Modeling Gaze Behavior
as a Function of Discourse Structure

Obed E. Torres, Justine Cassell, Scott Prevost
The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street
Cambridge, MA 02139
{obed,justine,prevost}@media.mit.edu

Abstract

This paper addresses the problem of designing conversational agents that exhibit appropriate gaze behavior during dialogues with human users. Previous research on gaze behavior has concentrated on its relationship to turn-taking phenomena [3,4,5]. Recent work has incorporated some of these findings into the design of autonomous human-like conversational agents and interactive communicative humanoids [1,13]. However, none of this research has examined the relationship between information structure and gaze behavior. In this paper we discuss why turn-taking is not an adequate explanation for gaze behavior in conversation and why information structure should be integrated with turn-taking as an explanation for this behavior. We then examine the relationship of gaze behavior to information structure and turn-taking through an empirical analysis of discourse transcripts for several dyadic conversations. A simple algorithm for assigning gaze behavior is proposed on the basis of the findings of this empirical analysis. We describe work in progress on implementing this algorithm in an existent communicative humanoid agent [14] with the goal of producing more natural gaze behavior related to propositional content in human-computer conversations.

Introduction

The Turing test has always been conceived of as a test of the *content* of a computer's contribution to a conversation. That is, from typed output, we are able to tell whether text was generated by a human or a computer. Recent advances in speech technology have led us to conceive of a Turing test taken over the phone. What about a face-to-face Turing test? What kinds of behaviors would a computer have to exhibit to convince us that it was not a grey box but a living, breathing body? We are perhaps not ready today for such a competition, but we may be one day. This paper attempts to move the field of human-computer conversation in that direction -- in the direction of *embodied* dialogue with computers. In other work, we have concentrated on hand gestures, intonation, head movement, and gaze. The current work revisits the question of gaze and attempts to reconcile two competing approaches to the general question of generating nonverbal behaviors.

Although there has been substantial research dedicated to the study of nonverbal communicative behaviors (including gaze behavior), such research has focused *either* on the interaction between conversational regulation (e.g. turn-taking) and non-verbal behaviors*or* the interaction between discourse structure and non-verbal behaviors. That is, there has been little research on the interaction between turn-taking and discourse structure, and even less research that takes both types of linguistic structures into account in investigating non-verbal behaviors. This lacuna is due to historical accidents of disciplinary boundaries rather than any lack of inherent theoretical interest. The current paper addresses the problem of designing conversational agents that exhibit appropriate gaze behavior through an approach that ties information structure to turn-taking. The exchange of looks between participants is related to information threads and the exchange of turns during the flow of conversation. Current approaches have failed to capture the relationship of gaze behavior to contextual boundaries in the incremental exchange of information. In order to account for this aspect of turn-taking, we employ information structure distinctions as a representation of coherence in the accumulation of information within and across turns in dyadic conversations. We examine gaze behavior in relation to propositional content. Semantic content is divided into thematic and rhematic constituents that allow propositions to be presented in a way that highlights the shared content between utterances.

After conducting an empirical analysis on experimental data, we propose these heuristics: the beginning of the thematic part of an utterance is frequently accompanied by gaze behavior that looks away from the hearer, while the beginning of the rhematic part is usually accompanied by gaze behavior that looks toward the hearer. In cases where the beginning of the theme coincides with the beginning of a turn, the speaker always looks away from the hearer. In cases where the beginning of the rheme coincides with the end of the turn, the speaker always looks toward the listener. A simple algorithm for assigning gaze behavior is proposed on the basis of these heuristics. We describe work in progress to implement a conversational agent with capabilities for generating gaze behavior related to propositional content in order to illustrate the application of this research to human-computer conversation.

Background

Gaze and Turn-Taking

When people engage in conversation, they take turns speaking. Turns almost always begin and end smoothly, with short lapses of time between them. Taking into account the dynamic and fast paced nature of conversations, it is remarkable that there are so few occasions when conversation breaks down through simultaneous speech or interruption.

In fact, the time between the exchange of turns is often too short to be explained as the result of the hearer's waiting for the speaker to finish before the hearer starts to speak. This is even more significant if one considers that pauses across turns are sometimes even shorter than pauses within a turn itself. Duncan [4,5] attempts to describe how a speaker signals a turn to hearers. He suggests several cues that indicate the end of a turn or invite the hearer to take a turn. These cues include falling pitch at the end of a sentence, the drawl of a syllable at the end of sentence, the termination of a gesture, specific phrases at the end of syntactic units, and changes in gaze direction, such as the speaker's looking away from the hearer as an utterance begins and toward the hearer as the utterance ends. Goodwin [5] elaborates on the role of gaze in turn-taking by also considering the gaze of the hearer and the coordination of the gaze of conversational participants. He claims that the speaker's look away at the beginning of turns occurs to avoid overloading information in the planning of an utterance. Because of research by Duncan, Goodwin, and others, gaze behavior has come to represent *the* cue to turn-organization and has been used as such in the design of embodied conversational agents.

Gaze Behavior of Conversational Agents

Takeuchi and Nagao [12] and others illustrate the communicative value of the face in human-computer conversation. Research by Cassell et al. and Pelachaud et al [1,8] on the design and implementation of autonomous human-like conversational agents incorporate some of the findings of Duncan and Goodwin to simulate the role of gaze and back channel feedback in turn-taking. They generated gaze behaviors of the sort correlated with the beginnings and ends of turns. For long turns, the speaker looks away from the hearer at the beginning of a turn and looks toward the hearer at the end of a turn. For short turns, the speaker looks toward the hearer from the beginning to the end of the turn. They also modeled rules and functions for the diverse types of feedback that take place within a turn. In their model of turn-taking behavior within a turn, the speaker looks at the hearer during grammatical pauses to obtain feedback. Then the hearer looks at the speaker and nods; this backchannel communication is followed by a speaker continuation signal consisting of a look away from the hearer, if the speaker intends to hold the turn.

Thórisson's research [13] on interactive communicative humanoids uses a situated model of turn-taking based on Sacks et al. [10,11]. According to Sacks et al., turn-taking is a phenomenon in which rules are subject to the control of the participants and emergent patterns arising from the interaction of the rules. Thórisson built an interactive communicative humanoid with turn-taking as one of its most relevant and robust features. He addresses the problem of real-time turn-taking by integrating turn-taking with a model of conversants' actions. In his model of conversant actions there are two roles: speaker and hearer. For each role, he defines different classes of behaviors, including perceptual, decision, and motor tasks. His efforts are mainly focused on defining the nature of the underlying perceptual mechanisms. User testing of Thórisson's conversational agent showed that presence of these nonverbal feedback behaviors (gaze, nods, beat gestures) diminished disfluency on the part of users and increased perceived efficiency of the humanoid agent [15].

Information Structure

Whittaker et al. [16] addresses how speakers signal information about discourse structure, beyond the level of the individual utterance, to hearers looking at the mechanism for shifts of control in conversations. The present research follows this discourse-related approach by concentrating on the flow of information between conversants and the informational threads that affect gaze behavior. One way of modeling such discourse phenomena is through information structure [6], which describes the relationship between the content of utterances (or clauses) and the emerging discourse context. Information structure allows the representation of information within an utterance to be connected with the knowledge of the speaker and hearer and the structure of their discourse. By employing such a model, we are attempting to formalize Goodwin's suggestion that gaze behaviors (and the consequent restarts, pauses, and hesitations) are indicative of "the speaker's attention to the construction of coherent sentences for his recipient" [5].

We follow Halliday [6] in using the terms "theme" and "rheme" to describe information structural components of an utterance. Other terms, such as "link" and "focus" have been widely used in the literature and are roughly synonymous (cf. [17]). The theme represents the part of the utterance that links it to the previous discourse and specifies what the utterance is about. The rheme, on the other hand, specifies what is contributed to the discourse with respect to the theme. That is, the rheme specifies what is new or interesting about the theme, and generally contains the information that the hearer could not have predicted from context. The linking of thematic threads in a discourse is part of what makes it coherent. In the sections below, we provide evidence, through the use of these information structural categories, that gaze behavior is directly related to discourse coherence .

Motivation

Research approaches to the study of conversational gaze behavior concentrate on providing descriptive models of the sentence planning and surface generation aspects of this phenomenon. Although the models proposed by Duncan and Sacks et al. served as the basis for the computational prototypes of Cassell et. al and Thórisson, they do not adequately address other issues involved in the predictive modeling and simulation of gaze behavior generation. First of all, Duncan's signaling approach examines surface linguistic phenomena to investigate what cues signal the end of a turn. None of the phenomena investigated actually correlate highly or predictably with turn-taking. That is, although looking toward the hearer is taken to be a reliable signal of giving over the turn, the majority of glances toward the listener are not found in the context of ends of turns. The employment of "turn constructional units" by Sacks et al. is useful for describing the fundamental units in the exchange of turns, but it does not adequately address how conversants recognize these units. A general theory of turn-taking should account for a consistent range of indicators that serve the very specific function of signaling the end or the beginning of a turn. A different approach for an empirical analysis could entail identifying first the boundaries of a turn and then attempting to explain how the turn exchange is signaled.

We chose to look at the distribution of gaze behavior for clues. That is, rather than looking at turns and all of the nonverbal behaviors that correlate with them, we chose to investigate the nonverbal behavior most popularly assumed to be indicative of turn-taking. In doing so, we wished to also repair a rift between two fields of study. The study of turn-taking has been the purview of conversational analysis (inter alia [3,5,10]), a field derived from sociology. The study of discourse structure (inter alia, [6]) has been the domain of linguists, who often neglect to talk to sociologists. And computational work that models language use perpetuates a similar divide (compare [7] with [9]). But theme and rheme (information structure) are, like turn-constructional-units, an account of the accumulation of information. The exchange of turns is related to information threads in the flow of the conversation. This can be intuitively validated, if one thinks of a chain of utterances in which new utterances are interpreted in the context of previous utterances. Current approaches to the study of turn-taking have failed to capture the relationship of turn-taking behavior to contextual boundaries in the incremental exchange of information. In order to account for this aspect of turn-taking, the research approach presented in this paper included conducting an empirical analysis that uses information structure distinctions as a representation of coherence in the accumulation of information within and across turns in dyadic conversations. Prevost's work [9] toward more natural spoken language generation demonstrates the benefits of using a representation of information structure to capture focal distinctions of importance in assigning intonational patterns to an utterance. Cassell and Prevost [2] employs the relationship between information structure and nonverbal behaviors to predict when intonation and gestures will occur in the stream of speech. Concretely, in terms of implementation goals, if information structure can be shown to predict gaze behavior, then all paraverbal behavior in autonomous humanoid conversational agents (intonation, hand gestures, and gaze) can be driven by a single underlying information structure representation. Such an implementation will facilitate investigating and modeling the interaction among these phenomena.

Experimental Data and Empirical Methodology

Our empirical analysis is based on transcripts of speech, gaze behavior, and head movements which occurred during the first three and a half minutes of three dyadic conversations recorded on videotape. Participants in each conversation were strangers to one another. All participants in the three different two-person conversations were given the same instructions: they were told to sustain a conversation on whatever topics and to whatever extent they liked for at least 20 minutes. All of them were native speakers of North American English. They were informed that the purpose of the data collection was to study several aspects of face-to-face interaction. All of them consented to be videotaped.

The conversations were videotaped using two cameras and a microphone placed so that the upper-body space of all participants was completely visible and their voices could be comfortably heard. The interaction was videotaped without altering the focus, zooming in or out, or increasing or decreasing the level of sound. The cameras and the microphone were set up in full view of the participants. The video camera, the microphone, and the video tape were running before participants started their conversations and were not stopped until the conversations ended. The positioning of the video cameras allowed a view of subtle details in the process of interaction, particularly in terms of head movements and gaze behavior.

The analysis was carried out on 100 turns. For each turn there were four steps in the transcription process, in order to ensure independence and consistency in transcribing verbal and nonverbal behaviors of the speaker and the hearer during and between gaze behaviors. In the first pass, we transcribed the verbal behavior of the speaker, mainly words and pauses. In the second pass, we transcribed the nonverbal behavior of the speaker, basically gaze behavior. In the third pass, we transcribed the verbal and paraverbal behaviors of the listener, mainly “hmm” and “uh-huh,” in alignment with the transcription of the speaker’s utterances. In the fourth pass, we transcribed the nonverbal behaviors of the listener, mainly head nods, also in alignment with the transcription of the speaker’s utterances.

An attempt was made to include only some types of nonverbal and verbal behavior and two different types of pauses: filled and unfilled. Nonverbal behaviors were mainly of three types: beginning of a look away from the hearer, beginning of a look toward the hearer, and the head nods of the hearer. Unfilled pauses were considered to be noticeable lapses of silence in the talk of speakers.

Three main units of empirical analysis were employed: turns, themes, and rhemes. A “turn” is the talk of the speaker delimited by the talk of the hearer, with the exception of ongoing communicative behavior by the hearer that lacks propositional content. The “beginning of a turn” was defined as the first word of a new turn. The “end of a turn” was defined as the last word +/- one word. The “theme” represents what the utterance is about. The “rheme” represents the contribution to the pool of knowledge in the conversation. The example below illustrates the theme/rheme annotations for the text of utterances, using BTh (beginning of theme) , ETh (end of theme), BRh (beginning of rheme) and ERh (end of rheme):

Q: What do you do?
A: BTh (I work with) ETh BRh (Mike B.) ERh

Discussion of Preliminary Results

Tables 1-4 display quantitative results.

Table 1: Look-Away (LA), Beginning of Turn (BT) and Beginning of Theme (BTh)

	BT	BTh	BTh at BT
LA	44%	70%	100%
No LA	56%	30%	0%

Table 2: Distribution of Look-Away (LA)

	LA when BT, not BTh	LA when BTh, not BT	LA when BTh and BT	Other LA
All of LA	28%	22%	10%	40%

Table 3: Look-Toward (LT), End of Turn(ET), Beginning of Rheme (BRh)

	ET	BRh	BRh at ET
LT	16%	73%	100%
No LT	84%	27%	0%

Table 4: Distribution of Look-Toward (LT)

	LT when ET, not BRh	LT when BRh, not ET	LT when ET and BRh	Other LT
All of LT	12%	40%	3%	45%

Our empirical analysis began by examining patterns that support the claims of already established research. We then moved toward findings of other observable and quantifiable patterns in the data. In this section we discuss both results. Two patterns previously investigated and formulated in the research literature are: the occurrence of a look away from the hearer at the beginning of a turn and a look toward the hearer by the end of a turn. We present as preliminary results the following: the occurrence of a "look-away" from the hearer at the beginning of a theme and a "look-toward" the hearer at the beginning of a rheme.

As described in the literature, the speaker does look away from the hearer at the beginning of a turn. Of all the turn beginnings in our data, 44% were accompanied by look-aways. In terms of how much gaze behavior is accounted for by turn-taking, these look-aways constituted 38% of all the look-aways in our data (see columns 2 and 4 in Table 2). 70% of the parts of utterances that were identified as thematic were accompanied by the speaker initially looking away from the hearer. These look-aways account for 32% of all the look-aways in the data (see columns 3 and 4 in Table 2). 40% of all the look-aways from the hearer were not associated with either the beginning of a turn or the beginning of thematic material. Most strikingly, however, when the beginning of a theme coincided with the beginning of a turn, speakers always looked away.

Thus, our results suggest that the information structural category of themes accounts for some gaze behavior, and that a co-temporaneous beginning of a theme and beginning of a turn always elicits a look-away.

According to the literature, the speaker looks toward the hearer at the end of a turn or at least is already looking toward the hearer by the end of a turn. This pattern is observable in the data, but leaves open the question of how close to the end of the turn this behavior occurs. In Tables 3 and 4, we counted look-towards that occurred within one word of the actual end of a utterance. Of all these ends of turns (given the one word window), 16% included a look-toward. These look-towards represented only 15% of all the look-towards in our data. A look-toward at the beginning of rhematic material occurred in 73% of the instances. These look-towards account for 43% of all the look-towards in our data. 45% of all the look-towards were not associated with either the end of a turn or the beginning rhematic material. Most strikingly, however, when the beginning of a rheme occurred within one word of the end of a turn, the speaker always looked at the listener. Thus, our results suggest that the information structural category of rhemes accounts for some gaze behavior, and that a co-temporaneous rheme and end of turn always elicits a look-toward.

It is clear that the association of turn-initial and turn-final units with information structure units is *very* predictive of gaze behavior. It is also clear that we still cannot account for the majority of gaze behavior (look-towards in particular) with the association of information structure and turn-taking. Additional analyses of the data (not reported here) suggest that back-channel and other kinds of feedback may be accounting for look-towards.

Gaze Behavior Assignment Algorithm

Some of the empirical results of this study are incorporated as weighted heuristics in a simple algorithm to assign suitable gaze behavior from propositional content. The two heuristics are: the beginning of a theme is accompanied by a look-away from the hearer; and the beginning of a rheme is accompanied by a look-toward the hearer. The following is the pseudocode of the algorithm, where *distribution(x)* is a randomized function that returns *true* with probability *x*:

```
for each proposition in the list of propositions to be realized sequentially by a language
generator
    if current proposition is thematic
        if beginning of turn or distribution(.70)
            attach a look-away from the hearer
        endif
    else if current proposition is rhematic
        if end of turn or distribution(.73)
            attach a look-toward the hearer
        endif
endfor
```

Multimodal Dialogue Generation for a Conversational Agent

The algorithm described above is being implemented in Gandalf, an animated agent originally developed by Thórisson [14] to interact with human users within a tightly constrained topic domain (the solar system). Gandalf interprets the user's speech, gaze, and gestures[1], and in return produces appropriate facial expressions, gestures, and spoken responses. Specifically, some of Gandalf's communicative behaviors include: blinking, raising the eyebrows, turning to and gazing at either the graphical model of the solar system or the user, offering nonverbal cues to show when it decides to take a turn, and producing beat and pointing gestures when appropriate. Augmenting Gandalf's turn-construction algorithm with an information structure algorithm enables us to drive all of Gandalf's nonverbal and paraverbal behaviors (gaze, hand gestures, intonation) from a single discourse structure representation of propositional content.

Currently, Gandalf produces short, canned responses with embedded intonational markings also defined in advance. Its turns are always co-extensive with a single utterance. We are currently enhancing the system by adding multimodal dialogue generation capabilities (speech, intonation, turn-taking) from the knowledge base and a discourse model. A consequence of these extensions is that utterances can be longer and contain more information. Additional work in our lab on intonation and turn-taking supports the existence of nonverbal behavior within utterances and, as such, is compatible with Gandalf's present nonverbal behavior generation capabilities across utterances. The work includes implementing the automatic assignment of suitable gaze behavior and generating intonational markings for the thematic (with a look-away and a rise-fall-rise tune respectively) and the rhematic (with a look-toward and rise-fall tune respectively) constituents of an utterance. Although these additional nonverbal behaviors are subtle, Thórisson et al. [15] confirms that such minute behaviors, supportive of ongoing communication, increase believability and effectiveness.

Conclusions and Future Work

Previous research on gaze behavior has focused primarily on its role in turn-taking. However, as our data shows, turn-taking only partially accounts for the gaze behavior in discourse. Although our preliminary findings are consistent with the conclusions drawn in turn-taking taking research, our data suggests that a better explanation for gaze behavior integrates turn-taking with the information structure of the propositional content of an utterance. 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—in fact, absolutely—predictive of gaze behavior.

[1] Currently users wear cybergloves and an eyetracker so that the system has access to their gestures and gaze. A computational vision approach to this issue is currently under development.

Why might there be such a link between gaze and information structure? The literature on gaze behavior and turn-taking suggests that speakers look toward hearers at the ends of turns to signal that the floor is “available” -- that hearers may take the turn. Our findings suggest that speakers look toward hearers at the begining of the rheme -- that is, when new information or the key point of the contribution is being conveyed. Gaze here may focus the attention of speaker and hearer on this key part of the utterance. And, of course, signaling the new contribution of the utterance and signaling that one is finished speaking are not entirely independent. Speakers may be more likely to give up the turn once they have conveyed the rhematic material of their contribution to the dialogue. In this case, gaze behavior is signaling a particular kind of relationship between information structure and turn-taking.

We are currently implementing the algorithm proposed here in Gandalf, an existing communicative humanoid agent [14], to provide it with capabilities for more natural gaze behavior related to propositional content. We would like to use a similar algorithm, along with information about intonation, to predict when rhematic units occur in *input* -- that is, when users have uttered the key contribution of their utterance. This would allow us to focus speech understanding efforts on this part of the utterance. And ultimately we hope that this research and other research along these lines will allow the Turing test to be taken face-to-face.

References

1. Cassell, J., Pelachaud, C., Badler, N.I., Steedman, M., Achorn, B., Beckett, T., Douville, B., Prevost, S. & Stone, M. (1994). “Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents.” *Computer Graphics 94*.

2. Cassell, J. and S. Prevost. (under review) "Embodied Natural Language Generation: A Framework for Generating Speech and Gesture."

3. Duncan, S. Jr. (1972). “Some Signals and Rules for Taking Speaking Turns in Conversations.” *Journal of Personality and Social Psychology*, 23(2), 283-292.

4. Duncan, S. Jr. (1974). “On the Structure of Speaker-Auditor Interaction during Speaking Turns.” *Language in Society*, 3(2), 161-180.

5. Goodwin, C. (1981) *Conversational Organization: Interaction between Hearers and Speakers*. New York, NY: Academic Press.

6. Halliday, M. (1967). *Intonation and Grammar in British English*. Mouton: The Hague.

7. Luff, P., Gilbert N., Frohlich, D., eds. (1990). *Computers and Conversation*. New York, NY: Academic Press.

8. Pelachaud, C., Badler N., & Steedman. (1996) “Generating Facial Expressions for Speech.” *Cognitive Science*, 20(1).

9. Prevost, S. (1996). "An Information Structural Approach to Spoken Language Generation." *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*.

10. Sacks, H. (1992). *Lectures on Conversation, Vol. I and II*. Cambridge, MA: Blackwell.

11. Sacks, H., Schegloff, E.A. & Jefferson, G. A. (1974). "A Simplest Systematics for the Organization of Turn-Taking in Conversation." *Language*, 50, 996-735.

12. Takeuchi, A. & Nagao, K. (1993) "Communicative Facial Displays as a New Conversational Modality." *Proceedings of InterCHI 93*.

13. Thórisson, K.R. (1996). "Communicative Humanoids: A Computational Model of Psychosocial Dialogue Skills." PhD thesis, Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

14. Thorisson, K.R. (1997). "Gandalf: An Embodied Humanoid Capable of Real-Time Multimodal Dialogue with People." *Autonomous Agents 97*.

15. Thorisson, K. & Cassell, J. (1996). "Why Put an Agent in a Body: The Importance of Communicative Feedback in Human-Humanoid Dialogue." *Lifelike Computer Characters 96*.

16. Whittaker, S. & Stenton, P. (1988). "Cues and Control in Expert-Client Dialogues." *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*.

17. Vallduvi, E. (1990). "The Informational Component." PhD thesis, University of Pennsylvania, Philadelphia, PA.