Cognitive Science 32 (2008) 921–935

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802222039

The Deep Versus the Shallow: Effects of Co-Speech Gestures in Learning From Discourse

Ilaria Cutica, Monica Bucciarelli

Department of Psychology, Center for Cognitive Science, University of Turin

Received 25 January 2006; received in revised form 17 October 2007; accepted 6 November 2007

Abstract

This study concerned the role of gestures that accompany discourse in deep learning processes. We assumed that co-speech gestures favor the construction of a complete mental representation of the discourse content, and we tested the predictions that a discourse accompanied by gestures, as compared with a discourse not accompanied by gestures, should result in better recollection of conceptual information, a greater number of discourse-based inferences drawn from the information explicitly stated in the discourse, and poorer recognition of verbatim of the discourse. The results of three experiments confirmed these predictions.

Keywords: Deep learning; Co-speech gestures

1. Introduction

The act of speaking is often accompanied by movements of the arms and hands (co-speech gestures). A well-established literature has already shown the influence of the speaker's gestures on the hearer's comprehension of a discourse. However, why gestures improve discourse comprehension is still not fully understood. In this article, we propose a tentative explanation.

According to the literature on text comprehension (e.g., Graesser, Millis, & Zwaan, 1997; McNamara, Miller, & Bransford, 1991; Zwaan, Magliano, & Graesser, 1995; Zwaan & Radvansky, 1998), deep learning involves the construction and manipulation of mental representations that reproduce the state of affairs described. The listener constructs such mental representations on the basis of the semantic and pragmatic information contained in the text, together with his or her prior knowledge, and any inferences that are drawn; they do not generally contain surface information (the linguistic form of sentences). According to

Correspondence should be sent to Ilaria Cutica, Department of Psychology, Center for Cognitive Science, Via Po, 14–10123, Turin, Italy. E-mail: cutica@psych.unito.it

different theoretical frameworks, such representations are referred to as the "mental model" (Johnson-Laird, 1983, 2006) or "situation model" (van Dijk & Kintsch, 1983; an extension is the Construction Integration Model of Comprehension—Kintsch, 1998). We consider the two terms to be equivalent, disregarding their different theoretical roots (see also Kaup, Kelter, & Habel, 1999). In ideal circumstances, in discourse comprehension people construct a model for each sentence, integrate such models also taking into account their prior knowledge, and consider what, if anything, follows (discourse-based inferences). Several factors may interfere with the comprehension process and its end; the aim of our study was to investigate the facilitating role of co-speech gestures.

2. Learning, diagrams, and gestures

Experimental studies have revealed that discourse comprehension benefits from co-speech gestures produced by the speaker (e.g., see Goldin-Meadow, 1999; Iverson & Goldin-Meadow, 2001; Kelly, Barr, Church, & Lynch, 1999; McNeil, Alibali, & Evans, 2000). However, there is still a need for studies that investigate the mechanisms and mental representations that may account for the observed facilitating effect of co-speech gestures. One relevant finding is that learning from text benefits from the visual presentation of material (e.g., Carney & Levin, 2002; Moreno & Mayer, 2002; Vekiri, 2002), and that information from visually presented material is incorporated together with information derived from the text into an integrated model (e.g., Glenberg & Langston, 1992). Indeed, we observed that co-speech gestures, as well as visually presented material, convey information in a non-discrete representational format (see Bucciarelli, 2007); as models are non-discrete mental representations (see Hildebrandt, Moratz, Rickheit, & Sagerer, 1999; Rickheit & Sichelschmidt, 1999), co-speech gestures may result in representations that are easily included together in the mental representation of discourse.

We assumed that such enriched models facilitate the retention of content information, and the drawing of correct inferences, at the expense of reducing memory for the surface code. Indeed, we know from the literature (Johnson-Laird, 1983; Johnson-Laird & Byrne, 1991) that individuals who have built an articulated model of a given material are more likely to draw correct inferences from the information explicitly contained in that material, compared to individuals who have built a less articulated model. For our purposes, we distinguished between discourse-based inferences and elaborative inferences, arguing that only the former are based on mental models. Discourse-based inferences make explicit that information which is originally implicit in the text; they may regard, for instance, the causal antecedent, the causal consequent, and the character's mental states (i.e., beliefs and intentions) with respect to the actions described (Graesser, Singer & Trabasso, 1994). Elaborative inferences (e.g., Singer, 1994) are instead a sort of enrichment of the original text.

We also assumed that enriched models of the discourse lead to a poor retention of the surface form of the text (see Garnham, Oakhill & Cain, 1998; Johnson-Laird & Stevenson, 1970). Indeed, we know from the literature that model representations do not generally contain surface information (the linguistic form of sentences; see Johnson-Laird, 1983).

3. Experiments

3.1. Experiment 1: A study on gestures and learning from discourse (recall)

We expected to find that a person listening to a discourse accompanied by gestures, compared to a person listening to a non-gesticulating speaker, (a) retains more information and (b) draws more discourse-based inferences. We had no predictions for the number of elaborative inferences because these do not depend upon model construction. Neither did we have predictions for the number of erroneous recollections because models do not prevent a person from making mistakes: If the hearer misunderstands some piece of information, the misunderstood information may be included in the mental model, thus supporting a wrong recollection.

3.1.1. Materials

The experimental material comprised two videotaped fictions (approximate length: 6 min) in which an actor delivered a discourse while a series of events occurred at a funfair (see Appendix A.1 for an excerpt). The discourse was produced under two different conditions: the *Gesture condition,* in which the actor accompanied the discourse with gestures (he was instructed to produce hand and arm movements as he felt appropriate with respect to the discourse flow); and the *No-Gesture condition,* in which the actor delivered the discourse without gesticulating. Examples of gestures are included in Appendix A.2. Two judges examined each gesture produced by the actor in the Gesture condition. They excluded the possibility that any of them conveyed information that was implicit or absent in the co-occurring speech. In particular, the two judges found that the actor never produced symbolic gestures (i.e., gestures with a widely recognized conventionalized meaning). This procedure was necessary to ascertain that the actor's gestures were not the source of any of the participants' discourse-based inferences.

3.1.2. Procedures

3.1.2.1. Pilot experiment: In order to make sure that the two videotaped fictions did not differ for any other factors such as, for instance, the intonation of the actor's voice, we carried out a pilot experiment involving 20 students reading psychology at Turin University. One half of them were presented with the audio recording pertaining to the fiction in the Gesture condition, and one half were presented with the audio recording pertaining to the No-Gesture condition. They were invited to listen to the audio recording; as soon as it ended, they were asked to recall as much information as they could. All of them were video recorded. To code the results, the discourse was divided into 54 semantic units, corresponding to as many main concepts that the hearer could recall. Two independent judges coded the participants' recollections individually; they reached a significant level of agreement on their first judgments (Cohen's K ranging from .93–.95; all ps $< .001$). For the final score, the judges discussed each item on which they disagreed, until reaching a full agreement. Each concept (i.e., semantic unit) recalled by the participants was evaluated according to the following coding schema:

- Correct recollection: a semantic unit recollected either in its literality or as a paraphrase.
- Discourse-based inference: a recollection in which the participant gave explicit information that was originally implicit in the semantic unit.

- Elaborative inference: a semantic unit recollected with the addition of plausible details.
- Erroneous recollection: a recollection with a meaning that was inconsistent with the semantic unit.

Consider, for instance, the following semantic unit in the discourse: "Night was falling, and the shadows were growing longer." According to the coding schema, the statement, "It was beginning to grow dark," was a correct recollection; the statement, "Since it was getting dark, the search became more difficult," was a discourse-based inference (because it refers to a causal consequent); the statement, "Suddenly everything [*the roundabouts*] closed down, and the shadows were growing longer," was an erroneous recollection. Now, consider the following semantic unit in the discourse: "She was clinging all alone to an apocalyptical beast"; according to the coding schema, "She was on the roundabout, her arms around a golden-headed monster" was an elaborative inference. The results of a series of t tests for independent samples showed no difference between performance by the participants assigned to the two conditions for any coding category, $t(18) = -0.14$ to 0.89, $p = .39$ to 1. We concluded that there was no difference in the two audio recordings pertaining to the different experimental conditions.

Experiment 1 participants, randomly assigned to one of the two experimental conditions, were invited to watch the videotaped fiction carefully, paying attention to what the actor said. They were told that at the end of the tape they would be asked to tell the experimenter all they could remember of the discourse. As soon as the video finished, they were asked to recall all they could remember (a free-recall task); all of them were video recorded.

3.1.3. Participants

Thirty-eight students from Turin University (32 women and 6 men; mean age = 23) participated. One half (balanced for age and gender) were assigned to the Gesture condition and one half to the No-Gesture condition.

3.1.4. Results

Considering the 54 semantic units, three independent judges coded the participants' recollections individually. The judges reached a significant level of agreement on their first judgments (Cohen's K ranging from .92–.94; all $ps < .001$). For the final score, the judges discussed each item on which they disagreed, until reaching a full agreement. Table 1 shows the mean scores for types of recollection in the two experimental conditions.

As a general result, and in both conditions, scores for correct recollections were higher than for other sorts of recollections. As predicted, there were more correct recollections and discourse-based inferences in the Gesture condition than in the No-Gesture condition (t test for independent samples): $t(36) = 5.60$, tied $p < .0001$ and $t(36) = 2.94$, tied $p = .004$, respectively. Elaborative inferences and erroneous recollections occurred to the same extent in the Gesture condition and in the No-Gesture condition (t test for independent samples): $t(36) = 1.63$, $p = .11$ and $t(36) = -0.82$, $p = .42$, respectively.[1]

Table 1
Mean types of recollections in the Gestures and No-Gestures conditions in Experiment 1

Condition	Correct Recollections	Discourse-Based Inferences	Elaborative Inferences	Errors
Gestures ($n = 19$)				
M	21.89	1.05	.95	.53
SD	3.54	1.17	.78	.51
No Gestures ($n = 19$)				
M	15.32	0.21	.58	.68
SD	3.69	0.42	.61	.67

3.1.5. Discussion

The results of Experiment 1 confirmed our expectations, suggesting that gestures facilitate the construction of a mental model from a discourse by favoring both the retention of correct information and the possibility of drawing discourse-based inferences. Our assumptions would be further supported by evidence that a person listening to a speaker who does not gesticulate recovers the literality of the discourse more easily than a listener who sees the speaker gesticulate. We thus devised a second experiment to explore this possibility.

3.2. Experiment 2: A study on gestures and learning from discourse

If gestures favor the construction of mental models, which do not contain the textbase (Johnson-Laird, 1983), then they should penalize the retention of the verbatim of the discourse. In Experiment 2, we aimed to verify whether a discourse delivered without gesturing, as compared with a discourse accompanied by gestures, results in better performance in recognizing verbatim of the sentences in the discourse.

3.2.1. Materials

The materials were the same as in Experiment 1; namely, the videotaped discourse in both the Gesture and the No-Gesture conditions.

3.2.2. Procedures

Participants, randomly assigned to one condition, were instructed to watch the video of the discourse carefully, paying attention to what the actor said. Further, they were told that at the end of the tape the experimenter would ask them some questions. Thus, they were not told that later on they would be required to recognize the actual utterances they had heard. As soon as the video finished, the participants were presented with a list of sentences, one by one in a random order, and were invited to consider whether or not the sentences were identical to those actually delivered by the actor (a recognition task). The sentences presented, as compared with those in the discourse, were of the following sorts: (a) identical (*literally correct*); (b) with the same meaning, but said with different words (*paraphrases*); (c) inconsistent in meaning (*wrong content*). We created 48 sentences, with 16 in each category (examples are in Appendix A.3). We coded responses of "Yes" to literally correct sentences, and responses of "No" to paraphrases and wrong content sentences as correct.

3.2.3. Participants

Thirty students at Turin University (24 women and 6 men, mean age = 23;4) participated. One half (balanced for age and sex) were assigned to the Gesture condition and one half to the No-Gesture condition.

3.2.4. Results

A fundamental requirement in order to interpret our results was to ensure that the 16 stimuli constituting each sentence category were homogeneous in difficulty. Thus, we conducted an analysis of variance (Cochran's Q test) on the stimuli pertaining to each sentence category. The results showed the stimuli to be comparable in difficulty in both the Gesture (Q value ranging from 16.17 to 22.82; p value ranging from .10 to. 14) and the No-Gesture (Q value ranging from 8.51 to 21.20; p value ranging from 0.52 to .48) conditions.

As predicted, a series of t tests for independent samples revealed that participants in the No-Gesture condition performed better than participants in the Gesture condition in recognizing sentences actually spoken by the actor ($t(28) = –2.71$, tied $p < .006$). Moreover, participants in the No-Gesture condition performed better than participants in the Gesture condition with paraphrases, thus tending not to endorse them ($t(28) = –2.36$, tied $p = .01$). There was no difference between the two conditions in terms of accuracy when dealing with wrong content sentences ($t(28) = 1.04$, $p = .31$). Table 2 illustrates the mean correct performance in both conditions.

Our recognition test was a yes–no task involving signal trials: Participants were presented with 16 signals (i.e., literally correct sentences) and 32 noise trials (16 paraphrases and 16 errors). On signal trials, yes responses were correct (*hits*); on noise trials, yes responses were incorrect (*false alarms*). Table 3 illustrates the proportions of the sort of recognition in both experimental conditions.

We applied the signal detection theory (SDT) to assess sensitivity. In the Gesture protocol, the results of a paired t test on d′ values for the comparison Literally correct–Paraphrases (the mean d′ was 0.42) and Literally correct–Wrong content (the mean d′ was 1.19) revealed greater sensitivity in the latter comparison than in the former ($t = –7.03$; $p < .001$), whereas in the No-Gesture protocol, the results of a paired t test on d′ values for the comparison Literally correct–Paraphrases (the mean d′ was 1.15) and Literally correct–Wrong content (the mean d′ was 1.31) suggested comparable sensitivity in both comparisons ($t = –1.567$; $p = .14$). Thus,

Table 2
Mean correct performance with the different sorts of sentences in the two experimental conditions in Experiment 2

Condition	Literally Correct ($n = 16$)	Paraphrases ($n = 16$)	Wrong Content ($n = 16$)
Gestures ($n = 15$)			
M	9.53	8.87	12.93
SD	3.14	3.14	2.31
No Gestures ($n = 15$)			
M	11.87	11.20	12.13
SD	1.13	2.21	1.89

Table 3
Proportions of sorts of recognition in the Gestures and No-Gestures conditions in Experiment 2

Condition	Hit	Miss	False Alarm Paraphrases	Correct Rejections Paraphrases	False alarm Errors	Correct Rejection Errors
Gestures ($n = 15$)	143/240	97/240	107/240	133/240	46/240	194/240
No Gestures ($n = 15$)	178/240	62/240	72/240	168/240	58/240	182/240

in the presence of the actor's gestures, Paraphrases were more often mistaken for Literally correct than for Wrong content. This result was consistent with our assumptions.

3.2.5. Discussion

The experimental results supported our assumption that deep learning from discourse is facilitated by the gestures that accompany the discourse. Within our theoretical framework, learning through models ought to benefit from gestures regardless of the content of the discourse: Spatial and non-spatial contents ought to be equally well represented in a mental model. However, studies on the function of gestures for the speaker (e.g., Alibali, Kita, Bigelow, Wolfman, & Klein, 2001; Feyereisen & Harvard, 1999; Krauss, 1998; Rauscher, Krauss, & Chen, 1996) have shown that speakers tend to produce more gestures when speaking about topics that involve visual or motor imagery (i.e., spatial and movement content). To exclude the possibility that the influence of gestures on learning as found in Experiments 1 and 2 may depend upon the fact that the discourse used in those experiments had a significant spatial and movement content, we replicated both experiments using an abstract and technical discourse, with little spatial and movement content.

3.3. Experiment 3: gestures and learning from a discourse with a low spatial and movement content

Experiment 3 consisted of a Recall task and a Recognition task; participants were randomly assigned to one of the two tasks. The predictions for Experiment 3 were the same as those for Experiment 1 (Recall task) and Experiment 2 (Recognition task).

3.3.1. Materials

The experimental material comprised two videotaped fictions (approximate length: 6 min) in which the same actor delivered a discourse with or without co-speech gestures. The discourse was concerned with color perception (see Appendix B.1 for an excerpt) and had low spatial and movement content. Examples of gestures produced by the actor are included in Appendix B.2. As in Experiment 1, two judges excluded the possibility that any gesture conveyed information that was implicit or absent in the co-occurring speech.

3.3.2. Procedures

3.3.2.1. Pilot experiment: As in Experiment 1, we carried out a pilot experiment on the audio recordings of the two conditions. We tested 20 students at Turin University; one half of them were presented with the audio recording of the Gesture condition and one half with the audio recording of the No-Gesture condition. To code the results, the discourse was divided into 35 semantic units, corresponding to as many main concepts that the hearer could recall. Two independent judges coded the participants' recollections individually, and reached a significant level of agreement on their first judgments (Cohen's K ranging from .93–.94; all $ps < .001$). For the final score, they discussed each item on which they disagreed, until reaching a full agreement. The same coding schema used in Experiment 1 was used to evaluate each concept (i.e., semantic unit) recalled by the participants (some instances are included in Appendix B.2.1). The results of a series of t tests for independent samples showed no difference between performance by the participants assigned to the two conditions for any coding category ($t(18) = 0.0$ to 0.76, $p = 1$ to $.46$).

Experiment 3 procedures for the Recollection task were the same as those for Experiment 1, and the procedures for the Recognition task were the same as those for Experiment 2.

3.3.3. Participants

Sixty students at Turin University (53 women and 7 men; mean age = 22;11) participated. One half (balanced for age and sex) were assigned to the Free-Recall Task (15 to the Gesture condition and 15 to the No-Gesture condition) and one half to the Recognition Task (15 to the Gesture condition and 15 to the No-Gesture condition).

3.3.4. Results

3.3.4.1. Free-recall task: Considering the 35 semantic units, three independent judges coded the participants' recollections individually; they reached a significant level of agreement on their first judgments (Cohen's K ranging from .88–.96; all $ps < .001$). For the final score, they discussed each item on which they disagreed, until reaching a full agreement. Table 4 illustrates the mean scores in the two experimental conditions.

The results of a series of t tests for independent samples revealed that, as predicted, the number of correct recollections was higher in the Gesture condition than in the No-Gesture condition ($t(28) = 3.12$, tied $p = .002$), and the number of discourse-based inferences was higher in the Gesture condition than in the No-Gesture condition ($t(28) = 2.59$, tied $p = .008$).

Table 4
Mean sorts of recollections in the Gestures and No-Gestures conditions in the recollection task in Experiment 3

Condition	Correct Recollections	Discourse-Based Inferences	Elaborative Inferences	Errors
Gestures ($n = 15$)				
M	10.73	.93	.60	.53
SD	3.11	.88	.51	.52
No Gestures ($n = 15$)				
M	7.60	.27	.87	.87
SD	2.35	.46	.83	.74

We found no difference between performance by the participants in the Gesture and No-Gesture conditions for production of elaborative inferences ($t(28) = -1.06$, $p = .3$). The same result held for erroneous recollections ($t(28) = -1.43$, $p = .16$).[2]

3.3.4.2. Recognition task: At the end of the videotaped fiction, participants were presented with 36 sentences (12 literally correct, 12 paraphrases, 12 wrong content) from among which they had to recognize the literally correct ones (examples are included in Appendix B.3). As in Experiment 2, we conducted an analysis of variance (Cochran's Q test) on the stimuli pertaining to each sentence category, and found them to be comparable in difficulty for participants in both the Gesture (Q value ranging from 5.94 to 13.70; p value ranging from .10 to .88) and No-Gesture (Q value ranging from 7.49 to 14.71; p value ranging from .06 to .76) conditions.

Our predictions were confirmed by the results of a series of t tests for independent samples: Participants in the No-Gesture condition performed better than participants in the Gesture condition in recognizing sentences actually spoken by the actor ($t(28) = 2.44$, tied $p = .01$), as well as with paraphrases, thus tending not to endorse them ($t(28) = -2.91$, tied $p = .004$). We also found a marginal difference in performance with errors: Participants in the Gesture condition were better at identifying wrong content sentences than participants in the No-Gesture condition, $t(28) = 2.04$, $p = .05$. Table 5 illustrates the mean correct performance in the two experimental conditions.

We also applied SDT on 12 signals (literally correct sentences) and 24 noise trials (12 paraphrases and 12 errors). Table 6 illustrates the proportions of types of recognition in the Gesture and No-Gesture protocols.

In the Gesture protocol, the results of a paired t test on d′ values for the comparison Literally correct–Paraphrases (the mean d′ was 1.12) and Literally correct–Wrong content (the mean d′ was 0.57) revealed greater sensitivity in the former comparison than in the latter ($t = 4.10$, $p = .001$). The same result held for the No-Gesture protocol (a mean d′ of 1.90 and 0.64, respectively; $t = 8.92$, $p < .001$). These results were quite different from those we obtained by applying SDT to the results of Experiment 1. We believe that for this technical discourse on color, the wrong items we constructed were more misleading than those we constructed for the narrative discourse. Indeed, in an attempt to maintain a good correspondence between the superficial form of the wrong items and the superficial form of

Table 5
Mean correct performance with the different sorts of sentences in the two experimental conditions in the recognition task in Experiment 3

Condition	Literally Correct ($n = 12$)	Paraphrases ($n = 12$)	Wrong Content ($n = 12$)
Gestures ($n = 15$)			
M	7.33	9.80	7.47
SD	2.09	0.94	1.64
No Gestures ($n = 15$)			
M	8.93	10.87	5.93
SD	1.44	1.06	2.40

Table 6
Proportions of sorts of recognition in the Gestures and No-Gestures conditions in the recognition task of Experiment 3

Condition	Hit	Miss	False Alarm Paraphrases	Correct Rejections Paraphrases	False Alarm Errors	Correct Rejection Errors
Gestures ($n = 15$)	106/180	74/180	33/180	147/180	68/180	112/180
No Gestures ($n = 15$)	134/180	46/180	17/180	163/180	90/180	90/180

the literal ones, we only modified the original technical sentence very slightly, thus creating wrong content sentences that were difficult to distinguish from the corresponding literal ones. Consistent with this explanation, although the number of correct rejections for the Paraphrases was the same for the two discourses, the number of correct rejections for Wrong content was higher for the funfair discourse.

4. Conclusion

Our results support the assumption that deep learning from discourse is favored by the possibility of having access to co-speech gestures by the speaker. In particular, the results of Experiment 1 and of the Free-Recall Task in Experiment 3 show that a discourse accompanied by gestures, as compared with a discourse not accompanied by gestures, favors a greater retention of correct information and a greater production of discourse-based inferences by the hearer. Both results indicate the possibility that the individual exposed to co-speech gestures built a more articulated mental model of the discourse she or he was exposed to. The results of Experiment 2 and of the Recognition Task in Experiment 3 showed that listeners who saw a speaker delivering a discourse without gesticulating achieved a better literal retention of the discourse compared to listeners who saw a speaker accompanying his discourse with co-speech gestures. This is consistent with the assumption that a listener who has access to the speaker's gestures, as compared with a listener who does not, constructs a richer mental model of the discourse, and thus tends to lose verbatim.[3]

A caveat to our conclusion is that if the goal in learning is also to remember the verbatim—and in school, as well as in other learning environments, this could be the case—then gestures may not improve learning. However, it should be noted that predictions on the role of gestures in discourse comprehension were validated in the ideal circumstances provided by our experimental settings. Therefore, further studies are necessary to understand the role of gestures in disrupted communicative contexts.

Notes

1. For explorative purposes, we analyzed the sorts of recollections as a function of the type of gesture with which the actor accompanied the relative semantic unit in the story.

As a starting point, we used a gesture type categorization (i.e., McNeill, 1992) that, although not allowing mutually exclusive categories to be defined, does establish some dimensions of gesture semiosis along which various types of gestures can be placed. The different dimensions are as follows: *iconicity, metaphoricity, rhythm,* and *deicticity.* Thirty-three of the 54 semantic units were accompanied by a single type of gesture (9 iconic, 7 metaphoric, 12 batonic, and 5 deictic). A series of analyses of variance revealed that the type of co-speech gesture only affected correct recollections, $F(3, 29) = 13.33$, $p < .0001$; in particular, metaphoric and deictic gestures supported correct recollections more than iconic and batonic gestures (Scheffe: p value varied from $<.0001$–$.031$).

2. For explorative purposes, we analyzed the sorts of recollections as a function of the type of gesture with which the actor accompanied the relative semantic unit in the story. Only 10 of the 35 semantic units were accompanied by a single type of gesture (3 iconic, 2 metaphoric, 5 batonic, and 0 deictic). A series of analyses of variance revealed that the type of co-speech gesture did not affect any sorts of recollections: $F(2, 7) = 2.8$ to 1.3, $p = .13$ to .33.
3. Explorative analyses of the sorts of recollections as a function of the type of gesture performed by the actor showed that, only for participants in Experiment 1, metaphoric and deictic gestures supported correct recollections more than iconic and batonic gestures did. This may suggest that various types of gestures could differ in terms of the likelihood of favoring correct recollections; there is also a possibility that the role of a specific type of gesture may differ, in part, according to the content of the discourse it accompanies. In any case, our study did not have the necessary granularity to investigate this issue.

Acknowledgments

This work was supported by Italian Ministry of University and Research of Italy, Italian Project of National Relevance Project #2007TNA9AA. We thank Bruno Bara, Marco Tamietto, Danielle McNamara, Joe Magliano, Art Graesser, and two anonymous reviewers for their helpful comments.

References

Alibali, M. W., Kita, S., Bigelow, L. J., Wolfman, C. M., & Klein, S. M. (2001). Gesture plays a role in thinking for speaking. In *Oralité et gestualité* (pp. 407–410). L'Harmattan, France.

Bucciarelli, M. (2007). How the construction of mental models improves learning. *Mind & Society, 6*, 67–89.

Carney, R. N., & Levin, J. R. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review, 14*, 5–26.

Feyereisen, P., & Harvard, I. (1999). Mental imagery and production of hand gestures while speaking in younger and older adults. *Journal of Nonverbal Behavior, 23,* 153–171.

Garnham, A., Oakhill J., & Cain K., (1998). Selective retention of information about the superficial form of text: Ellipses with antecedents in main and subordinate clauses. *The Quarterly Journal of Experimental Psychology: Section A, 51*, 19–39.

Glenberg, A. M., & Langston, W. E. (1992). Comprehension of illustrated text: Pictures help to build mental models. *Journal of Memory and Language, 31*, 129–151.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Science, 3*, 419–429.

Graesser, A. C., Millis K. K., & Zwaan R. A. (1997). Discourse comprehension. *Annual Review of Psychology, 48*, 163–189.

Graesser, A. C., Singer M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Hildebrandt, B., Moratz, R., Rickheit, G., & Sagerer, G. (1999). Cognitive modelling of vision and speech understanding. In G. Rickheit & C. Habel (Eds.), *Mental models in discourse processing and reasoning* (pp. 213–236). New York: Elsevier.

Iverson, J. M., & Goldin-Meadow, S. (2001). The resilience of gesture in talk: Gesture in blind speakers and listeners. *Developmental Science, 4*, 416–422.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, and consciousness*. Cambridge, England: Cambridge University Press.

Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). *Deduction*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Johnson-Laird, P. N., & Stevenson, R. (1970). Memory for syntax. *Nature, 227*, 412.

Kaup, B., Kelter, S., & Habel, C. (1999). Taking the functional aspect of mental models as a starting point for studying discourse comprehension. In G. Rickheit & C. Habel (Eds.), *Mental models in discourse processing and reasoning* (pp. 93–112). New York: Elsevier.

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7*, 54–60.

McNamara, T., Miller, D. L., & Bransford, J. D. (1991). Mental models and reading comprehension. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 490–511). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

McNeil, N. M., Alibali M. V., & Evans J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need them, now they don't. *Journal of Nonverbal Behavior, 24*, 131–150.

Moreno, R., & Mayer, R. E. (2002). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology, 94*, 156–163.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*, 226–231.

Rickheit, G., & Sichelschmidt, L. (1999). Mental models: Some answers, some questions, some suggestions. In G. Rickheit & C. Habel (Eds.), *Mental models in discourse processing and reasoning* (pp. 9–40). New York: Elsevier.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479–515). San Diego, CA: Academic.

van Dijk, I. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic.

Vekiri, I. (2002). What is the value of graphical displays in learning? *Educational Psychology Review, 14*, 261–312.

Zwaan, R. A., Magliano J. P., & Graesser, A C. (1995). Dimensions of situation model construction in narrative comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 386–397.

Zwaan, R. A., & Radvansky G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*, 162–185.

Appendix A

A.1. Excerpt from the discourse in Experiment 1 (semantic units are separated by slashes)

It was there, at the funfair, it was there that I found her, and it was at the funfair that I lost her. It was a vast funfair./A funfair with shooting-ranges and candy floss stalls and Japanese bagatelle tables, stalls with bottles of champagne and showmen's booths and roundabouts./And the roundabouts turned and creaked and the candy floss scented the air and the rifles shot./I was shooting at the target./I can shoot at the target very well and I am proud of it./No, wait a moment, I am wrong! I did not meet her at the shooting-range./I met her at the candy floss stall. Yes, it was at the candy floss stall that I found her./The candy floss scented the air,/and she was eating it and she blew on her candy and I was all covered with white powder./She started laughing/and I asked her: "What's your name?"/And she shouted to me: "I'll tell you later."/Later, we went to the shooting-range/and it was there that I lost her./I aimed at the target, breaking all the clay pipes/and every time she shouted to me: "Well done!"/And then, when there were no pipes left to break, I aimed at the egg, the one held up by the jet of water,/and while I was aiming I shouted to her: "What's your name?"/And she replied: "I'll tell you later."/I shot, the egg popped up./I turned aside and she wasn't there any more.

A.2. Examples of gestures produced by the actor

Iconic: While saying "turned" in delivering the sentence, " . . . the roundabouts *turned* . . . ," the actor raises his right hand until it is in front of him, parallel to the floor, palm down. He then turns it anti-clockwise, in front of him, using his elbow as a pivot point, making a full circle.

Metaphoric: While saying *the first* "later" in delivering the sentence, ". . . she shouted to me, -*Later,* I'll tell you later-," the actor holds his right hand in front of him, half closed, with his fingers half bent and turns it anti-clockwise, using his wrist as a pivot point, making $1\frac{1}{2}$ circles.

Batonic: While saying "found" in delivering the sentence, "I *found* her and I *lost* her," the actor first holds his right hand stretched out, palm up, in a line with his forearm. He turns his hand to the left until it is vertical, and quickly raises and lowers it, always keeping it in line with his forearm. He repeats the rapid movement on the word "lost."

Deictic: While saying "there" in delivering the sentence, " . . . and it was *there* that I lost her," the actor moves his left hand until it is in front of him, turned slightly to the left, holding his index finger straight and his other fingers bent so that they are resting against his thumb.

A.3. Examples of sentences used for the recognition task of Experiment 2

Literal: And I was looking for her answer in her lip movements.
Paraphrases: And I looked at her lips to understand her answer.
Wrong: And I understood her answer by looking at her eyes.

Literal: I shot, the egg popped up. I turned aside, she wasn't there any more.
Paraphrases: When I shot at the egg she disappeared.
Wrong: While I was going to shoot to the egg, I turned aside and she wasn't there any more.

Appendix B

B.1. Excerpt from the discourse in Experiment 3 (semantic units are separated by slashes)

It's beyond dispute that colors carry strong expressive components./Some attempts have been made to describe the specific expressive characters of the various colors and to draw some general conclusions from the symbolic use the different cultures have made of them./There is a very widespread belief that the expression of colors is based on association./Therefore, red should be considered exciting because it reminds us of the connotations of fire, blood and revolution./Green evokes the restorative thought of nature,/and blue is refreshing like water./However, the theory of association is not more interesting or prolific in this field than in others. The effects of colors are too direct and spontaneous to be simply the results of a interpretation given through knowledge./On the other hand, no hypothesis has been advanced so far on the kind of physiologic process which could help to explain the influence of colors on the organism./It is well known that extreme brightness, high saturation and shade of color corresponding to vibrations in wavelength breadth cause excitement./

B.2. Examples of gestures produced by the actor

Iconic: While saying "widespread" in delivering the sentence, "There is a very *widespread* belief," the actor quickly raises his right arm until it is parallel to the floor and turns it until his hand, palm down, is in line with his left shoulder. He then turns the palm of his hand slightly toward himself and slowly makes a semicircular movement with his arm.

Metaphoric: While saying "thought" in delivering the sentence, "Green evokes the *thought* . . . ," the actor raises his left hand, which is open, bringing it close to the side of his head, with the palm toward his face; he moves the fingers on his hand forward and backward, each one separately.

Batonic: While saying "observed" in delivering the sentence, "As a practicing neurologist he *observed* that . . . ," the actor raises his left hand from his leg up to his breastbone. His index finger is raised and pointing upward, his fingers are closed, and his hand is facing toward the right.

B.2.1. Examples of codification for Experiment 2

The discourse contains the sentence, "It's beyond dispute that colors carry strong expressive components"; the statement, "It is well known that color is a powerful expressive means," is a correct recollection; the statement, "Several researchers have demonstrated that color is an important expressive means," is a discourse-based inference (because it refers to a causal antecedent); and the statement, "The expressive characteristics and functions of color are the

subject of much debate," is an erroneous recollection. Now, consider the following statement in the original discourse: "Red should be considered exciting because it reminds us of the connotations of fire, blood, and revolution"; the statement, "Red can evoke passion, blood, and revolution," is an elaborative inference.

B.3. Examples of sentences used for the recognition task of Experiment 3

Literal: Green evokes the restorative thought of nature.
Paraphrases: The color evoking the restorative thought of nature is green.
Wrong: Green evokes the sense of tiredness of nature.

Literal: Goldstein's experimental attempts in this field are worthy of note.
Paraphrases: The experiments carried out by Goldstein are still interesting.
Wrong: Goldstein's experimental attempts in this field are not worth reporting here.