Deictic Believability: Coordinated Gesture, Locomotion, and Speech in Lifelike Pedagogical Agents

James C. Lester   Jennifer L. Voerman

Stuart G. Towns   Charles B. Callaway

Multimedia Laboratory
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206 USA

Phone: (919) 515-7534   Fax: (919) 515-7925
{lester, jlvoerma, sgtowns, cbcallaw}@eos.ncsu.edu

Abstract

Lifelike animated agents for knowledge-based learning environments can provide timely, customized advice to support students' problem solving. Because of their strong visual presence, they hold significant promise for substantially increasing students' enjoyment of their learning experiences. A key problem posed by lifelike agents that inhabit artificial worlds is *deictic believability*. In the same manner that humans refer to objects in their environment through judicious combinations of speech, locomotion, and gesture, animated agents should be able to move through their environment, and point to and refer to objects appropriately as they provide problem-solving advice. In this paper we describe a framework for achieving deictic believability in animated agents. A deictic behavior planner exploits a world model and the evolving explanation plan as it selects and coordinates locomotive, gestural, and speech behaviors. The resulting behaviors and utterances are believable, and the references exhibit a lack of ambiguity. This approach to spatial deixis has been implemented in a lifelike animated agent, Cosmo, who inhabits a learning environment for the domain of Internet packet routing. Cosmo provides realtime advice to students as they escort packets through a virtual world of interconnected routers. Results of an informal focus group study with the Cosmo agent suggest that the spatial deixis framework produces clear explanatory animated behaviors.

1 Introduction

Lifelike animated agents offer great promise for knowledge-based learning environments. Because of the immediate and deep affinity that children seem to develop for these agents, the potential pedagogical benefits they provide are perhaps even exceeded by their motivational benefits. By creating the illusion of life, animated agents may significantly increase the time that children seek to spend with educational software, and recent advances in affordable graphics hardware are beginning to make the widespread distribution of realtime animation technology a reality. Endowing animated agents with believable, lifelike qualities has been the subject of a growing body of research [André and Rist, 1996, Bates, 1994, Cassell et al., 1994a, Granieri et al., 1995, Blumberg and Galyean, 1995, Kurlander and Ling, 1995, Maes et al., 1995, Walker et al., 1997] and much interesting work has examined the social aspects of human-computer interaction and users' anthropomorphization of software [Nass et al., 1993, Nass et al., 1995, Reeves and Nass, 1992]. Animated pedagogical agents [Rickel and Johnson, 1997a, Stone and Lester, 1996] constitute an important category of animated agents whose intended use is educational applications. A recent large-scale empirical study suggests that these agents can be pedagogically effective [Lester et al., 1997b]. Moreover, it was determined that students perceived the agent as being very helpful, credible, and entertaining [Lester et al., 1997a].

A key problem posed by lifelike agents that inhabit artificial worlds is *deictic believability.* In the same manner that humans refer to objects in their environment through combinations of speech, locomotion, and gesture, animated agents should be able to move through their environment, point to objects, and refer to them appropriately as they provide problem-solving advice. Deictic believability in animated agents requires the design of an agent behavior planner that considers the physical properties of the world inhabited by the agent. The agent must exploit its knowledge of the positions of objects in the world, its relative location with respect to these objects, as well as its prior explanations to create deictic gestures, motions, and utterances that are both natural and unambiguous.

To address these issues, we have developed a spatial deixis framework for achieving deictic believability. Building on our previous work on dynamically sequencing animated pedagogical agents [Stone and Lester, 1996] and enhancing pedagogical agent believability [Lester and Stone, 1997] as well as on Cassell *et al.*'s foundational work on agent deixis [Cassell et al., 1994b], a deictic behavior planner exploits a world model and the evolving explanation plan as it selects and coordinates locomotive, gestural, and speech behaviors. The resulting behaviors are believable, and by considering the relative proximity of objects, the references are clear and exhibit a lack of ambiguity. This approach has been implemented in a lifelike animated agent, Cosmo, who inhabits a learning environment for the domain of Internet packet routing.

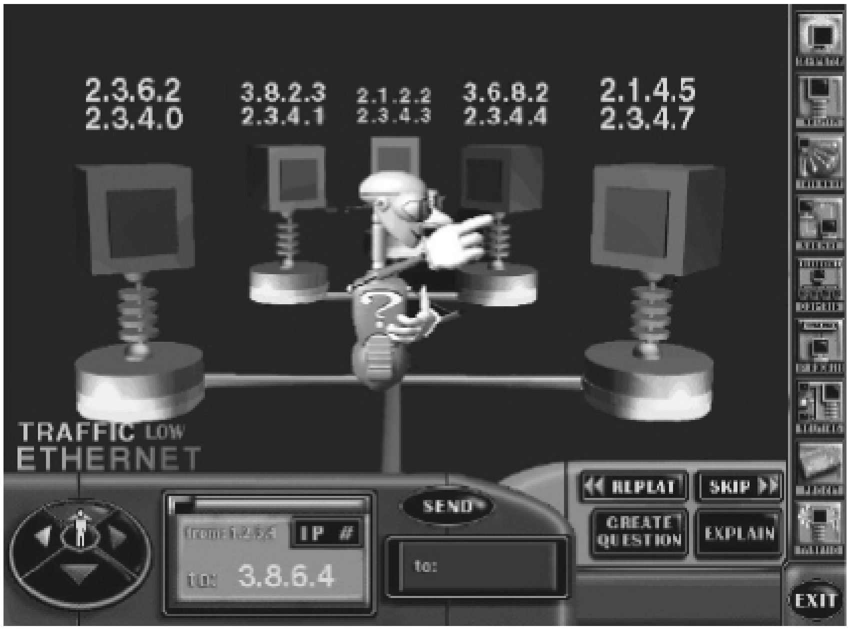


Figure 1: Cosmo and the Internet Advisor World

Cosmo is an impish, antenna-bearing creature who hovers about in a virtual world of routers and networks and provides advice to students as they decide how to ship packets through the network to specified destinations (Figure 1). His appearance, mannerisms, and behavior space of actions and utterances are the combined creation of a large multidisciplinary team of computer scientists, graphic artists, modelers, and animators. In response to students' problem-solving activities and questions, Cosmo interjects explanations that refer to specific routers, subnets and address labels in the environment. By carefully selecting and coordinating speech, gesture, and locomotion, his behavior planner creates deictic references that are natural and unambiguous. A focus group study with students interacting with Cosmo in the Internet Advisor learning environment is encouraging.

This article provides an account of the representations and computational mechanisms underlying the spatial deixis framework for achieving deictic believability. It is structured as follows. Section 2 sets forth design criteria for deictic believability in lifelike pedagogical agents and describes a learning environment that serves as a testbed for studying deictic believability. Section 3 presents the spatial deixis framework for coordinating deictic gesture, locomotion, and speech. This includes computational methods for ambiguity appraisal, gesture and locomotion planning, selecting deictic referring expressions, and coordinating all resulting behaviors. Section 4 describes an implemented lifelike agent, Cosmo, and provides a trace of deictic behavior sequencing generated to provide advice to a student solving problems in the Internet Advisor learning environment. Section 5 describes an informal focus group study of students interacting with Cosmo in the Internet Advisor learning environment. Section 6 concludes with a summary and a discussion of future directions.

2 Deictic Believability in Lifelike Pedagogical Agents

In the course of communicating with one another, interlocutors employ deictic techniques to create context-specific references. Hearers interpret linguistic events in concrete contexts. To understand a speaker's utterance, hearers must consider the physical and temporal contexts in which the utterance is spoken, as well as the identities of the speaker and hearer. Referred to as the *deictic center* of an utterance, the triple of location, time, and identities also plays an important role in generating linguistic events [Fillmore, 1975]. The first of these, location, is critical for achieving *spatial deixis*, a much studied phenomenon in linguistics which is used to create references in the physical world [Jarvella and Klein, 1982]. Speakers use spatial deixis to narrow hearers' attention to particular entities. In one popular psycho-social framework for analyzing spatial deixis, the *figure-ground* model [Roberts, 1993], the world is categorized into *ground*, which is the common physical environment shared by the speaker and hearer, and the *referent*, which is the aspect of the ground to which the speaker wishes to refer. Through carefully constructed referring expressions and well-chosen gestures, the speaker assists the hearer in focusing on the particular referent of interest.

The ability to handle *spatial deixis* effectively is especially critical for animated pedagogical agents that inhabit virtual worlds. To provide problem-solving advice to students who are interacting with objects in the world, the agent must be able to refer to objects in the world to clearly explain their function and to assist students in performing their tasks. Deictic mechanisms for animated pedagogical agents should satisfy three criteria:

- *Lack of Ambiguity:* In a learning environment, an animated agent's clarity of expression is of the utmost importance. To effectively communicate advice and explanations to students, the agent must be able to create deictic references that are unambiguous. Avoiding ambiguity is critical in virtual environments, where an ambiguous deictic reference can cause mistakes in problem solving and foster misconceptions. Ambiguity is particularly challenging in virtual environments housing a multitude of objects, especially when many of the objects are visually similar.

- *Immersivity:* People are frequently physically immersed in the environments in which they create spatial references to objects. Just as they gesture and move within it, e.g., by walking across a scene to a cluster of objects and pointing to one of them, to achieve believability, agents should behave accordingly.

- *Pedagogical Soundness:* Deictic mechanisms for agents that inhabit learning environments must support their central pedagogical intent. Rather than operating in a communicative vacuum, spatial deixis must support the ongoing advisory discourse and be appropriately situated in the problem-solving context.

The lack-of-ambiguity requirement implies that deictic planning mechanisms must make use of an expressive representation of the world. While unambiguous deictic references can be created with object highlighting or by employing a relatively stationary agent with a long pointer, e.g., [André and Rist, 1996], the immersivity requirement implies that lifelike agents should artfully combine speech, gesture, and locomotion. Finally, the pedagogical soundness requirement implies that all deictic utterances, speech, and movements must be integrated with explanation plans that are generated in response to student questions and problem-solving impasses.

In general (after [Bates, 1994]), we refer to the *believability* of lifelike agents as the extent to which users interacting with them come to believe that they are observing a sentient being with its own beliefs, desires, intentions, and personality. It has been shown that believable pedagogical agents in interactive learning environments can produce the *persona effect*, in which the very presence of a lifelike character in a learning environment can have a strong positive effect on learners' perception of their learning experience [Lester et al., 1997a]. In a study with 100 middle school students, it was found that when learners interact with a lifelike agent that is expressive, i.e., an agent that exhibits both animated and verbal advisory behaviors, students perceive it to be encouraging and of high utility.[1]

A critical but largely unexplored aspect of agents' believability for learning environments is deictic believability. We say that lifelike agents that make deictic references in a manner that simultaneously achieves a lack of ambiguity, does so in an immersive setting, and operates in a pedagogically sound manner exhibit *deictic believability.*

2.1 Related Work

Several aspects of spatial deixis have been addressed by the natural language generation and intelligent multimedia communities. Natural language researchers have studied reference generation, e.g., Dale's classic work on referring expressions [Dale, 1992], scene description generation [Novak, 1987], and spatial layout description generation [Sibun, 1992]. Work on intelligent multimedia systems [André et al., 1993, Feiner and McKeown, 1990, Maybury, 1991, Roth et al., 1991, Mittal et al., 1995] has produced techniques for dynamically incorporating highlights, underlines,

[1] The same study employed pre- and post-tests to evaluate learning effectiveness and found statistically significant gains in students' performance [Lester et al., 1997b].

and blinking [Neal and Shapiro, 1991]. However, none of these consider the orchestration of an agent's communicative behaviors in an environment.

Work on lifelike agents [André and Rist, 1996, Bates 1994, Cassell et al., 1994a, Granieri et al., 1995, Blumberg and Galyean, 1995, Kurlander and Ling, 1995, Maes et al., 1995, Walker et al., 1997] has yielded more sophisticated techniques for referring to onscreen entities. The EDWARD system [Claassen, 1992] employs a stationary persona that "grows" a pointer to a particular object in the interface and the PPP agent [André and Rist, 1996] is able to dynamically indicate various onscreen objects with a long pointer. While these techniques are effective for many tasks and domains, they do not provide a general solution for achieving deictic believability that deals explicitly with ambiguity by both selecting appropriate referring expressions and by producing lifelike gestures and locomotion.

Begun at the University of Pennsylvania's JACK project and continued at MIT, Cassell *et al.*'s work on conversational agents is perhaps the most advanced to date on agents that combine gesture, speech, and facial expression [Cassell et al., 1994a]. In addition to deictics, they also exhibit iconic, metaphoric, and beat gestures. However, this work neither provides a solution to the intricacies of detecting ambiguity in complex physical environments (and then addressing it with integrated speech, gesture, and locomotion) nor is its focus on pedagogical interactions.

Despite the promise of lifelike pedagogical agents, with the exception of work on the DESIGN-A-PLANT project [Lester et al., 1996, Stone and Lester, 1996, Lester and Stone, 1997, Lester et al., 1997c, Lester et al., in press] and the Soar Training Expert for Virtual Environments (STEVE) project [Rickel and Johnson, 1997a, Rickel and Johnson, 1997b], in which agents provide instruction about procedural tasks in a virtual reality environment, lifelike agents for pedagogy have received little attention. Neither the STEVE nor the DESIGN-A-PLANT projects address deictic believability.

2.2 A Deictic Believability Testbed

Features of environments, agents, and tasks that force spatial deixis issues to the forefront are threefold. (1) A world populated by a multitude of objects, many of which are similar, will require agents to plan speech, gesture, and locomotion carefully to avoid ambiguity. (2) We can select a domain and problem-solving task for learners that requires agents to provide advice and explanations that frequently refer to different objects in the world. (3) Problem-solving tasks that require students to make decisions based on factors physically present in the environment will induce clarity requirements on agents' communicative capabilities. In contrast to a more abstract domain such as algebra, we can select a domain that can be graphically represented with objects in perhaps idiosyncratic and complex spatial layouts, thereby requiring the agent to produce clear problem-solving advice that integrates spatial deixis with explanations of concepts and problem-solving strategies.

To investigate deictic believability in lifelike pedagogical agents, we have developed a testbed in the form of an interactive learning environment. Because it has each of the features outlined above, the INTERNET ADVISOR provides a "laboratory" in which to study the coordination of deictic speech, gesture, and locomotion. Designed to foster exploration of computational mechanisms for animation behavior sequencing of lifelike characters and realtime human-agent problem-solving interaction, the INTERNET ADVISOR consists of a virtual world populated by many routers and networks.[2]

Students interact with COSMO as they learn about network routing mechanisms by navigating through a series of subnets. Given a packet to escort through the Internet, they direct it through networks of connected routers. At each subnet, they may send their packet to a specified router or view adjacent subnets. By making decisions about factors such as address resolution and traffic congestion, they learn the fundamentals of network topology and routing mechanisms. Helpful, encouraging, and with a bit of attitude, COSMO explains how computers are connected, how routing is performed, what types of networks have different physical characteristics, how Internet address schemes work, and how network outages and traffic considerations come into play. Students' journeys are complete when they have successfully navigated the network and delivered their packet to its proper destination. The learning environment serves as an excellent testbed for exercising spatial deixis because each subnet has a variety of routers attached to it and the agent must refer unambiguously to them as it advises students about their problem-solving activities.

3 Coordinating Deictic Gesture, Locomotion, and Speech

The primary role of lifelike pedagogical agents is to serve as an engaging vehicle for communication. Hence, in the course of observing a student attempt different solutions in a learning environment, a lifelike pedagogical agent should clearly explain concepts and convey problem-solving strategies. It is in this context that spatial deixis arises. The spatial deixis framework guides the operation of the *deictic planner*, a key component of the agent behavior planning architecture (Figure 2). The interaction manager provides an interface between the learning environment and the agent that inhabits it. By monitoring a student's problem-solving activities in the learning environment, the interaction manager invokes the agent behavior planner in two situations: (1) when a student pauses for an extended period of time, which may signal a problem-solving impasse, and (2) when a student commits an error, which indicates a possible misconception.

The agent behavior planner consists of an explanation planner and a deictic planner. The

[2] In addition to the authors, the INTERNET ADVISOR was created by 9 graphic artists (environment designers, 3D modelers, and animators), as well as a musician, a voice actor, and several programmers.

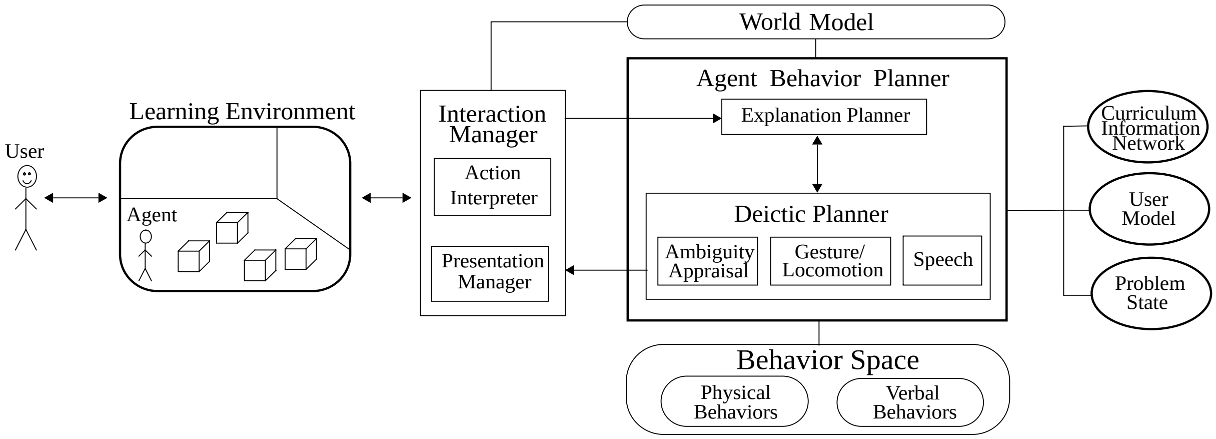


Figure 2: Lifelike pedagogical agent deictic behavior planning architecture

explanation planner serves a function that is analogous to the discourse planner of natural language generation systems [Suthers, 1991, Cawsey, 1992, Hovy, 1993, Mittal, 1993, Moore, 1995, Lester and Porter, 1997]. Natural language generation systems typically consist of a discourse planner that determines the content and structure of multisentential texts and a realization system that plans the surface structure of the resulting prose. Analogously, given a communicative goal, the explanation planner of the agent behavior planner determines the content and structure of an agent's explanations and then passes these specifications to the deictic planner, which realizes these specifications in speech, gesture, and locomotion. The explanation planner invokes the deictic planner by specifying a communicative act C, a topic T, a gestural referent R_g, and a spoken referent R_s (summarized in Figure 3).

To accomplish its task, the deictic behavior planner examines the representational structures in a world model, a curriculum information network, a user model, the current problem state, which includes both the student's most recently proposed solution and the learning environment's analysis of that solution, and two focus histories, one for gesture and one for speech. It then constructs a sequence of physical behaviors and verbal explanations that will collectively constitute the advice which the agent will deliver. For example, given a communicative goal, the explanation planner for Cosmo typically produces an explanation plan that calls for the agent to speak 6–10 utterances and perform several locomotive and gestural behaviors. These are then passed to the presentation manager which manipulates the agent persona in the learning environment. Problem-solving actions performed by the student are therefore punctated by customized explanations provided by the agent in a manner reminiscent of classic task-oriented dialogs.

Deictic planning comes into play when the behavior planner determines that an explanation must refer to an object in the environment. For each utterance that makes a reference to an environmental object, the explanation planner invokes the deictic system and supplies it with

Given:

- For each communicative act created by the explanation planner, the deictic planner is given:
  - *Communicative Act Category*: C
    Examples: `State-Correct`, `Give-Advice`
  - *Topic*: T
    Example: `Address-Resolution`
  - *Gestural Referent*: R_g
    Example: `Computer #15`
  - *Spoken Referent*: R_s
    Example: `SubNetwork #2`
- *World Model*: W (ontology, spatial knowledge, and physical characteristics of objects in environment)
  Example: Ontology of computers and networks, knowledge of the relative locations and proximity of computers to one another, and relative sizes of all environmental and interface objects.
- *Focus histories*: H
  - *Gestural focus history*: H_g
    Example: (`Computer #7`, `Traffic-Information-Label`)
  - *Spoken focus history*: H_s
    Example: (`Computer #12`, `SubNetwork #4`)
- *Current location of agent*: L_A (x-y coordinates)

Determine:

- *Gesture*: G
  Examples: `NIL-Gesture`, `left-across`, `right-up`
- *Locomotion*: L
  Examples: `NIL-Locomotion`, `left-up`, `right-down`
- *Speech*: S (referring expression)
  Example: "These computers"
- *Gaze*: Z (orientation of head, and consequently eyes)
  Example: `right-down`

Figure 3: The deictic behavior planning task specification

the intended referent R. The deictic system operates in the following phases to plan the agent's gestures, locomotion, and speech:

1. **Ambiguity Appraisal:** The deictic system first assesses the situation by determining whether a reference to R may be ambiguous. By examining the evolving *explanation plan*, which contains a record of the objects the agent has referred to during utterances spoken so far in the current explanation sequence, the deictic planner evaluates R's initial potential for ambiguity. This assessment will contribute to gesture, locomotion, and speech planning decisions.

2. **Gesture and Locomotion Planning:** To determine the agent's physical actions, the deictic

system uses the specification of the relative positions of the objects in the scene of the world model, as well as the previously made ambiguity assessment, to plan the agent's deictic gestures and movement. By considering the proximity of objects in the world, the deictic system determines whether the agent should point to R, and if so, whether it should move to R.

3. **Utterance Planning and Coordination:** To determine what the agent should say to refer to R, the deictic system considers focus information, the ambiguity assessment, and the world model. Utterance planning pays particular attention to the relative locations of the referent and the agent, taking into account its planned locomotion from the previous phase. The result of utterance planning is a referring expression consisting of the appropriate proximal/non-proximal demonstratives and pronouns. Finally, the behavior planner coordinates the agent's spoken, gestural, and locomotive behaviors,[3] orchestrates their exhibition by the agent in the learning environment, and returns control to the student.[4]

The deictic behavior planning algorithm is summarized in Figure 4. After briefly discussing the primary knowledge sources available to the behavior planner, the computational methods underlying ambiguity appraisal, gesture and locomotion planning, deictic referring expression planning, and speech-behavior coordination are described in detail below.

3.1 Knowledge Sources for Deictic Behavior Planning

The explanation planner invokes the deictic planner by specifying a communicative act C, a topic T, a gestural referent R_g, and a spoken referent R_s. The *communicative act C* indicates the category of speech act to be performed, such as giving advice or stating which aspects of the student's proposed solution are correct. The *topic T* indicates the content of the communicative action. For example, address resolution and congestion are topics in the INTERNET ADVISOR. The *gestural referent R_g* specifies the object to which the agent may point. For example, if the explanation planner determined that the agent should explain why the learner chose an inappropriate computer during problem solving, it would indicate to the deictic planner that the agent may need to point to that particular computer when referring to it. The *spoken referent R_s* specifies the concept to which the agent should refer in speech. R_s is typically the same as R_g, but sometimes they differ. For example, when the agent needs to refer verbally to an object that is not the subject of his utterance, the two will differ. This phenomenon of *multiple intrasentential deixis* is discussed in Section 3.5.

[3] It also coordinates gaze.

[4] In fact, a bypass mechanism in the behavior planner enables the student to interrupt the agent's explanation in midstream if she prefers to proceed with problem solving.

1. Appraise ambiguity:
   For $R \leftarrow R_g$ and R_s, determine if R is in focus by examining previous utterances U_{i-1} and U_{i-2} in H_g and H_s (U_i is the utterance currently being planned)

   (a) Novel reference assessment: $R \neq U_{i-1}$ and $R \neq U_{i-2}$.

   (b) Unique focus assessment:
   ($R = U_{i-1}$ and U_{i-2} = NIL) or ($R = U_{i-2}$ and U_{i-1} = NIL) or ($R = U_{i-1}$ and U_{i-2})

   (c) Multiple foci assessment:
   ($R = U_{i-1}$) and ($U_{i-2} \neq$ (NIL or R)) or ($R = U_{i-2}$) and ($U_{i-1} \neq$ (NIL or R))

2. Plan gesture and locomotion:

   (a) If Step (1) determines that a novel reference or a multiple focus is in effect:

      i. Multiple proximal foci assessment: If W indicates R_g is near objects found in U_{i-1} or U_{i-2} in H_g, then
         Locomotion-Recommended? $\leftarrow$ True, Gesture-Required? $\leftarrow$ True

      ii. Multiple proximal similarity assessment: If W indicates R_g is near objects of the same ontological type as R_g, then
         Locomotion-Recommended? $\leftarrow$ True, Gesture-Required? $\leftarrow$ True

      iii. Diminutiveness assessment: If W indicates R_g is small relative to other objects in the world, then
         Locomotion-Recommended? $\leftarrow$ True, Gesture-Required? $\leftarrow$ True

      iv. else [unique focus is in effect]
         Locomotion-Recommended? $\leftarrow$ False, $L \leftarrow$ NIL-Locomotion
         Gesture-Required? $\leftarrow$ False, $G \leftarrow$ NIL-Gesture

   (b) If (Locomotion-Recommended? = True) and (Gesture-Required? = True) then determine precise values for G and L (motion path from L_A to L_R) of agent.

      i. Determine coordinates of deictic target R_g.

      ii. Compute gestural direction G and locomotion L between L_A and R_g (e.g., right-up).

      iii. Determine position of agent's finger if it were extended with the agent in L_A in the direction found in (ii).

      iv. Determine coordinates of agent's body L_R for it to point to coordinates found in (i) with direction in (ii).

      v. Compute vector from L_A to L_R.

3. Plan and coordinate utterances:

   (a) Select referring expressions:

      i. Unique focus assessed in Step (1.b):

         A. If R_s occurs in U_{i-1}, pronominalize.

         B. If R_s occurs in U_{i-2}, definite article.

      ii. Novel or multiple foci assessed in Steps (1.a, 1.c):

         A. If G and L are not NIL then proximal demonstrative (handling number as appropriate for R_s).

         B. If G and L are NIL then non-proximal demonstrative (handling number as appropriate for R_s).

   (b) Introduce gaze: $Z \leftarrow G$.

   (c) Coordinate speech with gesture, locomotion, and gaze by initiating exhibition of G, L, and Z before onset of S and completing their exhibition before proceeding to behaviors generated for next communicative act.

Figure 4: The deictic behavior planning algorithm summary

Because deictic behavior planning is both a physical task and a pedagogical task, the behavior planner requires access to knowledge about the world as well as knowledge about pedagogy and communication. Hence, to accomplish its task, it examines six knowledge sources: the world model, the curriculum information network, the user model, the current problem state, and the two focus histories.

The *world model* represents both domain knowledge and spatial knowledge about the objects in the environment. The former includes ontological knowledge about objects, including knowledge of their physical properties such as size, and the latter includes knowledge about the spatial layout of the environment so the agent can draw inferences about the relative location of objects. For example, the Internet Advisor's world model includes routers, subnets, computers, and network topology.

The *curriculum information network* houses a representation of the topics and problem-solving skills for the given domain and task of the learning environment [Wescourt et al., 1981]. Imposed on these nodes is a prerequisite structure in the form of a partial ordering. A concept C_j that occupies a position in the partial order after another C_i will not be explained (or, in problem solving, exercised) until both C_i and the others before it in a topological sort of the network have also been explained. For example, in the Internet Advisor's CIN, the concept of IP address resolution precedes that of the effect of network type on packet routing decisions. In general, imposing only those relations that are clearly mandated by the domain retains greater flexibility in explanation generation.

Research on user modeling and plan recognition has explored different approaches to computational methods for representing and inferring users' beliefs, goals, and plans. The behavior planner employs the very simple approach of *overlay user models* [Carr and Goldstein, 1977], which represent users' skills in the same formalism as the domain model and marks skills as they are demonstrated in the course of problem solving. For example, the Internet Advisor marks packet routing skills in the user model as students successfully solve problems in the learning environment. While overlay models do not offer the inferential precision of more sophisticated techniques such as Bayesian approaches [Conati et al., 1997], they offer the advantages of simplicity of design and construction.

The behavior planner employs a tripartite *problem state* representation. First, it includes features of the world model that bear on the current problem being attempted by the student. For example, in the Internet Advisor, the problem state specifies the amount of traffic on all of the subnets. Second, the problem state includes the student's proposed solution. For example, the Internet Advisor monitors the student's decision about the next router to which they wish to direct their packet. Third, the problem state includes a diagnostic evaluation of the student's proposed solution. In the INTERNET ADVISOR, each time the student makes a routing decision, the learning environment assesses both the correctness and the optimality of the student's proposed solution and notes its diagnosis in the problem state.

Finally, in addition to the knowledge sources noted above, the deictic planner exploits two focus histories, a *gestural focus history* H_g and a *spoken focus history* H_s. H_g and H_s are stacks of recent gestural and speech referents that are pushed with each new communicative act. By design, the n topmost referents are the most critical in deictic behavior planning because they represent the entities that are currently in focus; empirical evidence in our domain suggests that a value of 2 for n is most effective. All referents are popped upon the completion of the final communicative act of a sequence of acts created to satisfy a single communicative goal.

3.2 Ambiguity Appraisal

The first phase of deictic planning consists of evaluating the potential for ambiguity. For each utterance in the evolving explanation plan that makes a reference to an object in the environment, the explanation planner invokes the deictic system. Deictic decisions depend critically on an accurate assessment of the discourse context in which the reference will be communicated. To correctly plan the agent's gestures, movements, and utterances, the deictic system determines whether the situation has the potential for ambiguity within the current explanation.[5] Because focus indicates the prominence of the referent at the current juncture in the explanation, the deictic system uses focus as the primary predictor of ambiguity: potentially ambiguous situations can be combatted by combinations of gesture and locomotion.

A referent R has the potential for ambiguity if it is currently not in focus or if it is in focus but is one of multiple objects in focus. To determine if the referent is in focus, the deictic system examines the evolving explanation plan and inspects it for previous deictic references to R. Suppose the explanation planner is currently planning utterance U_i. It examines utterances U_{i-1} and U_{i-2} for preceding deictic references to R. There are three cases to consider:

- *Novel Reference:* If the explanation planner locates no deictic reference to R in U_{i-1} or U_{i-2},then R is ambiguous, and is therefore deserving of greater deictic emphasis. For example, if a student interacting with the INTERNET ADVISOR chooses to send a packet to a particular router which does not lie along the optimal path to the packet's destination, COSMO interrupts the student and makes an initial reference to that router. He should therefore introduce the referent into the discourse.

[5] This initial phase of ambiguity assessment considers only discourse issues; spatial considerations are handled in the following two phases.

- *Unique Focus:* If the explanation planner locates a reference to R in U_{i-1} and U_{i-2} but not to other entities, then R has already been introduced and the potential for ambiguity is less. For example, when COSMO's explanation consists of multiple utterances about a particular router, a reference to that router will be in unique focus. Consequently, the need for special deictic treatment is reduced.

- *Multiple Foci:* If the explanation planner locates a reference to R but also to other entities in U_{i-1} and U_{i-2}, then the situation is potentially ambiguous. For example, if COSMO points to one router and subsequently points to another which the student has just selected, but he now needs to refer to the first router again for purposes of comparison, multiple referents are in focus and he must therefore take precautions against making an ambiguous reference.

The result of this determination is recorded for use in the following two phases of gesture and locomotion planning and referring expression planning.

3.3 Gesture and Locomotion Planning

When potential ambiguities arise, endowing the agent with the ability to point and move to objects to which it will be referring enables it to increase its clarity of reference. The deictic system plans two types of physical behaviors: gestures and locomotion. In each case, it first determines whether a behavior of that type is warranted. If so, it then computes the behavior.

To determine whether the agent should exhibit a pointing gesture to physically designate the referent within the environment, the behavior planner inspects the conclusion of the ambiguity computation in the previous phase. If the referent was deemed ambiguous or potentially ambiguous, the system will plan a pointing gesture for the agent.

In addition to pointing, the agent can also move from one location to another to clarify a deictic reference which otherwise might be ambiguous. If the referent has been determined to be unambiguous, i.e., it is in a unique focus, the agent will remain stationary.[6] In contrast, if the referent is ambiguous, i.e., if it is a novel reference, the deictic system instructs the agent to move towards the object specified by the referent as the agent points at it. For example, if COSMO is discussing a router which has not been previously mentioned in the last two utterances, he will move to that router as he points to it. If the referent is potentially ambiguous, i.e., it is a reference to one of the concurrently active foci, then the Deictic Planer must decide if locomotion is needed. If no locomotion is needed, the agent will point at R without moving towards it. In contrast, if any of the following three conditions hold, the agent will move towards R as it points:

[6] More precisely, the agent will not perform a locomotive behavior. In fact, for purposes of believability, the agent is always in subtle but constant motion. COSMO, for example, typically performs his "anti-gravity bobbing" and blinking behaviors.

- *Multiple Proximal Foci:* If the object specified by R is near another object that is also in focus, the agent will move to the object specified by R. For example, if two nearby routers are being compared, Cosmo will move to the router to which he is referring to ensure that his reference is clear.

- *Multiple Proximal Similarity:* Associated with each object is an ontological category. If the object specified by R is near other objects of the same category, the agent will move to the object specified by R. For example, if Cosmo were referring to a router and there were several routers nearby, he would move to the intended router.

- *Diminutiveness:* If the object specified by R is unusually small, the agent will move to the object specified by R. Small objects are labeled as such in the world model. For example, many interface control buttons are relatively small compared to objects in the environment. If Cosmo needs to make a clear reference to one of them, he will move toward that button.

After a sequence of high-level gesture and locomotive behaviors are computed, they must be interpreted within the learning environment. For example, the current implementation of Cosmo provides for six basic pointing gestures: `left-up`, `left-across`, `left-down`, `right-up`, `right-across` and `right-down`. To enable the agent to correctly point to the object specified by the referent, the behavior planner first consults the world model. It obtains the location of the agent (L_A) and the referent (L_R) in the environment. It then determines relative orientation of the vector from (L_A) to (L_R). For example, Cosmo might be hovering in the lower-left corner of the environment and need to point to a router in the upper-right corner. In this case, he will point up and to his left towards the router using his `left-up` gesture (Figure 5, Step 1).

The behavior planner must then determine whether or not the agent really needs to move based on his current location. If it determines that locomotion is called for, the interaction manager must first determine if the agent is already near the object, which would obviate the need for moving towards it. Nearness of two objects is computed by measuring the distance between them and ascertaining whether it is less than a *proximity bound.* If the distance between the agent and the intended object is less than the proximity bound, then there is no need for the agent to move because it can already clearly point to the object, and so it will remain in its current position.

If locomotion is appropriate, the behavior planner computes a direct motion path from the agent's current location to the object specified by R. To do so, it first determines the *deictic target*, which is the precise location in the world at which the agent will point (Figure 5, Step 1). To avoid ambiguity, the agent will move its finger (or, more generally, its deictic pointer) toward the center of referent.[7] It then computes the direction of the vector defining the agent's direction of

[7] This is determined by computing the center of the referent's bounding box.

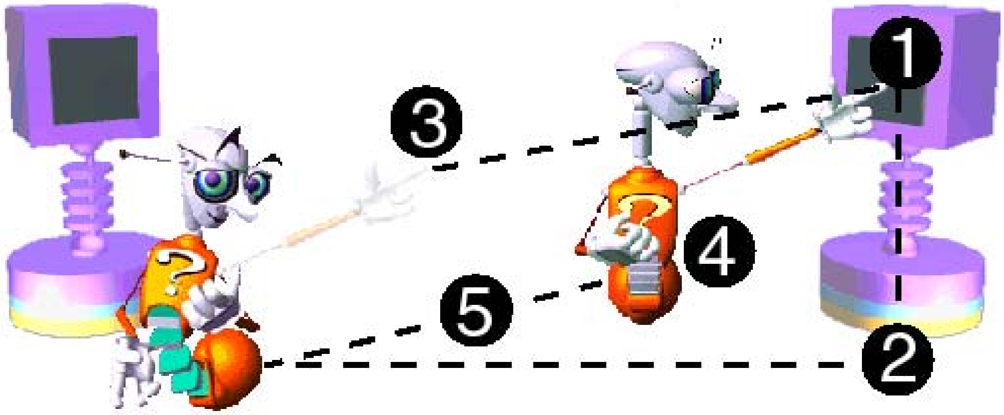


Figure 5: Determining vectors of locomotion

travel from L_A and to the deictic target (Figure 5, Step 2). To do so, it first determines the position of its finger if it were extended in the direction computed in Step 2 with the agent in L_A (Figure 5, Step 3). It then determines the ideal location of the agent's body position if its outstretched finger were to touch the deictic target (Figure 5, Step 4) in the final position. Finally, it traverses the resulting motion path connecting L_A and its final body position and location (Figure 5, Step 5).

3.4 Deictic Referring Expression Planning and Coordination

To effectively communicate the intended reference, the deictic system must combine gesture, locomotion, and speech. Having completed gesture and locomotion planning, the deictic planner turns to speech. To determine an appropriate referring expression for the agent to speak as it performs the deictic gestures and locomotion, the deictic system first examines the results of the ambiguity appraisal. If it was determined that R is in unique focus, there is no potential for ambiguity because R has already been introduced and no other entities are competing for the student's attention. It is therefore introduced with a simple referring expression using techniques similar to those outlined in [Dale, 1992]. For example, "the router" will be pronominalized to "it."

In contrast, if R is ambiguous or potentially ambiguous, i.e., R is a novel reference or is one of multiple foci, the deictic planner makes three assessments: (1) it determines the demonstrative category called for by the current situation; (2) it examines the ontological type of R and the other active foci; and (3) it considers the number of R. It first categorizes the situation into one of two deictic families:

- *Proximal Demonstratives:* If the deictic planner determined that the agent must move to R or that it would have moved to R if it were not already near R, then employ a proximal demonstrative such as "this" or "these."

- *Non-Proximal Demonstratives:* If the deictic planner determined that R was not nearby but that the agent did not need to move to R, then employ a non-proximal demonstrative such as "that" or "those."

After it has determined which of the demonstrative categories to use, the deictic planner narrows its selection further by considering the ontological type of R and the previous two utterances in the evolving explanation plan. If R belongs to the same ontological type as the other entities which are in focus, then the deictic planner selects the phrase, "This one ...." For example, suppose the system has determined that a proximal demonstrative should be used and that the preceding utterance referred to one router, e.g., "This router has more traffic." To refer to a second router in the current utterance, rather than saying, "This router has less traffic," it will say, "This one has less traffic." Finally, it uses the number of R to make the final lexical choice. If R is singular, it uses "this" for proximal demonstratives and "that" for non-proximals. If R is plural, it uses "these" and "those." The resulting referring expression is then passed onto the behavior planner for the final phase.

To integrate the agent's physical behaviors and speech, the behavior planner then coordinates the selected utterances, gestures, and locomotion. Three types of coordination must be achieved. (1) Each utterance may be accompanied by a deictic gesture, and it is critical that the agent's referring expressions be tightly coupled to its corresponding pointing movements. (2) Pointing and locomotion should be carefully coordinated so that they occur in a natural manner, where "natural" suggests that the agent should perform its pointing gesture *en route* to the referent and arrive at the referent at precisely the same time that it reaches the apex of the pointing gesture. (3) When the agent exhibits a sequence of speech, gestural, and locomotive behaviors to communicate an explanation, the behavior planner must ensure that each cluster of utterances, gestures, and possible agent movements are completed before the next is initiated. The behavior planner enacts the coordination by specifying that the utterance be initiated when the agent reaches the apex of its pointing gesture. In contrast, if the speech were initiated at the same time as the gesture and locomotion, the utterance would seem to complete prematurely, thereby producing both ambiguity and the appearance of incongruous behavior.

Finally, to underscore the deictic gestures, the behavior planner introduces gaze into the final behavior. As demonstrated by Cassell's incorporation of a gaze generator in her conversational agents [Cassell et al., 1994a], gaze offers an important communication medium for acknowledgements and turn taking. In addition, gaze can play an important role in deixis. For example, when COSMO refers to a particular computer on a subnet by moving towards it and pointing at it as he speaks about it, he should also look at it. The behavior planner enacts gaze via specifications for the agent to "look" at the referent by moving its head in precisely the direction in which it is pointing.[8]

The behavior planner combines the speech, gesture, locomotion, and gaze specifications and directs the agent to perform them in the order dictated by the explanation plan. The agent's behaviors are then assembled and sequenced in the learning environment in realtime to provide students with clear advice that couples full deictic expression with integrated lifelike locomotion, gesture, speech, and gaze.

3.5 Handling Advanced Deictic Phenomena

In addition to the deictic behavior planning techniques discussed above, the framework also supports (1) virtual deixis and (2) multiple intrasentential deixis. *Virtual deixis* is the phenomenon of combining locomotion, gesture, speech, and gaze to indicate a referent that is not visible in the current environment but exists elsewhere in the world and can be referred to via an intermediate artifact. For example, in performing an Assistance act, COSMO frequently must refer to an object that is not onscreen but exists elsewhere and is accessible by using the navigation tool in the lower left corner of the interface. For example, he might suggest that, "We want to choose a subnet with low traffic," while moving down and pointing to a quadrant on the navigation spinner that represents the subnet to which he refers (Figure 6). The student can then click on the specified quadrant of the spinner and go the recommended subnet.

By exploiting knowledge about the relationship between the offscreen entities (e.g., subnets) and their onscreen representations (e.g., quadrants on the navigation spinner), the behavior planner enables the agent to employ virtual deixis in the same manner that it coordinates other deictic behaviors. Computationally, this is accomplished by introducing a "Step (0)" to the algorithm. In this step, if R_g is not currently onscreen, a representational substitution for R_g is made whereby an onscreen object replaces the original R_g. This substitute object is determined by annotations in the world model which indicate legal representational substitutions. To illustrate, in the above example, quadrants of the navigational spinner in the interface represent offscreen subnets, so when

[8] The direction in which an agent's eyes focus play an important role in signaling its interest. In the implementation, the behavior planner accomplishes this not through runtime inference of eye control but by exploiting agent head rendering in which the eyes were crafted by the animators to look in the direction in which the head is pointing, e.g., if the head is turned toward the right, the eyes look towards the right. The USC/ISI animated agents group has been successfully experimenting with similar gaze techniques such as "leading with the eyes" [Johnson and Rickel, 1997].

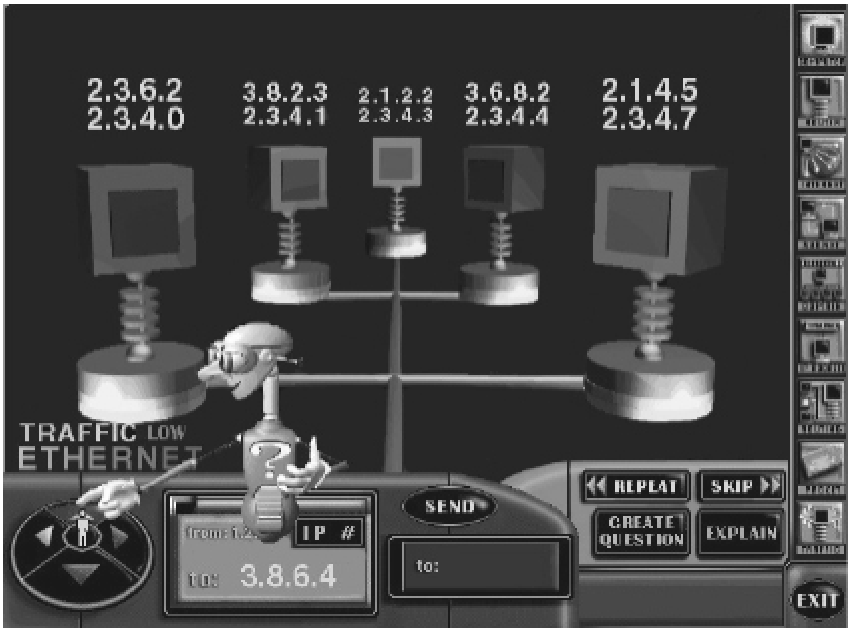


Figure 6: Cosmo achieving virtual deixis

a particular subnet requires a reference but does not appear onscreen, a virtual deictic reference to it can be constructed in which the agent gestures to the appropriate spinner quadrant (Figure 6).

The deictic planner also supports *multiple intrasentential deixis*. For example, when Cosmo explains, "This subnet has high traffic," he can refer in speech to the particular subnet (the first referent) and then point to the traffic information associated with that subnet (the second referent). To handle multiple intrasentential deictics, the initial, single representation of the focus history was extended to include both a gestural focus history and an utterance focus history. The gestural focus history is used to track objects to which the agent has recently pointed, while the utterance focus history is used to track objects about which the agent has recently spoken. In the example above, the utterance focus history is updated to record the spoken deictic reference (i.e., the particular subnet), and the gestural history is updated to record the gestural deictic reference (i.e., traffic information). By maintaining the dual histories, the behavior planner avoids ambiguities which would otherwise arise. The twin focus histories and modes of intrasentential deictic expression increase agents' deictic flexibility and permits them to generate more lifelike natural behaviors and language.

4 An Implemented Lifelike Pedagogical Agent

Cosmo (Figure 1) is a realtime implementation of a lifelike animated agent that has a head with movable antennae and expressive blinking eyes, arms with bendable elbows, hands with a large number of independent joints, and a body with an accordion-like torso. The student interacts with Cosmo and the learning environment via the Internet Advisor's interface (Figure 1). As she attempts to route her packet to a given destination, she makes a series of routing decisions to direct the packet's hops through the network. At each router, she is given four different subnets, each with five possible computers with unique addresses from which to choose. She is also provided information about the type of the subnet and the amount of traffic on the subnet. In the lower left hand corner of the interface, she can click on different quadrants of a spinner to navigate between the four possible attached subnets. When she has found what she believes to be a reasonable computer to send her packet towards, she clicks on the address of the computer. Cosmo then comments on the correctness and optimality of her decision. If it is either incorrect or sub-optimal, he provides assistance on how to improve it. If her decision was deemed optimal, he congratulates her, and she clicks on the "Send" button to send her packet to the next subnet in the network.

Cosmo's deictic planner is implemented in the Clips production system language [NASA, 1993]. His explanation planner and the Internet Advisor learning environment are implemented in C++, and the interaction manager employs the Microsoft Game Software Developer's Kit (SDK). Cosmo's behaviors run at 15 frames/second with 16 bits/pixel color on a Pentium Pro 200 Mhz PC with 80 MB of RAM.[9] His speech was created by a trained voice actor and an audio engineer.

Given a request to explain a concept or to provide a hint, the behavior planner selects the explanatory content by examining the world model, the curriculum information network, the user model, the problem state (which encodes knowledge about the current packet's destination address, subnet type, IP numbers for the computers and routers on the current subnet, and network congestion), and the two focus histories. When invoked, the planner first consults the knowledge sources noted above to select a sequence of communicative acts. These acts include the following:

- State-Correct: Provides information about the factors of the student's selection which were correct, e.g., showing which fields of an address match. A State-Correct act may require deictic behaviors to identify objects in the learning environment that play a role in a correct student solution.

- State-Incorrect: Provides information about the factors of the student's selection which were incorrect, e.g., showing that a selected subnet has high traffic. A State-Incorrect act may require deictic behaviors to identify objects in the learning environment about which the student may have misconceptions or have failed to recognize or consider.

- Cause: Poses a rhetorical question to the student with regards to what would happen if her choice were implemented. By explaining the theoretical background of the domain in terms

[9] The additional memory is employed to avoid load delays for the agent's images by keeping them resident.

of a concrete problem-solving situation, the agent grounds the principles that govern the domain.

- Effect: Answers the rhetorical question just posed by the agent in the Cause utterance with regards to the current problem, e.g., telling the student that the packet will move through the network slowly.
- Rationale: Provides detail on why the student's choice was incorrect, e.g. showing the student that the current subnet has high traffic.
- Give-Background: Provides foundational information about entities in the domain, e.g., an explanation of a router's function.
- Assistance: Gives a hint, e.g., to try a subnet with lower traffic. After a student has repeatedly demonstrated difficulty with a particular concept, the agent provides advice which is accompanied by deictic gestures, locomotion, and gaze to unambiguously refer to objects in the optimal solution.

Cosmo can perform a variety of behaviors including pointing, blinking, leaning, clapping, locomotion, and raising and bending his antennae. His verbal behaviors include 200 utterances ranging in duration from 1–20 seconds. He was modeled and rendered in 3D on SGIs with Alias/Wavefront. The resulting bitmaps were subsequently post-edited with Photoshop and AfterEffects on Macintoshes and transferred to PCs where users interact with them in a $2\frac{1}{2}$D environment. Cosmo's behaviors are assembled in realtime as directed by the behavior planner. Each action is annotated with the number of frames and transition methods. Actions are of two types: *full-body* behaviors, in which the agent's entire body is depicted, and *compositional* behaviors that represent various body parts individually. To sequence non-deictic behaviors such as clapping and leaning, the behavior planner employs the full-body images. To sequence deictic behaviors, including both the gesture and gaze, the behavior planner combines compositional behaviors of torsos, left and right arms, and heads (Figure 7).[10]

To illustrate how the behavior planner produces deictic gesture, motion, and speech as it provides problem-solving assistance in realtime, consider the following situation in an Internet Advisor learning session. Suppose a student has just routed her packet to a fiber optic subnet with

[10] To produce the highest quality onscreen presence for the agent, the presentation manager first draws each image of the agent, whether it is full-body or composed from atomic components, to its own offscreen surface. It also employs both a primary (visible) surface and a secondary (hidden) surface for displaying images to the screen. Double buffering is implemented by blitting images from the offscreen surfaces to the secondary surface. Then the secondary and primary surfaces are exchanged so the next frame is displayed. In practice, this technique significantly decreases the amount of flicker during animations.

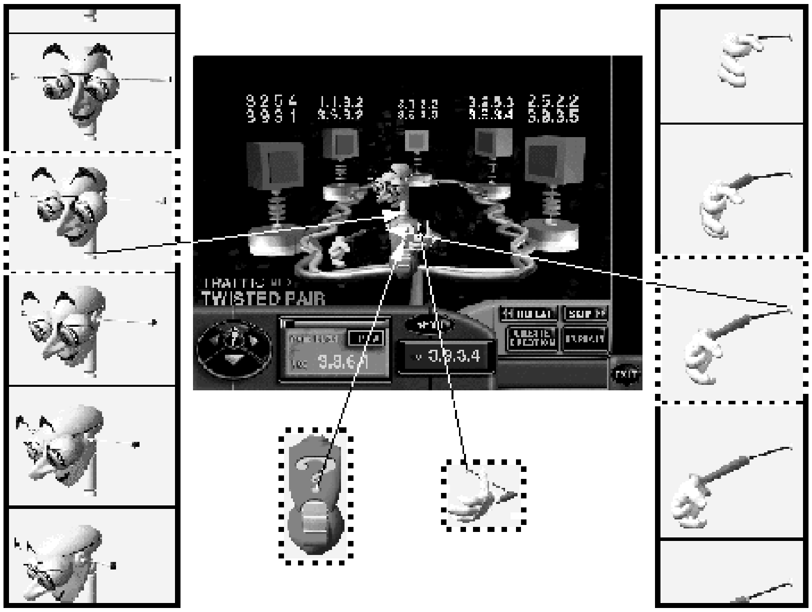


Figure 7: Realtime composition of agent body components

low traffic. She surveys the connected subnets and selects a router which she believes will advance it one step closer to the packet's intended destination. Although she has chosen a reasonable subnet, it is suboptimal because of non-matching addresses, which will slow her packet's progress. She has made a couple of mistakes on address resolution already, and so the explanation is fairly detailed. The behavior planner selects and sequences the following communicative acts and orchestrates the agent's gestural, locomotive, and speech behaviors as indicated in (Figure 8 and shown in detail in the Appendix):

1. `State-Correct(Subnet-Type)`: The explanation planner determines that the agent should interject advice and invokes the deictic planner. Since nothing is in focus because it is planning the first utterance of a new explanation and Cosmo currently occupies a position on the screen far from information about the subnet, i.e., the distance from his current location to the subnet information exceeds the proximity bound, he moves towards and points at the onscreen subnet information and says, "You chose the fastest subnet."

2. `State-Correct(Traffic)`: Cosmo then tells the student that the choice of a low traffic subnet was also a good one. The gesture focus history indicates that, while the type of subnet has already been the subject of a deictic reference, the traffic information has not. Cosmo therefore moves to the onscreen congestion information and points to it. However, the utterance focus history indicates that he has mentioned the subnet in a recent utterance, he pronominalizes the subnet as "it" and says, "Also, it has low traffic. Fabulous!"

3. `Cause()`: Because Cosmo wants the student to rethink her choice, he scratches his head and poses the question, "But more importantly, if we sent the packet here, what will happen?" Since this is a non-deictic act, no modifications are made to the deictic focus histories other than pushing `nil`s onto each of the stacks.[11]

4. `Effect(Address-Resolution)`: Cosmo tells the student of the ill-effect of chosing that router, and saying dejectedly, "If that were the case, we see it doesn't arrive at the right place." Because this does not impact the deictic context, `nil`s are pushed onto each of the focus histories.

5. `Rationale(Address-Resolution)`: To explain the reason why the packet won't arrive at the correct destination, Cosmo adds, "This computer has no parts of the address matching." Because the computer that serves as the referent is currently not in the focus histories and Cosmo is far from that computer, the behavior planner sequences deictic locomotion and a gesture to accompany the utterance.

6. `Background(Address-Resolution)`: To emphasize the previous remark, the behavior planner adds a background utterance: "Addresses are used by networked computers to tell each other apart." The deictic planner is not invoked so `nil`s are pushed onto each of the focus histories.

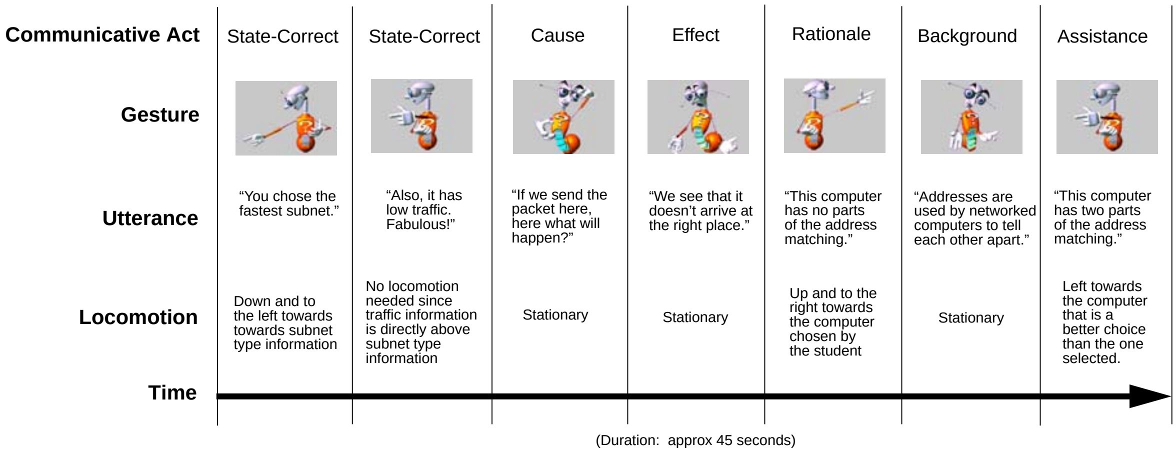
7. `Assistance(Address-Resolution)`: Finally, Cosmo assists the student by making a suggestion about the next course of action to take. Because the student has committed several mistakes on address resolution problems, Cosmo provides advice about correcting her decision by pointing to the location of the optimal computer—it has not been in focus—and stating, "This router has two parts of the address matching."

5 Evaluation and Discussion

5.1 Focus Group Study

To gauge the effectiveness of the spatial deixis framework for deictic believability in animated pedagogical agents, an informal focus group study was conducted with students interacting with Cosmo in the Internet Advisor learning environment. The exploratory study was designed to investigate (1) how well the spatial deixis approach produces explanations that are clear and helpful and (2) how an agent-based approach to deixis in learning environments compares with

[11] The computational framework for sequencing non-deictic (in this case, *emotive*) behaviors is discussed in [Towns et al., in press].



Communicative Act	State-Correct	State-Correct	Cause	Effect	Rationale	Background	Assistance
Gesture							
Utterance	"You chose the fastest subnet."	"Also, it has low traffic. Fabulous!"	"If we send the packet here, here what will happen?"	"We see that it doesn't arrive at the right place."	"This computer has no parts of the address matching."	"Addresses are used by networked computers to tell each other apart."	"This computer has two parts of the address matching."
Locomotion	Down and to the left towards towards subnet type information	No locomotion needed since traffic information is directly above subnet type information	Stationary	Stationary	Up and to the right towards the computer chosen by the student	Stationary	Left towards the computer that is a better choice than the one selected.
Time							

(Duration: approx 45 seconds)

Figure 8: Coordinating deictic gesture, locomotion, and speech

a non-agent-based approach. Isolating modal deictic phenomena and their communicative effects poses a significant challenge in studies of deixis. In particular, the investigators sought to tease apart the effects of verbal deictics from physical deictics, i.e., to separate the effects of utterances from those of gesture, locomotion, and gaze. To address the multiple modes of expression, two versions of the Internet Advisor were created:

- *Agent-based learning environment*: In the agent-based version, as students solved Internet routing problems, the agent's behavior planner selected and coordinated locomotive, gestural, and speech behaviors according to the spatial deixis framework.
- *Agent-free learning environment*: In the agent-free version, students solved the same type of Internet routing problems, but no agent was present. Rather, a disembodied advisor spoke the same advice as in the agent-based version. Because the agent was absent from the environment, no deictic gesture, locomotion, or gaze were employed. However, to create a situation that was more comparable to the agent-based version by compensating for the missing agent's functionalities, the agent-free system flashed a blinking red line under the referent in the environment each time a deictic reference was created.

The subjects of the study were 7 men and 3 women. To obtain a broad spectrum of responses, subjects with a broad range of ages (14–54) were chosen. On average, each subject interacted with the Internet Advisor for 30 minutes. To avoid an ordering bias, approximately half the subjects first interacted with the agent-based environment and then interacted with the agent-free environment; the other half of the subjects first interacted with the agent-free and then with the agent-based environment.

Bearing in mind that the study was very informal, the results suggest that the spatial deixis framework produces clear explanations. Based on the subjects' comments and actions in the course of problem solving, it appears that most participants understood the agent's advice most of the time. Although some subjects expressed a desire for an agent that was less dramatic and some suggested alternative organizations for the communicative acts, the agent's clarity of expression was positively received.

5.2 Discussion

In interpreting the results of the study, it is important to note the limitations of both the spatial deixis framework in general and the implemented deictic behavior planner in particular. First, the deictic planner does not deal with integrating multiple types of gestures, e.g., metaphoric and beat gestures, as does Cassell's framework [Cassell et al., 1994a], so no conclusions can be drawn about accommodating the full repertoire of gestural behaviors. Second, the deictic behavior planner selects referring expressions, but it doesn't address issues of prosody. Because the planner has no means of reasoning about the contours of intonation patterns, the evaluation does not address the issues of marking spoken referring expressions for phrase boundaries or emphasis. Finally, the implementation operates in a monitoring/interjection mode but does not support interrogation by the learner. Because learners cannot pose questions to the system, communication is more limited than is desirable.

The two most salient findings of the study pertain to agents' clarity of communication and their compelling presence. Based on subjects' successful interactions with COSMO in the INTERNET ADVISOR learning environment, it appears that agent-based environments clearly communicate advice, though not necessarily more clearly than agent-free environments. This is consistent with the findings of André *et al.*'s study which revealed no significant differences in the comprehension of technical material between an agent-based and an agent-free study [André et al., 1998]. Interestingly, some subjects suggested that a combination of agent gestures with the blinking indicators might be more effective than either in isolation, a preference supported by a growing body of HCI evidence on multimodal interfaces, e.g., [Oviatt, 1997].

Perhaps most telling was the subjects' unanimous preference for the agent-based version over the agent-free version. In general, especially given the age of the subjects, the agent's motivating role was surprisingly strong. This finding is consistent with the *persona effect* [Lester et al., 1997a], in which the very presence of a lifelike character in an interactive learning environment can have a strong positive effect on learners' perception of their learning experience.

6 Conclusions and Future Work

Because of their strong lifelike presence, animated pedagogical agents can capture students' imaginations and play a critical motivational role in keeping them deeply engaged in a learning environment's activities. Believability is a key feature of lifelike pedagogical agents, and deictic believability is an important characteristic of animated agents that inhabit artificial worlds. To dynamically sequence lifelike pedagogical agents in a manner that promotes deictic believability, agent behavior planners can employ the spatial deixis framework for coordinating gesture, locomotion, and speech. This framework has been used to implement Cosmo, a lifelike pedagogical agent for a testbed learning environment.

In this framework, an agent behavior planner evaluates potential ambiguities and exploits a world model representing position and orientation in the virtual world to plan the agent's deictic actions and utterances. To do so, it first examines the evolving discourse plan to ascertain the focus status of the referent. It then inspects the world model to determine the referent's proximity to similar objects and to the agent itself. In this manner, it determines whether to move and point to the referent and what type of demonstrative utterance the agent should use to indicate it. Finally, the behavior planner integrates the gestures, locomotion, and speech into a communicative acts specification that produces a seamless sequence of utterances and actions that are unambiguous and believable. The net result of the behavior planning is a lifelike character who provides clear problem-solving advice in realtime with a strong visual presence.

Deictic behavior planning is a critical component of lifelike pedagogical agents, and this work represents a promising first step towards the goal of creating enchanting characters for learning environments. Nevertheless, the field of lifelike pedagogical agents is still in its infancy and four lines of investigation appear particularly worthwhile and challenging. First, endowing pedagogical agents with sophisticated models of emotion [Elliott, 1992, Velasquez, 1997] may yield important communication and motivational benefits. Second, integrating state-of-the-art work on computational models of conversation-based, task-oriented dialogue [Walker, 1993, Smith and Hipp, 1994, Traum, 1994, Guinn, 1995, Freedman, 1996] will significantly broaden the bandwidth of student-agent communication. Third, providing pedagogical agents with full-scale realtime natural generation will yield important increases in communicative flexibility. Finally, developing techniques for marrying lifelike pedagogical agent technologies with 3D explanation generators [Karp and Feiner, 1993, Bares and Lester, 1997, Butz, 1997, Zhou and Feiner, 1997] may create a qualitatively better form of learning environment. We will be exploring these directions in our future work.

Acknowledgements

Thanks to Dorje Bellbrook, Tim Buie, Mike Cuales, Jim Dautremont, Amanda Davis, Rob Gray, Mary Hoffman, Alex Levy, Will Murray, and Roberta Osborne of the North Carolina State University IntelliMedia Initiative for their work on 3D modeling, animation, sound, and environment design for the Internet Advisor. Special thanks to Patrick FitzGerald of North Carolina State University's School of Design for leading this effort. The authors gratefully acknowledge the anonymous reviewers for their detailed and helpful suggestions. Support for this work was provided by the following organizations: the National Science Foundation under grants CDA-9720395 (Learning and Intelligent Systems Initiative) and IRI-9701503 (CAREER Award Program); the North Carolina State University IntelliMedia Initiative; Novell, Inc.; and equipment donations from Apple and IBM.

References

[André et al., 1993] André, E., Finkler, W., Graf, W., Rist, T., Schauder, A., and Wahlster, W. (1993). WIP: The automatic synthesis of multi-modal presentations. In Maybury, M. T., editor, *Intelligent Multimedia Interfaces*, chapter 3. AAAI Press.

[André and Rist, 1996] André, E. and Rist, T. (1996). Coping with temporal constraints in multimedia presentation planning. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 142–147.

[André et al., 1998] André, E., Rist, T., and Müller, J. (1998). Integrating reactive and scripted behaviors in a life-like presentation agent. In *Proceedings of the Second International Conference on Autonomous Agents*, pages 261–268. Minneapolis.

[Bares and Lester, 1997] Bares, W. H. and Lester, J. C. (1997). Realtime generation of customized 3D animated explanations for knowledge-based learning environments. In *AAAI-97: Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 347–354, Providence, Rhode Island.

[Bates, 1994] Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125.

[Blumberg and Galyean, 1995] Blumberg, B. and Galyean, T. (1995). Multi-level direction of autonomous creatures for real-time virtual environments. In *Computer Graphics Proceedings*, pages 47–54.

[Butz, 1997] Butz, A. (1997). Anymation with CATHI. In *Proceedings of the Ninth Innovative Applications of Artificial Intelligence Conference*, pages 957–62.

[Carr and Goldstein, 1977] Carr, B. and Goldstein, I. P. (1977). Overlays: A theory of modelling for computer aided instruction. Technical Report AI Memo 406, Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

[Cassell et al., 1994a] Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., and Stone, M. (1994a). Modeling the interaction between speech and gesture. In *SIGGRAPH '94*.

[Cassell et al., 1994b] Cassell, J., Steedman, M., Badler, N., Pelachaud, C., Stone, M., Douville, B., Prevost, S., and Achorn, B. (1994b). Modelling the interaction between speech and gesture. Technical report, University of Pennsylvania.

[Cawsey, 1992] Cawsey, A. (1992). *Explanation and Interaction: The Computer Generation of Explanatory Dialogues.* MIT Press.

[Claassen, 1992] Claassen, W. (1992). Generating referring expressions in a multimodal environment. In Dale, R., Hovy, E., Rosner, D., and Stock, O., editors, *Aspects of Automated Natural Language Generation*, pages 247–62. Springer-Verlag, Berlin.

[Conati et al., 1997] Conati, C., Gertner, A., VanLehn, K., and Druzdzel, M. (1997). On-line student modeling for coached problem solving using Bayesian networks. In *Proceedings of the Sixth International Conference on User Modeling*, pages 231–242.

[Dale, 1992] Dale, R. (1992). *Generating Referring Expressions.* MIT Press.

[Elliott, 1992] Elliott, C. (1992). *The Affective Reasoner: A Process Model of Emotions in a Multi-agent System.* PhD thesis, Northwestern University.

[Feiner and McKeown, 1990] Feiner, S. K. and McKeown, K. R. (1990). Coordinating text and graphics in explanation generation. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 442–449, Boston, MA.

[Fillmore, 1975] Fillmore, C. (1975). *Santa Cruz Lectures on Deixis 1971.* Available from Indiana University Linguistics Club.

[Freedman, 1996] Freedman, R. K. (1996). *Interaction of Discourse Planning, Instructional Planning and Dialogue Management in an Interactive Tutoring System.* PhD thesis, Northwestern University.

[Granieri et al., 1995] Granieri, J. P., Becket, W., Reich, B. D., Crabtree, J., and Badler, N. I. (1995). Behavioral control for real-time simulated human agents. In *Proceedings of the 1995 Symposium on Interactive 3D Graphics*, pages 173–180.

[Guinn, 1995] Guinn, C. I. (1995). *Meta-Dialogue Behaviors: Improving the Efficiency of Human-Machine Dialogue – A Computational Model of Variable Initiative and Negotiation in Collaborative Problem-Solving.* PhD thesis, Duke University.

[Hovy, 1993] Hovy, E. H. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63:341–385.

[Jarvella and Klein, 1982] Jarvella, R. J. and Klein, W., editors (1982). *Speech, Place, and Action: Studies in Deixes and Related Topics.* John Wiley & Sons.

[Johnson and Rickel, 1997] Johnson, L. and Rickel, J. (1997). Personal communication.

[Karp and Feiner, 1993] Karp, P. and Feiner, S. (1993). Automated presentation planning of animation using task decomposition with heuristic reasoning. In *Proceedings of Graphics Interface '93*, pages 118–127.

[Kurlander and Ling, 1995] Kurlander, D. and Ling, D. T. (1995). Planning-based control of interface animation. In *Proceedings of CHI '95*, pages 472–479.

[Lester et al.,] Lester, J., Stone, B., and Stelling, G. Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction*. In press.

[Lester et al., 1997a] Lester, J. C., Converse, S. A., Kahler, S. E., Barlow, S. T., Stone, B. A., and Bhogal, R. (1997a). The persona effect: Affective impact of animated pedagogical agents. In *Proceedings of CHI'97 (Human Factors in Computing Systems)*, pages 359–366, Atlanta.

[Lester et al., 1997b] Lester, J. C., Converse, S. A., Stone, B. A., Kahler, S. E., and Barlow, S. T. (1997b). Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation. In *Proceedings of Eighth World Conference on Artificial Intelligence in Education*, pages 23–30, Kobe, Japan.

[Lester et al., 1997c] Lester, J. C., FitzGerald, P. J., and Stone, B. A. (1997c). The pedagogical design studio: Exploiting artifact-based task models for constructivist learning. In *Proceedings of the Third International Conference on Intelligent User Interfaces*, pages 155–162, Orlando, Florida.

[Lester and Porter, 1997] Lester, J. C. and Porter, B. W. (1997). Developing and empirically evaluating robust explanation generators: The KNIGHT experiments. *Computational Linguistics*, 23(1):65–101.

[Lester and Stone, 1997] Lester, J. C. and Stone, B. A. (1997). Increasing believability in animated pedagogical agents. In *Proceedings of the First International Conference on Autonomous Agents*, pages 16–21, Marina del Rey, California.

[Lester et al., 1996] Lester, J. C., Stone, B. A., O'Leary, M. A., and Stevenson, R. B. (1996). Focusing problem solving in design-centered learning environments. In *Proceedings of the Third International Conference on Intelligent Tutoring Systems*, pages 475–483, Montreal.

[Maes et al., 1995] Maes, P., Darrell, T., Blumberg, B., and Pentland, A. (1995). The ALIVE system: Full-body interaction with autonomous agents. In *Proceedings of the Computer Animation '95 Conference*.

[Maybury, 1991] Maybury, M. T. (1991). Planning multimedia explanations using communicative acts. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, pages 61–66, Anaheim, CA.

[Mittal et al., 1995] Mittal, V., Roth, S., Moore, J. D., Mattis, J., and Carenini, G. (1995). Generating explanatory captions for information graphics. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 1276–1283.

[Mittal, 1993] Mittal, V. O. (1993). *Generating Natural Language Descriptions with Integrated Text and Examples*. PhD thesis, University of Southern California.

[Moore, 1995] Moore, J. D. (1995). *Participating in Explanatory Dialogues*. MIT Press.

[NASA, 1993] NASA (1993). CLIPS reference manual. Technical report, Software Technology Branch, Lyndon B. Johnson Space Center, NASA.

[Nass et al., 1995] Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, D. C. (1995). Can computer personalities be human personalities. *International Journal of Human-Computer Studies*, 43:223–239.

[Nass et al., 1993] Nass, C., Steuer, J., Henriksen, L., and Reeder, H. (1993). Anthropomorphism, agency and ethopoeia: Computers as social actors. In *Proceedings of the International CHI Conference*.

[Neal and Shapiro, 1991] Neal, J. G. and Shapiro, S. C. (1991). Intelligent multi-media interface technology. In Sullivan, J. W. and Tyler, S. W., editors, *Intelligent User Interfaces*, pages 11–43. Addison-Wesley, New York.

[Novak, 1987] Novak, H.-J. (1987). Strategies for generating coherent descriptions of object movements in street scenes. In Kempen, G., editor, *Natural Language Generation*, pages 117–132. Martinus Nijhoff, Dordrecht, The Netherlands.

[Oviatt, 1997] Oviatt, S. (1997). Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction*, 12:93–129.

[Reeves and Nass, 1992] Reeves, B. and Nass, C. (1992). Information characteristics of media technologies that give a sense of “being there”. In *Annual Meeting of the International Communication Association*.

[Rickel and Johnson, 1997a] Rickel, J. and Johnson, L. (1997a). Integrating pedagogical capabilities in a virtual environment agent. In *Proceedings of the First International Conference on Autonomous Agents*, pages 30–38.

[Rickel and Johnson, 1997b] Rickel, J. and Johnson, L. (1997b). Intelligent tutoring in virtual reality: A preliminary report. In *Proceedings of the Eighth World Conference on AI in Education*, pages 294–301.

[Roberts, 1993] Roberts, L. D. (1993). *How Reference Works: Explanatory Models for Indexicals, Descriptions, and Opacity.* State University of New York Press.

[Roth et al., 1991] Roth, S. F., Mattis, J., and Mesnard, X. (1991). Graphics and natural language as components of automatic explanation. In Sullivan, J. W. and Tyler, S. W., editors, *Intelligent User Interfaces*, pages 207–239. Addison-Wesley, New York.

[Sibun, 1992] Sibun, P. (1992). Generating text without trees. *Computational Intelligence*, 8(1):102–122.

[Smith and Hipp, 1994] Smith, R. W. and Hipp, D. R. (1994). *Spoken Natural Language Dialog Systems.* Oxford University Press, Cambridge, Massachusetts.

[Stone and Lester, 1996] Stone, B. A. and Lester, J. C. (1996). Dynamically sequencing an animated pedagogical agent. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 424–431, Portland, Oregon.

[Suthers, 1991] Suthers, D. D. (1991). A task-appropriate hybrid architecture for explanation. *Computational Intelligence*, 7(4):315–333.

[Towns et al.,] Towns, S. G., FitzGerald, P. J., and Lester, J. C. Visual emotive communication in lifelike pedagogical agents. In *Proceedings of the Fourth International Conference on Intelligent Tutoring Systems.* In press.

[Traum, 1994] Traum, D. (1994). *A Computational Theory of Grounding in Natural Language Generation.* PhD thesis, University of Rochester.

[Tu and Terzopoulos, 1994] Tu, X. and Terzopoulos, D. (1994). Artificial fishes: Physics, locomotion, perception, and behavior. In *Computer Graphics Proceedings*, pages 43–50.

[Velasquez, 1997] Velasquez, J. D. (1997). Modeling emotions and other motivations in synthetic agents. In *AAAI-97: Proceedings of the Fourteenth National Conference on Artificial Intelligence*, pages 10–15.

[Walker, 1993] Walker, M. (1993). *Informational Redundancy and Resource Bounds in Dialogue.* PhD thesis, University of Pennsylvania.

[Walker et al., 1997] Walker, M. A., Cahn, J. E., and Whittaker, S. J. (1997). Improvising linguistic style: Social and affective bases of agent personality. In *Proceedings of the First International Conference on Autonomous Agents.*

[Wescourt et al., 1981] Wescourt, K. T., Beard, M., and Barr, A. (1981). Curriculum information networks for CAI: Research on testing and evaluation by simulation. In Suppes, P., editor, *University-Level Computer-Assisted Instruction at Stanford: 1968–1980*, pages 817–839. Stanford University, Stanford, California.

[Zhou and Feiner, 1997] Zhou, M. X. and Feiner, S. K. (1997). Top-down hierarchical planning of coherent visual discourse. In *Proceedings of the Third International Conference on Intelligent User Interfaces.*

Appendix

The following trace shows the results of the deictic behavior planner's activity as it creates gesture, locomotion, speech, and gaze in response to the single communicative goal of correcting a misconception. The structures associated with each of the seven communicative acts it generates are expressed below in a syntactic sugar that demonstrates the effects of behavior planning for that particular communicative act. The structures shown includes both the relevant knowledge that is given to the deictic planner, as well as the structures produced as a result of firing the set of CLIPS operators representing each of the communicative acts.

The `topic`, `cat-name`, `point`, and `talk` sub-structures represent the content topic, communicative act category, gesture referent, and speech referent, respectively. These constitute the information given to the deictic planner. Next, structures such as `point-ambig` and `talk-ambig` represent the results of ambiguity appraisals of the gesture and speech referents. The `move` sub-structures represent the results of locomotion decision making, and `number` and `focus` sub-structures are used in utterance planning. Finally, the `focus-stacks` pairs each have a `point` stack, representing the current contents of the gestural focus history, and a `talk` stack, representing the current contents of the speech focus history. Utterances planned for both deictic and non-deictic acts are also shown.

```
(1) FIRE:  get-state-correct

=> (deixes
     (topic subnet-type) (cat-name state-correct)
     (point subnet-information) (talk subnet-1)
     (point-ambig novel) (talk-ambig novel)
     (move movepoint) (number singular) (focus yes)
     (says "you chose the fastest subnet"))
=> (focus-stacks
     (point subnet-information)
     (talk subnet-1))

(2) FIRE:  get-state-correct

=> (deixes
     (topic cong) (cat-name state-correct)
     (point subnet-information) (talk subnet-1)
     (point-ambig novel) (talk-ambig unique)
     (move movepoint) (number singular) (focus yes)
     (says "also, it has low traffic"))
```

```
=> (focus-stacks
     (point traffic-information subnet-information)
     (talk subnet-1 subnet-1))

(3) FIRE:  get-cause

=> (non-deixes
     (says "suppose we sent the packet here,
          what will happen?")
     (action scratch-head))
=> (focus-stacks
     (point nil traffic-information subnet-information)
     (talk nil subnet-1 subnet-1))

(4) FIRE:  get-effect

=> (non-deixes
     (says "if that were the case, we see that it
           doesn't arrive at the right place")
     (action disappointed))
=> (focus-stacks
     (point nil nil traffic-information subnet-information)
     (talk nil nil subnet-1 subnet-1))


(5) FIRE: get-background

=> (non-deixes
     (says "addresses are used by networked computers to
          tell each other apart")
     (action hand-wave))
=> (focus-stacks
     (point nil nil nil traffic-information subnet-information)
     (talk nil nil nil subnet-1 subnet-1))

(6) FIRE:  get-rationale

=> (deixes
     (topic address-resolution) (cat-name rationale)
     (p-ambig novel) (t-ambig novel)
     (point computer-6) (talk computer-6)
     (move movepoint) (number singular) (focus no)
     (says "this computer has no parts of the address
          matching"))
=> (focus-stacks
     (point computer-6 nil nil nil traffic-information
          subnet-information)
     (talk computer-6 nil nil nil subnet-1 subnet-1))

(7) FIRE:  get-detailed-hint

=> (deixes
     (topic addr) (cat-name detailed-hint)
     (point-ambig multi) (talk-ambig multi)
     (point computer-8) (talk computer-8)
     (move movepoint) (number singular) (focus yes)
     (says "this one does have one part of the address
          matching"))
=> (focus-stacks
     (point computer-8 computer-6 nil nil nil traffic-information
          subnet-information)
     (talk computer-8 computer-6 nil nil nil subnet-1 subnet-1))
```