Title
Focusing Attention with Deictic Gestures and Linguistic Expressions

Permalink
https://escholarship.org/uc/item/201422tj

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 27(27)

ISSN
1069-7977

Authors
Bangerter, Adrian
Louwerse, Max M.

Publication Date
2005

Peer reviewed

Focusing Attention with Deictic Gestures and Linguistic Expressions

Max M. Louwerse (mlouwers@memphis.edu)
Department of Psychology / Institute for Intelligent Systems
Memphis, TN 38152 USA

Adrian Bangerter (adrian.bangerter@unine.ch)
Groupe de Psychologie Appliquée
CH - 2000 Neuchâtel, Switzerland

Abstract

Comprehension and production of text and discourse do not solely depend on linguistic expressions, but also on the physical context. The questions addressed in this study are 1) whether deictic gestures are substitutable for deictic expressions, and 2) whether deictic gestures establish joint attention. An eye tracking experiment investigated the effect of referring expressions and gestures on various aspects of attention. Results indicated that deictic gestures substitute for location descriptions. Furthermore, the manipulation of the synchrony between gesture and speech showed that hearers benefit from focusing of visual attention.

Introduction

Understanding and production of naturally occurring language does not solely rely on linguistic modalities like content, prosody and text or dialog structure. It also very much relies on non-linguistic modalities like eye gaze, facial expressions, body posture and gestures. It seems that gestures fulfill an important supportive role in bringing about the communicative project: everybody uses them, whether they are pointing out directions on the map, emphasizing a point they are trying to make, whether they are in a face-to-face argument or chatting on a cell phone.

So why do we gesture? There are several explanations that are not necessarily mutually exclusive. According to one account, we gesture to facilitate lexical access (Butterworth & Beattie, 1978; Rime & Schiaratura, 1991). The timing gap between a gesture and an unfamiliar word is larger than between a gesture and a familiar word (Morrel-Samuels & Krauss, 1992). Furthermore, gesture is associated with fluent speech: when the speech is disrupted, like in stuttering, the gesture is halted (Mayberry & Jaques, 2000). According to a second account, gestures facilitate thinking (Goldin-Meadow, 2003; McNeill, 1992). Gesture and speech are coexpressive manifestations of one integrated system. They form complementary components of one underlying process and thereby help organizing thought. Indeed, children's performance on counting tasks improves when they gesture (Alibali & DiRusso, 1999). According to both of these accounts, gestures help speakers but not necessarily hearers. By a third account, gestures support communicative joint activities, that is, they are informative for hearers (Clark, 1996). The speaker and hearer are participating in the joint project of communication. Gesture is thereby part of the language use. Evidence for this account comes for instance from Özyürek (2002) who showed that speakers change orientation of their gestures dependent on their hearers. However, these three different accounts investigate the production of gestures. An important question that remains is how addressees perceive these gestures. According to the first two accounts the effect gestures have on the addressee is irrelevant, according to the third account there is an immediate impact.

There is strong evidence that gesture is intrinsically related to (symbolic) language processing (Butterworth and Morrisette, 1996). For instance, 90% of all gestures occur when we speak (McNeill, 1992). Furthermore, there are close ties between gesture and language development (Butcher & Goldin-Meadow, 2000). Also, humans are the only species that gesture (Butterworth, 2003; Povinelli, Bering, Giambrone, 2003). At the same time gestures are very different than linguistic cues. For instance, despite the fact that interlocutors rely on cues from gestures, particularly when the speech is ambiguous (Thompson & Massaro, 1996) or when the environment is noisy (Rogers, 1978), they are often unable to remember what hand gestures they have seen (Krauss, Morrell-Samuels, et al. 1991). Gestures thus seem to fulfill an important but subtle function: They have close ties to the linguistic system and seem to be intrinsically integrated with it, at the same time there are differences in production and understanding. An important research question is therefore what the relation is between gestures and linguistic expressions.

It is important to keep in mind that we have a large range of gestures available. Kendon (1988) nicely places hand gestures along a continuum from 1) gesticulation, 2) language-like gestures and 3) emblems to 4) sign languages. Moving from left to right along the continuum gestures are replacing the role of speech; hardly so in gesticulation, very much so in sign language. This paper uses gesture to solely refer to gesticulation. Within gestures different types can be identified (Ekman & Friesen, 1969; Goldin-Meadow, 2003; McNeill, 1992). Generally, four categories are distinguished: 1) iconic gestures that mimic the object being represented through the gesture (making sawing movements when talking about sawing a tree), 2) concrete deictic gestures (pointing at a painting when talking about the Rembrandt's *Nightwatch*), 3) abstract deictic gestures (gesturing from left to right saying "from the beginning to the end"); 4) beat movements (used in the rhythm of the speech or to mark important intonational boundaries). In this paper we will focus on concrete deictic gestures

Deictic Gestures

We have earlier argued that gestures and language are intrinsically linked. In particular, concrete deictic gestures nicely map onto deictic expressions like “this” and “that”, “these” and “those”, “here” and “there”. Thus, they seem to substitute particularly well for certain linguistic expressions, especially spatial expressions. At the same time, deictic gestures form indices to individual things. Clark (2003) distinguishes two kinds of indicating: directing-to and positioning-for. The first kind is what is generally considered as pointing and serves to move the hearer’s attention from the speaker to the approximate region of the referent (Marlsen-Wilson, Levy, & Tyler, 1982). If speaker and hearer both know that their attention is focused on a similar region, then this facilitates reference resolution (Hanna & Tanenhaus, 2004).

The purpose of the present study is two-fold. First, it aims to answer the question whether pointing helps the hearer in the communicative process. That is, gestures may be used to organize thoughts or support lexical access for the speaker, but may not facilitate the joint communicative activity. In contrast, the hypothesis investigated here is that deictic gestures help hearers identify the target indirectly, by guiding their gaze to its region. By this hypothesis, pointing helps establish a joint focus of attention between speaker and hearer (joint-attention hypothesis). This in turn facilitates processing on the part of the hearer. Second, the study aims at determining whether deictic gestures are substitutable for certain linguistic spatial expressions. Contrary to the prediction that gestures add information to the communicative act, our hypothesis is that gestures can substitute for language functions (substitution hypothesis).

In other words, we suggest that the effect of pointing on the addressee is similar to that of a verbal description of an approximate region of space, e.g., “the upper right corner” (Bangerter, 2004). At least three different kinds of strategies for referring to objects in shared visual space can be identified, one gestural and two linguistic:

1. Pointing (e.g. hand pointing to target while saying “John is right there”)
2. Feature description (e.g. saying “John is the man with the hat”)
3. Location description (e.g. saying “John is the one on the top right”).

Unambiguous feature descriptions (i.e. that specify a unique referent among possible competitors) have been discussed as the way people typically identify referents (Olson, 1970). But with a large set of potential referents, such a strategy may not be feasible, nor pragmatically appropriate. In real conversational situations, people typically try to create joint attention by circumscribing the domain of reference (Beun & Cremers, 1998). By the substitution hypothesis, this can be done either by pointing or by location descriptions. By the joint-attention hypothesis, focusing attention in this way should facilitate reference resolutions. The effects of pointing and linguistic feature and location descriptions were investigated using eye-tracking methodology.

Experiment

The current study investigates the effects of referring expressions and pointing gestures on the addressee’s attention. Participants viewed a video clip of a person describing and/or pointing to an array of objects on a computer monitor while their eye movements were recorded.

Method

Participants

The participants were 30 undergraduate students at a southern urban university. The participants received extra credit in an undergraduate course for participating in this experiment.

Materials

Participants saw 30 short movies (5 seconds each). Each movie consisted of 12 smiley faces differing in props (e.g., hat, moustache, glasses) and emotion (happy, sad) and dependent on the condition a human pointer, pointing out and/or describing the target. The position of the faces (three columns, four rows), the position of the pointer’s arm and hand and the movement of the pointing, the feature description of the smiley faces using two distinctive features at a time (emotion and additional feature), and the location description (left and right versus top and bottom dimension) all remained constant. In a Latin Square-like design, each participant cycled through each condition 5 times in a random order, totaling 30 trials. In addition, they completed 10 filler trials (an unrelated text comprehension task).

Six conditions were used. The factorial combination of the presence versus absence of a location description and the presence versus absence of gestural pointing in combination to the feature description resulted in four conditions (Table 1).

Table 1: Overview of pointing x description conditions

	location description	no location description
pointing	Pointing + *John is on the top left with a hat and bow tie*	Pointing + *John has a hat and bow tie*
no pointing	*John is on the top left with a hat and bow tie*	*John has a hat and bow tie*

Two additional conditions were created by manipulating the time and order of linguistic expressions and pointing: in a fifth condition, pointing preceded the linguistic expressions (feature description only), but with an inserted pause of two seconds. In the final condition the feature description followed the pointing after a two-second pause.

Apparatus

Participants’ eye movements were tracked using a Model 501 Applied Science Laboratory eye tracker. A magnetic head tracker with a head mounted apparatus was used so that participants could move their head during data collection. Computer software recorded the eye movements. Participants were calibrated throughout the session to insure reliable data. During calibration, participants viewed nine points on a 1024 x 768 computer monitor and the eye tracker recorded corresponding x-y coordinates. The temporal resolution of the Model 501 eye tracker was 60 Hz. The spatial resolution was a .50 degree angle horizontally and a .40 degree angle vertically.

Procedure

Participants were asked to watch each clip with the 12 smiley faces and click the mouse button as soon as they had identified the target face. During the process of the participant identifying the target, their eye movements were recorded. After they clicked the mouse button, they were presented with 12 circles that represented the positions of the 12 faces and were asked to identify the target. Accuracy of the target identification was recorded.

Results

The results of the experiments consisted of two datasets, one containing the participants' answers and one the eye tracking data. The accuracy of identifying the target showed an effect for pointing ($F_1(1, 29) = 34.66$, $p < .01$, $MSE = .02$; F_2 $(1, 29) = 14.80$, $p < .01$, $MSE = .04$), location description ($F_1(1, 29) = 54.10$, $p < .01$, $MSE = .02$; F_2 $(1, 29) = 29.23$, $p < .01$, $MSE = .03$) as well as their interaction ($F_1(1, 29) = 44.62$, $p < .01$, $MSE = .01$; F_2 $(1, 29) = 18.27$, $p < .01$, $MSE = .03$). The number of correct answers was higher when pointing or the location description was present (Figure 1). The selection of the correct target was facilitated when the instructions helped the hearer in identifying the target region, either by means of a location description or gesture. This shows that both the use of pointing and the use of location description increase accuracy. Interestingly, when *both* pointing and location description were presented no additional facilitation was found. Though this may be attributed to a ceiling effect, it may well suggest that pointing can substitute for the location description, providing support for the substitution hypothesis. It also supports the joint-attention hypothesis.

The joint-attention hypothesis was also tested by comparing the natural pointing condition with the asynchronous pointing conditions where pointing either preceded or succeeded the feature description with an inserted pause of two seconds. No evidence was found for this hypothesis: The accuracy of answers was the same when pointing preceded the speech and when it followed speech ($p > .6$).

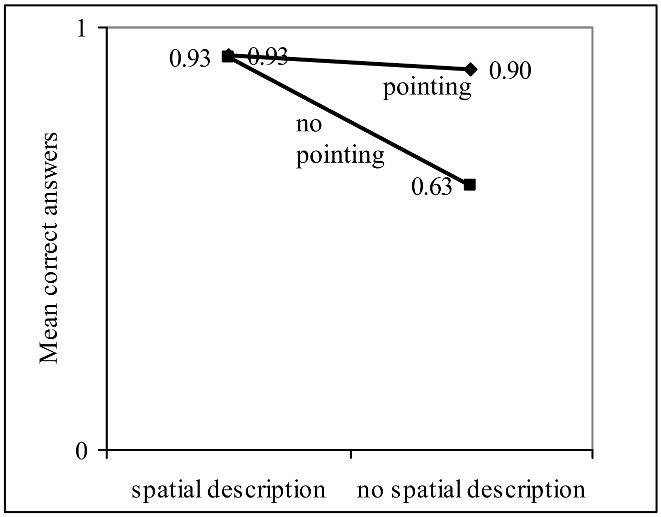

Figure 1: Accuracy of answers for pointing x location description.

However, it should be noted that the total time spent on task could be a confounding variable in the accuracy results. For instance, the presence of spatial information (pointing or location descriptions) necessarily allowed participants to consider the target longer than in the absence of this information. The fact that participants had slightly more time on making their choice in the location description condition could explain the accuracy findings. Eye tracking data can rule out this explanation by considering the total time on task. The total fixation time on all items in both correctly and incorrectly answered items did not show significant differences between the pointing, location description, or pointing and location description conditions ($p > .3$), upholding the evidence for the substitution hypothesis. Whereas we did not find differences for the accuracy of answers between the asynchrony conditions, the time on task did yield a significant difference. When pointing preceded the location description, the total fixation time was less ($M = 1.22$, $SD = .21$) than in the reverse situation ($M = 1.51$, $SD = .64$), but not significantly different from the natural pointing condition ($M = 1.14$, $SD = .31$), ($F_1(2, 58) = 13.27$, $p < .01$, $MSE = .08$; F_2 $(2, 58) = 13.5$, $p < .01$, $MSE = .09$). In other words, when pointing preceded speech, regardless of the delay, the accuracy of answers was not affected, but participants were able to get to the right answer faster. This provides evidence for the joint-attention hypothesis in that deictic gestures guide the hearers gaze to the target. When this guidance follows the target identification, it results in confusion.

In the remainder of the eye tracking analyses, we removed those items that were answered incorrectly. Accordingly, 14% of the data was removed, but these cases were distributed over all stimuli and all participants. When the total fixation time was considered on all faces for which the target was identified correctly, no differences were found for the three pointing and location description conditions). The advantage we found for the accuracy of items with the presence of pointing or the presence of a location description did not reflect the total amount of time spent on the task between conditions. On the other hand, for the asynchrony conditions, we found the same patterns as before. Not only does it take less time to answer an item, but it also takes less time to answer an item correctly (*M*= 1.20, *SD* = .15 vs. *M* = 1.51, *SD* = .65; (F_1(1, 29) = 9.87, *p* < .01, *MSE* = .45; F_2 (1, 29) = 17.36, *p* < .01, *MSE* = .08). Not surprisingly, the number of regressive eye movements is also significantly less when pointing precedes speech (*M* = 3.47, *SD* = .55 vs. *M*= 4.34, *SD* = .91; F_1(1, 29) = 21.55, *p* < .01, *MSE* = .62; F_2 (1, 29) = 25.57, *p* < .01, *MSE* = .45). Interestingly, the number of regressive eye movements is lower on the targets as well as the non-targets in the condition where pointing precedes speech. In all cases the insertion of the delay had no effect on fixation times and regressions, as long as pointing preceded the feature description.

As Figure 2 shows, pointing resulted in more regressive eye movements on the correct targets than did the no pointing condition (F_1(1, 29) = 4.59, *p* = .04, *MSE* = .26; F_2 (1, 29) = 3.26, *p* = .07, *MSE* = .49), and so did the presence of the location description (F_1(1, 29) = 3.92, *p* = .06, *MSE* = 3.78; F_2 (1, 29) = 14.73, *p* < .01, *MSE* = 2.02). as before, no interaction was found between pointing and the location description.

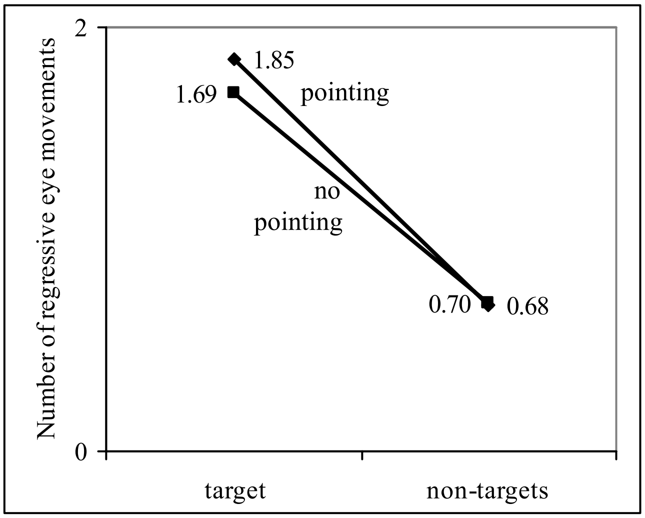

Figure 2: Effects of pointing on regressions

The problem with the results of regressive eye movements in this task is that they can explain cognitive processes in two directions. One could argue that more regressions on targets are an indication that the task is easier: the hearer verifies the information by considering alternatives, each time moving back to the target strengthening the choice. An alternative explanation is that regressions on the target are an indication of confusion: the hearer is not certain of the choice and needs to move back and forth. Because of this ambiguity, we performed a more subtle analysis. We counted the number of non-targets that were considered more than once. In either a verification or falsification process, each non-target may have to be considered, but should be eliminated after being considered once. Therefore, if a non-target item was considered more than once it indicated confusion in the listener.

The presence of pointing indeed reduced the number of items considered (F_1(1, 29) = 18.25, *p* < .01, *MSE* = 14.64; F_2 (1, 29) = 17.38, *p* < .01, *MSE* = 21.44). The presence of a location description had a similar effect, though marginally significant (F_1(1, 29) = 3.68, *p* < .065, *MSE* = .15.99; F_2 (1, 29) = 4.23, *p* < .049, *MSE* = 13.87). Again, no interaction was found for pointing and location description, providing evidence for the substitution hypothesis (Table 2).

Table 2: Number of non-targets considered

	presence	absence
location description	5.49	5.76
pointing	5.33	5.92

The asynchrony condition provided evidence for the joint-attention hypothesis, showing that pointing preceding speech resulted in a significantly smaller number of items being considered (F_1(1, 29) = 50.15, *p* < .01, *MSE* = 14.38 ; F_2 (1, 29) = 55.47, *p* < .01, *MSE* = 12.99) (Figure 3).

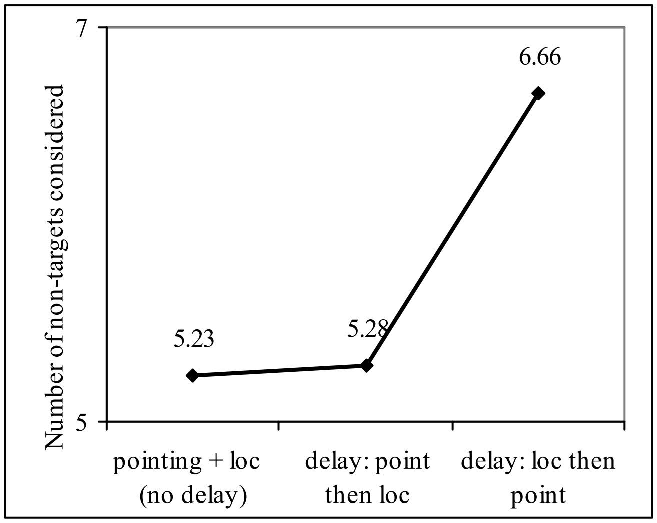

Figure 3: Synchrony and asynchrony in location description and pointing

Discussion

The findings of this study support the view that deictic gestures can substitute for language functions. That is, when a feature description is accompanied by either a deictic gesture or a deictic expression, accuracy in target identification increases. However, when both the deictic gesture and the deictic expression are present, no additional gains are found in accuracy. This pattern was also found in the number of regressive eye movements. Participants spent more time on the correct target when pointing was present or when the location description was present, but not when they were combined. This provides support for what we have called the substitution hypothesis for deictic gestures.

The results also provide evidence that gestures support communicative joint activities, as stated by the joint-attention hypothesis. Eye tracking data show that pointing helps establish a joint focus of attention between speaker and hearer, as predicted by the joint-attention hypothesis. If the joint focus of attention is identified after the target identification, it results in confusion, as indicated by more regressive eye movements, higher fixation times to identify the target and more non-targets being considered before the correct target is identified. Whether the visual guidance precedes with a two second delay or whether it naturally co-occurs with the linguistic expression does not affect fixations.

This paper has focused on concrete deictic gestures (pointing). Other gestures, including iconic, abstract deictic and beat gestures may not support the substitution and joint-attention hypotheses. Although there is some evidence that they do (see Goldin-Meadow, 2003; Özyürek, 2002), further research is needed. Concrete deictic gestures for instance have the special relationship with linguistic expressions that they can be substituted. That relationship is less direct in the other gesture types.

In addition to having limited ourselves to concrete deictic gestures, we have only considered a limited set of linguistic expressions in English. The role of typical deictic expressions like “here”, “there”, “this” and “that” and deictic gestures in the comprehension process have not been tested here. Moreover, for practical reasons we have only considered linguistic expressions in English. To generalize the relationship between gestures and linguistic expressions, a cross-linguistic analysis taking into account different morphological and syntactic constructions may have to be considered in the future.

The findings presented here have implications for a number of research areas. For instance, it suggests that in building intelligent systems, gestures should not be ignored, since they support the joint visual attention with the user. Moreover, if the alignment of gesture to speech is not in synchrony, this could have an important impact on the user, for instance in intelligent tutoring systems (Louwerse, et al., 2004).

These findings also have implications for the answer why we gesture. The alignment results support the notion that gesture and speech are indeed coexpressive manifestations of one integrated system. Disintegrating the two, for instance by changing their order, results in confusion. But regardless of whether we gesture to facilitate lexical access or to organize thoughts, our findings at least show that we gesture to support communicative joint activities. That function can also be fulfilled by specific linguistic expressions, as long as the description is as specific as the gesture.

Acknowledgments

This research was supported by grant NSF-IIS-0416128 to the first author and by an Advanced Researchers Fellowship from the Swiss National Science Foundation (8210-061238) to the second. 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 institutions. Correspondence concerning this article should be addressed to Max M. Louwerse, Institute for Intelligent Systems / Department of Psychology, University of Memphis, 202 Psychology Building, Memphis, Tennessee 38152-3230.

References

Alibali, M. W., & DiRusso, A. A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development, 14*, 37-56.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*, 415-419.

Beun, R.J. & Cremers, A.H.M. (1998) Object Reference in a Shared Domain of Conversation. *Pragmatics and Cognition, 6,* 111-142.

Butcher, C., & Goldin-Meadow, S. (2000). Gesture and the transition from one to two-word speech: When hand and mouth come together. In D. McNeill (Ed.), *Language and gesture* (pp. 235-257). New York: Cambridge University Press.

Butterworth, G. (2003). Pointing is the royal road to language for babies. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 9-33). Mahwah, NJ: Erlbaum.

Butterworth B. L. & Beattie G. W. (1978). Gestures and silence as indicator of planning in speech. In Campbell R. N., Smith P. T. (Eds.), *Recent Advances in the Psychology of Language: Formal and Experimental Approaches* (pp. 347-360). New York: Olenum Press.

Butterworth, G., & Morissette, P. (1996). Onset of pointing and the acquisition of language in infancy. *Journal of Reproductive and Infant Psychology, 14,* 219-231.

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

Clark, H.H. (2003). Pointing and placing. In S. Kita (Ed.), Pointing. *Where language, culture, and cognition meet* (pp. 243–268). Hillsdale NJ: Erlbaum.

Ekman, P. & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica, 1*, 49- 98.

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

Hanna, J. E., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science, 28*, 105-115.

Kendon, A. 1988. How gestures can become like words. In F. Poyatos (ed.), *Crosscultural perspectives in nonverbal communication* (pp. 131-41). Toronto: Hogrefe.

Krauss, R. M., Morrel-Samuels, P., Colasante, C. (1991). Do conversational hand gestures communicate? *Journal of Personality and Social Psychology, 61*, 743-754.

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.

Marslen-Wilson, W., Levy, E., & Tyler, L. K. (1982). Producing interpretable discourse: The establishment and maintenance of reference. In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action. Studies in deixis and related topics*. (pp. 339-378). Chichester: John Wiley.

Mayberry, R.I. & Jaques, J. (2000). Gesture production during stuttered speech: Insights into the nature of gesture-speech integration. In D. McNeill (Ed.), *Language and Gesture: Window into Thought and Action* (pp. 199-213). Cambridge: Cambridge University Press.

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

Morrel-Samuels, P. & Krauss, R.M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18, 615-623.

Olson, D. R. (1970). Language and thought: Aspects of a cognitive theory of semantics. *Psychological Review, 77*, 257-273.

Özyürek, A. (2002). Do speakers design their co-speech gestures for their addressees? The effects of addressee location on representational gestures. *Journal of Memory and Language, 46*, 688-704.

Povinelli, D. J., Bering, J. M., & Giambrone, S. (2003). Chimpanzee' "pointing": Another error of the argument by analogy? In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 35-68). Mahwah, NJ: Erlbaum.

Rime, B. and Schiaratura, L. (1991) Gesture and speech, In: R.S. Feldman and B. Rime (Eds.) *Fundamentals of Nonverbal Behavior* (pp. 239-284). Cambridge: Cambridge University Press.

Rogers, W. (1978). The contribution of kinesic illustrators towards the comprehension of verbal behavior within utterances. *Human Communication Research, 5*, 54-62.

Thompson, L. & D. Massaro (1986). Evaluation and integration of speech and pointing gestures during referential understanding. *Journal of Experimental Child Psychology 42*, 144- 168.