Annotating and Generating Posture from Discourse Structure in Embodied Conversational Agents

Justine Cassell†, Yukiko I. Nakano†, Timothy W. Bickmore†, Candace L. Sidner‡, and Charles Rich‡

†MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139
{justine, yukiko, bickmore}@media.mit.edu

‡Mitsubishi Electric Research Laboratories (MERL)
201 Broadway
Cambridge, MA 02139
{sidner, rich}@merl.com

ABSTRACT
This paper addresses the problem of designing embodied conversational agents that exhibit appropriate posture shifts during dialogues with human users. Previous research has noted the importance of hand gestures, eye gaze and head nods in conversations between embodied agents and humans. However, this research has neglected the role of other body movements, in particular postural shifts. We present an analysis of human monologues and dialogues that suggests that postural shifts can be predicted as a function of discourse state in monologues, and discourse state and conversation state in dialogues. On the basis of these findings, we have implemented an embodied conversational agent that uses a dialogue manager called Collagen in such a way as to generate postural shifts.

Keywords
posture shift, discourse structure, conversation structure.

1. INTRODUCTION
This paper provides empirical support for the relationship between posture shifts and discourse structure, and then derives an algorithm for generating posture shifts in an animated embodied conversational agent from discourse states produced by the middleware architecture known as Collagen [16]. Other nonverbal behaviors have been shown to be correlated with the underlying conversational structure and information structure of discourse. For example, gaze shifts towards the listener correlate with a shift in conversational turn (from the conversational participants' perspective, they can be seen as a signal that the floor is available). Gestures correlate with rhematic content in accompanying language (from the conversational participants' perspective, these behaviors can be seen as a signal that accompanying speech is of high interest). A better understanding of the role of nonverbal behaviors in conveying discourse structures enables improvements in the naturalness of embodied dialogue systems, such as embodied conversational agents, as well as contributing to algorithms for recognizing discourse structure in speech-understanding systems. Previous work, however, has not addressed major body shifts during discourse, nor has it addressed the nonverbal correlates of topic shifts.

2. Background
Only recently have computational linguists begun to examine the association of nonverbal behaviors and language. In this section we both review research by non-computational linguists, discuss how this research has been employed to formulate algorithms for natural language generation or understanding.

About three-quarters of all clauses in descriptive discourse are accompanied by gestures of one kind or another [17], and within those clauses, the most effortful part of gestures tends to co-occur with or just before the phonologically most prominent syllable of the accompanying speech [13]. Of course, communication is still possible without gesture. But it has been shown that when speech is ambiguous [21] or in a speech situation with some noise [19], listeners do rely on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). Similarly, [17] established that listeners rely on information conveyed only in gesture as they try to comprehend a story. Gesture and speech do not always manifest the same information, but what they convey is virtually always compatible. For example, gesture may depict the way in which an action was carried out when this aspect of meaning is not depicted in speech. Even when the gestural content overlaps with speech (reported to be the case in roughly 50% of utterances, for descriptive discourse), gesture often emphasizes information that is also focused pragmatically by mechanisms like prosody in speech. In fact, the semantic and pragmatic compatibility seen in the gesture-speech relationship recalls the interaction of words and graphics in multimodal presentations [10].

On the basis of results such as these, several researchers have built animated embodied conversational agents that ally synthesized speech with animated hand gestures. For example, Lester [15] generate deictic gestures and choose referring expressions as a function of the potential ambiguity and proximity of objects referred to. Rickel and Johnson [18]'s pedagogical agent produces a deictic gesture at the beginning of explanations about objects. André et al. [1] generate pointing gestures as a sub-action of the rhetorical action of labeling, in turn a sub-action of elaborating. Cassell and Stone [3] generate either speech, gesture, or a combination of the two, as a function of the information structure status and surprise value of the discourse entity.

Head and eye movement has also been examined in the context of discourse and conversation. Looking away from one's interlocutor has been correlated with the beginning of turns. From the speaker's point of view, this look away may prevent an overload of visual and linguistic information. On the other hand, during the execution phase of an utterance, speakers look more often at listeners. Head nods and eyebrow raises are correlated with emphasized linguistic items – such as words accompanied by pitch accents [7]. Some eye movements occur primarily at the ends of utterances and at grammatical boundaries, and appear to function as synchronization signals. That is, one may request a response from a listener by looking at the listener and suppress the listener's response by looking away. Likewise, in order to offer the floor, a speaker may gaze at the listener at the end of the utterance. When the listener wants the floor, s/he may look at and slightly up at the speaker [9]. It should be noted that turn taking only partially accounts for eye gaze behavior in discourse. A better explanation for gaze behavior integrates turn taking with the information structure of the propositional content of an utterance [5]. Specifically, the beginning of themes are frequently accompanied by a look-away from the hearer, and the beginning of rhemes are frequently accompanied by a look-toward the hearer. When these categories are co-temporaneous with turn construction, then they are strongly—in fact, absolutely—predictive of gaze behavior.

Results such as these have led researchers to generate eye gaze and head movements in animated embodied conversational agents. Takeuchi and Nagao, for example, [20] generate gaze and head nod behaviors in a "talking head." Cassell et al. [2] generate eye gaze and head nods as a function of turn taking behavior, head turns just before an utterance, and and eyebrow raises as a function of emphasis.

To our knowledge, until now research on posture shifts and other gross body movements, have not been used in the design or implementation of computational systems. In fact, although a number of conversational analysts and ethnomethodologists have described posture shifts in conversation, their studies have been purely qualitative in nature, and difficult to reformulate as the basis of algorithms for the generation of language and posture. Nevertheless, researchers in the non-computational fields have discussed posture shifts extensively. Kendon [13] reports a hierarchy in the organization of movement such that the smaller limbs such as the fingers and hands to engage in more frequent movements, while the trunk and lower limbs change relatively rarely.

A number of researchers have noted that changes in physical distance during interaction seem to accompany changes in the topic or in the social relationship between speakers. For example Condon and Osgton [8] have suggested that in a speaking individual the changes in these more slowly changing body parts occur at the boundaries of the larger units in the flow of speech. Scheflen (1973) also reports that posture shifts and other general body movements appear to mark the points of change between one major unit of communicative activity and another. Blom & Gumperz (1972) identify posture changes and changes in the spatial relationship between two speakers as indicators of what they term "situational shifts" -- momentary changes in the mutual rights and obligations between speakers accompanied by shifts in language style. Erickson (1975) concludes that proxemic shifts seem to be markers of 'important' segments. In his analysis of college counseling interviews, they occurred more frequently than any other coded indicator of segment changes, and were therefore the best predictor of new segments in the data. Unfortunately, in none of these studies are statistics provided, and their analyses rely on intuitive definitions of discourse segment or "major shift". For this reason, we carried out our own empirical study.

3. Empirical Study

Videotaped monologues and dialogues were used as the basis for the current study. In monologues, subjects were asked to describe each of the rooms in their home, then give directions between four pairs of locations they knew well (e.g., home and the grocery store). The experimenter acted as a listener, only providing backchannel feedback (head nods, smiles and paraverbals such as "uh-huh"). For dialogues, two subjects were asked to generate an idea for a class project that they would both like to work on, including: 1) what they would work on; 2) where they would work on it (including facilities, etc.), and; 3) when they would work on it. Subjects stood in both conditions and were told to perform their tasks in 5-10 minutes.

The video data was transcribed and coded for three features: discourse segment boundaries, turn boundaries, and posture shifts. In this study we chose initially to look at high-level discourse segmentation phenomena rather than those discourse segments embedded deeper in the discourse. Thus, the time points at which the assigned task topics were started served as segmentation points. Turn boundaries were coded (for dialogues only) as the point in time in which the start or end of an utterance co-occurred with a change in speaker, but excluding backchannel feedback. Turn overlaps were coded as open-floor time. Posture shifts were coded with start and end time of occurrence and an estimated energy level. Energy level was normalized per subject by taking the largest posture shift observed for each subject as 100% and coding all other posture shift energies relative to the 100% case. Posture shifts which occurred as part of gesture or were clearly intentionally generated (e.g., turning one's body while giving directions) were not coded. For the purpose of this study we focused primarily on changes in gross leg, hip, arm, and shoulder motion. The exact surface form of each posture shift was only coded informally to facilitate data analysis.

4. Results

Data from seven monologues and five dialogues were transcribed, and then coded and analyzed independently by two raters. A total of 70.5 minutes of data was analyzed (42.5 minutes of dialogue and 29.2 minutes of monologue). A total of 67 discourse segments were identified (25 in the dialogues and 42 in the monologues), along with a total of 407 turns in the dialogue data.

For the current study, as described above, we used the instructions given to subjects concerning the topics to discuss as segmentation boundaries. In future research, we will address the thorny question of inter-rater reliability for hierarchical discourse segmentation. Posture shifts also pose a challenge to inter-coder reliability, as the form of these major body shifts turns out to be quite idiosyncratic. For this reason, raters coded all posture shifts independently, and then employed the

conservative strategy of only analyzing those instances that were judged to be posture shifts by both raters.

4.1 Analysis

Posture shifts were observed to occur regularly throughout the data. This, together with the fact that the majority of time was spent within discourse segments and within turns (rather than between segments), led us to normalize our posture shift data for comparison purposes. For relatively brief intervals (inter-discourse-segment and inter-turn) normalization by number of inter-segment occurrences was sufficient, however, for long intervals (intra-discourse segment and intra-turn) we needed to normalize by time to obtain meaningful comparisons. This resulted in metrics of posture-shifts-per-interval (ps/int) and posture-shifts-per-second (ps/s). Thus, in the tables below, posture shifts that occurred during short spans of time (such as inter-turns) are described in terms of posture-shift-per-second (ps/s), and also posture-shift-per-interval (ps/int). Posture shifts that occurred during long intervals (such as within turns) are described in terms of the numbers of posture-shifts-per-second (ps/s).

Our initial analysis compared posture shifts made by the current speaker within discourse segments (intra-dseg) to those produced at the boundaries of discourse segments (inter-dseg). It can be seen (in Table 4.1.1) that posture shifts occur an order of magnitude more frequently at discourse segment boundaries than within discourse segments in both monologues and dialogues. Inter-segment posture shifts also occur more frequently in monologues than in dialogues. Posture shifts also tend to be more energetic at discourse segment boundaries within monologues.

Table 4.1.1. Spkr Posture WRT Discourse Segments

	Monologues			**Dialogues**		
	ps/s	ps/int	energy	ps/s	ps/int	energy
inter-dseg	0.254	0.633	0.778	0.143	0.233	0.636
intra-dseg	0.026		0.619	0.024		0.683

Listeners are also observed to perform posture shifts when the speaker changes the topic. As Table 4.1.2 shows, they are roughly ten times more likely to perform a posture shift when the speaker shifts discourse segments than within the speaker's discourse segments.

Table 4.1.2 Listener Posture WRT Discourse Segments

	ps/s	ps/int	energy
inter-dseg	0.122	0.240	0.666
intra-dseg	0.009		0.717

Initially, we classified data as being inter- or intra-turn. Table 4.1.3 shows that turn structure does have an influence on posture shifts; subjects were five times more likely to exhibit a shift at a boundary than within a turn.

Table 4.1.3 Speaker Shifts WRT Turns

	ps/s	ps/int	energy
inter-turn	0.063	0.120	0.678
intra-turn	0.010		0.681

An interaction exists between turns and discourse segments such that discourse segment boundaries are ten times more likely to co-occur with turn changes than within turns (see Table 4.1.4). Both turn and discourse structure exhibit an influence on posture shifts, with discourse having the most predictive value. Starting a turn while starting a new discourse segment is marked with a posture shift roughly 10 times more often than when starting a turn while staying within discourse segment.

Table 4.1.4 Spkr Posture by Discourse and Turn Breakdown

	ps/s	ps/int
inter-dseg/start-turn	0.265	0.259
inter-dseg/mid-turn	0.000	0.000
inter-dseg/end-turn	0.000	0.000
intra-dseg/start-turn	0.038	0.078
intra-dseg/mid-turn	0.015	
intra-dseg/end-turn	0.021	0.042

It is clear from these results that posture is indeed correlated with discourse state, such that speakers generate a posture shift when initiating a new discourse segment, which is often at the boundary between turns.

5. System

5.1 System Architecture

Rea is an embodied conversational agent that interacts with a user in the real estate agent domain [2]. The system architecture of Rea is shown in Figure 1. Rea takes input from a microphone and two cameras. The UM interprets and integrates this multimodal input and outputs the semantic representation (using pattern matching). The UM then sends the output to Collagen as the Dialogue Manager (DM). Collagen, as further discussed

below, maintains the state of the dialogue as shared between a user and an agent. The ReaAgent decides the next action of Rea based on the discourse state maintained by Collagen. It also assigns the information structure [12] of the Utterance content so that gestures can be appropriately generated. The semantic

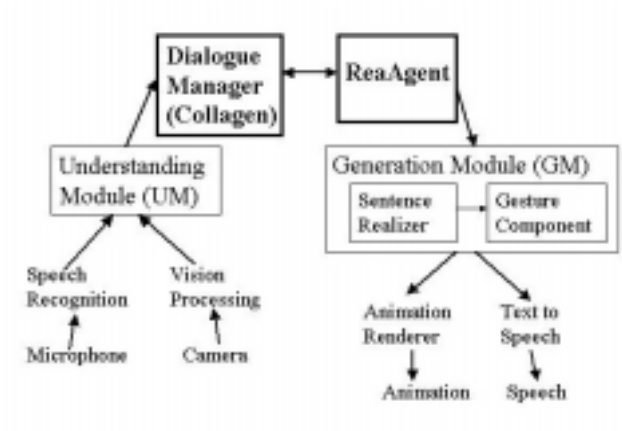

Figure5.1: System architecture

representation of the action, including verbal and non-verbal behaviors, is sent to the Generation Module (GM) which receives the representation of the action and generates surface linguistic expressions and gestures. The output from the GM is a set of instructions to achieve synchronization between animation and speech. This instruction is executed by a 3D animation renderer and a text-to-speech system. Table 5.1 shows the associations between discourse and conversational state that Rea is currently able to handle. In other work we have discussed information structure that Rea can deal with [6]. In the following sections, we focus on ReaAgent's generation of posture shifts.

5.2 The Collagen dialogue manager

Collagen™ is JAVA middleware for building COLLAborative interface AGENts to work with users on interface applications. Collagen is designed with the capability to participate in collaboration and conversation, based on [11], [16]. Collagen updates a model of the discourse state (focus stack and recipe tree) using a combination of the discourse interpretation algorithm of Lochbaum [16] and plan recognition algorithms [14]. It takes as input both user and system utterances and user and system interface actions, and accesses a library of recipes describing actions in the domain. After updating the discourse state, Collagen makes three resources available to the interface agent: the focus of attention (using the focus stack), the segmented interaction history (a record of closed parts of the conversation) and an agenda of next possible actions created from the focus stack and recipe tree. When ReaAgent is not present, a default interface agent using Collagen can communicate with the user in natural language utterances produced by straightforward template generation from the internal agent language.

5.3 ReaAgent as an interface agent

ReaAgent works as a content planner in the Rea architecture, and also plays the role of an interface agent in Collagen. It has access to the discourse state and the agenda using APIs provided by Collagen. Based on the results we reported above, we describe here how ReaAgent plans Rea's next nonverbal actions using the resources that Collagen maintains.

Table 5.1: Discourse functions and non-behavior cues

Discourse level info.	**functions**	**non-behavior cues**
Discourse structure	new segment	Posture_shift
Conversation structure	turn giving	eye_gaze & (stop_gesturing hand_gesture)
	turn keeping	(look_away keep_gesture)
	turn taking	eye_gaze posture_shift
Information structure	emphasize information	eye_gaze & (beat_gesture other_hand_gestures)

The empirical study revealed that posture shifts help to indicate discourse segment boundaries and turn boundaries. As in Table 4.1.4, a posture shift most frequently occurs when both the discourse segment and the turn are changed. About 26% of all the discourse boundaries that coincide with a change of speaker are accompanied by a posture shift. On the other hand, posture shifts occur in only 8% of all the turn boundaries that are not discourse boundaries. Therefore, a posture shift decision rule that covers these two cases can be defined as follows:

if the next turn is for Rea

if Rea's next utterance does not directly contribute to the current discourse purpose

then use a posture shift in 26% of cases

else use a posture shift in 8% of cases

In order to implement this rule in the Collagen framework, ReaAgent needs to know the current discourse purpose and to judge whether the next utterance that the ReaAgent plans must generate contributes to the current purpose. Collagen provides APIs that access the agenda and get the next agent action. ReaAgent accesses the focus stack and gets the current discourse purpose, which is shared between the user and Rea. By comparing the current purpose and the purpose of the next agent action, ReaAgent can judge whether the Rea's next action contributes to the current discourse purpose or not. For example, if the current discourse purpose is to find a preferred house (FindHouse), and the next utterance that the ReaAgent plans to say is as follows;

(1) (Ask.What (agent Propose.What (user FindHouse <storage ?>)))

Rea says: "What kind of storage do you need?"

ReaAgent uses Collagen APIs to compare the current discourse purpose (FindHouse) to the purpose of utterance (1). The purpose of this utterance is to ask the value of the storage parameter of FindHouse. Thus, ReaAgent judges that this utterance contributes to the current discourse purpose. ReaAgent decides to change the posture in 8% of these cases. On the other hand, if Rea's next utterance is about showing a house:

Figure5.2: Rea demonstrating a posture shift

(2) (Propose.Should (agent ShowHouse (joint 123ElmStreet))

Rea says: "Let's look at 123 Elm Street."

This utterance does not directly contribute to the current discourse purpose because it does not ask a parameter of FindHouse. Instead, it introduces a new discourse purpose ShowHouse. In 26% of such case, ReaAgent changes Rea's posture. Rea illustrates a posture shift in Figure 5.2.

6. Example

This section describes an example dialogue between Rea and the user, and shows how ReaAgent decides where to generate posture shifts. Figure 6.1 shows an example dialogue between Rea and the user. This dialogue consists of three major segments; greetings, finding a preferred house, and farewell. Based on this task structure, we defined plan recipes for Collagen. The first shared discourse purpose **[goal: have a conversation]** is introduced by the user in utterance (1). Also, in utterance (3), the user introduces the main part of the conversation **[goal: find house]**. The next goal in the agenda is **[goal: identify preferred city]** which is a subgoal for accomplishing **[goal: find house]**. This goal is directly contributes to the current purpose, **[goal: find house]**. Thus, this case is judged to be a turn boundary within a discourse segment. In this case, a posture shift is generated in 8% of the time. In utterance (25), Rea introduces new discourse purpose **[goal : show house]**. Rea, using a default rule, decides to take the initiative on this goal. At this point Rea accessesses the discourse state and confirms that a new goal is about to start. ReaAgent judges this case as a discourse boundary and also a turn boundary. In 25% of the time, ReaBrain sends a command to the Generation Module for a posture shift. This is illustrated in Figure 5.2.

7. Conclusion and Further work

We have demonstrated a clear relationship between nonverbal behavior and discourse state, and demonstrated how this finding can be incorporated into the generation of language and nonverbal behaviors for an embodied conversational agent.

Speakers produce posture shifts at 25% of discourse segment boundaries, and listeners also produce posture shifts when

(1) U: Hello
(2) R: Hello
(3) U: I'm looking for a house
(4) R: Where would you like to live?
(5) U: I like Boston
(6) R: what kind of transportation do you use?
(7) U: I need T access
…
(23) R: What kind of storage do you need?
(24) U: I will need extra storage in the basement
(25) R: Let's look at 123 Elm.
(26) U: OK

Figure 6.1: Example dialogue

speakers are initiating a new discourse segment, also at around 25% of discourse boundaries. Posture shifts occur much more frequently at discourse segment boundaries than discourse segment-internally. Furthermore, there is a relationship between discourse segmentation and conversational structure such that when speakers initiate a new segment at the same time as starting a turn, they produce a posture shift more frequently than when these categories do not co-occur.

Although this paper reports results from a limited number of monologues and dialogues, the findings are promising. They point the way, however, to a number of future directions. First, given the relationship between conversational structure and information structure in [5], we are led to examine the three-way relationship between discourse state, conversational structure (turns), and information structure. Next, we need to look at finer segmentations of the discourse. It is possible that smaller discourse segments are marked by similar posture shifts, performed with less energy. Finally, evaluation of the importance of such nonverbal behaviors to user interaction is essential. In a user study of our earlier Gandalf system [4], users rated the smoothness of the interaction and the agent's language skills significantly higher under test conditions in which Gandalf deployed conversational behaviors (gaze, turn-taking and limited gesture) than when these behaviors were disabled. Such an evaluation would also be informative for the Rea system. In addition, we would like to test whether generating posture shifts of this sort actually serves as a signal to listeners: do listeners remember more of the topics covered, when Rea generates

posture shifts at discourse segment boundaries? These evaluations are a part of our future research plans.

8. ACKNOWLEDGMENTS

This research reported here was supported by NSF (award IIS-9618939), MERL, AT&T, and the other generous sponsors of the MIT Media Lab. Thanks to the other members of the Gesture and Narrative Language Group for their contribution to this work, in particular Ian Gouldstone and Hannes Vilhjálmsson.

9. REFERENCES

[1] André, E., Rist, T., and Muller, J., Employing AI methods to control the behavior of animated interface agents, Applied Artficial Intelligence, vol. 13, pp. 415-448, 1999.

[2] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., and Yan, H., Embodiment in Conversational Interfaces: Rea, presented at CHI 99, Pittsburgh, PA, 1999.

[3] Cassell, J., Stone, M., and Yan, H., Coordination and context-dependence in the generation of embodied conversation, presented at INLG 2000, Mitzpe Ramon, Israel, 2000.

[4] Cassell, J. and Thorisson, K. R., The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents, Applied Artificial Intelligence, vol. 13, pp. 519-538, 1999.

[5] Cassell, J., Torres, O., and Prevost, S., Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation., in Machine Conversations, Y. Wilks, Ed. The Hague: Kluwer, 1999, pp. 143-154.

[6] Cassell, J., Vilhjálmsson, H., and Bickmore, T., BEAT: The Behavior Expression Animation Toolkit, Computer Graphics (Proceedings of SIGGRAPH), Los Angeles, CA, 2001.

[7] Chovil, N., Discourse-Oriented Facial Displays in Conversation, Research on Language and Social Interaction, vol. 25, pp. 163-194, 1992.

[8] Condon, W. S. and Osgton, W. D., Speech and body motion synchrony of the speaker-hearer, in The perception of language, D. H. Horton and J. J. Jenkins, Eds. New York: Academic Press, 1971, pp. 150-184.

[9] Duncan, S., On the structure of speaker-auditor interaction during speaking turns, Language in Society, vol. 3, pp. 161-180, 1974.

[10] Green, N., Carenini, G., Kerpedjiev, S., and Roth, S. F., A Media-Independent Content Language for Integrated Text and Graphics Generation, presented at Workshop on Content Visualization and Intermedia Representations at COLING and ACL '98, 1998.

[11] Grosz, B. and Sidner, C., Attention, Intentions, and the Structure of Discourse, Computational Linguistics, vol. 12, pp. 175-204, 1986.

[12] Halliday, M. A. K., Explorations in the Functions of Language. London: Edward Arnold, 1973.

[13] Kendon, A., Some Relationships between Body Motion and Speech, in Studies in Dyadic Communication, A. W. Siegman and B. Pope, Eds. Elmsford, NY: Pergamon Press, 1972, pp. 177-210.

[14] Lesh, N., Rich, C., and Sidner, C., Using Plan Recognition in Human-Computer Collaboration, presented at Proceedings of the Conference on User Modelling, Banff, Canada, 1999.

[15] Lester, J., Towns, S., Callaway, C., Voerman, J., and FitzGerald, P., Deictic and Emotive Communication in Animated Pedagogical Agents, in Embodied Conversational Agents, J. Cassell, J. Sullivan, et al, Eds. Cambridge: MIT Press, 2000.

[16] Lochbaum, K., A Collaborative Planning Model of Intentional Structure, Computational Linguistics, vol. 24, pp. 525-572, 1998.

[17] McNeill, D., Hand and Mind: What Gestures Reveal about Thought. Chicago, IL/London, UK: The University of Chicago Press, 1992.

[18] Rickel, J. and Johnson, W. L., Task-Oriented Collaboration with Embodied Agents in Virtual Worlds, in Embodied Conversational Agents, J. Cassell, J. Sullivan, et al, Eds. Cambridge: MIT Press, 2000.

[19] Rogers, W., The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior within Utterances, Human Communication Research, vol. 5, pp. 54-62, 1978.

[20] Takeuchi, A. and Nagao, K., Communicative facial displays as a new conversational modality, presented at InterCHI '93, Amsterdam, Netherlands, 1993.

[21] Thompson, L. and Massaro, D., Evaluation and Integration of Speech and Pointing Gestures during Referential Understanding, Journal of Experimental Child Psychology, vol. 42, pp. 144-168, 1986.