Coordination and context-dependence in the generation of embodied conversation

Justine Cassell* **Matthew Stone†** **Hao Yan***

*Media Laboratory
MIT
E15-315
20 Ames, Cambridge MA
{justine,yanhao}@media.mit.edu

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

Abstract

We describe the generation of communicative actions in an implemented embodied conversational agent. Our agent plans each utterance so that multiple communicative goals may be realized opportunistically by a composite action including not only speech but also coverbal gesture that fits the context and the ongoing speech in ways representative of natural human conversation. We accomplish this by reasoning from a grammar which describes gesture declaratively in terms of its discourse function, semantics and synchrony with speech.

1 Introduction

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. Research on *embodied conversational agents* aims to imbue interactive dialogue systems with the same nonverbal skills and behaviors (Cassell, 2000a).

There is good reason to think that nonverbal behavior will play an important role in evoking from users 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 dialogues with other people. For example, (Cassell and Thórisson, 1999) show 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. This argument takes on particular importance given that users repeat themselves needlessly, mistake when it is their turn to speak, and so forth when interacting with voice dialogue systems (Oviatt, 1995). In life, noisy situations like these provoke the non-verbal modalities to come into play (Rogers, 1978).

In this paper, we describe the generation of communicative actions in an implemented embodied conversational agent. Our generation framework adopts a goal-directed view of generation and casts knowledge about communicative action in the form of a grammar that specifies how forms combine, what interpretive effects they impart and in what contexts they are appropriate (Appelt, 1985; Moore, 1994; Dale, 1992; Stone and Doran, 1997). We expand this framework to take into account findings, by ourselves and others, on the relationship between spontaneous coverbal hand gestures and speech. In particular, our agent plans each utterance so that multiple communicative goals may be realized opportunistically by a composite action including not only speech but also coverbal gesture. By describing gesture declaratively in terms of its discourse function, semantics and synchrony with speech, we ensure that coverbal gesture fits the context and the ongoing speech in ways representative of natural human conversation. The result is a streamlined implementation that instantiates important theoretical insights into the relationship between speech and gesture in human-human conversation.

2 Exploring the relationship between speech and gesture

To generate embodied communicative action requires an architecture for embodied conversation; ours is provided by the agent REA (“Real Estate Agent”), a computer-generated humanoid that has an articulated graphical body, can sense the user passively through cameras and audio input, and supports communicative actions realized in speech with intonation, facial display, and animated gesture. REA currently offers the reasoning and display capabilities to act as a real estate agent showing users the features of various models of houses that appear on-screen behind her. We use existing features of REA here as a research platform for implementing models of the relationship between speech and spontaneous hand gestures during conversation. For more details about the functionality of REA see (Cassell, 2000a).

Evidence from many sources suggests that this relationship is a close one. About three-quarters of all clauses in narrative discourse are accompanied by gestures of one kind or another (McNeill, 1992), and within those clauses, the most effortful part of gestures tends to co-occur with or just before the phonologically most prominent syllable of the accompanying speech (Kendon, 1974).

Of course, communication is still possible without gesture. 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). Similarly, Cassell et al. (1999) 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, but what they convey is virtually always compatible. *Semantically*, speech and gesture give a consistent view of an overall situation. For example, gesture may depict the way in which an action was carried out when this aspect of meaning is not depicted in speech. *Pragmatically*, speech and gesture mark information about this meaning as advancing the purposes of the conversation in a consistent way. Indeed, gesture often emphasizes information that is also focused pragmatically by mechanisms like prosody in speech (Cassell, 2000b). 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 fact, some suggest (McNeill, 1992), that 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 theoretical perspective on speech and gesture involves two key claims with computational import: that gesture and speech reflect a common conceptual source; and that the content and form of a gesture is tuned to the communicative context and the actor's communicative intentions. We believe that these characteristics of the use of gesture are universal, and see the key contribution of this work as providing a general framework for building dialogue systems in accord with them. However, a concrete implementation requires more than just generalities behind its operation; we also need an understanding of the precise ways gesture and speech are used together in a particular task and setting.

To this end, we collected a sample of real-estate descriptions in line with what REA might be asked to provide. To elicit each description, we asked one subject to study a video and floor plan of a particular house, and then to describe the house to a second subject (who did not know the house and had not seen the video). During the conversation, the video and floor plan were not available to either subject; the listener was free to interrupt and ask questions.

The collected conversations were transcribed, yielding 328 utterances and 134 referential gestures, and coded to describe the general communicative goals of the speaker and the kinds of semantic features realized in speech and gesture.

Analysis of the data revealed that for roughly 50% of the gesture-accompanied utterances, gestural content was redundant with speech; for the other 50% gesture contributed content that was different, but complementary, to that contributed by speech. In addition, the relationship between content of gesture, content of speech and general communicative functions in house descriptions could be captured by a small number or rules; these rules are informed by and accord with our two key claims about speech and gesture. For example, one rule describes dialogue contributions whose general function was what we call *presentation*, to advance the description of the house by introducing a single new object. These contributions tended to be made up of a sentence that asserted the existence of an object of some type, accompanied by a non-redundant gesture that elaborated the shape or location of the object. Our approach casts this extended description of a new entity, mediated by two compatible modalities, as the speaker's expression of one overall function of presentation.

(1) is a representative example.

(1) It has [a nice garden]. (right hand, held flat, traces a circle, indicating location of the garden surrounding the house)

Six rules account for 60% of the gestures in the

Figure 1: Interacting with REA

transcriptions (recall) and apply with an accuracy of 96% (precision). These patterns provide a concrete specification for the main communicative strategies and communicative resources required for REA. A full discussion of the experimental methods and analysis, and the resulting rules, can be found in (Yan, 2000).

3 Framing the generation problem

In REA, requests for the generation of speech and gesture are formulated within the dialogue management module. REA's utterances reflect a coordination of multiple kinds of processing in the dialogue manager – the system recognizes that it has the floor, derives the appropriate communicative context for a response and an appropriate set of communicative goals, triggers the generation process, and realizes the resulting speech and gesture. The dialogue manager is only one component in a *multithreaded* architecture that carries out hardwired reactions to input as well as deliberative processing. The diversity is required in order to exhibit appropriate interactional and propositional conversational behaviors at a range of time scales, from tracking the user's movements with gaze and providing nods and other feedback as the user speaks, to participating in routine exchanges and generating principled responses to user's queries. See (Cassell, 2000a) for description and motivation of the architecture, as well as the conversational functions and behaviors it supports.

REA's design and capabilities reflect our research focus on allying conversational content with conversation management, and allying nonverbal modalities with speech: how can an embodied agent use all its communicative modalities to contribute new content when needed (propositional function), to signal the state of the dialogue, and to regulate the overall process of conversation (interactional function)? Within this focus, REA's talk is firmly delimited. REA's utterances take a question-answer format, in which the user asks about (and REA describes) a single house at a time. REA's sentences are short; generally, they contribute just a few new semantic features about particular rooms or features of the house (in speech and gesture), and flesh this contribution out with a handful of meaningful elements (in speech and gesture) that ground the contribution in shared context of the conversation.

Despite the apparent simplicity, the dialogue manager must contribute a wealth of information about the domain and the conversation to represent the communicative context. This detail is needed for REA to achieve a theoretically-motivated realization of the common patterns of speech and gesture we observed in human conversation. For example, a variety of changing features determine whether marked forms in speech and gesture are appropriate in the context. REA's dialogue manager tracks the changing status of such features as:

- *Attentional prominence,* represented (as usual in natural language generation) by setting up a *context set* for each entity (Dale, 1992). Our model of prominence 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, following (Steedman, 1991), *themes*, which describe the salient questions currently at issue in the discourse (Prince, 1986). 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 $theme(e)$.

In fact, both speech and gesture depend on the same kinds of features, and access them in the same way; this specification of the dialogue state crosscuts distinctions of communicative modality.

Another component of context is provided by a domain knowledge base, consisting of facts explicitly labeled with the *kind of information* they represent. This defines the common ground in the conversation in terms of sources of information that speaker and hearer *share*. Modeling the discourse as a shared source of information means that new semantic features REA imparts are added to the common ground as the dialogue proceeds. Following results from (Kelly et al., 1999) which show that information from both speech and gesture is used to provide context for ongoing talk, our common ground may be updated by both speech and gesture.

The structured domain knowledge also provides a resource for specifying communicative strategies. Recall that REA's communicative strategies are formulated in terms of functions which are common in naturally-occurring dialogues (such as "presentation") and which lead to distinctive bundles of content in gesture and speech. The knowledge base's *kinds of information* provide a mechanism for specifying and reasoning about such functions. The knowledge base is structured to describe the relationship between the system's *private* information and the questions of interest that that information can be used to settle. Once the user's words have been interpreted, a layer of production rules constructs *obligations* for response (Traum and Allen, 1994); then, a second layer plans to meet these obligations by deciding to present a specified kind of information about a specified object. This determines some concrete communicative goals—facts of this kind that a contribution to dialogue could make. *Both* speech and gesture can access the whole structured database in realizing these concrete communicative goals. For example, a variety of facts that bear on where a residence is—which city, which neighborhood or, if appropriate, where in a building—all provide the same kind of information, and would therefore fit the obligation to specify the location of a residence. Or, to implement the rule for presentation described in connection with (1), we can associate an obligation of presentation with a cluster of facts describing an object's type, its location in a house, and its size, shape or quality.

The communicative context and concrete communicative goals provide a common source for generating speech and gesture in REA. The utterance generation problem in REA, then, is to construct a complex communicative action, made up of speech and coverbal gesture, that achieves a given constellation of goals and tightly fits the context specified by the dialogue manager.

4 Generation and linguistic representation

We model REA's communicative actions as composed of a collection of atomic elements, including both lexical items in speech and clusters of semantic features expressed as gestures; 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 semantic and pragmatic coordination between speech and gesture. The information conveyed 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.

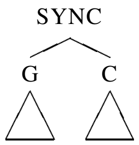
We use the SPUD generator ("Sentence Planning Using Description") introduced in (Stone and Doran, 1997) to carry out this task for REA. SPUD builds the utterance element-by-element; at each stage of construction, SPUD's representation of the current, incomplete utterance specifies its syntax, semantics, interpretation and fit to context. This representation both allows SPUD to determine which lexicalized descriptors are available at each stage to extend the utterance, and 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. (We have developed a suite of guidelines for the design of syntactic structures, semantic and pragmatic representations, and the interface between them so that SPUD's greedy search, which is necessary for real-time performance, succeeds in finding concise and effective utterances described by the grammar (Stone et al., 2000).)

As part of the development of REA, we have constructed a new inventory of lexicalized descriptors. REA's descriptors consist of entries that contribute to coverbal gestures, as well as revised entries for spoken words that allow for their coordination with gesture under appropriate discourse conditions. 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 both REA's gesture and the relationship between gesture and speech vary as a function of pragmatic context in the same way as natural gestures and speech do. More abstractly, these entries enable SPUD to realize the concrete goals tied to common communicative functions with same distribution of speech and gesture observed in natural conversations.

To explain how these entries work, 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 its utterance. 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 coordination with linguistic phrases. Our device for this is a construction SYNC which pairs a description of a gesture G with the syntactic structure of a spoken constituent C:

(2)



The temporal interpretation of (2) mirrors the rules for surface synchrony between speech and gesture presented in (Cassell et al., 1994). That is, the preparatory phase of gesture G is set to begin before the time constituent C begins; the stroke of gesture G (the most effortful part) co-occurs with the most phonologically prominent syllable in C; and, except in cases of coarticulation between successive gestures, by the time the constituent C is complete, the speaker must be relaxing and bringing the hands out of gesture space (while the generator specifies synchrony as described, in practice the synchronization of synthesized speech with graphics is an ongoing challenge in the REA project). In sum, the production of gesture G is *synchronized* with the production of speech C. (Our representation of synchrony in a single tree conveniently allows modules downstream to describe embodied communicative actions as marked-up text.)

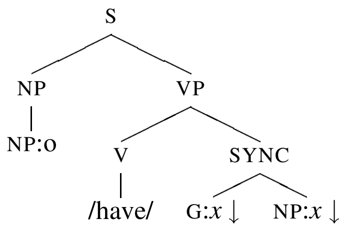
The syntactic description of the gesture itself indicates the *choices* the generator must make to produce a gesture, but does not analyze a gesture literally as a hierarchy of separate movements. Instead, these choices specify independent semantic features which we can associate with aspects of a gesture (such as handshape and trajectory through space). Our current grammar does not undertake the final step of associating semantic features to choice of particular handshapes and movements, or gesture *morphology*; we reserve this problem for later in the research program. We allow gesture to accompany alternative constituents by introducing alternative syntactic entries; these entries take on different pragmatic requirements (as described below) to capture their respective discourse functions.

So much for syntax. 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 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 using the entry. Unlike the semantics, however, the pragmatics does not achieve specific communicative goals like identifying referents. Instead, the pragmatics establishes a general fit between the entry and the context.

The entry schematized in (3) illustrates these three components; the entry also suggests how these components can define coordinated actions of speech and gesture that respond coherently to the context.

(3) a syntax:



b semantics: $have(o, x)$

c pragmatics: $hearer\text{-}new(x) \wedge theme(o)$

(3) describes the use of *have* to introduce a new fea-

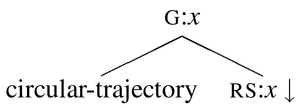
ture of (a house) o. The feature, indicated throughout the entry by the variable x, is realized as the object NP of the verb *have*, but x can also form the basis of a gesture G coordinated with the noun phrase (as indicated by the SYNC constituent). The entry asserts that o has x.

(3) is a presentational construction; in other words, it coordinates non-redundant paired speech and gesture in the same way as demonstrated by our house description data. To represent this constraint on its use, the entry carries two pragmatic requirements: first, x must be new to the hearer; moreover, o must link up with the open question in the discourse that the sentence responds to.

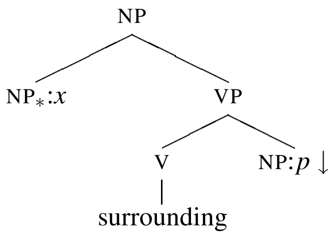
The pragmatic conditions of (3) help support our theory of the discourse function of gesture and speech. A similar kind of sentence could be used to address other open questions in the discourse—for example, to answer *which house has a garden?* This would not be a presentational function, and (3) would be infelicitous here. In that case, gesture would naturally coordinate with and elaborate on the answering information—in this case the house. So the different information structure would activate a different entry, where the gesture would coordinate with the subject and describe o.

Meanwhile, alternative entries like (4a) and (4b)—two entries that both convey (4c) and that both could combine with (3) by LTAG operations—underlie our claim that our implementation allows gesture and speech to draw on a single conceptual source and fulfill similar communicative intentions.

(4) a syntax:



b syntax:



c semantics: $surround(x, p)$

(4a) provides a structure that could substitute for the G node in (3) to produce semantically and pragmatically coordinated speech and gesture. (4a) specifies a right hand gesture in which the hand traces out a circular trajectory; a further decision must determine the correct handshape (node RS, as a function of the entity x that the gesture describes). We pair (4a) with the semantics in (4c), and thereby model that the gesture indicates that one object, x, surrounds another, p. Since p cannot be further described, p must be identified by an additional presupposition of the gesture which picks up a reference frame from the shared context.

Similarly, (4b) describes how we could modify the VP introduced by (3) (using the LTAG operation of *adjunction*), to produce an utterance such as *It has a garden surrounding it.* By pairing (4b) with the same semantics (4c), we ensure that SPUD will treat the communicative contribution of the alternative constructions of (4) in a parallel fashion. Both are triggered by accessing background knowledge and both are recognized as *directly* communicating specified facts.

5 Solving the generation problem

We now sketch how entries such as these combine together to account for REA's utterances. Our example is the dialogue in (5):

(5) a User: Tell me more about the house.
   b REA: It has [a nice garden]. (right hand, held flat, traces a circle)

REA's response indicates both that the house has a nice garden and that it surrounds the house.

As we have seen, (5b) represents a common pattern of description; this particular example is motivated by an exchange two human subjects had in our study, cf. (1). (5b) represents a solution to a generation problem that arises as follows within REA's overall architecture. The user's directive is interpreted and classified as a directive requiring a deliberative response. The dialogue manager recognizes an obligation to respond to the directive, and concludes that to fulfill the function of presenting the garden would discharge this obligation. The presentational function grounds out in the communicative goal to convey a collection of facts about the garden (type, quality, location relative to the house). Along with these goals, the dialogue manager supplies its communicative context, which represents the centrality of the house in attentional prominence, cognitive status and information structure.

In producing (5b) in response to this NLG problem, SPUD both calculates the applicability of and determines a preference for the lexicalized descriptors involved. Initially, (3) is applicable; the system knows the house has the garden, and represents the garden as new and the house as questioned. The entry can be selected over potential alternatives based on its interpretation—it achieves a communicative goal, refers to a prominent entity, and makes a relatively specific connection to facts in the context. Similarly, in the second stage, SPUD evaluates and selects (4a) because it communicates a needed fact in a way that helps flesh out a concise, balanced communicative act by supplying a gesture that by using (3) SPUD has already realized belongs here. Choices of remaining elements—the words *garden* and *nice*, the semantic features to represent the garden in the gesture—proceed similarly. Thus SPUD arrives at the response in (5b) just by reasoning from the declarative specification of the meaning and context of communicative actions.

6 Related Work

The *interpretation* of speech and gesture has been investigated since the pioneering work of (Bolt, 1980) on deictic gesture; recent work includes (Koons et al., 1993; Bolt and Herranz, 1992). Systems have also attempted *generation* of gesture in conjunction with speech. Lester et al. (1998) generate deictic gestures and choose referring expressions as a function of the potential ambiguity of objects referred to, and their proximity to the animated agent. Rickel and Johnson (1999)'s pedagogical agent produces a deictic gesture at the beginning of explanations about objects in the virtual world. André et al. (1999) generate pointing gestures as a sub-action of the rhetorical action of labeling, in turn a sub-action of elaborating.

Missing from these prior systems, however, is a representation of communicative action that treats the different modalities on a par. 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. While many insights from these formalisms are relevant in embodied conversation, spontaneous gesture requires a distinct analysis with different emphasis. For example, we need some notion of discourse pragmatics that would allow us to predict where gesture occurs with respect to speech, and what its role might be. Likewise, we need a model of the *communicative effects* of spontaneous coverbal gesture—one that allows us to reason naturally about the multiple goals speakers have in producing each utterance.

7 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 it informs the generation of communicative action in our artificial embodied conversational agent, REA. REA has a working implementation, which includes the modules described in this paper, and can engage in a variety of interactions including that in (5). Experiments are underway to investigate the extent to which REA's conversational capacities share the strengths of the human capacities they are modeled on.

Acknowledgments

The research reported here was supported by NSF (award IIS-9618939), Deutsche Telekom, AT&T, and the other generous sponsors of the MIT Media Lab, and a postdoctoral fellowship from RUCCS. Hannes Vilhjálmsson assisted with the implementation of REA's discourse manager. We thank Nancy Green, James Lester, Jeff Rickel, Candy Sidner, and anonymous reviewers for comments on this and earlier drafts.

References

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

Douglas Appelt. 1985. *Planning English Sentences*. Cambridge University Press, Cambridge England.

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

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

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

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

J. Cassell, D. McNeill, and K. E. McCullough. 1999. Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition*, 6(2).

Justine Cassell. 2000a. Embodied conversational interface agents. *Communications of the ACM*, 43(4):70–78.

Justine Cassell. 2000b. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors, *Embodied Conversational Agents*, pages 1–28. MIT Press, Cambridge, MA.

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

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

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

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

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

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

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

S. D. Kelly, J. D. Barr, R. B. Church, and K. Lynch. 1999. Offering a hand to pragmatic understanding: The role of speech and gesture in comprehension and memory. *Journal of Memory and Language*, 40:577–592.

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

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

James Lester, Stuart Towns, Charles Calloway, and Patrick FitzGerald. 1998. Deictic and emotive communication in animated pedagogical agents. In *Workshop on Embodied Conversational Characters*.

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

Johanna Moore. 1994. *Participating in Explanatory Dialogues*. MIT Press, Cambridge MA.

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

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

Ellen F. Prince. 1992. 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.

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

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

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

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

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

Matthew Stone, Tonia Bleam, Christine Doran, and Martha Palmer. 2000. Lexicalized grammar and the description of motion events. In *TAG+: Workshop on Tree-Adjoining Grammar and Related Formalisms*.

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

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

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

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

Hao Yan. 2000. Paired speech and gesture generation in embodied conversational agents. Master's thesis, Media Lab, MIT.