Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems

Justine Cassell
Media Laboratory
MIT
E15-315
20 Ames Street, Cambridge MA
justine@media.mit.edu

Matthew Stone
Department of Computer Science &
Center for Cognitive Science
Rutgers University
110 Frelinghuysen Road, Piscataway NJ 08854-8019
mdstone@cs.rutgers.edu

Abstract

In this paper we discuss the application of aspects of a psychological theory about the relationship between speech and gesture to the implementation of interactive dialogue systems. We first lay out some uncontroversial facts about the interaction of speech and gesture in conversation and describe some psychological theories put forth to explain those data, settling on one theory as being the most interesting for interactive dialogue systems. We then lay out our implementation of an interactive dialogue system that is informed by the theory—concentrating on two particular claims of the theory: that gesture and speech reflect a common conceptual source; and that the content and form of gesture is tuned to the communicative context and the actor's communicative intentions. We compare our work to some other similar interactive systems, and conclude with some thoughts about how both implementation and theory can benefit from this kind of close partnership.

Epigraph

Pantomime without discourse will leave you nearly tranquil, discourse without gestures will wring tears from you.
Essay on the Origin of Languages, Jean-Jacques Rousseau

Introduction

In this paper we discuss the application of aspects of a psychological theory about the relationship between speech and spontaneous speech-accompanying gesture to the implementation of interactive dialogue systems. We will concentrate on two particular claims of the theory. The first concerns the nature of the underlying representation that gives rise to the two channels of communication. The second concerns the function of gesture in communicative intent.

Why is it interesting, relevant, useful to apply such a psychological theory about speech and gesture to the development of interactive dialogue systems (that is, to systems that can carry on a conversation with a human user)? First, it makes sense to go looking for a theory that would explain how to integrate speech and gesture into interactive dialogue systems when we take a close look at what goes on in human-human dialogue. To be sure, we can speak on the telephone with one another and make ourselves understood perfectly well but, when we are face-to-face with another human, no matter what our language, cultural background, or age, we virtually all use our faces and hands as an integral part of our dialogue with others. So, if the metaphor of *computer as conversational partner* is to be taken seriously, then we might think that the computer should be imbued with the skills and behaviors that our human conversational partner has. This becomes increasingly important as interactive systems become more capable of true collaboration with their human users.

Second, a theory about the integration between verbal and nonverbal modalities may play a useful role when we reflect on the difficulties we have getting users to behave as they need to when interacting with pretty adequate spoken dialogue systems. Users repeat themselves needlessly, mistake when it is their turn to speak, and otherwise behave in ways that make dialogue systems less likely to function well (Oviatt 1995). It is in situations just like these in life that the nonverbal modalities come in to play: in noisy situations, humans increase their dependence on access to more than one modality (Rogers 1978).

Third, applying a theory of the relationship between speech and gesture seems relevant if we consider the benefits that might accrue to an interactive system that allows humans to communicate *naturally*. That is, while humans have long years of practicing communication with other humans (some might even say that the drive to practice the ability is innate (Trevarthen 1986)), communication with machines is learned. And yet, it has been shown that given the slightest chance, humans will attribute social responses, behaviors, and internal states to computers (Reeves and Nass 1996). If we can skillfully build on that social response to computers, channel it even into the kind of response that we give one another in human conversation, and build a system that gives back the response (verbal and nonverbal) that humans give, then we may evoke in humans the kinds of communicative dialogue behaviors they use with other humans, and thus allow them to use the computer with the same kind of efficiency and smoothness that characterizes their human dialogues. There is good reason to think that nonverbal behavior will play an important role in evoking these social communicative attributions. Our research (Cassell and Thórisson 1999) shows that humans are more likely to consider computers lifelike, and to rate their language skills

more highly, when those computers display not only speech but appropriate nonverbal communicative behavior.

Others have argued the contrary: that nonverbal behaviors may not play any useful role in collaborative systems. For example, Whittaker & O'Conaill (1997) tested whether video (video-conferencing) provided (a) cognitive cues that facilitate shared understanding; (b) process cues to support turn-taking, and (c) social cues and access to emotional information. Only the last kind of cue was found to be supported by video in communication. We believe that one factor contributing to their findings may be the fact that current implementations of video technology (even high quality video) have not been able to provide audio and video without significant time lags. This, of course, disrupts conversational process. We have found, similarly, that in embodied conversational agents (interactive dialogue systems embodied in graphical human figures) users attribute meaning to even small disruptions in timing between verbal and nonverbal behaviors. While results such as these underscore both the robustness of human-human dialogue and the stringent constraints that natural non-verbal behaviors respect, they remain consistent with our hypothesis that nonverbal cues are tightly linked to people's attribution of communicative intent in face-to-face interaction. In fact, because timing can be varied with graphical systems (to some extent), embodied conversational systems like the one presented below may provide a good testing ground for the role of these non-verbal behaviors as well as a fruitful context for their use.

In the next section we discuss some data about the relationship between speech and gesture, and lay out several theories to account for the data, finally settling on one theory that we have applied to the implementation of an interactive system.

A Theory of the Relationship between Speech and Gesture

Evidence from many sources suggests a close relationship between speech and spontaneous hand gestures during conversation. At the prosodic level, Kendon (1974) found that the stroke phase (the most effortful part) of speech-accompanying gestures tends to co-occur with or just before the phonologically most prominent syllable of the accompanying speech. Other evidence comes from the sheer frequency of gestures during speech. About three-quarters of all clauses in narrative discourse are accompanied by gestures of one kind or another (McNeill 1992), and, perhaps surprisingly, although the proportion of gesture types may change, all of these gesture types, and spontaneous gesturing in general, appear to be found in discourses by speakers of most languages.

Of course, communication is still possible without gesture. Information appears to be just about as effectively communicated in the absence of gesture—on the telephone, or from behind a screen (Short *et al.* 1976; Williams 1977). But it has been shown that when speech is ambiguous (Thompson and Massaro 1986) or in a speech situation with some noise (Rogers 1978), listeners do rely on gestural cues (and, the higher the noise-to-signal ratio, the more facilitation by gesture). And, gestures are still produced in situations where there is no listener, or the listener cannot see the speaker's hands (Rimé 1982), although more gestures may be produced when an addressee is present (Cohen 1977; Cohen and Harrison 1973). In addition, we know that when adults are asked to assess a child's knowledge, they use information that is conveyed in the child's gesture (and that is not the same as that conveyed by the child's speech) to make that assessment (Goldin-Meadow *et al.* 1992; Alibali *et al.* 1994). Similarly, Cassell et al. (1998) established that listeners rely on information conveyed only in gesture as they try to comprehend a story.

Most interesting in terms of building interactive dialogue systems is the semantic and pragmatic relationship between gesture and speech. The two channels do not always manifest the same information about an idea, but what they convey is virtually always compatible, both *semantically*, in that speech and gesture give a consistent view of an overall meaning to be conveyed, and *pragmatically*, in that speech and gesture mark information about this meaning as advancing the purposes of the conversation in a consistent way. For example, gesture may depict the way in which an action was carried out when this aspect of meaning is not depicted in speech. And even when the depiction of gesture overlaps with that of speech (a situation which we have found to be the case in roughly 50focused pragmatically by mechanisms like prosody in speech. For example, gesture may co-occur with lexical items that are more difficult for listeners to predict. It has been suggested (Kendon 1994) that those concepts difficult to express in language may be conveyed by gesture. Thus simultaneity of two events, or the respective locations of two objects may be expressed by the position of the two hands. The semantic and pragmatic compatibility seen in the gesture-speech relationship recalls the interaction of words and graphics in multimodal presentations (Feiner and McKeown 1991; Green *et al.* 1998; Wahlster *et al.* 1991). In storytelling, underlying narrative structure may be indexed by differential use of gesture: iconic gestures tend to occur with plot-advancing description of the action, deictic gestures with the introduction of new characters, and beat gestures at the boundaries of episodes (Cassell and McNeill 1991).

In thinking about the cause of this close relationship, several theories have been advanced. Some claim that gesture is a late occurring process in the production of language, either a translation of speech, or an outcome of the search for particular lexical items (Butterworth and Beattie 1978; Butterworth and Hadar 1989), hence the production of so-called word-finding gestures during pauses. These researchers believe that gesture is not integral to communicative intent because, so they claim, the production of gesture is epiphenomenal to the production of speech.

Some claim that the primary function of gesture is not to communicate to the listener but to support the speaker's encoding of information (Freedman 1972; Rimé 1982; Krauss *et al.* 1991)—a kind of continuation of the sensorimotor stage of knowing that Piaget described for infants. This would be why gesture is produced when, for example, speaking on the telephone.

Some researchers, however, reject the notion that either gesture or speech might be primary (Kendon 1972; McNeill 1992). According to McNeill, gesture and speech arise together from an underlying representation that has both visual and linguistic aspects, and so the relationship between gesture and speech is essential to the production of meaning and to its comprehension. This is said to explain why we find the strict temporal synchronization between the production of gesture and speech (Kendon 1972), the parallel semantic and pragmatic content of gesture and speech (McNeill 1992), the simultaneous acquisition in children (Riseborough 1982) and tendency for the two systems to break down in parallel ways in aphasics (Feyereisen 1983; Pedelty 1987). Those researchers who claim a tight coupling of speech and gesture at the earliest stages of production have tended to eschew linear encoding and information processing models where a pre-verbal message is passed on to the language production module. De Ruiter (de Ruiter to appear), however, gives us an instance where an IP model of this sort can help us evaluate the theory.

Two key aspects of this last theory that are maintainable in a computationally implementable model are the claim that gesture and speech arise from a common conceptual source, and the claim that gesture plays an intrinsic role in communicative intent. In what follows we discuss the demands that are placed on our interactive dialogue system as a function of these two claims. First, in the system that we discuss, one single underlying conceptual source must serve as the representation that, in the dialogue generation engine, gives rise to the form of both speech and gesture. Second, communicative intent must be specified in such a way that gesture and speech can both be clearly said to advance it.

Previous Interactive Systems that Use Speech and Gesture

Other systems have made attempts to integrate speech and spontaneous gesture, with more or less appeal to any theoretical underpinnings about the relationship between the two media.

The *interpretation* of speech and gesture has been the object of investigation since the pioneering work of (Bolt 1980) on deictic gesture; recent work includes (Koons *et al.* 1993; Bolt and Herranz 1992; Sparrell 1993). In all of these systems, interpretation is not carried out until the user has finished the utterance, and speech drives the analysis of the gestures. In studying the *generation* of gesture in artificial agents, researchers have primarily addressed nonverbal behaviors that stand on their own without accompanying speech, such as American Sign Language (Loomis *et al.* 1983; Lee and Kunii 1993), emblematic behaviors (Chen *et al.* 1993; Kurlander *et al.* 1996), and other believable postures and motions (Rijpkema and Girard 1991; Perlin and Goldberg 1996). Some researchers have generated gestures in conjunction with verbal behavior. Lester et al. (1998) generate deictic gestures and choose referring expressions as a function of the potential ambiguity of objects referred to, and the proximity of those objects to the animated agent. However, the generation of gestures and the choice of referring expressions are accomplished in two entirely independent (additive) processes. Similarly, Rickel and Johnson (1999) have their pedagogical agent move to objects in the virtual world and then generate a deictic gesture at the beginning of an explanation about that object. This system, then, does not deal with the issue of how to allocate communicative intentions across modalities. Andre et al. (1999) generate pointing gestures as a sub-action of the rhetorical action of labeling, in turn a sub-action of the action of elaborating.

Missing from all of these understanding and generation systems is a representation of communicative intent that treats the different modalities on a par with one another. Such representations have been explored in research on combining linguistic and *graphical* interaction. For example, multimodal *managers* have been described to allocate an underlying content representation for generation of text and graphics (Wahlster *et al.* 1991; Green *et al.* 1998). Meanwhile, (Johnston *et al.* 1997; Johnston 1998) describe a formalism for tightly-coupled interpretation which uses a grammar and semantic constraints to analyze input from speech and pen. But spontaneous gesture requires a distinct analysis. For example, we need some notion of discourse pragmatics or conversational structure (a notion of *speaking turn* or *shared information* for instance) that would allow us to predict where gesture occurs with respect to speech, and what its role might be. Likewise, we need a distinct account of the *communicative effect* of spontaneous gesture that might allow us to allocate certain communicative goals to speech and certain other ones to gesture. In the absence of such an analysis, the role of gesture cannot be analyzed at more than a sentence-constituent-replacement level, or in general terms. The result is that the spontaneous gesture recognition systems can only understand a gesture if its meaning fills a syntactic slot (such as reference resolution, in the case of deictic gestures and pronominal references). In the extant generation systems, gestures can only be generated (a) if a library of gestures and their distribution can be specified beforehand, (b) if the role of gesture is simply to enhance believability of the agent, rather than convey meaning, or (c) if the role of gesture is always additive and redundant.

In what follows we depend on just such notions of discourse structure and communicative effect to provide a reason to include gesture, as well as a way to implement its generation.

Application of the Theory

Thus far, we have laid out two psychological claims about gesture in natural dialogue. First, gesture and speech reflect a common conceptual source. Second, the content and form of gesture is tuned to the communicative context and the actor's communicative intentions. In this section, we describe our implemented dialogue agent, REA, and show that these claims describe the process that REA uses to construct communicative acts.

REA ("Real Estate Agent") is a computer-generated humanoid that has an articulated graphical body, can sense the user passively through cameras and audio input, and is capable of speech with intonation, facial display, and gestural output. REA's domain of expertise is real estate—she acts

as a real estate agent showing users the features of various models of houses that appear on-screen behind her. REA is designed to conduct a mixed-initiative conversation, pursuing the goal of describing the features of a house that fits the user's requirements, and the features of a house that might be considered generally attractive, while also responding to the user's verbal and non-verbal input that may lead in new directions. (Cassell *et al.* 1999) provides an extended overview of REA's design, implementation and capabilities.

REA (like her predecessors, Gandalf (Cassell and Thórisson 1999) and GestureJack (Cassell *et al.* 1994)) is the platform for a large-scale research program addressing many different aspects of the relationship between natural verbal and non-verbal behavior (in most detail, the relationship between speech and gesture) in order to implement this relationship in interactive systems. This research program can be seen as an attempt to understand and model the answers to a set of increasingly specific questions about the role played by non-verbal behavior in natural discourse and dialogue:

- What functions in discourse and conversation are played by the verbal and non-verbal modalities (Cassell in press; Cassell *et al.* 1999)?
- What theoretical approach to discourse and conversation allows us to specify the role played by both the verbal and non-verbal modalities (Cassell *et al.* in press; 1999)?
- How do we *evaluate* under what conditions generation of verbal and non-verbal behavior is useful to interactive systems (Cassell and Thórisson 1999)?
- Where does gesture occur in the discourse with respect to discourse structure (Cassell *et al.* 1994)?
- Where does gesture occur in the discourse with respect to the temporal or surface structure of the utterance (Cassell *et al.* in prep)?
- Where does gesture occur in the discourse with respect to semantic structure (that is, which semantic features or kinds of information tend to be conveyed by gesture and which by speech)?
- When do gestures and speech convey the same features (redundantly) and when do they convey different features (complementarity)?
- How do the hands convey the information that they do (what is the morphology of gesture)?

As can be seen by the citations listed above, in other work we have begun to address the more general questions. Here, we will focus in more detail on the specific questions about semantic structure and complementarity—those aspects that play a role in the generation of the form of REA's speech and gesture. We first explain how REA's dialogue manager combines static and dynamic information to create the overall conceptual representation from which both speech and gesture are derived. A contribution to dialogue is then assembled from this representation by repeatedly selecting and combining meaningful actions, using the SPUD generator (Stone and Doran 1997; Stone and Webber 1998). We continue by showing how the assembly process treats words and gestures on a par—allowing both to access components of the underlying conceptual representation, to convey information to the hearer, and to respond to the state of the dialogue.

Dialogue management and representation

In REA, requests for the generation of speech and gesture are formulated by a broad module for dialogue management. This module is charged with three general functions.

- It coordinates the interaction between the system and the user. This task involves such processes as taking and yielding speaking turns, and participating in rituals of interaction such as greetings and farewells. REA's model of these processes depends on an explicit distinction between interactional and propositional functions of communicative actions.
- It keeps a record of the ongoing state of the discourse. This discourse model maintains the entities that have been evoked in discourse and their salience in a model of attention. The discourse model also records the communicative acts that have taken place, so as to keep track of their effects on the interaction and to keep track of any propositional content that those acts may have conveyed. REA's present incarnation does not represent possible discrepancies between this model and the user's model of the conversation: the representation of the discourse is idealized as a shared construct that REA and the user agree on. This suffices for the limited flexibility of REA's dialogue but would have to be elaborated in order to address such phenomena as the detection and repair of misunderstanding (McRoy and Hirst 1995).
- It interprets the user's utterances and formulates communicative goals to respond to those utterances. The current system is predominantly reactive. The user's words are interpreted as specifying one of a small set of directives that REA understands. A first layer of production rules construct *obligations* that respond to these directives (Traum and Allen 1994); a second layer devises goals for communicative action to meet these obligations.

Each of REA's utterances represents a coordination of these three kinds of processing in the dialogue manager. In particular, the system recognizes that it is time to speak, formulates the appropriate set of communicative goals and communicative context for the SPUD generator, triggers the generation process, and realizes the resulting speech and gesture.

The breadth of the dialogue manager makes for a wealth of information about the domain and the conversation that the dialogue manager can provide to the generator. To start with, there is SPUD's background knowledge. SPUD is initialized with a structured body of knowledge about the domain that it can draw on for communicative content. This knowledge base is made up of facts explicitly labeled with the *kind of information* they represent.[1] *Both* speech and gesture access

[1]Our language for distinguishing kinds of information is first-order multi-modal logic. The formula $[i]p$—meaning that p is necessary according to $[i]$, or that p is information of kind $[i]$—appeals to an analogy between $[i]$ as a kind of information and the knowl-

the *whole* structured database; SPUD's kinds of information do *not* include a kind of information that gesture is drawn from and a kind of information that speech is drawn from. The organization of the knowledge base instead serves two purposes relevant to all communication modalities:

- It defines the common ground, in terms of the sources of information that speaker and hearer *share*. A fact is part of the common ground if the fact is a consequence of information from shared sources. This idea is of course inspired by Clark and Marshall's psychological model of *co-presence*, according to which people keep track of common ground using heuristics that identify mutually available bodies of information (Clark and Marshall 1981).
- It describes the relationship between the system's *private* information and the questions of interest that that information can be used to settle. For example, the variety of facts that bear on where a residence is—which city, which neighborhood, which position in a block or, if appropriate, where in a building—all provide the same kind of information. Likewise, a range of facts describing light, space, and decor are grouped together as providing information that bears broadly on what sort of environment the residence offers. Thus, REA's requests for utterances ask SPUD to present specified kinds of information about a specified object.

In addition to this background knowledge and specification of communicative goals, SPUD gets updates from REA's dialogue manager to keep on top of the changing state of conversation. The dialogue manager informs SPUD of the new facts that are added to the common ground as the dialogue proceeds. Once such a fact becomes shared, SPUD keeps it in the common ground indefinitely. (Note that this assumption adds an idealization of the user's memory to the idealization of understanding and agreement in the dialogue which we have already mentioned.)

The dialogue manager also provides SPUD with a characterization of the current attentional state of the discourse, which varies from utterance to utterance. Although relatively wide-ranging, this characterization is doubtless incomplete. We have focused only on those factors that license marked forms in the grammar of speech and gesture that we are constructing, and we have simplified the representation and dynamics of these factors where REA's simple dialogue strategies permit. We currently treat:

- *Attentional prominence*, represented (as usual in natural language generation) by setting up a *context set* for each entity (Dale 1992). The context set for an entity gives an (idealized) shared perspective on what entities are as salient (and as likely to be referenced) as it is. Our model of attention is a simple local one similar to (Strube 1998).
- *Cognitive status*, including whether an entity is *hearer-old* or *hearer-new* (Prince 1992), and whether an entity is *in-focus* or not (Gundel *et al*. 1993). We can assume that houses and their rooms are *hearer-new* until REA describes them; and that just those entities mentioned in the prior sentence are *in-focus*.
- *Information structure*, including the *open propositions* or *themes* which describe the salient questions currently at issue in the discourse (Prince 1986; Steedman 1991). In REA's dialogue, open questions are always general questions about some entity raised by a recent turn; although in principle such an open question ought to be formalized as $theme(\lambda P.Pe)$, REA can use the simpler $theme(e)$.

As with domain knowledge, the specification of the dialogue state crosscuts distinctions of communicative modality; as we shall see, both speech and gesture depend on the same kinds of contextual factors, and access those factors in the same way. Thus, despite the variety of information it contains, the generator's input represents a single overall conceptual representation; its content is relevant to the choice of communicative actions regardless of their realization in speech and gesture.

Generation

The utterance generation problem in REA, then, is to construct a communicative action that achieves given goals—to convey domain propositions that encode specified kinds of information about a specified object—and that fits the context specified by the dialogue manager, to the best extent possible. We assume that the communicative action is composed of a collection of atomic elements, including both lexical items in speech and constraints on movement in gesture; since we assume that any such item usually conveys a specific piece of content, we refer to these elements generally as lexicalized descriptors. The generation task in REA thus involves selecting a number of such lexicalized descriptors and organizing them into a grammatical whole that manifests the right syntactic relations within speech and the right synchrony between speech and gesture. The information conveyed by them must be enough that the hearer can identify the entity in each domain reference from among its *context set*. Moreover, the descriptors must provide a source which allows the hearer to recover any needed new domain proposition, either explicitly or by inference.

We use the SPUD generator ("Sentence Planning Using Description") introduced in (Stone and Doran 1997) to carry out this task for REA. SPUD builds the utterance element-by-element; at each stage of construction, SPUD's representation of the current, incomplete utterance specifies its syntax, semantics, interpretation and fit to context. This representation allows SPUD to determine which lexicalized descriptors are available at each stage to extend the utterance; it also allows SPUD to assess the progress towards its communicative goals which each extension would bring about. At each stage, then, SPUD selects the available option that offers the best immediate advance toward completing the utterance successfully.

As part of the development of REA, we have constructed a new inventory of lexicalized descriptors for SPUD to draw on in REA's utterances. (Previous experiments with SPUD have describe how SPUD can be used to generate contextually-appropriate syntactic variation (Stone and Doran 1997),

edge that would be available to an agent *i* with a limited view of a situation. (Stone 1998)

to generate concise referring expressions within sentences (Stone and Webber 1998), and to realize optional modifiers flexibly (Bourne 1998). However, all this research has been aimed at generating text.) The new inventory includes entries that contribute to gestures that can accompany speech as well as revised entries for spoken words that describe their possible synchrony with gesture. The organization of these entries assures that—using the same mechanism as with speech—REA's gestures draw on the single available conceptual representation and that REA's gesture varies as a function of pragmatic context in the same way as natural gestures do. To explain how these entries assure this, we need to consider SPUD's representation of lexicalized descriptors in more detail.

Each entry is specified in three parts. The first part—the *syntax* of the element—sets out what words or other actions the element contributes to an utterance that contains it. The syntax is a hierarchical structure, formalized using Feature-Based Lexicalized Tree Adjoining Grammar (LTAG) (Joshi *et al.* 1975; Schabes 1990). Syntactic structures are also associated with referential indices that specify the entities in the discourse that the entry refers to. For the entry to apply at a particular stage, its syntactic structure must combine by LTAG operations with the syntax of the ongoing utterance.

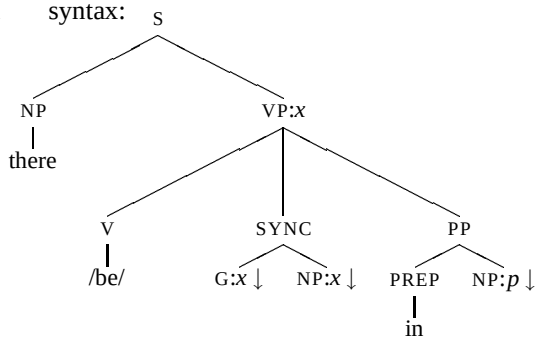
REA's syntactic entries combine typical phrase-structure analyses of linguistic constructions with annotations that describe the occurrence of gestures in synchrony with linguistic phrases (following rules described further in (Cassell *et al.* 1994)). Note that we regard components of gesture as *constraining* an overall movement. The hierarchical description of a gesture indicates the *choices* the generator must make to produce a gesture, but does not analyze a gesture literally as a hierarchy of separate movements.

The second part—the *semantics* of the element—is a formula that specifies the content that the element carries. Before the entry can be used, SPUD must establish that the semantics holds of the entities that the entry describes. If the semantics already follows from the common ground, SPUD assumes that the hearer can use it to help identify the entities described. If the semantics is merely part of the system's private knowledge, SPUD treats it as new information for the hearer.

Finally, the third part—the *pragmatics* of the element—is also a formula that SPUD looks to prove before the entry can be used. Unlike the semantics, however, the pragmatics does not achieve specific communicative goals like identifying referents or conveying new information to the hearer. Instead, the pragmatics establishes a general fit between the entry and the context. One representative use of pragmatic conditions is to test for the appropriate *cognitive status* for a referent before a special referring form is used; for example a pronoun can only be used when its referent is *in focus* (Gundel *et al.* 1993). Another is to test for appropriate *open propositions* before the use of a marked syntactic form (Prince 1986; Ward 1985).

The entry schematized in (1) illustrates these three components; the entry also suggests how these components can define synchronized actions of speech and gesture that respond coherently to the context.

(1) a syntax:



b semantics: $in(x, p)$

c pragmatics: $\textit{hearer-new}(x) \wedge \textit{theme}(p)$

(1) describes the use of a presentational *there*-sentence to introduce a new object to the discourse while relating the entity to a location it's *in*. The object, indicated throughout the entry by the variable *x*, is realized as the complement NP of the verb *be*. The other complement of *be* is a PP headed by the lexical item *in*. The location where *x* is, indicated throughout the entry by the variable *p*, is realized as the NP complement of *in*. The new object *x* can also form the basis of a gesture G synchronized with the noun phrase (as indicated by the SYNC constituent). The entry asserts quite simply that *x* is located in *p*. However, the entry carries two pragmatic requirements: in keeping with the presentational function of the construction, *x* must be new to the hearer; moreover, in keeping with the emphasis given to *x* by the simultaneous gesture, the object *p* must provide the theme of the utterance (so that *x* gives the information-structure *rheme*).

(1) is the entry that figures in utterance (2), produced as a response to the directive *tell me about the kitchen*:

(2) There's [a chimney] in it. (right hand is cupped cylindrically, touching left hand which is held out flat)

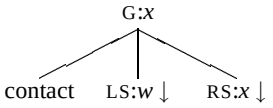
The utterance indicates not only the presence of the chimney in the kitchen but its site against a wall there. We assume that these two pieces of information respond to the explicit facts that SPUD is to communicate in describing the kitchen—facts that SPUD believes to be relevant to the listener's requirements, or generally attractive facts about a home. This latter condition is a stand-in for an analysis that we believe to be necessary of how gestures not only contribute to communicative intent, but also represent visually salient aspects of a scene being described.

The pragmatic conditions of (1) encode an important aspect of our theory of the discourse function of gesture and speech. The same preposition *in* might be appear with the same meaning in constructions that address other open questions in the discourse. This is true even for presentational sentences—imagine *where is there a chimney?* The differences are reflected both in the placement of associated gestures and in the concepts that those gestures realize. Gesture tends to synchronize with and elaborate on the answering information (the *rheme*) rather than the *theme*. In SPUD we define such alternatives by introducing separate entries with different pragmatic requirements—and different syntactic frames, to encode the gesture and intonation that would

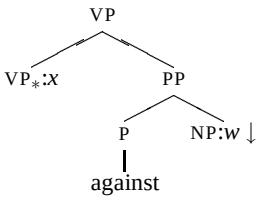
naturally be associated with them.

While (1) explains how we fit gesture and speech tightly to the context and to one another, it is alternative entries like (3a) and (3b)—two entries that both convey (3c) and that both could combine with (1) by LTAG operations—that underlie our claim that our implementation allows gesture and speech to draw on a single conceptual source and fulfill similar communicative intentions.

(3) a syntax:



b syntax:



c semantics: $against(x, w)$

(3a) provides a structure that could substitute for the G node in (1) to produce the synchronized speech and gesture of (2). The hierarchical structure of (3a) indicates that, in this gesture, the hands are to be positioned in contact in space, and that further decisions are required to determine the right handshape (node RS, as a function of the entity *x* that the gesture describes) and the left handshape (node LS, as a function of the reference entity *w*). We pair (3a) with the semantics in (3c), and thereby model that the gesture indicates that the chimney is against something.

Similarly, (3b) describes how we could modify the VP introduced by (1) (using the LTAG operation of *adjunction*), to produce an utterance such as *There's a chimney in it, against the wall.* By pairing (3b) with the same semantics (3c), we ensure that SPUD will treat the communicative contribution of the alternative constructions of (3) in a parallel fashion. Both are triggered by accessing background knowledge about the kitchen and both are recognized as *directly* communicating one of the specified facts needed to complete the description of the kitchen.

Other entries involved in (2) include the word *chimney*, which provides the noun phrase that evokes the entity *x*, and a *cupped handshape* constraint, which relates the right hand to the entity *x* that the hand portrays. These entries too involve different modalities, yet have a parallel structure to each other—and to (3)—and a parallel algorithmic treatment. SPUD again must access a single body of knowledge about the domain and the discourse state to determine whether the two entries are applicable. Moreover, both *cupped handshape* and *chimney* appear in the sentence through an *indirect* relationship to SPUD's explicit communicative goals. They respond to the choices (of NP and RS) required to complete the realization of constituents where SPUD's communicative goals can be fulfilled.

Conclusion

Research on the robustness of human conversation suggests that a dialogue agent capable of acting as a conversational partner would provide for efficient and natural collaborative dialogue. But human conversational partners display gestures that derive from the same underlying conceptual source as their speech, and which relate appropriately to their communicative intent. In this paper, we have summarized the evidence for this view of human conversation, and shown how we are using it to inform the design of our artificial conversational agent, REA. While REA is constantly evolving, REA has a working implementation, which includes the modules described in this paper, and can engage in interactions including that in (2) and many others.

The tight connection between theory and implementation has already both streamlined the design of the system and exposed gaps in the theory. The theoretical claim that gesture and speech reflect a common semantic representation and common dimensions of discourse function helps us leverage previous research in dialogue and computational linguistics to gesture. For example, it means that we can more easily maintain overall representations of discourse state and communicative action in REA's dialogue manager, and can more easily adapt natural language generation algorithms to produce REA's embodied communicative behavior.

Conversely, REA grows out of the theoretical questions raised by our previous research (Cassell *et al.* 1994): it was only in watching the final animation that we realized that too many nonverbal behaviors were being generated—the impression was of a bank teller talking to a foreigner and trying to enhance his speech with supplementary nonverbal cues. This problem arose because each nonverbal behavior was generated independently on the basis of its association with discourse and turn-taking structure and timed by intonation, but without reference to the other nonverbal phenomena present in the same clause. Our conclusion was that we lacked a function in our system: a multimodal *manager* that distributes meaning across the modalities but that is essentially modality-independent in its functioning. The SPUD generation system fills this function in the current implementation of REA.

Additionally, in this most recent implementation, we have discovered that we must have a way of predicting what gestural forms will convey what gestural meanings (what *handshapes and trajectories* of the hands will convey the meaning features that SPUD specifies). Achieving this precision in the theory, and in the corresponding implementation, is an important problem for future work.

Acknowledgments

The research reported here was supported by the National Science Foundation (award IIS-9618939), Deutsche Telekom, AT&T, and the other generous sponsors of the MIT Media Lab, and a postdoctoral fellowship from RUCCS. Hao Yan and Hannes Vilhjálmsson accomplished the implementation of REA's generation and discourse management modules. We thank Nancy Green, James Lester, Jeff Rickel, Candy Sidner, and two anonymous reviewers for comments that improved the paper.

References

M. W. Alibali, L. Flevares, and S. Goldin-Meadow. Going beyond what children say to assess their knowledge. Manuscript, University of Chicago, 1994.

Elisabeth André, Thomas Rist, and Jochen Müller. Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence*, 13:415–448, 1999.

R. A. Bolt and E. Herranz. Two-handed gesture in multi-modal natural dialog. In *UIST 92: Fifth Annual Symposium on User Interface Software and Technology*, 1992.

R. A. Bolt. Put-that-there: voice and gesture at the graphics interface. *Computer Graphics*, 14(3):262–270, 1980.

Juliet Bourne. Generating effective instructions: Knowing when to stop. PhD Thesis Proposal, Department of Computer & Information Science, University of Pennsylvania, July 1998.

B. Butterworth and G. Beattie. Gesture and silence as indicators of planning in speech. In R. N. Campbell and P. T. Smith, editors, *Recent Advances in the Psychology of Language: Formal and Experimental Approaches*. Olenum Press, New York, 1978.

B. Butterworth and U. Hadar. Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review*, 96:168–174, 1989.

J. Cassell and D. McNeill. Gesture and the poetics of prose. *Poetics Today*, 12(3):375–404, 1991.

J. Cassell and K. Thórisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(3), 1999.

Justine Cassell, Catherine Pelachaud, Norm Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *SIGGRAPH*, pages 413–420, 1994.

J. Cassell, D. McNeill, and K. E. McCullough. Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition*, 6(2), 1998.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. Embodiment in conversational characters: Rea. In *CHI 99*, 1999.

Justine Cassell, Scott Prevost, and Mark Steedman. Embodied generation: A framework for generating intonation, speech and gesture. manuscript, in prep.

Justine Cassell, Obed Torres, and Scott Prevost. Turn taking vs. discourse structure: how best to model multimodal conversation. In Yorick Wilks, editor, *Machine Conversations*. Kluwer, in press.

Justine Cassell. Embodied conversation: Integrating face and gesture into automatic spoken dialogue systems. In Susan Luperfoy, editor, *Spoken Dialogue Systems*. MIT Press, Cambridge, MA, in press.

D. Chen, S. Pieper, S. Singh, J. Rosen, and D. Zeltzer. The virtual sailor: in implementation of interactive human body modeling. In *VRAIS 93: 1993 Virtual Reality Annual International Symposium*, 1993.

Herbert H. Clark and Catherine R. Marshall. Definite reference and mutual knowledge. In Aravind K. Joshi, Bonnie Lynn Webber, and Ivan Sag, editors, *Elements of Discourse Understanding*, pages 10–63. Cambridge University Press, Cambridge, 1981.

A. A. Cohen and R. P. Harrison. Intentionality in the use of hand illustrators in face-to-face communication situations. *Journal of Personality and Social Psychology*, 28:276–279, 1973.

A. A. Cohen. The communicative functions of hand illustrators. *Journal of Communication*, 27(4):54–63, 1977.

Robert Dale. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes*. MIT Press, Cambridge MA, 1992.

Jan Peter de Ruiter. The production of speech and gesture. In David McNeill, editor, *Language and Gesture: Window into Thought and Action*. Cambridge University Press, Cambridge, UK, to appear.

S. Feiner and K. McKeown. Automating the generation of coordinated multimedia explanations. *IEEE Computer*, 24(10):33–41, 1991.

P. Feyereisen. Manual activity during speaking in aphasic subjects. *International Journal of Psychology*, 18:545–556, 1983.

N. Freedman. The analysis of movement behavior during the clinical interview. In A. Siegman and B. Pope, editors, *Studies in Dyadic Communication*. Pergamon, New York, 1972.

S. Goldin-Meadow, D. Wein, and C. Chang. Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, 9(3):201–219, 1992.

Nancy Green, Giuseppe Carenini, Spehan Kerpedjiev, Steven Roth, and Johanna Moore. A media-independent content language for integrated text and graphics generation. In *CVIR '98 – Workshop on Content Visualization and Intermedia Representations*, 1998.

Jeanette K. Gundel, Nancy Hedberg, and Ron Zacharski. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2):274–307, 1993.

M. Johnston, P. R. Cohen, D. McGee, J. Pittman, S. L. Oviatt, and I. Smith. Unification-based multimodal integration. In *ACL/EACL 97: Proceedings of the Annual Meeting of the Assocation for Computational Linguistics*, 1997.

Michael Johnston. Unification-based multimodal parsing. In *COLING/ACL*, 1998.

Aravind K. Joshi, L. Levy, and M. Takahashi. Tree adjunct grammars. *Journal of the Computer and System Sciences*, 10:136–163, 1975.

A. Kendon. Some relationships between body motion and speech. In A. W. Siegman and B. Pope, editors, *Studies in Dyadic Communication*. Pergamon, New York, 1972.

A. Kendon. Movement coordination in social interaction: somem examples described. In S. Weitz, editor, *Nonverbal Communication*. Oxford, New York, 1974.

A. Kendon. Do gestures communicate? A review. *Research on Language and Social Interaction*, 27(3):175–200, 1994.

D. B. Koons, C. J. Sparrell, and K. R. Thórisson. Integrating simultaneous input from speech, gaze and hand gestures. In M. T. Maybury, editor, *Intelligent Multi-media Interfaces*. MIT Press, Cambridge, 1993.

R. Krauss, P. Morrel-Samuels, and C. Colasante. Do conversational hand gestures communicate? *Journal of Personality and Social Psychology*, 61(5):743–754, 1991.

D. Kurlander, T. Skelly, and D. Salesin. Comic chat. In *Proceedings of SIGGRAPH 96, Computer Graphics Proceedings, Annual Conference Series*, 1996.

J. Lee and T. Kunii. Visual translation: from native language to sign language. In *Proceedings of IEEE Workshop on Visual Languages*, 1993.

James Lester, Stuart Towns, Charles Calloway, and Patrick FitzGerald. Deictic and emotive communication in animated ped-

agogical agents. In *Workshop on Embodied Conversational Characters*, 1998.

J. Loomis, H. Poizner, U. Bellugi, A. Blakemore, and J. Hollerbach. Computer graphic modeling of American Sign Language. *Computer Graphics*, 17(3):105–114, 1983.

David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, 1992.

Susan W. McRoy and Graeme Hirst. The repair of speech act misunderstandings by abductive inference. *Computational Linguistics*, 21(4):435–478, 1995.

S. L. Oviatt. Predicting spoken language disfluencies during human-computer interaction. *Computer Speech and Language*, 9(1):19–35, 1995.

L. L. Pedelty. *Gesture in Aphasia*. PhD thesis, University of Chicago, 1987. Department of Behavioral Sciences.

K. Perlin and A. Goldberg. Improv: a system for interactive actors in virtual worlds. In *Proceedings of SIGGRAPH 96, Computer Graphics Proceedings, Annual Conference Series*, pages 205–216, 1996.

Ellen Prince. On the syntactic marking of presupposed open propositions. In *Proceedings of the 22nd Annual Meeting of the Chicago Linguistic Society*, pages 208–222, Chicago, 1986. CLS.

Ellen F. Prince. The ZPG letter: Subjects, definiteness and information status. In William C. Mann and Sandra A. Thompson, editors, *Discourse Description: Diverse Analyses of a Fund-raising Text*, pages 295–325. John Benjamins, Philadelphia, 1992.

B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television and New Media like Real People and Places*. Cambridge University Press, Cambridge, 1996.

Jeff Rickel and W. Lewis Johnson. Animated agents for procedural training in virtual reality: Perception, cognition and motor control. *Applied Artificial Intelligence*, 13:343–382, 1999.

H. Rijpkema and M. Girard. Computer animation of hands and grasping. *Computer Graphics*, 25(4):339–348, 1991.

B. Rimé. The elimination of visible behavior from social interactions: effects of verbal, nonverbal and interpersonal variables. *European Journal of Social Psychology*, 12:113–129, 1982.

M. G. Riseborough. Meaning in movement: An investigation into the interrelationship of physiographic gestures and speech in seven-year-olds. *British Journal of Psychology*, 73:497–503, 1982.

W. T. Rogers. The contribution of kinesic illustrators towards the comprehension of verbal behavior within utterances. *Human Communication Research*, 5:54–62, 1978.

Yves Schabes. *Mathematical and Computational Aspects of Lexicalized Grammars*. PhD thesis, Computer Science Department, University of Pennsylvania, 1990.

J. Short, E. Williams, and B. Christie. *The Social Psychology of Telecommunications*. Wiley, New York, 1976.

C. J. Sparrell. Coverbal iconic gesture in human-computer interaction. Master's Thesis, MIT, 1993.

Mark Steedman. Structure and intonation. *Language*, 67:260–296, 1991.

Matthew Stone and Christine Doran. Sentence planning as description using tree-adjoining grammar. In *Proceedings of ACL*, pages 198–205, 1997.

Matthew Stone and Bonnie Webber. Textual economy through close coupling of syntax and semantics. In *Proceedings of INLG*, pages 178–187, 1998.

Matthew Stone. *Modality in Dialogue: Planning, Pragmatics and Computation*. PhD thesis, University of Pennsylvania, 1998.

Michael Strube. Never look back: An alternative to centering. In *Proceedings of COLING-ACL*, 1998.

L. A. Thompson and D. W. Massaro. Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology*, 42:144–168, 1986.

David R. Traum and James F. Allen. Discourse obligations in dialogue processing. In *ACL*, pages 1–8, 1994.

C. Trevarthen. Sharing makes sense: intersubjectivity and the making of an infant's meaning. In R. Steele and T. Threadgold, editors, *Language Topics: Essays in Honour of M. Halliday*, volume 1, pages 177–200. J. Benjamins, Amsterdam, 1986.

W. Wahlster, E. André, W. Graf, and T. Rist. Designing illustrated texts. In *Proceedings of EACL*, pages 8–14, 1991.

Gregory Ward. *The Semantics and Pragmatics of Preposing*. PhD thesis, University of Pennsylvania, 1985. Published 1988 by Garland.

S. Whittaker and B. O'Conaill. The role of vision in face-to-face and mediated communication. In A. J. Sellen K. E. Finn and S. B. Wilbur, editors, *Video-Mediated Communication*, pages 23–49. Lawrence Erlbaum, Hillsdale, NJ, 1997.

E. Williams. Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin*, 84:963–976, 1977.