Non-Verbal Cues for Discourse Structure

Justine Cassell†, Yukiko I. Nakano†, Timothy W. Bickmore†,

Candace L. Sidner‡, and Charles Rich‡

†MIT Media Laboratory
20 Ames Street
Cambridge, MA 02139
{justine, yukiko, bickmore}@media.mit.edu

‡Mitsubishi Electric Research Laboratories
201 Broadway
Cambridge, MA 02139
{sidner, rich}@merl.com

Abstract

This paper addresses the issue of designing embodied conversational agents that exhibit appropriate posture shifts during dialogues with human users. Previous research has noted the importance of hand gestures, eye gaze and head nods in conversations between embodied agents and humans. We present an analysis of human monologues and dialogues that suggests that postural shifts can be predicted as a function of discourse state in monologues, and discourse and conversation state in dialogues. On the basis of these findings, we have implemented an embodied conversational agent that uses Collagen in such a way as to generate postural shifts.

1. Introduction

This paper provides empirical support for the relationship between posture shifts and discourse structure, and then derives an algorithm for generating posture shifts in an animated embodied conversational agent from discourse states produced by the middleware architecture known as Collagen [18]. Other nonverbal behaviors have been shown to be correlated with the underlying conversational structure and information structure of discourse. For example, gaze shifts towards the listener correlate with a shift in conversational turn (from the conversational participants' perspective, they can be seen as a signal that the floor is available). Gestures correlate with rhematic content in accompanying language (from the conversational participants' perspective, these behaviors can be seen as a signal that accompanying speech is of high interest). A better understanding of the role of nonverbal behaviors in conveying discourse structures enables improvements in the naturalness of embodied dialogue systems, such as embodied conversational agents, as well as contributing to algorithms for recognizing discourse structure in speech-understanding systems. Previous work, however, has not addressed major body shifts during discourse, nor has it addressed the nonverbal correlates of topic shifts.

2. Background

Only recently have computational linguists begun to examine the association of nonverbal behaviors and language. In this section we review research by non-computational linguists and discuss how this research has been employed to formulate algorithms for natural language generation or understanding.

About three-quarters of all clauses in descriptive discourse are accompanied by gestures [17], 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 [13]. It has been shown that when speech is ambiguous or in a speech situation with some noise, listeners rely on

gestural cues [22] (and, the higher the noise-to-signal ratio, the more facilitation by gesture). Even when gestural content overlaps with speech (reported to be the case in roughly 50% of utterances, for descriptive discourse), gesture often emphasizes information that is also focused pragmatically by mechanisms like prosody in speech. In fact, the semantic and pragmatic compatibility in the gesture-speech relationship recalls the interaction of words and graphics in multimodal presentations [11].

On the basis of results such as these, several researchers have built animated embodied conversational agents that ally synthesized speech with animated hand gestures. For example, Lester et al. [15] generate deictic gestures and choose referring expressions as a function of the potential ambiguity and proximity of objects referred to. Rickel and Johnson [19]'s pedagogical agent produces a deictic gesture at the beginning of explanations about objects. André et al. [1] generate pointing gestures as a sub-action of the rhetorical action of labeling, in turn a sub-action of elaborating. Cassell and Stone [3] generate either speech, gesture, or a combination of the two, as a function of the information structure status and surprise value of the discourse entity.

Head and eye movement has also been examined in the context of discourse and conversation. Looking away from one's interlocutor has been correlated with the beginning of turns. From the speaker's point of view, this look away may prevent an overload of visual and linguistic information. On the other hand, during the execution phase of an utterance, speakers look more often at listeners. Head nods and eyebrow raises are correlated with emphasized linguistic items – such as words accompanied by pitch accents [7]. Some eye movements occur primarily at the ends of utterances and at grammatical boundaries, and appear to function as synchronization signals. That is, one may request a response from a listener by looking at the listener, and suppress the listener's response by looking away. Likewise, in order to offer the floor, a speaker may gaze at the listener at the end of the utterance. When the listener wants the floor, s/he may look at and slightly up at the speaker [10]. It should be noted that turn taking only partially accounts for eye gaze behavior in discourse. A better explanation for gaze behavior integrates turn taking with the information structure of the propositional content of an utterance [5]. Specifically, the beginning of themes are frequently accompanied by a look-away from the hearer, and the beginning of rhemes are frequently accompanied by a look-toward the hearer. When these categories are co-temporaneous with turn construction, then they are strongly predictive of gaze behavior.

Results such as these have led researchers to generate eye gaze and head movements in animated embodied conversational agents. Takeuchi and Nagao, for example, [21] generate gaze and head nod behaviors in a "talking head." Cassell et al. [2] generate eye gaze and head nods as a function of turn taking behavior, head turns just before an utterance, and eyebrow raises as a function of emphasis.

To our knowledge, research on posture shifts and other gross body movements, has not been used in the design or implementation of computational systems. In fact, although a number of conversational analysts and ethnomethodologists have described posture shifts in conversation, their studies have been qualitative in nature, and difficult to reformulate as the basis of algorithms for the generation of language and posture. Nevertheless, researchers in the non-computational fields have discussed posture shifts extensively. Kendon [13] reports a hierarchy in the organization of movement such that the smaller limbs such as the fingers and hands engage in more frequent movements, while the trunk and lower limbs change relatively rarely.

A number of researchers have noted that changes in physical distance during interaction seem to accompany changes in the topic or in the social relationship between speakers. For example Condon and Osgton [9] have suggested that in a speaking individual the changes in these more slowly changing body parts occur at the boundaries of the larger units in the flow of speech. Scheflen (1973) also reports that posture shifts and other general body

movements appear to mark the points of change between one major unit of communicative activity and another. Blom & Gumperz (1972) identify posture changes and changes in the spatial relationship between two speakers as indicators of what they term "situational shifts" -- momentary changes in the mutual rights and obligations between speakers accompanied by shifts in language style. Erickson (1975) concludes that proxemic shifts seem to be markers of 'important' segments. In his analysis of college counseling interviews, they occurred more frequently than any other coded indicator of segment changes, and were therefore the best predictor of new segments in the data. Unfortunately, in none of these studies are statistics provided, and their analyses rely on intuitive definitions of discourse segment or "major shift". For this reason, we carried out our own empirical study.

3. Empirical Study

Videotaped "pseudo-monologues" and dialogues were used as the basis for the current study. In "pseudo-monologues," subjects were asked to describe each of the rooms in their home, then give directions between four pairs of locations they knew well (e.g., home and the grocery store). The experimenter acted as a listener, only providing backchannel feedback (head nods, smiles and paraverbals such as "uh-huh"). For dialogues, two subjects were asked to generate an idea for a class project that they would both like to work on, including: 1) what they would work on; 2) where they would work on it (including facilities, etc.), and 3) when they would work on it. Subjects stood in both conditions and were told to perform their tasks in 5-10 minutes. The pseudo-monologue condition (pseudo- because there was in fact an interlocutor, although he gave backchannel feedback only and never took the turn) allowed us to investigate the relationship between discourse structure and posture shift independent of turn structure. The two tasks were constructed to allow us to identify exactly where discourse segment boundaries would be placed.

The video data was transcribed and coded for three features: discourse segment boundaries, turn boundaries, and posture shifts. A discourse segment is taken to be an aggregation of utterances and sub-segments that convey the discourse segment purpose, which is an intention that leads to the segment initiation [12]. In this study we chose initially to look at high-level discourse segmentation phenomena rather than those discourse segments embedded deeper in the discourse. Thus, the time points at which the assigned task topics were started served as segmentation points. Turn boundaries were coded (for dialogues only) as the point in time in which the start or end of an utterance co-occurred with a change in speaker, but excluding backchannel feedback. Turn overlaps were coded as open-floor time. We defined a posture shift as a motion or a position shift for a part of the human body, excluding hands and eyes (which we have dealt with in other work). Posture shifts were coded with start and end time of occurrence (duration), body part in play (for this paper we divided the body at the waistline and compared upper body vs. lower body shifts), and an estimated energy level of the posture shift. Energy level was normalized for each subject by taking the largest posture shift observed for each subject as 100% and coding all other posture shift energies relative to the 100% case. Posture shifts that occurred as part of gesture or were clearly intentionally generated (e.g., turning one's body while giving directions) were not coded.

4. Results

Data from seven monologues and five dialogues were transcribed, and then coded and analyzed independently by two raters. A total of 70.5 minutes of data was analyzed (42.5 minutes of dialogue and 29.2 minutes of monologue). A total of 67 discourse segments were identified (25 in the dialogues and 42 in the monologues), which constituted 407 turns in the dialogue data.

We used the instructions given to subjects concerning the topics to discuss as segmentation boundaries. In future research, we will address the smaller discourse segmentation. For posture shift coding, raters coded all posture shifts independently, and then calculated reliability on the transcripts of one monologue (5.2 minutes) and both speakers from one dialogue (8.5

minutes). Agreement on the presence of an upper body or lower body posture shift in a particular location (taking location to be a 1-second window that contains all of or a part of a posture shift) for these three speakers was 89% (kappa = .64). For interrater reliability of the coding of energy level, a Spearman's rho revealed a correlation coefficient of .48 ($p<.01$).

4.1 Analysis

Posture shifts occurred regularly throughout the data (an average of 15 per speaker in both pseudo-monologues and dialogues). This, together with the fact that the majority of time was spent within discourse segments and within turns (rather than between segments), led us to normalize our posture shift data for comparison purposes. For relatively brief intervals (inter-discourse-segment and inter-turn) normalization by number of inter-segment occurrences was sufficient (ps/int), however, for long intervals (intra-discourse segment and intra-turn) we needed to normalize by time to obtain meaningful comparisons. For this normalization metric we looked at posture-shifts-per-second (ps/s). This gave us a mean average of .06 posture shifts/second (ps/s) in the monologues (SD=.07), and .07 posture shifts/second in the dialogues (SD=.08).

Table 4.1.1. Posture WRT Discourse Segments

	Monologues			**Dialogues**		
	ps/s	ps/int	energy	ps/s	ps/int	energy
inter-dseg	0.340	0.837	0.832	0.332	0.533	0.844
intra-dseg	0.039		0.701	0.053		0.723

Our initial analysis compared posture shifts made by the current speaker within discourse segments (intra-dseg) to those produced at the boundaries of discourse segments (inter-dseg). It can be seen (in Table 4.1.1) that posture shifts occur an order of magnitude more frequently at discourse segment boundaries than within discourse segments in both monologues and dialogues. Posture shifts also tend to be more energetic at discourse segment boundaries ($F(1,251)=10.4$; $p<0.001$).

Table 4.1.2 Posture Shifts WRT Turns

	ps/s	ps/int	energy
inter-turn	0.140	0.268	0.742
intra-turn	0.022		0.738

Initially, we classified data as being inter- or intra-turn. Table 4.1.2 shows that turn structure does have an influence on posture shifts; subjects were five times more likely to exhibit a shift at a boundary than within a turn.

Table 4.1.3 Posture by Discourse and Turn Breakdown

	ps/s	ps/int
inter-dseg/start-turn	0.562	0.542
inter-dseg/mid-turn	0.000	0.000
inter-dseg/end-turn	0.130	0.125
intra-dseg/start-turn	0.067	0.135
intra-dseg/mid-turn	0.041	
intra-dseg/end-turn	0.053	0.107

An interaction exists between turns and discourse segments such that discourse segment boundaries are ten times more likely to co-occur with turn changes than within turns. Both turn and discourse structure exhibit an influence on posture shifts, with discourse having the most predictive value. Starting a turn while starting a new discourse segment is marked with a posture shift roughly 10 times more often than when starting a turn while staying within discourse segment. We noticed, however, that posture shifts appeared to congregate at the beginnings or ends of turn boundaries, and so our subsequent analyses examined start-turns, mid-turns and end-turns. It is clear from these results that posture is indeed correlated with discourse state, such that speakers generate a posture shift when initiating a new discourse segment, which is often at the boundary between turns.

In addition to looking at the occurrence and energy of posture shifts we also analyzed the distributions of upper vs. lower body shifts and the duration of posture shifts. Speaker upper body shifts were found to be used more frequently at the start of turns (48%) than at the middle of turns (36%) or end of turns (18%) ($F(2,147)=5.39$; $p<0.005$), with no significant

dependence on discourse structure. Finally, speaker posture shift duration was found to change significantly as a function of both turn and discourse structure (see Figure 4.1.3). At the start of turns, posture shift duration is approximately the same whether a new topic is introduced or not (2.5 seconds). However, when ending a turn, speakers move significantly longer (7.0 seconds) when finishing a topic than when the topic is continued by the other interlocutor (2.7 seconds) ($F(1,148)=17.9$; $p<0.001$).

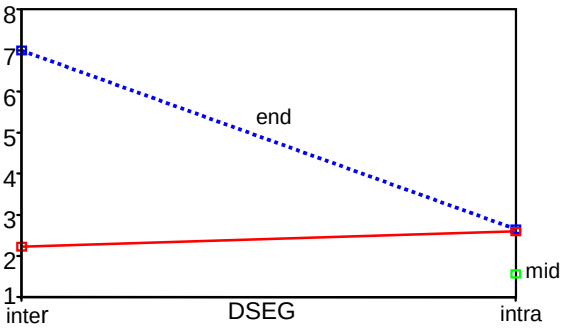

Figure 4.1.1 Posture Shift Duration by DSeg and Turn

5. System

In the following sections we discuss how the results of the empirical study were integrated along with Collagen into our existent embodied conversational agent, Rea.

5.1 System Architecture

Rea is an embodied conversational agent that interacts with a user in the real estate agent domain [2]. The system architecture of Rea is shown in Figure 5.1. Rea takes input from a microphone and two cameras in order to sense the user's speech and gesture. The UM interprets and integrates this multimodal input and outputs a unified semantic representation. The Understanding Module then sends the output to Collagen as the Dialogue Manager.

Collagen, as further discussed below, maintains the state of the dialogue as shared between Rea and a user. The Reaction Module decides Rea's next action based on the discourse state maintained by Collagen. It also assigns information structure to output utterances so that gestures can be appropriately generated. The

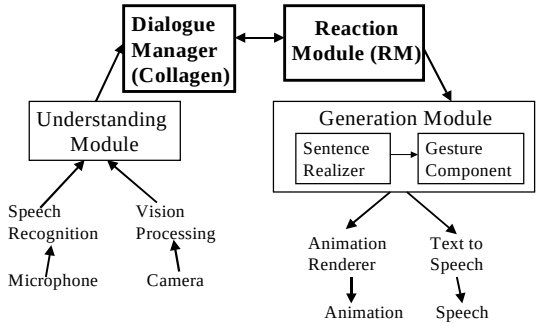

Figure5.1: System architecture

semantic representation of the action, including verbal and non-verbal behaviors, is sent to the Generation Module which generates surface linguistic expressions and gestures, including a set of instructions to achieve synchronization between animation and speech. These instructions are executed by a 3D animation renderer and a text-to-speech system. Table 5.1 shows the associations between discourse and conversational state that Rea is currently able to handle. In other work we have discussed how Rea deals with the association between information structure and gesture [6]. In the following sections, we focus on Rea's generation of posture shifts.

Table 5.1: Discourse functions & non-verbal behavior cues

Discourse level info.	**Functions**	**non-verbal behavior cues**
Discourse structure	new segment	Posture_shift
Conversation structure	turn giving	eye_gaze & (stop_gesturing hand_gesture)
	turn keeping	(look_away keep_gesture)
	turn taking	eye_gaze & posture_shift
Information structure	emphasize information	eye_gaze & beat_and other_hand_gsts

5.2 The Collagen dialogue manager

Collagen™ is JAVA middleware for building COLLAborative interface AGENts to work with users on interface applications. Collagen is designed with the capability to participate in collaboration and conversation, based on [12], [16]. Collagen updates the focus stack and recipe tree using a combination of the discourse interpretation algorithm of [16] and plan recognition algorithms of [14]. It takes as input user and system utterances and interface actions, and accesses a library of recipes describing actions in the domain. After updating the discourse state, Collagen makes three resources available to the interface agent: focus of attention (using the focus stack), segmented interaction history (of completed segments) and an agenda of next possible actions created from the focus stack and recipe tree.

5.3 Output Generation

The Reaction Module works as a content planner in the Rea architecture, and also plays the role of an interface agent in Collagen. It has access to the discourse state and the agenda using APIs provided by Collagen. Based on the results reported above, we describe here how Rea plans her next nonverbal actions using the resources that Collagen maintains.

The empirical study revealed that posture shifts are distributed with respect to discourse segment and turn boundaries, and that the form of a posture shift differs according to these co-determinants. Therefore, generation of posture shifts in Rea is determined according to these two factors, with Collagen contributing information about current discourse state.

5.3.1 Discourse structure information

Any posture shift that occurs between the end of one discourse segment and the beginning of the next is defined as an inter-discourse segment posture shift. In order to elaborate different generation rules for inter- vs. intra-discourse segments, Rea judges **(D1)** whether the next utterance starts a new topic, or contributes to the current discourse purpose, **(D2)** whether the next utterance is expected to finish a segment. First, **(D1)** is calculated by referring to the focus stack and agenda. In planning a next action, Rea accesses the goal agenda in Collagen and gets the content of her next utterance. She also accesses the focus stack and gets the current discourse purpose that is shared between her and the user. By comparing the current purpose and the purpose of her next utterance, Rea can judge whether the her next utterance contributes to the current discourse purpose or not. For example, if the current discourse purpose is to find a house to show the user (FindHouse), and the next utterance that Rea plans to say is as follows,

(1) (Ask.What (agent Propose.What (user FindHouse <city ?>)))

Rea says: "What kind of transportation access do you need?"

then Rea uses Collagen APIs to compare the current discourse purpose (FindHouse) to the purpose of utterance (1). The purpose of this utterance is to ask the value of the transportation parameter of FindHouse. Thus, Rea judges that this utterance contributes to the current discourse purpose, and continues the same discourse segment (**D1** = continue). On the other hand, if Rea's next utterance is about showing a house,

(2) (Propose.Should (agent ShowHouse (joint 123ElmStreet))

Rea says: "Let's look at 123 Elm Street."

then this utterance does not directly contribute to the current discourse purpose because it does not ask a parameter of FindHouse, and it introduces a new discourse purpose ShowHouse. In this case, Rea judges that there is a discourse segment boundary between the previous utterance and the next one (**D1** = topic change).

In order to calculate **(D2)**, Rea looks at the plan tree in Collagen, and judges whether the next utterance addresses the last goal in the current discourse purpose. If it is the case, Rea expects to finish the current discourse segment by the next utterance (**D1** = finish topic). As for conversational structure, Rea needs to know; **(T1)** whether Rea is taking a new turn with the next utterance, or keeping her current turn for the next utterance, **(T2)** whether Rea's next utterance requires that the user respond.

Place of a posture shift	Case	Discourse structure information		Conversation structure information		Posture shift decision probability	Posture shift selection		
							energy	duration	body part
beginning of the utterance	a	**D1**	topic change	**T1**	take turn	0.54/int	high	default	upper & lower
	b		topic change		keep turn	0	-	-	-
	c		continue		take turn	0.13/int	low	default	upper or lower
	d		continue		keep turn	0.14/sec	low	short	lower
End of the utterance	e	**D2**	finish topic	**T2**	give turn	0.04/int	high	long	lower
	f		continue		give turn	0.11/int	low	default	lower

Table 5.3.1:Posture Decision Probabilities for Dialogue

First, **(T1)** is judged by referring to the dialogue history[1]. The dialogue history stores both system utterances and user utterances that occurred in the dialogue. In the history, each utterance is stored as a logical form based on an artificial discourse language [20]. As shown above in utterance (1), the first argument of the action indicates the speaker of the utterance; in this example, it is "agent". The turn boundary can be estimated by comparing the speaker of the previous utterance with the speaker of the next utterance. If the speaker of the previous utterance is not Rea, there is a turn boundary before the next utterance (**T1** = take turn). If the speaker of the previous utterance is Rea, that means that Rea will keep the same turn for the next utterance (**T1** = keep turn).

Second, **(T2)** is judged by looking at the type of Rea's next utterance. For example, when Rea asks a question, as in utterance (1), Rea expects the user to answer the question. In this case, Rea must convey to the user that the system gives up the turn (**T2** = give up turn).

5.3.2 Deciding and selecting a posture shift

Combining information about discourse structure (**D1**, **D2**) and conversation structure (**T1**, **T2**), the system decides on posture shifts for the beginning of the utterance and the end of the utterance. Rea decides to do or not to do a posture shift by calling a probabilistic function that looks up the probabilities in Table 5.3.1.

A posture shift for the beginning of the utterance is decided based on the combination of **(D1)** and **(T1)**. For example, if the combined factors match Case (a), the system decides to generate a posture shift with 54% probability for the beginning of the utterance. Note that in Case (d), that is, Rea keeps the turn without changing a topic, we cannot calculate a per interval posture shift rate. Instead, we use a posture shift rate normalized for time. This rate is used in the GenerationModule, which calculates the utterance duration and generates a posture shift during the utterance based on this posture shift rate. On the other hand, ending posture shifts are decided based on the combination of **(D2)** and **(T2)**.

For example, if the combined factors match Case (e), the system decides to generate a posture shift with 0.04% probability for the ending of the utterance. When Rea does decide to activate a posture shift, she then needs to choose which posture shift to perform. Our empirical data indicates that the energy level of the posture shift differs depending on whether there is a discourse segment boundary or not. Moreover the duration of a posture shift differs depending on the place in a turn: start-, mid-, or end-turn.

[1] We currently maintain a dialogue history in Rea even though Collagen has one as well. This is in order to store and manipulate the information to generate hand gestures and assign intonational accents. This information will be integrated into Collagen in the near future.

Based on these results, we define posture shift selection rules for energy, duration, and body part. The correspondence with discourse information is shown in Table 5.3.1. For example, in Case (a), the system selects a posture shift with high energy, using both upper and lower body. After deciding whether or not Rea should shift posture and (if so) choosing a kind of posture shift, Rea sends a command to the Generation Module to generate a specific kind of posture shift within a specific time duration.

Table 5.3.2: Posture Decision Probabilities: Monologue

Case	Discourse structure information		Posture shift decision probability	Posture shift selection
				energy
g	**D1**	change topic	0.84/int	high
h		continue	0.04/sec	low

Posture shifts for pseudo-monologues can be decided using the same mechanism as that for dialogue, but omitting conversation structure information. The probabilities are given in table Table 5.3.2. For example, if Rea changes the topic with her next utterance, a posture shift is generated 84% of the time with high-energy motion. In other cases, the system randomly generates low-energy posture shifts 0.04 times per second.

6. Example

Figure 6.1 shows a dialogue between Rea and the user, and shows how Rea decides to generate posture shifts. This dialogue consists of two major segments: finding a house (dialogue), and showing a house (pseudo-monologue). Based on this task structure, we defined plan recipes for Collagen. The first shared discourse purpose **[goal: HaveConversation]** is introduced by the user before the example. Then, in utterance (1), the user introduces the main part of the conversation **[goal: FindHouse]**.

The next goal in the agenda, **[goal: IdentifyPreferredCity],** should be accomplished to identify a parameter value for **[goal: FindHouse]**. This goal directly contributes to the current purpose, **[goal: FindHouse]**. This case is judged to be a turn boundary within a discourse segment (**Case (c)**), and Rea decides to generate a posture shift at the beginning of the utterance with 13% probability. If Rea decides to shift posture she selects a low energy posture shift using either upper or lower body. In addition to a posture shift at the beginning of the utterance, Rea may also choose to generate a posture shift to end the turn. As utterance (2) expects the user to take the turn, and continue to work on the same discourse purpose, this is **Case (f)**. Thus, the system generates an end utterance posture shift 11% of the time. If generated, a low energy posture shift is chosen. If a beginning and/or ending posture shifts are generated, they are sent to the GM, which calculates the schedule of these multimodal events and generates them.

In utterance (25), Rea introduces a new discourse purpose **[goal : ShowHouse]**. Rea, using a default rule, decides to take the initiative on this goal. At this point, Rea accesses the discourse state and confirms that a new goal is about to start. Rea judges this case as a discourse segment boundary and also a turn boundary (**Case (a)**). Based on this information, Rea selects a high energy posture shift. An example of Rea's high energy posture shift is shown on the right in Figure 5.2.

As a subdialogue of showing a house, in a discourse purpose **[goal : DiscussFeature]**, Rea keeps the turn and continues to describe the house. We handle this type of interaction as a pseudo-monologue. Therefore, we can use table Table 5.3.2 for deciding on posture shifts here. In utterance (27), Rea starts the discussion about the house, and takes the initiative. This is judged as **Case (g)**, and a high energy body motion is generated 84% of the time.

[**Finding a house**] < dialogue>

(1) U: I'm looking for a house.
(2) R: (c) Where do you want to live? (f)
(3) U: I like Boston.
(4) R: (c) (d) What kind of transportation access do you need? (f)
(5) U: I need T access.

....

(23) R: (c) (d) How much storage space do you need? (f)
(24) U: I need to have a storage place in the basement.

[Showing a house] <Pseudo-monologue>

(25) R: (a) (d) Let's look at 123 Elm Street. (f)
(26) U: OK.
[Discuss a feature of the house]
(27) R: (g) Let's discuss a feature of this place.
(28) R: (h) Notice the hardwood flooring in the living room.
(29) R: (h) Notice the jacuzzi.
(30) R: (h) Notice the remodeled kitchen

Figure 6.1: Example dialogue

7. Conclusion and Further work

We have demonstrated a clear relationship between nonverbal behavior and discourse state, and shown how this finding can be incorporated into the generation of language and nonverbal behaviors for an embodied conversational agent.

Speakers produce posture shifts at 53% of discourse segment boundaries, more frequently than they produce those shifts discourse segment-internally, and with more motion energy. Furthermore, there is a relationship between discourse structure and conversational structure such that when speakers initiate a new segment at the same time as starting a turn (the most frequent case by far), they are more likely to produce a posture shift; while when they end a discourse segment and a turn at the same time, their posture shifts last longer than when these categories do not co-occur.

Although this paper reports results from a limited number of monologues and dialogues, the findings are promising. In addition, they point the way to a number of future directions, both within the study of posture and discourse, and more generally within the study of non-verbal behaviors in computational linguistics.

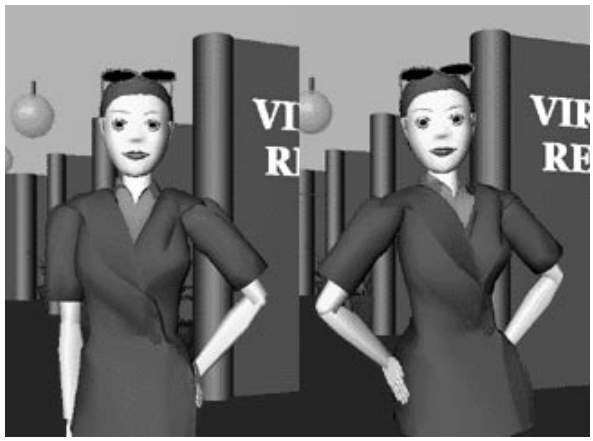

Figure 6.2: Rea demonstrating a low and high energy posture shift

First, given the relationship between conversational and information structure in [5], a natural next step is to examine the three-way relationship between discourse state, conversational structure (turns), and information structure (theme/rheme). For the moment, we have demonstrated that posture shifts may signal *boundaries* of units; do they also signal the information content of units? Next, we need to look at finer segmentations of the discourse, to see whether larger and smaller discourse segments are distinguished through non-verbal means. Third, the question of listener posture is an important one. We found that a number of posture shifts were produced by the participant who was not speaking. More than half of these shifts were produced at the same time as a speaker shift, suggesting a kind of mirroring. In order to interpret these data, however, a more sensitive notion of turn structure is required, as one must be ready to define when exactly speakers and listeners shift roles. Also, of course, evaluation of the importance of such nonverbal behaviors to user interaction is essential. In a user study of our earlier Gandalf system [4], users rated the agent's language skills significantly higher under test conditions in which Gandalf deployed conversational behaviors (gaze, head movement and limited gesture) than when these behaviors were disabled. Such an evaluation is also necessary for the Rea-posture system. But, more generally, we need to test whether generating posture shifts of this sort actually serves as a signal to listeners, for example to initiative

structure in task and dialogue [8]. These evaluations form part of our future research plans.

8. Acknowledgements

This research was supported by MERL, France Telecom, AT&T, and the other generous sponsors of the MIT Media Lab. Thanks to the other members of the Gesture and Narrative Language Group, in particular Ian Gouldstone and Hannes Vilhjálmsson.

9. REFERENCES

[1] Andre, E., Rist, T., & Muller, J., Employing AI methods to control the behavior of animated interface agents, Applied Artificial Intelligence, vol. 13, pp. 415-448, 1999.

[2] Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjalmsson, H., & Yan, H., Embodiment in Conversational Interfaces: Rea, Proc. of CHI 99, Pittsburgh, PA, ACM, 1999.

[3] Cassell, J., Stone, M., & Yan, H., Coordination and context-dependence in the generation of embodied conversation, Proc. INLG 2000, Mitzpe Ramon, Israel, 2000.

[4] Cassell, J. and Thorisson, K. R., The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents, Applied Art. Intell., vol. 13, pp. 519-538, 1999.

[5] Cassell, J., Torres, O., & Prevost, S., Turn Taking vs. Discourse Structure: How Best to Model Multimodal Conversation., in Machine Conversations, Y. Wilks, Ed. The Hague: Kluwer, 1999, pp. 143-154.

[6] Cassell, J., Vilhjálmsson, H., & Bickmore, T., BEAT: The Behavior Expression Animation Toolkit, Proc. of SIGGRAPH, ACM Press, 2001.

[7] Chovil, N., Discourse-Oriented Facial Displays in Conversation, Research on Language and Social Interaction, vol. 25, pp. 163-194, 1992.

[8] Chu-Carroll, J. & Brown, M., Initiative in Collaborative Interactions - Its Cues and Effects, Proc. of AAAI Spring 1997 Symp. on Computational Models of Mixed Initiative, 1997.

[9] Condon, W. S. & Osgton, W. D., Speech and body motion synchrony of the speaker-hearer, in The perception of language, D. Horton & J. Jenkins, Eds. NY: Academic Press, 1971, pp. 150-184.

[10] Duncan, S., On the structure of speaker-auditor interaction during speaking turns, Language in Society, vol. 3, pp. 161-180, 1974.

[11] Green, N., Carenini, G., Kerpedjiev, S., & Roth, S, A Media-Independent Content Language for Integrated Text and Graphics Generation, Proc. of Workshop on Content Visualization and Intermedia Representations at COLING and ACL '98, 1998.

[12] Grosz, B. & Sidner, C., Attention, Intentions, and the Structure of Discourse, Computational Linguistics, vol. 12, pp. 175-204, 1986.

[13] Kendon, A., Some Relationships between Body Motion and Speech, in Studies in Dyadic Communication, A. W. Siegman and B. Pope, Eds. Elmsford, NY: Pergamon Press, 1972, pp. 177-210.

[14] Lesh, N., Rich, C., & Sidner, C., Using Plan Recognition in Human-Computer Collaboration, Proc. of the Conference on User Modelling, Banff, Canada, NY: Springer Wien, 1999.

[15] Lester, J., Towns, S., Callaway, C., Voerman, J., & FitzGerald, P., Deictic and Emotive Communication in Animated Pedagogical Agents, in Embodied Conversational Agents, J. Cassell, J. Sullivan, et. al, Eds. Cambridge: MIT Press, 2000.

[16] Lochbaum, K., A Collaborative Planning Model of Intentional Structure, Computational Linguistics, vol. 24, pp. 525-572, 1998.

[17] McNeill, D., Hand and Mind: What Gestures Reveal about Thought. Chicago, IL/London, UK: The University of Chicago Press, 1992.

[18] Rich, C. & Sidner, C. L., COLLAGEN: A Collaboration Manager for Software Interface Agents, User Modeling and User-Adapted Interaction, vol. 8, pp. 315-350, 1998.

[19] Rickel, J. & Johnson, W. L., Task-Oriented Collaboration with Embodied Agents in Virtual Worlds, in Embodied Conversational Agents, J. Cassell, Ed. Cambridge, MA: MIT Press, 2000.

[20] Sidner, C., An Artificial Discourse Language for Collaborative Negotiation, Proc. of 12th Intnl. Conf. on Artificial Intelligence (AAAI), Seattle, WA, MIT Press, 1994.

[21] Takeuchi, A. & Nagao, K., Communicative facial displays as a new conversational modality, Proc. of InterCHI '93, Amsterdam, NL, ACM, 1993.

[22] Thompson, L. and Massaro, D., Evaluation and Integration of Speech and Pointing Gestures during Referential Understanding, Journal of Experimental Child Psychology, vol. 42, pp. 144-168, 1986.