To appear in Dialog with Robots, AAAI Fall Symposium,
Washington, DC, November 2010.

Collaborative Discourse, Engagement and Always-On Relational Agents

Charles Rich and Candace L. Sidner
Worcester Polytechnic Institute
Worcester, Massachusetts, USA
{rich|sidner}@wpi.edu

Abstract

We summarize our past, present and future research related to human-robot dialogue, starting with its foundations in collaborative discourse theory, continuing to our current research on recognizing and generating engagement, and concluding with an outline of new work we are beginning on the modeling of long-term relationships between humans and robots.

This paper is a summary of our past, present and future research related to human-robot dialogue. In the first section below, we describe the main elements of collaborative discourse theory and the architecture of a collaborative interaction manager, which serve as the foundation of all of our later work. In the second section, we discuss current work on engagement, which underlies and supports collaboration and dialogue. Finally, in the third section, we outline new research we are beginning, which focuses on robots (and other agents) that are "always on" and therefore need to build long-term relationships with humans in their environment. More details on each of these topics can be obtained by referring to the cited publications.

Collaborative Discourse

The common thread through almost all of our research for the past many years has been the view that, whenever there is interaction—and especially communication—between two intelligent agents, collaborative discourse theory provides important insights into what is going on and helpful guidance in designing computer tools to support the interaction. Human-robot dialogue is the most recent example of where we have applied this approach.

Collaboration is a process in which two or more participants coordinate their actions toward achieving shared goals. Most collaboration between humans involves communication. *Discourse* is a technical term for an extended communication between two or more participants in a shared context, such as a collaboration. Collaborative discourse theory thus refers to a body of empirical and computational research about how people communicate in the context of a collaboration.

SharedPlans

Grosz and Sidner's SharedPlans (Grosz and Sidner 1986; 1990; Grosz and Kraus 1996) provide a general computational model of how collaborative, coordinated activity emerges from the individual beliefs and intentions (goals and plans) of the collaborators. Two collaborators have a SharedPlan for a given goal when they mutually believe that: (a) they have a common (shared) goal; (b) they have agreed on a recipe to accomplish the goal; (c) they are each capable of performing their respective actions; (d) each intends to perform their respective actions; and (e) they are both committed to the overall success of the collaboration.

In a typical collaboration, not all of the five conditions above are satisfied at the start. Instead, SharedPlans typically start in a partial state, e.g., having a shared goal, and incremental refinement of the SharedPlan is interleaved with performing actions that contribute toward the goal. Along the way, much of the communication between the collaborators often has to do with refining the SharedPlan, such as negotiating who should do which action.

Collaborative Interaction Manager

We have implemented two collaborative interaction managers based on SharedPlans, Collagen (Rich and Sidner 1998; Rich, Sidner, and Lesh 2001) and its recent successor, Disco. Collagen has been used to build more than a dozen human-computer collaborative systems. Unlike most so-called "dialogue managers," Collagen and Disco manage *both* the conversational and the task structure of an interaction. This is because, according to SharedPlan theory, these two structures are deeply intertwined. Both Collagen and Disco currently support only two-participant discourse (dialogue), although we have experimented with multi-participant extensions to both of them.

Disco differs from Collagen mainly in using the ANSI/CEA-2018 standard, whose development was led by Rich (2009), for representing task models. Disco also lacks the logical inference and truth maintenance facilities included in Collagen. Disco is written in Java and is distributed under the MIT open-source license; a copy may be obtained by sending email to Rich.

Figure 1 shows the architecture of both Collagen and Disco. The two key data structures in this architecture are the task model and the discourse state. The task model is

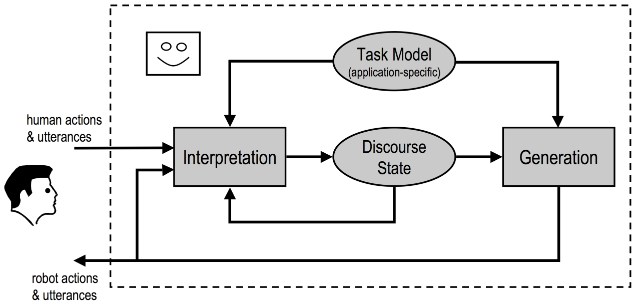


Figure 1: Architecture of Collaborative Interaction Manager

an abstract, hierarchical, partially ordered representation of the actions typically performed to achieve goals in the application domain. The discourse state contains instances of classes/types defined in the task model. The two key algorithms, discourse interpretation and generation, are essentially application-independent (small procedural extensions can be provided to support application-specific heuristics).

Engagement

Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake (Sidner et al. 2005). To elaborate,

> ...when people talk, they maintain conscientious psychological connection with each other and each will not let the other person go. When one is finished speaking, there is an acceptable pause and then the other *must* return something. We have this set of unspoken rules that we all know unconsciously but we all use in every interaction. If there is an unacceptable pause, an unacceptable gaze into space, an unacceptable gesture, the cooperating person will change strategy and try to reestablish contact. Machines do none of the above, and it will be a whole research area when people get around to working on it. (Biermann, invited talk at User Modeling Conference, 1999)

Nonverbal behaviors, i.e., movements (gestures) of the head, eyes, limbs and body, such as looking, nodding, pointing and stance, are thus an essential part of the engagement process. In our work to date, we have concentrated on understanding how looking, pointing and head nods and shakes contribute to maintaining engagement.

The relationship between engagement and collaboration is illustrated in Figure 2. Generally speaking, engagement is a "lower level" process; it is closer to the "hardware" and has shorter real-time constraints. Collaboration is a higher level cognitive function, with a correspondingly slower real-time action rate. In general, engagement supports collaboration. For example, a collaborator relies on the engagement state to know when it is appropriate to continue with the collaboration.

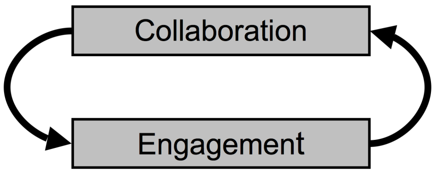


Figure 2: Engagement and Collaboration

However, engagment and collaboration are not strictly layered. The state of the collaboration can also affect how engagement behaviors are interpreted. For example, whether or not to interpret breaking eye contact (looking away) as an attempt at disengagement depends on whether the next action in the collaboration requires looking at a shared artifact—if it does, then looking away does not signal disengagement.

We believe that engagement is a fundamental process that underlies all human interaction and has common features across a very wide range of interaction circumstances. At least for humanoid robots, this implies that modeling engagement is crucial for constructing robots that can interact effectively with humans without special training. The main goal of our current research is therefore to develop a Robot Operating System (ROS) engagement module that can be reused across different robots and applications. ROS (see ros.org) is an open-source multi-platform robotics software framework, whose goal is to increase code reuse in the robotics research and development community.

Connection Events

Our most recent theoretical contribution in the area of engagement has been to identify and codify, based on our own (Rich et al. 2010) and others' studies of human behavior, four types of what we call *connection events*, i.e., events involving gesture and speech that contribute to the perceived connection between humans: directed gaze, mutual facial gaze, conversational adjacency pairs and backchannels (see Figure 3). Our hypothesis is that these events, occuring at some minimum frequency, are the process mechanism for maintaining engagement.

Figures Figure 4(a) through (d) show time lines for these four types of connection events. In the discussion below, we describe the observeable behaviorial components of each event type and hypothesize regarding the accompanying intentions of the participants. Dotted lines indicate optional behaviors. Also, note that gesture and speech events often overlap.

Directed Gaze In directed gaze (Kendon 1967), one person (the *initiator*) looks and optionally points at some object or group of objects in the immediate environment, following which the other person (the *responder*) looks at the same object(s). We hypothesize that the initiator intends to bring the indicated object(s) to the responder's attention, i.e., to make the object(s) more salient in the interaction. This event is often synchronized with the initiator referring to the object(s) in speech, as in "now spread the *cream cheese* on the cracker." By turning his gaze where directed, the responder intends to be cooperative and thereby signals his desire to continue the interaction (maintain engagement).

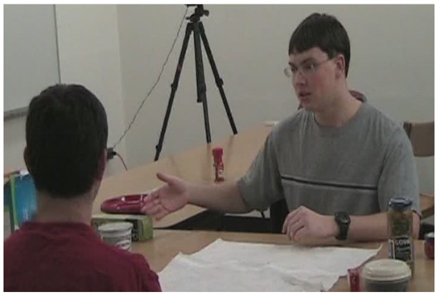
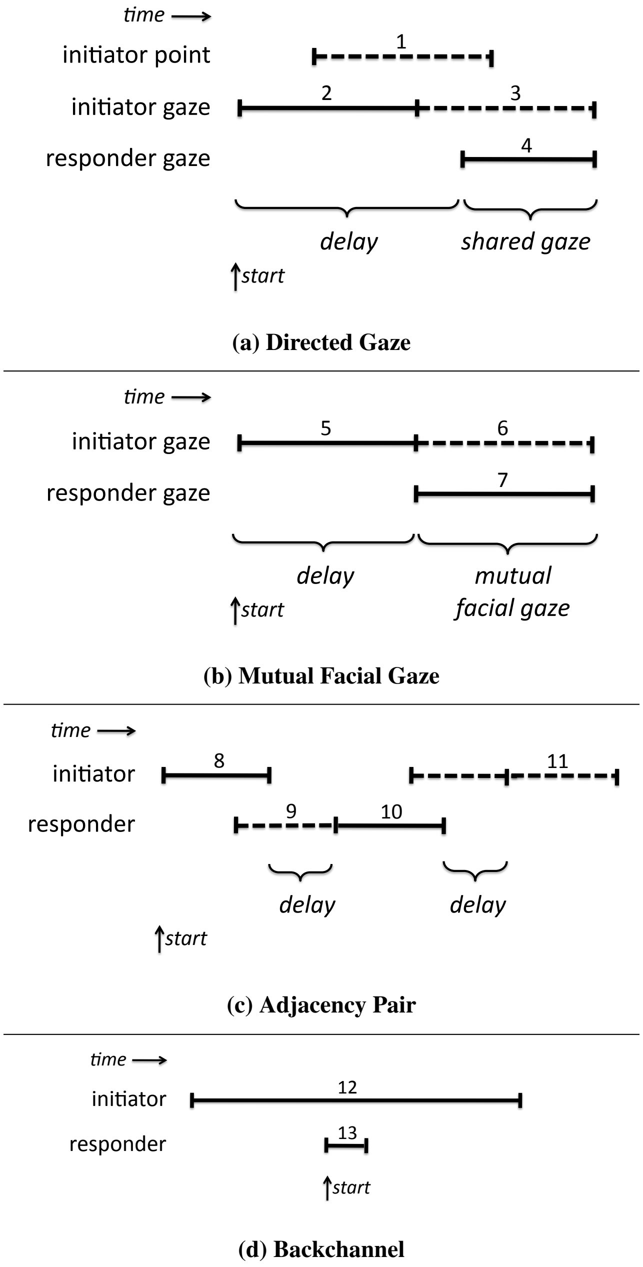


Figure 3: Human Engagement Study



(a) Directed Gaze

(b) Mutual Facial Gaze

(c) Adjacency Pair

(d) Backchannel

Figure 4: Time Lines for Connection Events (numbers for reference in text)

In more detail (see Figure 4(a)), notice first that the pointing behavior (1), if it is present, begins after the initiator starts to look (2) at the indicated object(s). This is likely because it is hard to accurately point at something without looking to see where it is located.[1] Furthermore, we observed several different configurations of the hand in pointing, such as extended first finger, open hand (palm up or palm down—see Figure 3), and a circular waving motion (typically over a group of objects). An interesting topic for future study (that will contribute to robot generation of these behaviors) is to determine which of these configurations are individual differences and which serve different communicative functions.

After some delay, the responder looks at the indicated object(s) (4). The initiator usually maintains the pointing (1), if it is present, at least until the responder starts looking at the indicated object(s). However, the initiator may stop looking at the indicated object(s) (2) before the responder starts looking (4), especially when there is pointing. This is often because the initiator looks at the responder's face, assumedly to check whether the responder has directed his gaze yet. (Such a moment is captured in Figure 3.)

Finally, there may be a period of shared gaze, i.e., a period when both the initiator (3) and responder (4) are looking at the same object(s). Shared gaze has been documented (Brennan 1999) as an important component of human interaction.

Mutual Facial Gaze Mutual facial gaze (Argyle and Cook 1976) has a time line (see Figure 4(b)) similar to directed gaze, but simpler, since it does not involve pointing. The event starts when the initiator looks at the responder's face (5). After a delay, the responder looks at the initiator's face, which starts the period of mutual facial gaze (6,7). Notice that the delay can be zero, which occurs when both parties simultaneously look at each other.

The intentions underlying mutual facial gaze are less clear than those for directed gaze. We hypothesize that both the initiator and responder in mutual facial gaze engage in this behavior because they intend to maintain the engagement process. Mutual facial gaze does however have other interaction functions. For example, it is typical to establish mutual facial gaze at the end of a speaking turn. Mutual facial gaze can also be affected by the social relationship between the parties (Emery 2000).

Finally, what we are calling mutual facial gaze is often referred to informally as "making eye contact." This latter term is a bit misleading since people do not normally stare continously into each other's eyes, but rather their gaze roams around the other person's face, coming back to the eyes from time to time.

Adjacency Pair In linguistics, an adjacency pair consists of two utterances by two speakers, with minimal overlap or

[1] It is usually possible to creatively imagine an exception to almost any rule such as this. For example, if a person is standing with his back to a mountain range, he might point over his shoulder to "the mountains" without turning around to look at them. We will not bother continuing to point out the possibility of such exceptions.

gap between them, such that the first utterance provokes the second utterance. A question-answer pair is a classic example of an adjacency pair. We generalize this concept slightly to include both verbal (utterances) and non-verbal communication acts. So for example, a nod could be the answer to a question, instead of a spoken "yes." Adjacency pairs, of course, often overlap with the gestural connection events, directed gaze and mutual facial gaze.

The simple time line for an adjacency pair is shown in Figure 4(c). First the initiator communicates what is called the *first turn* (8). Then there is a delay, which could be zero if the responder starts talking before the the initiator finishes (9). Then the responder communicates what is called the *second turn* (9,10). In some conversational circumstances, this could also be followed by a *third turn* (11) in which the initiator, for example, repairs the responder's misunderstanding of his original communication or comments on what the responder said.

Backchannel A backchannel is an event (see Figure 4(d)) in which one party (the responder) directs a brief verbal or gestural communication (13) back to the initiator *during* the primary communication (12) from the initiator to the responder. Typical examples of backchannels are nods and/or saying "uh, huh." Backchannels are typically used to communicate the responder's comprehension of the initiator's communication (or lack thereof, e.g., a quizzical facial expression) and/or desire for the initiator to continue. Unlike the other three connection event types, the start of a backchannel event is defined as the start of the responder's behavior and this event has no concept of delay.

Engagement Recognition and Generation

Based on the analysis of connection events above, we have implemented an engagement recognition module (Rich et al. 2010), and are working on an engagement generation module for human-robot interaction.

The engagement recognition module is organized as four parallel finite state machines (recognizers), each of which recognizes the time line of a given connection event type (more than one recognizer may be active at a time). The inputs to these recognizers is information (e.g., from the robot's cameras) about where the human is looking and pointing and when the human nods or shakes his head. The recognizers also need to know where the robot is looking and pointing and when it nods or shakes its head (this information is available from the robot's control system).

In addition to providing real-time feedback to the robot on the successful or unsuccessful completion of specific connection events, the engagement module also computes ongoing statistics on the overall engagement process, such as the mean time between connection events (MTBCE), which we hypothesize captures something of what is informally called the "pace" of an interaction (Dix 1992):

$$\text{pace} \quad \propto \quad \frac{1}{\text{MTBCE}}$$

In other words, the faster the pace, the less the time between connection events.

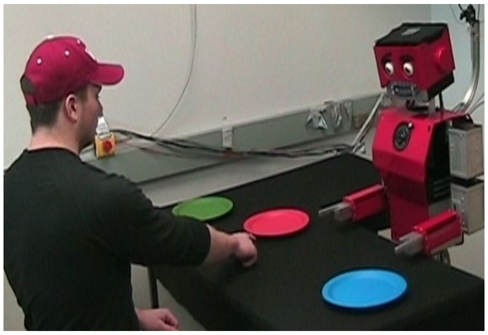


Figure 5: Human-Robot Interaction

The engagement generation module receives communicative intentions from the higher cognitive functions of a robot and is responsible for implementing them by adding nonverbal behaviors, such as looking, pointing, nodding, etc., to the given verbal material. For example, this component may decide to initiate a directed gaze toward the peanut butter jar at the appropriate point in the utterance, "please put the peanut butter on the round cracker." We are using the Behavior Markup Language (Kopp et al. 2006) as a tool in this process. The engagement generation module is also responsible for maintaining what might be called the engagment "heartbeat," i.e., making sure that there is some kind connection event, such a mutual facial gaze, at some minimum frequency.

As a preliminary validation of our engagment recognition model, we developed a simple human-robot demonstration, which we call the "pointing game" (see Figure 5), that naturally includes the three of the connection event types above (no backchannels). Several plates of different colors are place randomly on the table between the human and robot. The robot starts the game by saying "Please point at a plate." The human is then expected to respond by pointing at any plate. The robot identifies the chosen plate by pointing to it and saying, for example, "You pointed at the red plate." If the human does not respond within a certain amount of time, the robot asks "Do you want to stop now?" If the human nods yes, the robot says "Thank you for playing"; if he shakes no, then the robot repeats its last request. We are currently working toward on a systematic evaluation of both the recognition and generation modules, using a more complex interaction involving collaboratively solving a tabletop tangram puzzle.

Always-On Relational Agents

The new research we are about to embark upon together with Tim Bickmore at Northeastern University is motivated by the following question:

> *What should your robot do when it is not busy obeying your commands?*

With the rapid drive of technology towards placing increasingly capable robots into human home, work and play environments, it is time to start asking this question. Obviously the answer depends a lot on what kind of robot we are talking about. For example, if it's a turtle-like floor cleaning robot, it probably should wait silently out of sight until it's time to clean the floor. But how about a humanoid household robot? According to some science fiction visions of the future, even such robots should stand silently at attention until commanded to do something.

Our answer is that a successful robot should spend its time building and maintaining long-term social relationships with the humans in its environment. Because humans are deeply and fundamentally social beings, they cannot help but expect a continuously present artificial being, especially if humanoid, to become part of their network of relationships. Furthermore, human-to-non-human social relationships can have value both in and of themselves, as demonstrated by the effectiveness of animal pet companions for isolated older adults (Banks and Banks 2002), and also contribute to the success of more instrumental interaction goals, as demonstrated by Bickmore's research on social dialogue in exercise coaches (Bickmore and Schulman 2009).

Figure 6 summarizes the architecture of the agents we will be building.

SharedPlans Relationship Theory

You cannot create a relationship in isolation. Relationships require the active participation and commitment of both (all of) the participants. As a starting point for the theoretical foundations of relational agents, we therefore look to theories of collaboration, specifically the SharedPlans theory discussed above.

All of the applications of SharedPlans thus far have involved what might be called *instrumental goals*. To develop the theoretical foundations for long-term always-on relational agents, we will apply SharedPlans theory to *relational goals*. We will also extend the theory to include default rules for how to refine SharedPlans for relational goals. Finally, we need to account for where the agents' relational goals come from.

Relational Goals. Psychological theories of relationship (Brehm 1992) indicate that partners initiate, build, maintain, modify, repair and terminate relationships. In order to model these activities as collaborative undertakings between a relational agent and a human, we will need to answer the following questions: How can these activities be expressed as shared goals? What is the taxonomy of relational goals? What types of actions contribute towards relational goals? What underlying beliefs and intentions need to be modeled to express individual and shared relational goals? What types of recipes do agents need to achieve relational goals?

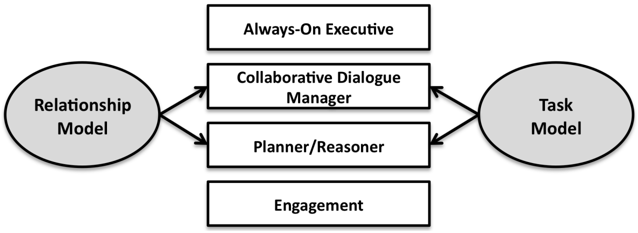


Figure 6: Architecture for Always-On Relational Agent

Psychologists have also observed (Duck 1998) that relational and instrumental goals and actions are deeply interwoven in social interactions. For example, an instrumental action, such as helping someone with their homework, also contributes toward the goal of building a relationship between the collaborators. Conversely, social relationship dialogue, such as "Where do you live?" and "How are you doing today?," contributes to the effectiveness of instrumental goals, as demonstrated by Bickmore's virtual exercise coach (Bickmore and Schulman 2009). This synergy is a key reason why we believe that domestic robots will inevitably become social members of the household.

However, in current dialogue systems the social dialogue, if any, is ad hoc and hand-coded into the same task model with the instrumental parts of the collaboration. In fact, some researchers, e.g., (Wilks et al. 2010), have argued that social dialogue cannot be handled with task models at all. SharedPlans Relationship Theory will provide a systematic treatment of both instrumental and relational goals and their interleaving. In addition to providing a clearer scientific understanding, a practical advantage of this approach is that it will allow independent development and reuse of relational and instrumental recipes, thereby reducing the cost of building always-on relational agents.

Relationship Model. In order to model relational goals adequately for use with planning algorithms, we need to represent the intended effects (postconditions) of relational actions, which requires a model of (the current state of) the social relationship between the relational agent and each of its human partners. As a starting point, we plan to use Bickmore's relationship model (Bickmore 2003), which is based on the notion of accommodation (Thomason 1990), in which a collaborator infers its partner's goals and takes action to help without being explicitly requested to do so. A relationship is then characterized by a set of accommodations, i.e., a set of implicitly agreed upon collaborative task commitments. These in turn create expectations for future collaborations, should the need arise, and reflect provision-based models of relationship in social psychology (Bickmore and Picard 2005). For example, the accommodation set between two friends might come to include all the activities they have done and enjoyed together in the past (and therefore expect in the future), such as playing cards, sharing a meal, and so on. We will develop methods for determining how to initialize and update the set of accommodations based on theories from the social psychology of personal relationships.

Default Rules. The default rules for refining instrumental goals are not adequate for refining relational goals. For example, instrumental collaborations are usually negotiated explicitly, whereas relational goals are by default tacit. It's perfectly natural for adults to say "Let's fix the faucet," but only children think to initiate a relationship by saying "Let's be friends." Adults use a range of tactics to start or move a relationship forward, but few are explicit. We will investigate such relationship-building tactics in detail and develop corresponding default rules for refining SharedPlans using them.

Default rules will also provide general strategies for adding a relational contribution to any instrumental task. One such rule is to reflect upon performance. The difference between a robot companion that simply knows how to play Gin Rummy, and a relational robot that can participate in a social game of cards is having default rules for appropriately generating comments such as "You've gotten a lot better!" or "I really blew that hand." The engagement behaviors discussed above, such recognizing when someone wants to initiate or terminate a conversation and knowing how to signal these intentions yourself, are also important relational skills.

Desire for Relationship. In addition to the beliefs and intentions of SharedPlans theory, we plan to imbue our relational agents with a permanent desire to establish and maintain social relationships with the humans in their environments. This desire gives rise to new and persistent relational goals, such as when a new person enters the environment. Other goals, such as playing cards, may come into play as ways of achieving a relationship goal, or from an agent-specific desire, such as helping the human to change his or her health behavior. The relationship model can also serve as a source for persistent goals; one of the tasks in the accommodation set can be chosen as a goal to maintain the relationship.

Acknowledgments This work was supported in part by the National Science Foundation under award IIS-0811942.

References

Argyle, M., and Cook, M. 1976. *Gaze and Mutual Gaze*. New York: Cambridge University Press.

Banks., M., and Banks, W. 2002. The effects of animal-assisted therapy on loneliness in an elderly population in long-term care facilities. *J Geronol Med Sci* 57A:M428–M432.

Bickmore, T., and Picard, R. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Trans. on Computer Human Interaction* 12(2):293–327.

Bickmore, T., and Schulman, D. 2009. A virtual laboratory for studying long-term relationships between humans and virtual agents. In *Proc. Autonomous Agents and Multi-Agent Systems*.

Bickmore, T. 2003. *Relational Agents: Effecting Change through Human-Computer Relationships*. Ph.D. Dissertation, MIT Media Laboratory.

Brehm, S. 1992. *Intimate Relationships*. New York: McGraw-Hill.

Brennan, S. 1999. How conversation is shaped by visual and spoken evidence. In Trueswell, J., and Tanenhaus, M., eds., *Approaches to Studying World-Situated Language Use*. Cambridge, MA: MIT Press. 95–129.

Dix, A. 1992. Pace and interaction. In *Proc. of HCI'92: People and Computers VII*, 193–207. Cambridge University Press.

Duck, S. 1998. *Human Relationships*. London: SAGE Publications.

Emery, N. J. 2000. The eyes have it: The neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews* 24:145–146.

Grosz, B. J., and Kraus, S. 1996. Collaborative plans for complex group action. *Artificial Intelligence* 86(2):269–357.

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

Grosz, B. J., and Sidner, C. L. 1990. Plans for discourse. In Cohen, P. R.; Morgan, J. L.; and Pollack, M. E., eds., *Intentions and Communication*. Cambridge, MA: MIT Press. 417–444.

Kendon, A. 1967. Some functions of gaze direction in two person interaction. *Acta Psychologica* 26:22–63.

Kopp, S.; Krenn, B.; Marsella, S.; Marshall, A. N.; Pelachaud, C.; Pirker, H.; Thrisson, K.; and Vilhjlmsson, H. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Proc. Conf. on Intelligent Virtual Agents*.

Rich, C., and Sidner, C. 1998. Collagen: A collaboration manager for software interface agents. *User Modeling and User-Adapted Interaction* 8(3/4):315–350.

Rich, C.; Lesh, N.; Rickel, J.; and Garland, A. 2002. A plug-in architecture for generating collaborative agent responses. In *Proc. 1st Int. J. Conf. on Autonomous Agents and Multiagent Systems*.

Rich, C.; Ponsler, B.; Holroyd, A.; and Sidner, C. 2010. Recognizing engagement in human-robot interaction. In *Proc. ACM Conf. on Human-Robot Interaction*.

Rich, C.; Sidner, C.; and Lesh, N. 2001. Collagen: Applying collaborative discourse theory to human-computer interaction. *AI Magazine* 22(4):15–25.

Rich, C. 2009. Building task-based user interfaces with ANSI/CEA-2018. *IEEE Computer* 42(8):20–27.

Sidner, C. L.; Lee, C.; Kidd, C.; Lesh, N.; and Rich, C. 2005. Explorations in engagement for humans and robots. *Artificial Intelligence* 166(1-2):104–164.

Thomason, R. 1990. Accommodation, meaning, and implicature: Interdisciplinary foundations for pragmatics. In Cohen, P. R.; Morgan, J. L.; and Pollack, M. E., eds., *Intentions and Communication*. Cambridge, MA: MIT Press. 325–364.

Wilks, Y.; Catizone, R.; Worgan, S.; and Turunen, M. 2010. Some background on dialogue management and conversational speech for dialogue systems. *Computer Speech and Language*. Forthcoming.