Multimodal Communication in Computer-Mediated Map Task Scenarios

Max M. Louwerse (mlouwerse@ memphis.edu)[a]
Patrick Jeuniaux (pjeuniau@memphis.edu)[a]
Mohammed E. Hoque (mhoque@memphis.edu)[c]
Jie Wu (wjie@mail.psyc.memphis.edu)[b]
Gwyneth Lewis (glewis@mail.psyc.memphis.edu)[a]

Department of Psychology / Institute for Intelligent Systems[a]
Department of Computer Science / Institute for Intelligent Systems[b]
Department of Electrical and Computer Engineering / Institute for Intelligent Systems[c]
Memphis, TN 38152 USA

Abstract

Multimodal communication involves the co-occurrence of different communicative channels, including speech, eye gaze and facial expressions. The questions addressed in this study are how these modalities correlate and how they are aligned to the discourse structure. The study focuses on a map task scenario whereby participants coordinate a route on a map, while their speech, eye gaze, face and torso are recorded. Results show that eye gaze, facial expression and pauses correlate at certain points in the discourse and that these points can be identified by the speaker's intentions behind the dialog moves. This study thereby sheds light on multimodal communication in humans and gives guidelines for implementation in animated conversational agents.

Introduction

Most communicative processes require multiple channels, both linguistic and paralinguistic (Clark, 1996). For instance, we talk on the phone while gesturing, we seek eye contact when we want to speak, we maintain eye contact to ensure that the dialog participant comprehends us, and we express our emotional and cognitive states through facial expressions. These different communicative channels play an important role in the interpretation of an utterance by the dialog partner. For instance, the interpretation of "Are you hungry?" depends on the context (e.g. just before going to a restaurant, during dinner), depends on eye gaze (staring somebody in the eyes or looking away), depends on prosody (e.g. stress on 'you' or 'hungry'), facial expressions (e.g. surprised look, disgusted look) and gestures (e.g. rubbing stomach, pointing at a restaurant). While multimodal communication is easy to comprehend for dialog participants, it is hard to monitor and analyze for researchers.

Despite the fact that we know linguistic modalities (e.g. dialog move, intonation, pause) and paralinguistic modalities (e.g. facial expressions, eye gaze, gestures) co-occur in communication, the exact nature of their interaction remains unclear (Louwerse, Bard, Steedman, Graesser & Hu, 2004). There are two primary reasons why an insight in the interaction of modalities in the communicative process is beneficial.

First, from a psychological point of view it helps us understand how communicative processes take shape in the minds of dialog participants. Under what psychological conditions are different channels aligned? Does a channel add information to the communicative process or does it merely co-occur with other channels? Research in psychology has shed light on the interaction of modalities, for instance comparing eye gaze (Argyle & Cook, 1976; Doherty-Sneddon, et al. 1997), gestures (Goldin-Meadow, 2003; Louwerse & Bangerter, 2005; McNeill, 1992) and facial expressions (Ekman, 1979) but many questions regarding multiple – i.e., more than pairs of – channels and their alignment remain unanswered.

Second, insight in multimodal communication is beneficial from a computational point of view, for instance in the development of animated conversational agents (Louwerse, Graesser, Lu, & Mitchell, 2004). The naturalness of the human-computer interaction can be maximized by the use of animated conversational agents, because of the availability of both linguistic (semantics, syntax) and paralinguistic (pragmatic, sociological) features. These animated agents have anthropomorphic, automated, talking heads with facial features and gestures that are coordinated with text-to-speech-engines (Cassell & Thórisson, 1999; Massaro & Cohen, 1994; Picard, 1997). Examples of these agents are Baldi (Massaro & Cohen, 1994), COSMO (Lester, Stone & Stelling, 1999), STEVE (Rickel & Johnson, 1999), Herman the Bug (Lester, Stone, Stelling, 1999) and AutoTutor (Graesser, Person, et al., 2001). Though the naturalness of these agents is progressively changing, there is room for improvement. Current agents for instance incessantly stare at the dialog partner, use limited facial features rather randomly, or produce bursts of unpaused speech. Both psycholinguistics and computational linguistics would thus benefit from answers to questions regarding the interaction of multimodal channels.

A specific and related question concerns the mapping of these channels onto the discourse structure. Research has shown that the structure of the dialog can often predict these modalities. For instance, Taylor, King, Isard, & Wright (1998) and Hastie-Wright, Poesio, and Isard (2002) have

shown that speech recognition can be improved by taking into account the sequence of dialogue moves (for example, answers following questions) and Flecha-Garcia (2002) has shown that eye brow movements can partly be explained by linguistic communicative functions related to dialogue structure.

In this exploratory study, we investigate the mappings of modalities onto the dialog structure as well as their correlations. The current paper reports on some initial findings of a study investigating these questions in computer mediated discourse between two dialog participants. The study reported here is part of a series conducted in a project investigating multimodal communication in humans and agents.

Intelligent Map Task Agent

In an ongoing project on multimodal communication in humans and agents, we are investigating the interaction between linguistic modalities, like prosody and dialog structure, and non-linguistic modalities, like eye gaze and facial expressions. The project aims to determine how these modalities are aligned, whether, and if so, when these modalities are observed by the interlocutor and whether the correct use of these channels actually aids the interlocutor's comprehension. Answers to these questions should provide a better understanding of the use of communicative resources in discourse and can subsequently aid the development of more effective animated conversational agents.

With so many variables in multimodal communication, it is desirable to control for genre, topic, and goals of unscripted dialogs. We therefore used the Map Task scenario (Anderson, et al., 1991). The Map Task is a restricted-domain route-communication task which makes clear to experimenters exactly what each participant knows at any given time. The Instruction Giver (Giver) coaches the Instruction Follower (Follower) through a route on the map. By way of instructions, participants are told that they and their interlocutors have maps of the same location but drawn by different explorers and so potentially different in detail. They are not told where or how the maps differ. The maps are of fictional locations and participants have only three sources of knowledge in their initial encounter with a map: 1) the instructions, 2) what appears on his/her map (cartoon landmarks, their labels, an in the case of the giver, the location of the route) and 3) what has been expressed during the map task dialogue, in terms of language, speech, eye contact, facial expressions, gestures. The route on the Giver's map is designed, however, to maintain not only the alternation of matching and mismatching landmarks, but also an 'easy stages' rule: the next landmark critical to the route will almost always be one of the landmarks closest to the current landmark.

Method

Two dialog partners collaborated in a Map Task scenario whereby their communication was recorded with camcorders, speech recorders and eye trackers. The analysis reported below focuses on the alignment of cognitive states as expressed through facial expressions, the eye gaze and pauses to the dialog structure.

Participants

Eight undergraduate students at the University of Memphis participated in this study, six females and two males. The participants received extra credit in an undergraduate course for participating in this study.

Materials

Four different maps were created based on the maps used in the HCRC Map Task corpus (Anderson, et al., 1991). The maps used different types of objects (cars, churches, houses, trees, lakes) and different colors in order to elicit dialog with various contrasts. The average number of object types was 7.8 (SD = 3.98) and 53 (SD = 5.79) object tokens. For the analysis reported below, the four maps were considered random variables.

Apparatus

Participants' communication was recorded by five Panasonic camcorders, two capturing the faces of each dialog participant (PV-GS31), two capturing the upper torsos of each participant (PV-GS150) and one capturing both participants from a bird's eye view (PV-GS150). Eye gaze was recorded for the Giver only using an SMI iView RED remote eye tracker. Speech was recorded using a Marantz PMD670 recorder whereby Giver and Follower were recorded on two separate (left and right) channels using two AKG C420 headset microphones.

Procedure

Participants were seated adjacent to each other, separated by a divider to ensure that they could not see each other. They communicated through microphones and headphones, while they could see the upper torso of the dialog partner and the map both on a computer monitor in front of them. Participants were randomly assigned to a Giver or Follower role.

Before the conversation started, all equipment was calibrated. The five camcorders were positioned and focused in order to best capture facial expressions and upper torso. The eye gaze of the Information Giver was calibrated using nine calibration points on the screen. To avoid interruption of the dialog, calibration only occurred once per map. If calibration was lost in the experiment, unbeknownst to the participant recording of eye tracking data was discontinued. The experiment started with a flash of light, in order to ensure alignment of the different channels for the data analysis.

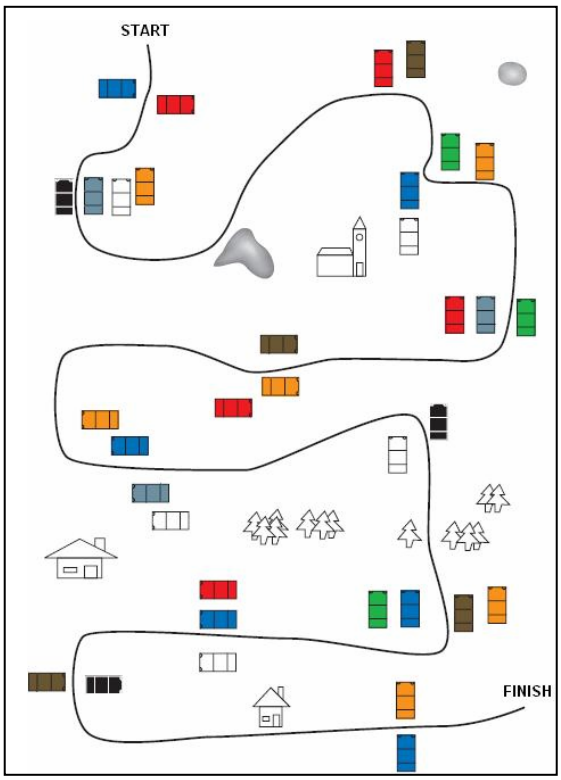

Figure 1. Example of map

The Giver was presented with a map in color similar to the one presented in Figure 1, with a route drawn on it. The Follower had a slightly different map without the route drawn on it. Followers were able to draw a route on the map using a stylus or a mouse.

Results and Discussion

The average duration of the four dialogs was 354.5 seconds (*SD* = 81.51). These durations are comparable with the average duration of Map Task dialogs, 407.75 seconds (Branigan, Lickley, & McKelvie, 1999).

All modalities were aligned at 250 milliseconds accurate time stamp. Two types of analyses were conducted: 1) a correlation analysis between the different modalities of cognitive state, eye gaze, eye blinks and pause; 2) an analysis of difference between dialog moves for each of the modalities variables.

Dialog moves

Three students in the Institute for Intelligent Systems transcribed the audio of the interactions between Information Giver and Information Follower. Next they classified and time-stamped (250ms intervals) these utterances in the 12 conversational game moves described by Carletta et al. (1997). An overview of these dialog moves is presented in Table 1. Kappas between the coders were acceptable at .53, higher than Kappa scores of .39 between human coders obtained in a different study (Louwerse & Crossley, 2006).

Eye gaze

Two dependent measures were recorded for the Giver's eye gaze. Total fixation times were computed for the two areas of interest, the Follower and the Giver's map. In addition, number of blinks was computed.

Pause

Pause was analyzed using the upper intensity limit and duration. In measurement of intensity, minimum pitch specifies the minimum periodicity frequency in any signal. In our case, 100 Hz for minimum pitch yielded a sharp contour for the intensity. Audio frames with intensity values consistently less than 50 dB and for the duration of longer than .5 seconds were classified as pauses. The 50 dB for intensity threshold was chosen to be the optimal value based on experimentation to capture all the utterances while rejecting the absolute silence during the interaction. The speech processing software *Praat* (Boersma & Weenink, 2006) was used to perform all the calculations to identify the pause regions.

Table 1. The 12 move types used in the Map Task, their frequency in percentages, a description and an example
Note: 'IG' and 'IF' refer to the frequency of dialog moves in Givers and Followers.

Dialog Act	IG	IF	Description	Example
INSTRUCT	38	0	Commands partner to carry out action	*Start at the top and go down between the blue and the red car.*
EXPLAIN	28	8	States information not directly elicited by partner	*Ok I went the wrong way.*
CHECK	9	12	Requests partner to confirm information	*So, between the black and the grey one?*
ALIGN	15	3	Checks attention, readiness, agreement of partner	*Ok, do you see those two blue cars?*
QUERY-YN	4	8	Yes/no question that is not CHECK or ALIGN	*Do you see that?*
QUERY-W	7	17	Any query not covered by the other categories	*If I'm at the red car what do I do there?*
ACKNOWL	14	7	Verbal response minimally showing understanding	*Uh huh.*
REPLY-Y	11	107	Reply to any yes/no query with yes-response	*Yeah, start at the top.*
REPLY-N	4	24	Reply to any yes/no query with no-response	*No, go like above the puddle.*
REPLY-W	2	1	Reply to any type of query other than 'yes or 'no'	*It goes below.*
CLARIFY	13	3	Reply to question over and above what was asked	*So you'll be between the blue and red car.*
READY	28	8	Preparing conversation for new dialog game	*Alright. We're going to move to the left.*

Cognitive states

Cognitive states were coded by three judges using the facial expression video footage. Standard emotion coding schemes, like Ekman, Friesen, Wallace and Hager's (2002) facial action coding scheme are problematic for Map Task scenarios, because negative cognitive states like disgust, anger or sadness do not occur in these interactions. Instead, we used a coding scheme inspired by Craig, Graesser, Sullins, and Gholson's (2004) and Kort, Reilly and Picard's (2001) schemes for affect in learning. Nine cognitive states were distinguished (Table 2). To account for degrees of these cognitive states, the average monothetic ratings of the cognitive states were used in the analysis.

Table 2. Cognitive states determined by facial expressions

Cognitive state	Description
Distracted	Directing attention away from the task; broken concentration
Uncertain	Hesitation or doubt; lack of assurance
Confused	Lack of understanding
Frustrated	Annoyance or irritation
Confident	Expression of assurance and certainty
Engaged	Heavy and uninterrupted involvement in activity
Encouraged	Inspiration and motivation
Interested	Expression of attention
Bored	Lack of interest in the activity

Correlations between modalities

Cognitive states, eye gaze and pauses were compared within both Giver and Follower, except for eye gaze, which was only recorded for the Giver.

Correlations between these modalities are presented in Table 3 (Giver) and 4 (Follower). The Givers' eye gaze on the Follower correlated significantly with the cognitive states of Engagement, Uncertainty and Boredom. In fact, Engagement and Uncertainty also correlated with the fixations on the map. Givers heavily involved in the task seemed to pay more attention to both the dialog partner and the map in front of them, either because they are absorbed in the task or because they are uncertain about an aspect of that task. The same patterns can also be found for the Follower. Though the Follower's eye gaze was not recorded, the coding for the Follower moving their eyes away from the map gives an adequate approximation of fixation on the Giver. Again, Uncertainty and Engagement are the cognitive states during which this happens most frequently.

Blinks co-occur with many of the cognitive states. It is noteworthy that, in addition to cognitive states like Engagement and Uncertainty, they correlate with the cognitive state of Confusion. On the contrary, when Givers feel confident, they pause less and look less at Follower, as suggested by the significant negative correlations.

Table 3. Correlations modalities Giver
Notes: $^{**}p < .01$; $^{*}p < .05$

	Looking Away	Fixation on person	Fixation on map	Blink	Pause
Distracted	-.065	-.071	-.083	-.075	-.078
Uncertain	.338**	.249*	.168*	.712**	.703**
Confused	.172*	.086	-.027	.538**	.443**
Confident	-.143*	-.067	-.112	-.103	-.155*
Engaged	.163*	.314**	.438**	.893**	.676**
Encouraged	.172*	-.034	.006	.258**	.147*
Interested	-.081	.006	-.020	-.028	-.097
Bored	-.026	.212**	.062	.023	-.003
Pause	.413*	.447**	.659**	.819**	

Table 4. Correlations modalities Follower
Notes: $^{**}p < .01$; $^{*}p < .05$

	Looking away	Pause
Distracted	-.037	-.059
Uncertain	.153*	-.086
Confused	.151*	.091
Confident	-.050	.079
Engaged	.229**	.095
Encouraged	.129	.146
Interested	-.038	.134
Bored	.336	-.139
Pause	.025	

While the correlations for looking away and cognitive states are similar between Giver and Follower, correlations between pause and cognitive states differ considerably. The most likely explanation for this is the difference in average length of an utterance between Givers and Followers. A Giver's turn is on average 5.87 seconds, while for Followers the average lies at .88 seconds. It is therefore important to normalize the utterances for length. This is what was done in the next analysis.

Differences between dialog moves

To allow for a comparison between dialog moves, the duration of the move must be taken into account. For instance, the INSTRUCT move that describes an action to the dialog partner necessarily takes more time than a REPLY-Y move that simply states "yes". The likelihoods of pausing in speech, of changing eye gaze or to express cognitive states are therefore necessarily more frequent in longer moves. The dependent variables were therefore normalized by the duration of the dialog move.

Main effects for dialog move were found in eye gaze of the Giver on the Follower, a 500 ms. pause as well as the cognitive states for Confused, Engaged, Encouraged and Distracted for the Giver and Confused, Confident, Distracted and Interested for the Follower. No effects were found between dialog moves for the cognitive states for Bored or Uncertain.

Table 5. Main effects for dialog move.
Notes: $^{**}p < .01$; $^{*}p < .05$

Dependent variable	Role	F	MS_e	Salient dialog move
Gaze on follower	Giver	1.89^{*}	.02	REPLY-Y, ALIGN
2500 ms Pause	Giver	1.83^{*}	.01	ACKNOWLEDGE
Confident	Follower	2.39^{**}	.02	CHECK, QUERY-YN, REPLY-Y
Confused	Giver	13.62^{**}	.01	QUERY-W
Confused	Follower	3.26^{**}	.02	REPLY-W
Distracted	Giver	3.65^{**}	.01	QUERY-W
Distracted	Follower	1.98^{**}	.01	READY
Encouraged	Giver	2.93^{**}	.02	INSTRUCT
Engaged	Giver	2.89^{**}	.02	INSTRUCT
Interested	Follower	2.70^{**}	.01	CHECK, QUERY-YN

An overview of these main effects is given in Table 5, as well as the salient dialog moves significant from the other dialog moves in a Bonferroni post-hoc test. These post-hoc tests revealed the exact nature of the differences between the dialog moves. The differences for Giver's eye gaze on the follower were primarily due to the REPLY-Y and the ALIGN move. Apparently Givers look more at the Followers when they confirm a YN-QUESTION or – not surprisingly – when they check the attention, readiness and agreement of the dialog partner.

The expression of Distraction in the Giver's cognitive states can particularly be found in QUERY-W. This may be explained by the Giver moving attention away from the conversation to formulate a question. The QUERY-W move differed from the other dialog moves in that it had higher frequencies of the cognitive state Confusion. This effect was both found in Givers and Followers. In a QUERY-W move the speaker asks for information that goes beyond a request for a confirmation from the dialog partner. It is therefore not surprising that the QUERY-W move is frequent in a cognitive state of confusion. On the other hand, where the speaker has the information to provide to the dialog partner, as is the case in the INSTRUCT dialog move, the opposite is true: the Giver expresses cognitive states of Encouragement and Engagement. Similarly, the Follower merely needs a confirmation of information the speaker has and expresses cognitive states of Confidence in dialog moves like CHECK, QUERY-YN, or confirms this information as in REPLY-Y. Finally, ACKNOWLEDGMENT moves likely mark the higher frequency of pauses in the speaker in the line of the partial acknowledgement, whereby the speaker has some hesitations, as described in Louwerse and Mitchell (2004).

Conclusion

The study reported in this paper is the first in a series investigating multimodal communication in humans and agents. This study is very much exploratory in nature, but it allows us to have a closer look at the complex interplay between different modalities in conversation. Studies like this elicit many research questions. For instance, what is the effect of the map on multimodal communication? What role do individual differences (personalities, sex, culture) play? What are the effects of modalities not discussed here, like eye brow movement, intonation contours or gestures? What are the best coding schemes for the various modalities in specific tasks, like the Map Task? What are the differences in our findings between computer-mediated interaction and direct face-to-face interaction?

Although studies investigating multimodal communication generate many research questions and are far from easy to conduct, they also provide a wealth of information on how humans communicate. The present study has provided an insight into how eye gaze, facial expression and pauses correlate with each other at certain points in the discourse and that these points can be identified by dialog moves. From a psycholinguistic perspective, using dialog moves as the unit of analysis helps to align the various modalities across various channels. From a computational linguistic perspective, using dialog moves helps to generate these modalities, which can in turn be studied for their alignment with discourse structure. Interdisciplinary research will bring us progressively closer to answers to questions that dialog participants generally do not consider, because for them multimodal communication is so natural and seemingly easy.

Acknowledgments

This research was supported by grant NSF-IIS-0416128. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding institution. We would like to thank Ellen Bard and Markus Guhe for the creation of the maps and their help in setting up the study, and Dominique Crocitto, Fatema Julia, Fang Yang and Megan Zirnstein for their help in the data collection and analyses.

References

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G. M., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S. & Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech, 34*, pp. 351-366.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Boersma, P. & Weenink, D. (2006). *Praat: doing phonetics by computer* (Version 4.4.06) [Computer program]. Retrieved January 30, 2006, from http://www.praat.org/

Branigan, H. Lickley, R., & McKelvie, D. (1999). Non-Linguistic Influences on Rates of Disfluency in Spontaneous Speech. *Proceedings of the 14th International Conference of Phonetic Sciences*, pp. 387–390.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23*, 13-31.

Cassell, J. & Thórisson, K. R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence, 13,* 519-538.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media, 29*, 241-250.

Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance. *Journal of Experimental Psychology: Applied, 3*, 105-125.

Ekman, P. (1979). About brows: Emotional and conversational signals. In M. von Cranach, K. Froppa, W. Lepenies, & D. Ploog (Eds.), *Human ethology: Claims and limits of a new discipline: Contributions to the colloquium* (pp. 169-248). Cambridge: Cambridge University Press.

Ekman, P., Friesen, Wallace V., & Hager, J.C. (2002). *Facial Action Coding System (FACS)*. CD-ROM.

Flecha-Garcia, M.L. (2002). Eyebrow raising and communication in map task dialogues. *Gesture: The Living Medium*. University of Texas at Austin.

Graesser, A., Person, N., Harter D., & the Tutoring Research Group. (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education, 12*, 23-39.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Harvard University Press

Hastie-Wright, H., Poesio, M., & Isard, S. (2002). Automatically predicting dialogue structure using prosodic features. *Speech-Communication, 36,* 63-79.

Kort, B., Reilly, R., & Picard, R.W. (2001). An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy-Building a Learning Companion. In *Proceedings of International Conference on Advanced Learning Technologies (ICALT 2001)*, Madison Wisconsin, August 2001.

Lester, J., Stone, B., & Stelling, G. (1999). Lifelike pedagogical agents for mixed initiative problem solving in constructive learning environments. *User Modeling User-Adapted Interaction, 9,* 1-44.

Louwerse, M.M., Bangerter, A. (2005). Focusing attention with deictic gestures and linguistic expressions. *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*.

Louwerse, M.M., Bard, E.G., Steedman, M., Hu, X., & Graesser, A.C. (2004). *Tracking multimodal communication in humans and agents*. Technical report, Institute for Intelligent Systems, University of Memphis, Memphis, TN.

Louwerse, M.M. & Crossley, S. (2006). Dialog act classification using n-gram algorithms. In *Proceedings of the Florida Artificial Intelligence Research Society International Conference (FLAIRS)*. Menlo Park, CA: AAAI Press.

Louwerse, M.M., Graesser, A.C., Lu, S., Mitchell, H. H. (2005). Social cues in animated conversational agents. *Applied Cognitive Psychology, 19*, 1-12.

Louwerse, M.M., & Mitchell, H.H. (2003). Towards a taxonomy of a set of discourse markers in dialog: a theoretical and computational linguistic account. *Discourse Processes, 35*, 199-239.

Massaro, D. W., & Cohen, M. M. (1994). Visual, orthographic, phonological, and lexical influences in reading. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1107- 1128.

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

Picard, R. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Rickel,J. & Johnson, W.L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence, 13,* 343-382.

Taylor, P., King, S., Isard, S., & Wright, H. (1998). Intonation and dialogue context as constraints for speech recognition. *Language and Speech, 41*, 493-512.