Unspoken Rules of Spoken Interaction

Timothy W. Bickmore, Boston University School of Medicine

Our face-to-face interactions with other people are governed by a complex set of rules, of which we are mostly unaware. For decades now, social scientists have been unraveling the threads of face-to-face interaction, investigating everything from descriptions of body posture used to indicate interest in starting a conversation, to eye gaze dynamics used to convey liking or disliking, to the myriad ways that language can convey attitude, social status, relationship status and affective state. Even though we are not always aware of them, these rules underpin how we make sense of and navigate in our social world. These rules may seem uninteresting and irrelevant to many computer scientists but, to the extent that a given interaction rule is universally followed within a user population, it can be profitably incorporated into a human-machine interface in order to make the interface more natural and intuitive to use. Computers without anthropomorphic faces and bodies can (and already do) make use of only a limited range of such rules—such as rules for conversational turn-taking in existing interfaces—but one kind of interface has the potential to make explicit, maximal use of these rules: embodied conversational agents.

Embodied conversational agents (ECAs) are animated humanoid computer characters that emulate face-to-face conversation through the use of hand gestures, facial display, head motion, gaze behavior, body posture, and speech intonation, in addition to speech content [5]. The use of verbal and nonverbal modalities gives ECAs the potential to fully employ the rules of etiquette observed in human face-to-face interaction. ECAs have been developed for research purposes, but there are also a growing number of commercial ECAs, such as those developed by Extempo, Artificial Life, and the Ananova newscaster. These systems vary greatly in their linguistic capabilities, input modalities (most are mouse/text/speech input only), and task domains, but all share the common feature that they attempt to engage the user in natural, full-bodied (in some sense) conversation.

Conversational Functions vs. Conversational Behaviors

Social scientists have also long recognized the utility of making a distinction between conversational behaviors (surface form, such as head nodding) and conversational function (the role played by the behavior, such as acknowledgement). This distinction is important if general interactional rules are to be induced that capture the underlying regularities in conversation, enabling us to build ECA architectures that have manageable complexity, and that have the potential of working across languages and cultures. This distinction is particularly important given that there is usually a many-to-many mapping between functions and behaviors (e.g., head nodding can also be used for emphasis and acknowledgement can also be indicated verbally).

Although classical linguistics has traditionally focused on the conveying of propositional information, there are actually many different kinds of conversational function. To gain

an understanding of the range of conversational functions and their associated behaviors, the list below reviews some of the functions most commonly implemented in ECAs.

Propositional Functions The propositional function of a conversational behavior involves representing a thought to be conveyed to a listener. In addition to the role played by speech, hand gestures are used extensively to convey propositional information that is either redundant with, or complementary to, the information delivered in speech. In ECA systems developed to date, the most common kind of hand gesture implemented is the deictic, or pointing gesture. Steve [10], the DFKI Persona [1], and pedagogical agents developed by Lester, et al, [7] all use pointing gestures which have the function of referencing objects in the agent's immediate (virtual or real) environment.

Interactional Functions Interactional functions are those that serve to regulate some aspect of the flow of conversation (also called "envelope" functions). Examples include turn-taking functions, such as signaling intent to take or give up a speaking turn, and conversation initiation and termination functions, such as greetings and farewells (used in REA, see sidebar). Other examples are "engagement" functions, which serve to continually verify that one's conversational partner is still engaged in and attending to the conversation, as implemented in the MEL robotic ECA [11]. Framing functions (enacted through behaviors called "contextualization cues") serve to signal changes in the kind of interaction that is taking place, such as problem-solving talk vs. small talk vs. joke-telling, and are used in the FitTrack Laura ECA (see sidebar).

Attitudinal Functions Attitudinal functions signal liking, disliking or other attitude directed towards one's conversational partner (as one researcher put it, "you can barely utter a word without indicating how you feel about the other"). One of the most consistent findings in this area is that the use of nonverbal "immediacy behaviors"--close conversational distance, direct body and facial orientation, forward lean, increased and direct gaze, smiling, pleasant facial expressions and facial animation in general, head nodding, frequent gesturing and postural openness--projects liking for the other and engagement in the interaction, and is correlated with increased solidarity [2]. Attitudinal functions were built into the FitTrack ECA so that it could signal liking when attempting to establish and maintain working relationships with users, and into the Cosmo pedagogical agent to express admiration or disappointment when students experienced success or difficulties [7].

Affective Display Functions In addition to communicating attitudes about their conversational partners, people also communicate their overall affective state to each other using a wide range of verbal and nonverbal behaviors. Although researchers have widely differing opinions about the function of affective display in conversation, it seems clear that it is the result of both spontaneous readouts of internal state and deliberate communicative action. Most ECA work in implementing affective display functions has focused on the use of facial display, such as the work by Poggi and Pelachaud [8].

Relational Functions Relational functions are those that either indicate a speaker's current assessment of his/her social relationship to the listener ("social deixis"), or serve

to move an existing relationship along a desired trajectory (e.g., increasing trust, decreasing intimacy, etc.). Explicit management of the ECA-user relationship is important in applications in which the purpose of the ECA is to help the user undergo a significant change in behavior or cognitive or emotional state, such as in learning, psychotherapy or health behavior change [3]. Both REA and Laura were developed to explore the implementation and utility of relational functions in ECA interactions.

While it is easiest to think of the occurrence (vs. non-occurrence) of a conversational behavior as achieving a given function, conversational functions are often achieved by the *manner* in which a given behavior is performed. For example, a gentle rhythmic gesture communicates a very different affective state or interpersonal attitude compared to a sharp exaggerated gesture. Further, while a given conversational behavior may be used primarily to effect a single function, it can usually be seen to achieve functions from several (if not all) of the categories listed above. A well-told conversational story can communicate information, transition a conversation into a new topic, convey liking and appreciation of the listener, explicate the speaker's current emotional state, and serve to increase trust between the speaker and listener.

The Rules of Etiquette

Within this framework, rules of etiquette can be seen as those conversational behaviors that fulfill certain conversational functions. Emily Post would have us believe that the primary purpose of etiquette is the explicit signaling of "consideration for the other"—that one's conversational partner is important and valued [9]—indicating these behaviors enact a certain type of attitudinal function. Etiquette rules also often serve as coordination devices (e.g., ceremonial protocols) in which case they can be seen as enacting an interactional function. They can also be used to explicitly signal group membership or to indicate a desire to move a relationship in a given direction, in which case they are fulfilling a relational function. Each of these functions has been (partially) explored in existing ECA systems.

Is etiquette—especially as enacted in nonverbal behavior—important in all kinds of human-computer interactions? Probably not. However, for tasks that are more fundamentally social in nature, the rules of etiquette and the affordances of nonverbal behavior can certainly have an impact. Several studies of mediated human-human interaction have found that the additional nonverbal cues provided by video-mediated communication do not affect performance in task-oriented interactions, but in interactions of a more relational nature, such as getting acquainted, video is superior [12]. These studies have found that for social tasks, interactions were more personalized, less argumentative and more polite when conducted via video-mediated communication, that participants believed video-mediated (and face-to-face) communication was superior, and that groups conversing using video-mediated communication tended to like each other more, compared to audio-only interactions. The importance of nonverbal behavior is also supported by the intuition of businesspeople who still conduct most important business meetings face-to-face rather than on the phone. It would seem that when a user is

performing these kinds of social tasks with a computer that an ECA would have a distinct advantage over non-embodied interfaces.

Conclusion

Will users willingly engage in a social chat with an animated real estate agent or tell their troubles to a virtual coach? Evidence to date indicates that the answer is, for the most part, yes. In the commercial arena, people appear willing to engage artifacts such as Tamagotchis, Furbies and robotic baby dolls in ever more sophisticated and encompassing social interactions. Experience in the laboratory also indicates that not only will users readily engage in a wide range of social behavior appropriate to the task context, but that the social behaviors have the same effect on them as if they had been interacting with another person [3-5]. This trend seems to indicate a human readiness, or even need, to engage computational artifacts in deeper and more substantive social interactions.

Unfortunately, there is no cookbook yet defining all of the rules for human face-to-face interaction that human-computer interface practitioners can simply implement. However, many of the most fundamental rules have been codified in work by linguists, sociolinguists and social psychologists (e.g., [2]), and exploration that makes explicit use of these rules in work with ECAs and robotic interfaces has begun. By at least being cognizant of these rules, and at most by giving them explicit representation in system design, developers can build systems that are not only more natural, intuitive and flexible to use, but which result in better outcomes for many kinds of tasks.

Sidebar: **REA the Polite Real Estate Agent**

REA is a virtual real estate agent who conducts initial interviews with potential home buyers, then shows them virtual houses that she has for sale [4]. In these interviews—based on studies of human real estate agent dialogue—REA is capable of using a variable level of etiquette, which in turn conveys varying levels of sensitivity to users' "face needs" (needs for acceptance and autonomy). If the etiquette gain is turned up, she starts the conversation with small talk, gradually eases into the real estate conversation, and sequences to more threatening topics, like finance, towards the end of the interview. If the etiquette gain is turned down, her conversational moves are entirely driven by task goals, resulting in her asking the most important questions first (location and finance) and not conducting any small talk whatsoever. The amount of etiquette required at any given moment is dynamically updated each speaking turn of the conversation based on an assessment of the relationship between REA and the user, and how it changes as different topics are discussed.

Figure 1. REA Interviewing a User

Rea's dialogue planner is based on an activation network that integrates information from the following sources to choose her next conversational move:

- Task goals – REA has a list of prioritized goals to find out about the user's housing needs in the initial interview. Conversational moves that directly work towards satisfying these goals (such as asking interview questions) are preferred (given activation energy).
- Logical preconditions – Conversational moves have logical preconditions (e.g., it makes no sense for REA to ask users how many bedrooms they want until she has established that they are interested in buying a house), and are not selected for execution until all of their preconditions are satisfied. Activation energy flows through the network to prefer moves that are able to be executed ("forward chaining") or that support (directly or indirectly) REA's task goals ("backward chaining").
- Face threat – Moves that are expected to cause face threats to the user, including threats due to overly invasive topics (like finance) are dispreferred.
- Face threat avoidance – Conversational moves that advance the user-agent relationship in order to achieve task goals that would otherwise be threatening (e.g., small talk and conversational storytelling to build trust) are preferred.
- Topic Coherence – Conversational moves that are somehow linked to topics currently under discussion are preferred.
- Relevance – Moves that involve topics known to be relevant to the user are preferred.
- Topic enablement – REA can plan to execute a sequence of moves that gradually transition the topic from its current state to one that REA wants to talk about (e.g., from talk about the weather, to talk about Boston weather, to talk about Boston real estate). Thus, energy is propagated from moves whose topics are not currently active to moves whose topics would cause them to become current.

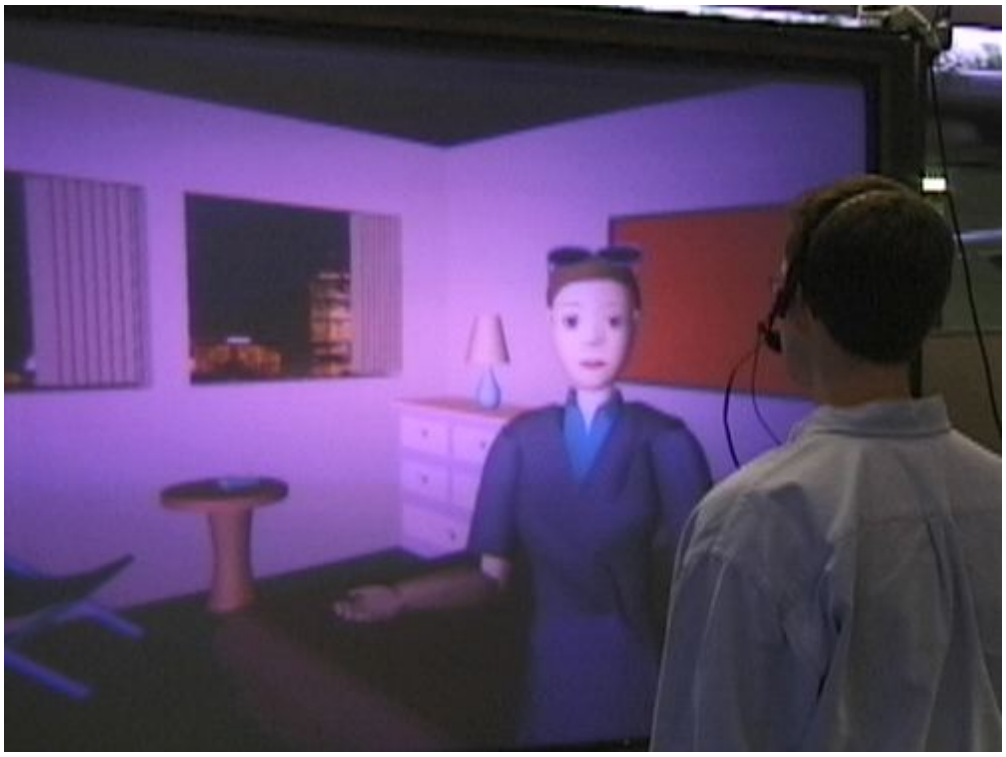

Figure 1. REA Interviewing a Buyer

Sidebar: **Automatic Generation of Nonverbal Behavior in BEAT**

Although the nonverbal behavior exhibited by an embodied conversational agent can play a significant role in enacting rules of etiquette, the correct production of these behaviors can be a very complex undertaking. Not only must the form of each behavior be correct, but the timing of the behavior's occurrence relative to speech must be precise if the behavior is to have the intended effect on the user.

The BEAT system simplifies this task, by taking the text to be spoken by an animated human figure as input, and outputting appropriate and synchronized nonverbal behaviors and synthesized speech in a form that can be sent to a number of different animation systems [6]. The nonverbal behaviors are assigned on the basis of linguistic and contextual analysis of the text, relying on rules derived from research into human conversational behavior. BEAT can currently generate hand gestures, gaze behavior, eyebrow raises, head nods and body posture shifts, as well as intonation commands for a text-to-speech synthesizer.

Figure 2. BEAT Annotated Parse Tree and Its Performance

The BEAT system was designed to be modular, to operate in real-time and to be easily extensible. To this end, it is written in Java, is based on an input-to-output pipeline approach with support for user-defined extensions, and uses XML as its primary data structure. Processing is decomposed into modules that operate as XML transducers; each taking an XML object tree as input and producing a modified XML tree as output. The first module in the pipeline operates by reading in XML-tagged text representing the character's script and converting it into a parse tree. Subsequent modules augment this XML tree with suggestions for appropriate nonverbal behavior while filtering out suggestions that are in conflict or do not meet specified criteria. Figure 2 shows an example XML tree at this stage of processing, with annotations for speech intonation (SPEECH-PAUSE, TONE, and ACCENT tags), gaze behavior (GAZE-AWAY and GAZE-TOWARDS, relative to the user), eyebrow raises (EYEBROWS), and hand gestures (GESTURE). In the final stage of processing, the tree is converted into a sequence of animation instructions and synchronized with the character's speech by querying the speech synthesizer for timing information.

BEAT provides a very flexible architecture for the generation of nonverbal conversational behavior, and is in use on a number of different projects at different research centers, including the FitTrack system, described on page XX.

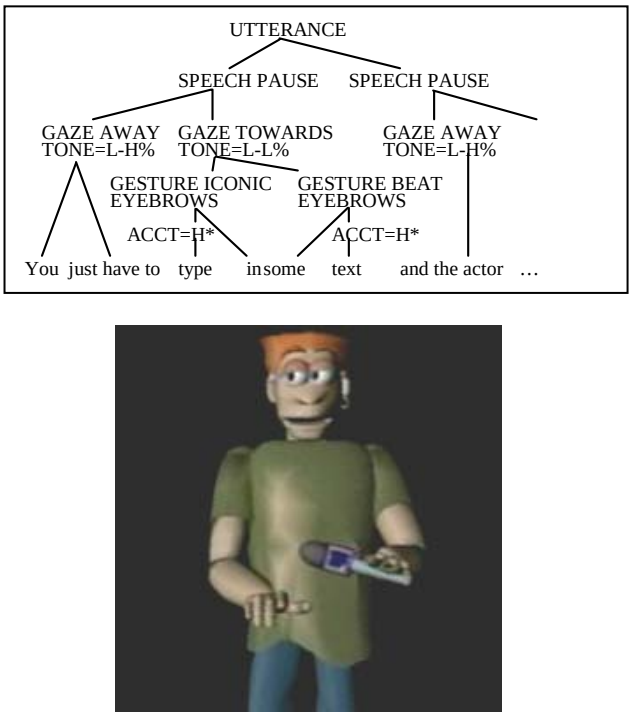

Figure 2. BEAT Annotated Parse Tree and Its Performance for "You just have to type in some text and the actor… "

Sidebar: **Managing Long-Term Relationships with Laura**

The effective establishment and maintenance of relationships requires the use of many subtle rules of etiquette that change over time as the nature of the relationship changes. The FitTrack system was developed to investigate the ability of embodied conversational agents to establish and maintain long-term, social-emotional relationships with users, and to determine if these relationships could be used to increase the efficacy of health behavior change programs delivered by the agent [3]. The system was designed to increase physical activity in sedentary users through the use of conventional health behavior change techniques combined with daily conversations with Laura, a virtual, embodied exercise advisor.

Laura's appearance and nonverbal behavior were based on a review of the health communication literature and a series of pre-test surveys (see Figure 3). BEAT (see page XX) was used to generate nonverbal behavior for Laura, and was extended so that it would generate different baseline nonverbal behaviors for high or low immediacy (liking or disliking of one's conversational participant demonstrated through nonverbal behaviors such as proximity and gaze) and different conversational frames (health dialogue, social dialogue, empathetic dialogue and motivational dialogue). In addition to the nonverbal immediacy behaviors, verbal relationship-building strategies used by Laura include: empathy dialogue, social dialogue, meta-relational communication (talk about the relationship), humor, reference to past interactions and future together, inclusive pronouns, expressing happiness to see the user, use of close forms of address (user's name) and appropriate politeness strategies.

Figure 3. Laura and the MIT FitTrack System

The exercise-related portion of the daily dialogues that Laura had with users was based on a review of the health behavior change literature, input from a cognitive-behavioral therapist, and observational studies of interactions between exercise trainers and MIT students. These interventions were coupled with goal-setting and self-monitoring, whereby users would enter daily pedometer readings and estimates of time in physical activity, and were then provided with graphs plotting their progress over time relative to their goals.

In a randomized trial of the FitTrack system, 60 users interacted daily with Laura for a month on their home computers, with one group interacting with the fully "relational" Laura and the other group interacting with an identical agent that had all relationship-building behaviors disabled. Users who interacted with the relational Laura reported significantly higher scores on measures of relationship quality, liking of Laura, and desire to continue working with Laura, compared with users in the non-relational group, although no significant effects of relational behavior on exercise were found. Most users seemed to enjoy the relational aspects of the interaction (though there were definitely exceptions). As one user put it: "I like talking to Laura, especially those little conversations about school, weather, interests, etc. She's very caring. Toward the end, I found myself looking forward to these fresh chats that pop up every now and then. They make Laura so much more like a real person."

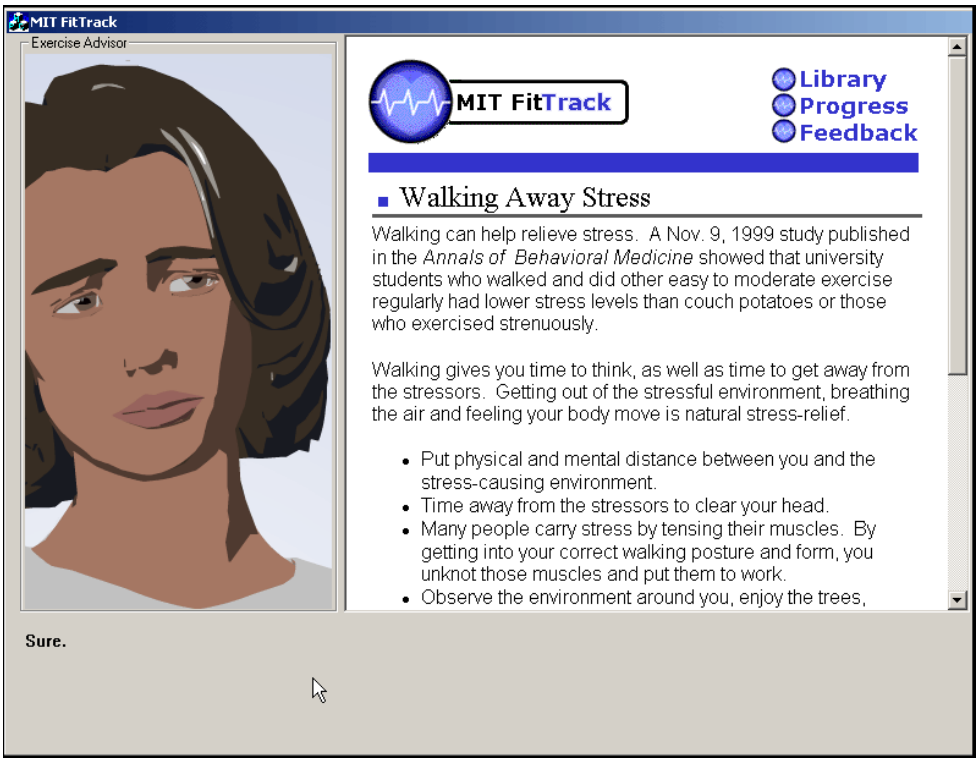

Figure 3. Laura and the MIT FitTrack System

References

1. Andre, E., Muller, J. and Rist, T., The PPP Persona: A Multipurpose Animated Presentation Agent. in *Advanced Visual Interfaces*, (1996).
2. Argyle, M. *Bodily Communication*. Methuen & Co. Ltd, New York, 1988.
3. Bickmore, T. Relational Agents: Effecting Change through Human-Computer Relationships *Media Arts & Sciences*, MIT, Cambridge, MA, 2003.
4. Cassell, J. and Bickmore, T. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intellient Agents. *User Modeling and Adaptive Interfaces, 13* (1-2). 89-132.
5. Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. (eds.). *Embodied Conversational Agents*. The MIT Press, Cambridge, MA, 2000.
6. Cassell, J., Vilhjálmsson, H. and Bickmore, T., BEAT: The Behavior Expression Animation Toolkit. in *SIGGRAPH '01*, (Los Angeles, CA, 2001), 477-486.
7. Lester, J., Towns, S., Callaway, C., Voerman, J. and Fitzgerald, P. Deictic and Emotive Communication in Animated Pedagogical Agents. in Cassell, J. ed. *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000.
8. Poggi, I. and Pelachaud, C. Performative Facial Expressions in Animated Faces. in Cassell, J., Sullivan, J., Prevost, S. and Churchill, E. eds. *Embodied Conversational Agents*, MIT Press, Cambridge, MA, 2000, 155-188.
9. Post, E. *Etiquette in Society, in Business, in Politics and at Home*. Funk and Wagnalls, New York, 1922.
10. Rickel, J. and Johnson, W.L. Animated Agents for Procedural Traning in Virtual Reality: Perception, Cognition and Motor Control. *Applied Artificial Intelligence*.
11. Sidner, C., Lee, C. and Lesh, N., Engagement Rules for Human-Computer Collaborative Interactions. in *IEEE International Conference on Systems, Man & Cybernetics*, (2003).
12. Whittaker, S. and O'Conaill, B. The Role of Vision in Face-to-Face and Mediated Communication. in Finn, K., Sellen, A. and Wilbur, S. eds. *Video-Mediated Communication*, Lawrence Erlbaum Associates, Inc., 1997, 23-49.