Title

When Input and Output Diverge: Mismatches in Gesture, Speech, and Image

Permalink

https://escholarship.org/uc/item/9g00h8h5

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 26(26)

ISSN

1069-7977

Authors

Melinger, Alissa
Kita, Sotaro

Publication Date

2004

Peer reviewed

When Input and Output Diverge: Mismatches in Gesture, Speech, and Image

Alissa Melinger (melinger@coli.uni-sb.de)
Department of Computational Psycholinguistics,
Saarland University, Saarbrücken, 66041, Germany

Sotaro Kita (sotaro.kita@bristol.ac.uk)
Department of Experimental Psychology, 8 Woodland Road
University of Bristol, Bristol BS8 1TN, United Kingdom

Abstract

The goal of the current paper is to investigate the behavior of gesture when the information conveyed by speech and the information conveyed by the image being described conflict as a result of perspective taking. To construct a corpus of speech-image mismatches, we designed a picture description elicitation procedure using path-like networks of colored circles. The results of our analysis demonstrate that gestures can be mismatched to both speech, as has been previously observed, and to the image, which has not been previously reported. The results provide insights into the nature of the representations that give rise to gestures.

The Origin of Gesture

This paper investigates the underlying cognitive processes involved in relating spatial information from a visual input to two separate output modalities, namely speech and gesture. Three theoretical possibilities have been proposed for how these three modalities of representation, visual-spatial, verbal and gestural, are related: The Lexical Semantic Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984), the Free Imagery Hypothesis (Krauss, Chen, & Chawla, 1996; Krauss, Chen, & Gottesman, 2000; but see de Ruiter, 1998, 2000 for another version of this hypothesis), and the Interface Hypothesis (Kita & Özyürek, 2003).

The Lexical Semantic Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984) proposes that gestures are generated from the semantics of the lexical items chosen to express the desired message. It predicts that gestures should always correspond to the meaning expressed by specific lexical items. In contrast, the Free Imagery Hypothesis (Krauss et al., 1996, 2000) claims that gestures are generated on the basis of pre-linguistic non-propositional representations; the strong reading of this proposal implies that the information conveyed by gesture should be unaffected by the specific lexical items selected during formulation and by the 'thinking for speaking' (Slobin, 1987, 1996) processes that convert the imagistic representation into propositional content (however, see below for alternative readings of this proposal). The Interface Hypothesis (Kita & Özyürek, 2003) claims that gestures originate from a mediating representation connecting spatio-motoric representations in memory and linguistic representations. According to this view, gestures are generated from the imagistic representation, but they can also be influenced by 'thinking for speaking' operations on this representation.

Discriminating between these theoretical alternatives is difficult because there is usually a close isomorphism between the semantic content of speech and the imagistic content of the representation speech is describing.

Bearing on this discussion are recent studies demonstrating that gestures can convey complementary information to what is expressed in speech. For example, when describing their solutions to the Tower of Hanoi problem, speakers' gestures sometimes corresponded to possible strategies that were not mentioned in the concurrent speech rather than to the strategy that was mentioned in speech. (Garber & Goldin-Meadow, 2002). The non-isomorphism between the content of speech and gesture has been referred to as *speech-gesture mismatches*. High rates of speech-gesture mismatches have also been reported for children who are in the transitional stage of acquiring the ability to correctly respond to the Piagetian conservation task (Church & Goldin-Meadow, 1986).

Speech-gesture mismatches appear to contradict the claims of the Lexical Semantic Hypothesis in that they express information not included in speech. However, the information expressed by gesture in the speech-gesture mismatches does not actually conflict with either the linguistic or the imagistic representation; instead they provide complementary information. Thus, they do not provide a strong test of the competing theories. In this paper we employed perspective taking to create situations in which what was said *conflicted* with what was seen. Examining the behavior of gesture in these cases should discriminate between the competing hypotheses regarding gesture generation. The Lexical Semantics Hypothesis predicts that gestures will always align with the speech. The Free Imagery Hypothesis predicts that the gestures will always align with the image. The Interface Hypothesis predicts that gesture alignment will be influenced by 'thinking for speaking' processes and therefore the alignment of gesture could be to either or both representations, depending on the specific situation.

Perspective Taking

Perspective taking is a critical step required to express spatial relations in speech (cf. Miller & Johnson-Laird, 1976). Spatial representations, which are inherently relative, must be grounded to some referent in a scene. The choice of the grounding referent impacts the linguistic terms that can be selected to express the relationship. Thus, perspective taking necessarily precedes linguistic formulation. It forms part of the 'thinking for speaking' conceptualizing process (cf. Levelt, Roelofs & Meyer, 1999) that abstracts away from the visual imagery and maps the relations onto propositional representations. Choice of perspective can be influenced by, among other things, language/culture specific resources (Levinson, 2003), the specific task at hand (Tversky, 1991) and/or pragmatic concerns (Levelt, 1996).

Consider the image in Figure 1. In describing the relationship between the ball and the car, the speaker can select himself as the grounding referent, describing the relationship from his own personal orientation, as in (1). This perspective will be referred to as the *deictic* perspective. Alternatively, he can select one of the objects in the scene as the grounding referent, such as the car, and describe the relation with respect to the car's inherent orientation, producing the *intrinsic* description in (2).

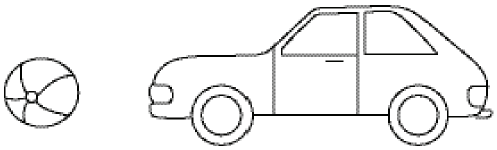

Figure 1

(1) The ball is to the left of the car.
(2) The ball is in front of the car.

When speakers choose to describe the relationship between the ball and car as in (2), a special situation arises; namely, the characteristics of the visual input, pre-abstraction, do not match the linguistic terms used to describe them. On the two dimensional representation of the image, nothing is *in front* of the car; the notion of *front* used in (2) is only relevant with respect to the orientation of the car — only within the perspectivized mental representation of the image. This contrast between the pre-abstraction visual input and the perspectivized mental representation provides the gesture researcher with the opportunity to contrast the content of the image with the content of speech in a unique way. Specifically, the content of the input imagistic representation and the output linguistic representation can be pitted against each other. How gesture behaves when the input and output representations conflict will reveal the underlying representation from which gesture was generated, thus discriminating between the three hypotheses.

Mismatch Corpus

To compile a corpus of speech-image-gesture mismatches, we presented speakers with networks of colored circles arrayed along a path. The images were very similar to networks used previously by Levelt (1996) to investigate perspective taking in speech production. As with other spatial relations, adopting different linguistic perspectives to describe an image, such as in Figure 2, results in the use of different linguistic terms to express the same spatial relations, as seen in examples (3a) and (3b).

Deictic Sample descriptions:
(3a) You begin with a yellow circle. *Above* that you see a blue circle. To the *right* you see a red circle and *above* the red circle you see another red circle. *Right* of the second red circle is the yellow circle and *right* of that is a blue circle.

Intrinsic Sample descriptions:
(3b) You begin with a yellow circle. You go *straight ahead* to a blue circle. Then you go to the *right* to a red circle and then *left* to another red circle. From the second red circle, go to the *right* again to a yellow circle and then *straight ahead* to a blue circle.

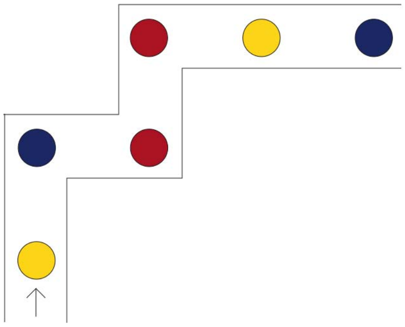

Figure 2

Notice that the term *straight ahead* in description (3b) is used to refer to two different directions of transition. First, it is used for the vertical transition from the first (yellow) circle to the second (blue) circle. Later, it is used again to refer to the lateral transition from the second from the last (yellow) circle to the last (blue) circle. In contrast, the terms used in description (3a) hold a constant relationship to a particular axis on the paper. For the deictic description, there is perfect isomorphism between the input image and the output description. In contrast, for the intrinsic descriptions, there is non-isomorphism, which allows for a further investigation of how gesture is related to the two representations.

If gesture is generated on a faithful memory representation of the image, as proposed by the strong version of the Free Imagery Hypothesis, then, in cases of speech-image mismatch, gesture should align to the image and conflict with speech. If gesture is generated from the lexical semantics of the words used to encode the message, as suggested by the Lexical Semantics Hypothesis, then gesture should always match the speech and conflict with the image. If gesture is generated from an interface representation that results from 'thinking for speaking' processes, as suggested by the Interface Hypothesis, then gestures may match preferentially either the image or the speech, depending on the needs of the speaker at any given moment. In this case, characteristics of the image, the lexical item, or the situation could affect to which representation the gesture is aligned.

Constructing the Corpus

Speakers. Sixteen native speakers of Dutch from the Max Planck Institute for Psycholinguistics' subject pool were paid for their participation.

Pictures. Sixteen path-like images depicting networks of colored circles were constructed. Each image consisted of an explicit start point as well as red, yellow, and blue circles arrayed along a path. Half of the pictures had branching paths while the other half did not. All speakers saw all pictures in the same presentation order. In sum, 256 picture descriptions were collected.

Procedure. Speakers were seated across from their interlocutor separated by a visual block. Their task was to describe the pictures to the interlocutor, who was, in fact, a confederate.

Speakers were given approximately 15 seconds to study the image, which was placed on the table by the experimenter. After this memorization period, the picture was removed and the speaker began to describe the image. Speakers were free to describe the routes in any way that was natural to them; they were not given any linguistic examples to bias their description strategy. The listener was instructed not to ask any specific questions that might bias the content of the descriptions. She was free, however, to ask the speaker to repeat portions or even the entire description of an image. All sessions were video recorded.

Coding system. A native Dutch speaker familiar with gesture transcription systems but blind to the hypotheses under investigation used the videotapes to create a transcription of the speech as well as a record of all gestures. Several types of linguistic information were identified, including directional information (e.g., *right*, *left*, *straight ahead*), destination information (e.g., *a red circle*, *a blue circle*), landmark information (e.g., *you arrive at an intersection*), and shape information (e.g., *you will travel in a big circle*). In this paper, we will focus exclusively on directional information and accompanying directional gestures.

Gestures were either produced with the head or hands. Both were coded for several features, most crucially for the direction of the stroke but also for handedness. Once speech and gesture were fully transcribed, they were coded for three binary features: Speech matches image, gesture matches speech, and gesture matches image. These codes, together with codes for which directional term was produced and unperspectivized direction of transition, form the bases of the mismatch analysis.

Speech-image mismatches were identified as any transition in the network for which the verbal description provided in the intrinsic perspective did not match the actual direction of the transition in the network. For example, any transition labeled *right* that did not progress rightward on the page was a mismatch. Likewise, any use of *straight ahead* that did not correspond to an upward transition was coded as a mismatch. (Note that as the image was placed on the table in front of the speaker, the upward transition in the image was in the forward direction for the speaker, for which *straight ahead* is felicitous.)

Every picture provided multiple mismatch opportunities. For example, 11 networks included an upwards transition, similar to the transition from circle 3 to 4 in Figure 2, which intrinsic speakers described as *right* or *left*. Eight networks included lateral transitions, similar to the final movement in Figure 2, linguistically described as *straight ahead.* Four networks included downward transitions, linguistically described as *right* and three networks including lateral transitions that followed downward transitions. These transitions leftwards or rightwards were described with the opposite directional term, namely, rightward turns were described as *left* and vise versa.

Corpus Analysis

In this section we first give some descriptive details of the corpus before we turn to the crucial questions under investigation.

Characteristics of speech and gesture varied greatly between speakers. Six speakers produced almost no gestures at all. Of the ten gesturers, three produced predominantly deictic descriptions and seven produced predominantly intrinsic descriptions. While the deictic speakers are generally orthogonal to the issue of speech-image mismatch, two produced some mixed perspective descriptions, producing mismatch opportunities. Only speakers who adopted the intrinsic frame of reference AND who gestured are of relevance to our investigation.

In total, the corpus of directional terms consisted of 1440 directional tokens, 389 of which were produced with a co-expressive gesture. Table 1 presents lexical, gesture, and mismatch frequencies for each of the directional terms found in our corpus.[1]

[1] We will present English translations for the Dutch directional terms found in our corpus. With respect to the description of our images, there are no critical differences in how directions and spatial relations are lexicalized.

Table 1. For each directional term, the total number of tokens, the percentage of tokens produced with a gesture, and the number of speech-image mismatches.

Lexemes	Total Tokens	% With Gesture	Speech- Image Mismatches
Right	494	27%	100
Left	317	26%	79
Straight ahead	321	15%	62
Up	120	8%	--
Down	19	10%	--
Back	106	37%	--
Further	63	13%	--

The three directional terms that are relevant for mismatches are *right, left,* and *straight ahead.* In 241 instances, these words did not match the direction in the image. 58 of these speech-image mismatches were produced with a gesture that could either align with the linguistic term or the direction in the image. The terms *up* and *down* were only used by deictic speakers and therefore always matched the input image. The terms *back* and *further* can only be interpreted in the context of prior movements, and therefore the question of whether they match the picture is not applicable

The term *back* received the highest proportion of co-expressive gestures. The terms *left* and *right* were each produced with co-expressive gestures over 25% of the time, *further* and *straight* were produced with intermediate gesture rates and *up* and *down* had the lowest gesture rates.

We now turn to the central question of the paper. The crucial data from the corpus are gestures produced when speech and image are mismatched. The critical question is whether these gestures reflect the direction represented in the image, in speech, or both. The number of gestures that matched the image or the speech for each directional term is presented in Table 2.

Table 2. Number of speech-image mismatches for which the gesture matches either the speech or the image, for the three relevant directional terms.

Lexeme	Gesture = Image	Gesture = Speech
Right	8	15
Left	9	10
Straight ahead	13	3

The distribution of cases where the gesture matches the image compared to when it matches speech is different for the three directional terms, *right, left,* and *straight ahead*, χ^2 (2) = 7.13, $p < .05$. Two by-speakers comparisons were also carried out to assess this relationship. First, for the cases in which gesture aligned with speech, the proportion of speech-image mismatches for each of the three lexemes was calculated for each speaker by dividing the number of speech-image mismatches for the lexeme divided by the total speech-image mismatches for all three lexemes. The proportions differed significantly between the lexemes, Friedman's χ^2 (2) = 7.3, N=8, $p < .05$. Second, for the cases in which gesture aligned with image, the proportion of speech-image mismatches for each of the three lexemes was calculated for each speaker in the analogous way to the previous analysis. There is no evidence that proportions differed between the lexemes, Friedman's χ^2 (2) = 0.9, N=8, $p > .1$.

Table 2 shows that gesture alignment patterned differently for different lexemes. Table 3 further breaks down the information for different directions of transition within the network.

Table 3. Number of speech-image mismatches in descriptions of either upwards, downwards or lateral transition in which gesture aligned to either the image or to speech.

Transition Direction	Lexeme	Gesture = Image	Gesture = Speech
Up	Right or left	5	12
Down	Right or left	5	11
Laterally	Right or left	7	2
Laterally	Straight ahead	13	3

The alignment pattern for vertical (up and down) transitions was significantly different compared to lateral transitions, χ^2 (1) = 12.15, $p < .001$. Speakers preferred to align with the image when the transition was lateral but preferred to align with speech when the transition was vertical. One possible interpretation of this pattern is that speakers generally prefer to gesture laterally.

In speech-image mismatch cases, gestures sometimes aligned with the image and sometimes with the speech. This split was quite even for gestures produced for the lexemes *left* and *right* but not for *straight ahead*. In the latter case, speakers preferred to align their gestures with the image. The different alignment patterns to different lexical items may also be interpreted in terms of a general preference to gesture laterally rather than vertically.

What the data from the corpus clearly indicate, however, is that there is no strong tendency to align gestures to the image at the expense of speech or vise versa. When the information conveyed in speech conflicts with the information presented in the visual input, gesture can align with either. The decision as to whether a gesture aligns with the con-current speech is mediated by a spatial factor (lateral vs. vertical transitions). This result was not predicted by the Free Imagery Hypothesis or the Lexical Semantic Hypothesis, as we will discuss in more details in the next section. The result is, however, compatible with the Interface Hypothesis.

Discussion

By using images consisting of path-like networks of circles, we succeeded in constructing a corpus of picture descriptions in which the content of speech and the content of the to-be-described image often conflicted. Our aim was to see whether gestures produced in these instances would be co-expressive with the lexical affiliate, as predicted by the Lexical Semantics Hypothesis (Butterworth & Hadar, 1989; Schegloff, 1984), with the characteristics of the image, as predicted by the strong reading of the Free Imagery Hypothesis (Krauss et al., 1996, 2000), or whether the alignment to one representation or another would be influenced by 'thinking for speaking' processes, as proposed by the Interface Hypothesis (Kita & Özyürek, 2003).

What our corpus analysis reveals is that gesture alignment behavior in speech-image mismatches was not driven solely by either the characteristics of the input image or the characteristics of speech. Rather, the gestural content seemed to be co-determined by the lexeme choice and the type of spatial representation. Specifically, when the lexemes *left* and *right* were used to express the spatial representations *upwards* and *downward*, gesture tended to align with speech rather than with the spatial representation. When the lexeme *straight ahead* was used to express the spatial concepts *leftwards* and *rightwards*, gestures tended to align with the spatial representation of the image. When the lexemes *left* and *right* were used to express the spatial representations *rightwards* and *leftwards*, respectively, gesture again tended to align with the spatial representation.

The fact that the gestural content was determined by the interplay between both lexical and spatial representations makes it difficult to maintain either the Lexical Semantics Hypothesis, which holds that gestures are generated from the semantic representations of lexical items that have been selected for speaking, or the strong version of the Free Imagery Hypothesis, which holds that gestures are generated from pre-linguistically generated imagery.

However, we need to recognize that there are different versions of the Free Imagery Hypothesis, which make different assumptions. In de Ruiter's (2000) version of the Free Imagery Hypothesis, gestures are generated in the Conceptualizer in Levelt's (1989) sense, which generates the (pre-linguistic) proposition to be linguistically formulated in the next utterance. According to de Ruiter, both gestural and linguistic perspectives are determined in the Conceptualizer. Similarly to the strong version of the Free Imagery Hypothesis, "the shape of the gesture [iconic gesture] will be largely determined by the content of the imagery" (de Ruiter, 2000: 293). As such, the results from the present study are problematic not only to the strong version of the Free Imagery Hypothesis, but also to de Ruiter's version. However, because in de Ruiter's model the shape of a gesture is determined in the Conceptualizer, which in principle has access to (pre-linguistic) propositions to be linguistically formulated, it might be possible to modify the model to account for the present results.

The gestural content is determined by the interplay between lexical choice and directions of the transition in the image. This result could be accounted for by the Interface Hypothesis (Kita & Özyürek, 2003), which proposes that gestures are generated from an interface representation, namely, a spatio-motoric representation that is in the process of being prepared for speech. According to this hypothesis, there is a general tendency for an interface representation to converge with the linguistic representation in the utterance being planned. The degree of convergence is determined by various contextual factors (Kita, 2000). In the case of this study, when the spatial representation of the transition is confusable, that is, when the transition is lateral (i.e., *leftwards* or *rightwards*), the convergence to the linguistic representation is weak, and thus gesture tends to match the spatial representation of the transition, rather than the linguistic representation. When the spatial representation of the transition is not confusable, that is, when the transition is vertical (i.e., *upwards* or *downwards*), the interface representation converges strongly to the linguistic representation. Note further that the idea that gestures help distinguish confusable spatial representations is compatible with theories of self-oriented functions, in particular, the theory that gestures help organize spatio-motoric information for speaking (Kita, 2000; Alibali, Kita, Yong, 2000; Kita, 2003).

The data also rule out the possibility that gestures can be randomly generated from either the input imagistic representation or the output lexical representations, alternating randomly between these two sources. This possibility, previously discussed in Kita and Özyürek (2003) predicts that speech-gesture mismatches should randomly align to the input or to speech, without a discernable pattern. This is not the observed pattern, as seen in Table 3.

One could object to our definition of speech-image mismatch. Consider for example the possibility that speakers mentally rotate the image in memory in order to calculate the correct directional term for the intrinsic perspective. In this case, apparent speech-image mismatches would in fact be matches. However, if this were true we would not have expected gesture alignment to ever conflict with speech, since speech would always match the perspectivized internal memory representation of the image for the speaker at that moment. This is not consistent with the observe data pattern, as seen in Table 1.

The gesture-speech mismatches reported in this study are of a different type from what have been attested in previous research. Gesture-speech mismatch has been observed in children's explanations for Piagetian conservation tasks (Church & Goldin-Meadow, 1986) and children's explanations for equivalence of an equation (Perry, Church, & Goldin-Meadow, 1988), and adult and children's description of the solution to the Tower of Hanoi puzzle (Garber & Goldin-Meadow, 2002). In these studies, gesture and speech refer to two distinct referents that are both relevant to the current goal of discourse, or two alternative strategies or solutions that might apply to the problem at hand. For example, in the explanation for a Piagetian conservation task, speech may indicate the height of a glass, *this one is tall*, and gesture may indicate the width of the same glass. By contrast, mismatches that result from perspective taking can be called same-referent mismatches. Speech and gesture have the same referent, namely a motion vector, but they map the vector to a gestural body movement under different perspectives. Using these same-referent mismatches allows the predictions of the competing theories to be properly tested.

To conclude, we have introduced a new source of evidence into the field of gesture research, namely *speech-image* mismatches with concomitant gestures. These speech-image mismatches allow the content of the linguistic and the imagistic representations to be separated and contrasted. An analysis of the behavior of these gestures revealed that gestures cannot be generated from a purely linguistic or purely imagistic representation. Rather, gestural content was determined by the interplay between the lexical items used in the description and the type of directional information in spatial representation of the transitions. Many issues in gesture research have had difficulty in finding clear evidence for or against specific proposals exactly because it is generally difficult to disentangle the independent contributions of linguistic and imagistic representations. The present paper uses perspective taking to avoid this problem. The present study is also significant in that the methodology affords reliable elicitation of same-referent mismatches from normal adult speakers.

Acknowledgments

We acknowledge useful comments from the members of the Gesture Project at the Max Plank Institute for Psycholinguistics. We greatly appreciated comments on an earlier version of this paper from Joana Cholin, Asifa Majid and Andrea Weber. We also benefited greatly from insightful comments by anonymous reviewers. Special thanks also go to Anne-Marie van Hoof and Esther Vrinzen for acting convincingly as our confederate and for transcribing the descriptions.

References

Alibali, M. W., Kita, S, & Young, A. J. (2000). Gesture and the process of speech production: we think, therefore we gesture. *Language and Cognitive Processes, 15,* 593-613.

Butterworth, B. & Hadar, U. (1989). Gesture, speech and computational stages: A reply to McNeill. *Psychological Review, 96*, 167-174.

Cassell, J., McNeill, D., & McCullough, K. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition, 7*, 1-33.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition 23*, 43-71.

de Ruiter, J.-P. (1998). *Gesture and Speech Production*. Ph.D. Dissertation, Max Planck Institute for Psycholinguistics, The Netherlands.

de Ruiter, J. P. (2000). The production of gesture and speech. In D. McNeill (Ed.), *Language and gesture* (pp. 284-311). Cambridge: Cambridge University Press.

Garber, P., & Goldin-Meadow, S. (2002). Gesture offers insight into problem solving in adults and children. *Cognitive Science, 26*, 817-831.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining match: Gesturing lightens the load. *Psychological Science, 12 (6)*, 516-522.

Kita, S. (2000). How representational gestures help speaking. In D. McNeill (Ed.), *Language and gesture* (pp. 162-185). Cambridge: Cambridge University Press.

Kita, S. (2003). Interplay of gaze, hand, torso orientation and language in pointing. In S. Kita (Ed.), *Pointing: where language, culture, and cognition meet* (pp.307-328). Mahwah, NJ: Lawrence Erlbaum

Kita, S. & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language, 48*, 16-32.

Krauss, R., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? *Advances in Experimental Social Psychology, 28*, 389-450.

Krauss, R., Chen, Y., & Gottesman, R. (2000). Lexical gestures and lexical access: a process model. In D. McNeill (Ed.), *Language and gesture: Window into thought and action.* Cambridge, UK: Cambridge University Press.

Levelt, W. (1996). Perspective taking and ellipsis in spatial descriptions. In P. Bloom, M. A. Peterson, M. F. Garrett, & L. Nadel (Eds.), *Language and space* (pp. 77-107). Cambridge: MIT Press.

Levelt, W.J.M., Roelofs, A., & Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*(1), 1-38.

Levinson, S. C. (2003). *Space in language and cognition: Exploration in cognitive diversity.* Cambridge: Cambridge University Press.

Melinger, A. & Kita, S. (Submitted). Conceptual load triggers gesture production.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and Perception.* Cambridge, MA: Harvard University Press.

Perry, M., Church, R.B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*, 359-400.

Schegloff, E. A. (1984). On some gestures' relation to speech. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversational analysis.* Cambridge: Cambridge University Press.

Slobin, D. I. (1987). Thinking for speaking. In J. Aske, N. Beery, L. Michaelis, & H. Filip (Eds.), *Proceedings of the 13th annual meeting of the Berkeley Linguistic Society* (pp. 435-445).

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96). Cambridge: Cambridge University Press.

Tversky, B. (1991) Spatial mental models. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory, Vol. 27* (pp.109-146). New York: Academic Press.