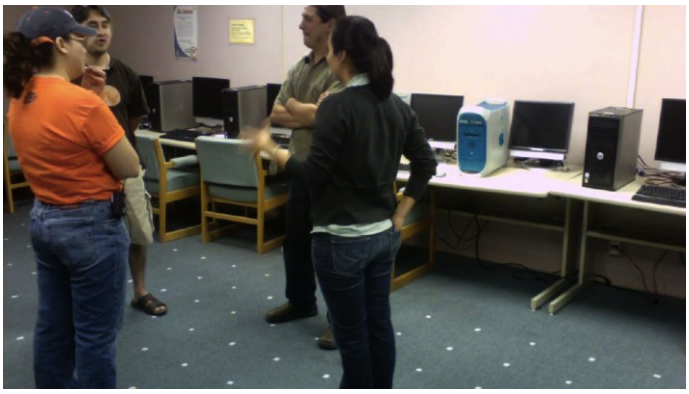

Figure 2. Speakers of American English, conversing, without artifact

4 Results

We begin with our analysis of the conversations conducted by the speakers of American English. In the conversation without the artifact, we observed that the relationship between gaze and nod differed markedly from that observed [5] in dyadic conversation. In dyadic conversation, the listener's nods are typically cued by gaze from the speaker. From a functional standpoint, this represents the listener's providing grounding feedback at points where the speaker can perceive it—and even if the speaker can see the listener peripherally the speaker in effect gives the listener the opportunity to ground (or not to ground) where, presumably, the speaker wants to check on the listener's understanding. Consequently, for speakers of American

English, conversants in dyadic conversations looked at each other less frequently than did conversants in multiparty conversations, presumably because higher rates of mutual gaze would be providing too-frequent cues for grounding feedback.

4.1 American Non-artifact Conversation

The (non-artifact) multiparty excerpt we studied here (Table 1 shows the first 12 seconds of the transcript) contains some paralinguistic behaviors associated with dyadic conversation. In particular, most of the listeners are looking at the speaker. This is true as the excerpt starts, where Conversant B is talking, and Conversants A, C and D are looking at B. And it is true again as C takes the floor (at 04:15), and A, then B, and then D look at C. Additionally, B looks away while talking.

Table 1. Excerpt of transcript of American English conversation, without artifact

Time Start	Time End	A Verbal	A Nonverbal	B Verbal	B Nonverbal	C Verbal	C Nonverbal	D Verbal	D Nonverbal
Initial			arms behind back, looking at B		arms crossed, looking away		hands in back pockets, looking at B		R arm crossed, L hand on chin, looking at B
00:00	04:10			when you're driving and you see people talking on the phone and driving	gestures phone with left hand and crosses arms again				
00:00	02:00		succession of small nods						
04:15	05:10					but they're driving like stupid			
04:25	09:60		looks at C						
05:50	07:50				looks at C	they're driving very slow and like	looks at A		
06:00	10:80								looks at C
08:00	10:75					they won't change lanes right	looks away, mimics changing lane		
08:15	08:40				glances at A, looks back at C				
09:40	11:90			-- go off or somethng or…					

But other behaviors differ markedly from those expected in dyadic conversation. When B ends the turn it is because C has grabbed the floor; B has not shifted gaze to a listener to check for grounding or to offer the turn. When C begins talking, she looks at A, who does not provide feedback, and then looks away. Although A does engage in a series of small nods while B is talking—but not looking at A, neither A nor the other conversants provide further grounding feedback in this excerpt. Relative to the dyadic conversants observed in [5], these four conversants do very little nodding. However, the conversants do use gestures to reinforce their verbal production. At 00:00 B mimics using a cell phone, and at 08:00 C gestures to indicate a lane change.

4.2 American Artifact Conversation

In the conversation among speakers of American English with a task that involved the plush-toy artifact (see Table 2), conversants tended to gaze more at the artifact than at each other. At the start of this excerpt, part-way into the conversation, the conversants are all looking at the artifact. And when Conversants B and C shift their gaze away from the artifact, at 04:15 and 06:50 respectively, they look primarily at non-speakers. None of the conversants nods. At 02:10 the turn transition between B and C has a brief overlap, and the transition is not coordinated with a gaze shift. Rather, the gaze shift lags the turn change.

Table 2. Excerpt of transcript of American English conversation, with artifact

Time Start	Time End	A Verbal	A Nonverbal	B Verbal	B Nonverbal	C Verbal	C Nonverbal	D Verbal	D Nonverbal
Initial			looking at toy on floor; arms crossed		looking at toy on floor; arms crossed		looking at toy on floor; arms crossed		looking at toy (on the floor); right arm crossed; left arm on chin
00:00	00:24					[laugh]			
00:24	02:17			the anonymous beast					
02:10	05:30							it has to be named or else how'll people tell other people what to buy	
04:15	06:20				looks at A, then C				
06:50	08:40						looks at A		

Relative to the non-artifact conversation, the conversants use far fewer gestures, perhaps because the conversants' common focus on the artifact takes their attention away from their conversational partners' possible gestural displays.

4.1 Mexican Artifact Conversation

We now turn from the excerpts of the American conversations to the excerpts of the Mexican conversations. In their conversation that included the plush-toy artifact (see Table 3), Mexican participants generally nodded when the speaker gaze was focused on the artifact. While nodding was below the overall frequency compared to conversations when no artifact was involved, when listeners did nod it was after a verbal consensus and agreement and rarely otherwise. For example, in the artifact task after speaker B proposes a toy name to the listeners at 16:50, A immediately takes the turn within a half a second and verbally agrees three times in succession with 1.5-2.0 second intervals. Meanwhile, C produces a succession of small nods between each verbal statement, even though the four conversants have their gaze focused on the artifact. This seems similar to the nodding behavior of Conversant A in the American non-artifact conversation. But these behaviors appear at odds with the explanation in [5], which suggested that if the speaker is not looking at you, it does not do much for you to nod because the speaker may not (cf. peripherally) see your action. It is important to note, though, that there were no artifacts involved in that study. One possible explanation for nodding, even when no one is looking, can be the need to express agreement while not wanting to take the floor to express it. The task asked for a group consensus on the naming of the artifact, which required all participants to agree on a name. Silence may be a weak agreement that is reinforced by non-verbal behavior to help achieve the task.

Table 3. Excerpt of transcript of Mexican Spanish conversation with artifact

Time Start	Time End	A Verbal	A Nonverbal	B Verbal	B Nonverbal	C Verbal	C Nonverbal	D Verbal	D Nonverbal
15:00	16:50		(looking at toy) small step backwards	blue punk no, algo asi?	(looking at toy) touches toy's hair		(looking at toy) quick glance at B		(holding toy, looking at toy)
16:50	17:00	blue punk, si	quick glance at B				looks at toy		
17:00	17:50			blue punk	quick glance at A				
17:50	18:00			es azul					
18:00	19:00			y luego trae el pelo aca					
19:00	19:50	si blue punk		y luego			nods		
19:50	20:00			son los punk					
20:00	20:50			blue punk			nods		
21:00	22:00	si blue punk							
22:00	23:00						nods	ey, blue punk	
23:00	24:00	blue punk							

We also note turn-taking differences for Mexican conversants in conversations with and without an artifact. When an artifact was involved, all the conversants looked at the plush toy through the majority of the conversation, looking away and at other participants only once (each) in the 20-second transcript. This occurred during seconds 15-17, when (B) suggested a toy name, but none retained the gaze more than a second. While not producing non-verbal behaviors, participants instead signaled turn taking by repeating a previous statement in as-a-matter-of-fact intonation or by adding to the collective description of the artifact with simple sentences (“es azul / it’s blue”, “tiene colmillos / it has fangs”, “tiene cuernos / it has horns”). In general, repetition seems to act as an acknowledgment and an invitation for someone else to take the floor, as the repeater rarely adds anything else to the conversation. Indeed, this seems like the “display” method of grounding described by Clark and Schaefer as the strongest form of acceptance of a contribution to discourse [16].

4.1 Mexican Non-artifact Conversation

In the Mexican non-artifact section of the corpus (see Table 4), conversants were relatively more inclined to gesture. This reinforces the effects of gaze, in which if you are being observed motivates the speaker to enhance his/her conversation with gestures.

Table 4. Excerpt of transcript of Mexican Spanish conversation without artifact

Time Start	Time End	A Verbal	A Nonverbal	B Verbal	B Nonverbal	C Verbal	C Nonverbal	D Verbal	D Nonverbal
06:50	08:00				looks at D				
07:30	08:10	si y luego como el barco	hand gestures indicating the sinking ship				locks hands at waist to her front		looks away from group
08:00	11:50				looks at A				
08:20	08:80	como osea							
08:30	14:00								looks at A
09:00	09:30	Si							
09:00	12:50		looks at B						
10:00	11:00	como va cambiand o y todo eso							
10:80	11:20			si	nods				
12:00	13:00							y luego que se quitan el collar y…	moves in her place and gestures a necklace with both hands
12:50	14:00		looks at D						

These gestures can be separated in two different types. The first type is analogous to the ones found in the artifact conversations, which are used for agreement or acknowledgement. One of the main differences is that without an artifact, the gesture tends to be done in conjunction to the verbal statement. This can be observed from 02:00 – 08:00 on (B), (C) and (D) in different statements, usually briefly following or in conjunction with “si” or “a mi tambien” (yes / me too). The second gesture type is exclusive to the non-artifact participants. In this case, conversants use gestures to enact the verbal part. In this particular conversation the conversants are critiquing the movie *Titanic*. Hand gestures conversants use include imitating the sinking ship (A) at 07:30, a necklace (D) at 12:00, and people drowning (B) at 15:00.

Overall, gestures appear to be significantly more frequent when there is no artifact. Without the artifact, conversants appeared to prefer overlapping for turn taking. However, conversants rarely used overlap to change topic but rather to elaborate on the current topic. During the first thirteen seconds of the conversation, verbal agreement was used to change turns, although no one kept the floor for long, and the ideas proposed after taking the floor were left unresolved. For example, at 12:00 conversant D takes the turn for the first time within seven seconds with a contribution but expresses only an unfinished sentence “… y luego se quitan el collar y …” (and then they take the necklace off and…).

5 Conclusion

Based on our observation of the four conversational excerpts discussed in Section 4, we now return to the four questions that motivated our study.

- Do grounding behaviors such as nodding get cued in ways similar to those in dyadic conversation? The evidence in the multiparty conversations we studied suggests that multiparty conversants nod less frequently and even these fewer nods are not being cued by the speaker’s gaze shift.
- How do the mechanisms of turn-transitions function? The evidence, which includes greater overlap at turn boundaries, suggests that conversants in multiparty conversation do not rely as much on gaze as a turn cue as do conversants in dyadic conversation. Rather, multiparty conversants repeatedly overlapped at turn boundaries, especially where one party grabbed the floor—possibly because the conversant could not engage the speaker’s gaze.
- Does the presence of an artifact lead to changes in grounding behaviors? The evidence suggests that the presence of an artifact draws the conversants’ gaze, thus reducing the amount of time that listeners gaze at the speaker. This may contribute to the phenomenon of speakers tending not to use gaze shifts to cue nodding as a grounding behavior. In one case (Mexican, artifact), the conversants seemed to substitute display for continued attention as a grounding behavior.
- How, if at all, do these behaviors differ across cultures? While we found some differences between the behaviors of speakers of American English and of Mexican Spanish, these differences likely reflect natural variation in conversation rather than clear cultural differences. Rather, comparison of the conversations across cultures revealed similarities, particularly with respect to differences in gaze patterns across the artifact/non-artifact condition and with respect to the lack of cueing of nods.

References

1. Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a model of face-to-face grounding, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, 553-561.
2. Bavelas, J. B., Coates, L. and Johnson, T. (2002), Listener Responses as a Collaborative Process: The Role of Gaze, Journal of Communication, 52: 566–580
3. Branigan, H. (2006). Perspectives on Multi-party Dialogue, Research on Language and Computation, 4:153–177
4. Traum, D. (2004). Issues in multiparty dialogues, Advances in agent communication, Springer, 201-211.
5. Novick, D. (2012). Paralinguistic behaviors in dialog as a continuous process, Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog, September 7-8, 2012, Stevenson, Washington, 54-57.
6. Otsuka, K., Takemae, Y., and Yamato, J. (2005, October). A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances, Proceedings of the 7th International Conference on Multimodal Interfaces, 191-198.
7. Takemae, Y., Otsuka, K., and Mukawa, N. (2004). Impact of video editing based on participants' gaze in multiparty conversation. In CHI'04 extended abstracts on Human Factors in Computing Systems, 1333-1336.
8. Battersby, S. A. (2011). Moving together: The organisation of non-verbal cues during multiparty conversation. Doctoral dissertation, Queen Mary, University of London.
9. Argyle, M. and Cook, M. (1976) Gaze and Mutual Gaze. Cambridge University Press, London.
10. Gu, E., and Badler, N. (2006). Visual attention and eye gaze during multiparty conversations with distractions, Intelligent Virtual Agents, Springer, 193-204.
11. Shriberg, E., Stolcke, A., and Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation, Proc. Eurospeech, Vol. 2, 1359-1362.
12. Aoki, P. M., Szymanski, M. H., Plurkowski, L., Thornton, J. D., Woodruff, A., and Yi, W. 2006. Where's the party in multi-party?: Analyzing the structure of small-group sociable talk, Proceedings of the 2006 20th Anniversary Conference on Computer-Supported Cooperative Work, 393-402.
13. Herrera, D., Novick, D., Jan, D., and Traum, D. (2011). Dialog behaviors across culture and group size, *Proceedings of HCI International 2011*, July 11-14, 2011, Orlando, FL, Lecture Notes in Computer Science, 2011, Volume 6766/2011, 450-459.
14. Herrera D., Novick D., Jan, D., and Traum D. (2010). The UTEP-ICT Cross-Cultural Multiparty Multimodal Dialog Corpus, Multimodal Corpora Workshop: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010), Valletta, Malta (2010).
15. Tacchetti, M. (2011). ELAN User Guide, version 4.1.0, http://www.mpi.nl/corpus/html/elan_ug/index.html.
16. Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. Cognitive science, 13(2), 259-294.