Linguistic Spatial Gestures

Leonard A. Breslow (len.breslow@nrl.navy.mil)
Anthony M. Harrison (anthony.harrison@nrl.navy.mil)
J. Gregory Trafton (Trafton@itd.nrl.navy.mil)
NCARAI, Naval Research Lab., Code 5515
Washington, DC 20375 USA

Abstract

We model the gestures accompanying spoken descriptions of spatial information and propose a conception of spatial gestures that differs from previous proposals by making a distinction between gestures used for thinking (cognitive gestures) and gestures used to help express predetermined ideas (linguistic gestures), and positing a tighter integration between gesture and language production in the latter than most previous researchers.

Keywords: gesture; spatial reasoning; language production

Introduction

Symbolic speech-accompanying gesture, representing spatial information, has lately been an area of active research (Alibali, 2005). Of particular interest is the relationship between gesture and the language it accompanies. In considering this relationship, we think it is useful to distinguish between gestures that help us determine *what* to communicate/express (*cognitive gestures*) and gestures that help us to express what we have determined to say, that is, gestures concerned with *how* to communicate (*linguistic gestures*). While these two functions clearly overlap in certain cases, we consider the distinction useful. Specifically, we argue that, in general, cognitive gestures lead language, whereas language leads gesture in the case of linguistic gestures.

Cognitive gesture leads language indirectly by facilitating thinking, thereby helping us determine what to say. Thus, they are used in situations with competing conceptual representations (Kita & Davies, 2009), high conceptual load (Melinger & Kita, 2007), mental rotation tasks (Chu & Kita, 2008), expert and novice scientific thinking (Trafton et al., 2006), and problem solving (Lozano & Tversky, 2006), among others. Such gestures are relatively independent of language, often expressing information different from that expressed in the accompanying language, and sometimes cognitively more advanced than the latter, e.g., in development (Alibali & Goldin-Meadow, 1993) or in problem-solving performance (Lozano & Tversky, 2006). While cognitive gestures sometimes aid communication (Lozano & Tversky, 2006), they are relatively independent of communication, as evidenced by their use when solving problems silently in solitude (Chu & Kita, 2008; Lozano & Tversky, 2006).

In contrast to cognitive gestures, linguistic gestures are more strongly tied to language and dependent upon language. They convey little or no information beyond what is expressed in the accompanying language (Beattie & Shovelton, 1999; So, Kita, & Goldin-Meadow, 2009), except where the respective roles of gesture and language are predetermined as in deixis ("Look at that!") or in language referring to gesture ("It was *this* big."). Neurological as well as behavioral evidence suggests the absence of priming of words by gestures in comprehension (Bernardis & Caramelli, 2007) or production (Beattie & Coughlan, 1999; Bernardis, Salillas, & Caramelli, 2008). On the contrary, language primes gesture comprehension (Bernardis & Caramelli, 2007) and cross-linguistic studies demonstrate that the grammatical organization of speech is predictive of the sequence and nature of symbolic gesturing (Kita & Ozyurek, 2003).

Also in contrast to cognitive gestures, linguistic gestures are typically associated with communication, as evidenced by the great reduction in gesturing when the listener cannot see the speaker (Alibali, Heath, & Myers, 2001) and the absence of gesturing outside of communication (e.g., in silence or solitude). However, we do not claim that linguistic gestures always *facilitate* communication, since people gesture even when speaking on the telephone (de Ruiter, 1995).

Note that the outward form of both cognitive and linguistic gestures may appear very similar – they are iconic gestures typically tied to a spatial representation of what is being thought or said. The types of gestures may be distinguished by the degree to which the gesturer has difficulty determining the spatial ideas he/she wishes to express, which may vary by population (e.g., child vs. adult) as well as by situation (e.g., problem-solving vs. simple description).

We will focus in the remainder of this paper on linguistic gestures. One question that researchers have considered is the extent to which the perceptual information being described by the speaker inputs directly into the generation of gestures, without the intermediary of language processing. Some argue that direct perception accounts for the few features of gestures that are not conveyed in the accompanying language (Kita & Ozyurek, 2003). Other theories (de Ruiter, 2007; Hostetter & Alibali, 2008) attempt to account for gesture solely on the basis of perception or imagery. Both types of theory are challenged in explaining the process by which perceptual features are selected for inclusion in gestural representation.

We propose a model of linguistic gestures that posits a tighter integration between gesture and language than most previous models (as does McNeill, 1992) by adopting a broader view of language representation than typically used.

Our approach draws on a recent linguistic theory proposed by Ray Jackendoff (2002), according to which language representation includes some irreducibly spatial components. It also draws on the construction grammar approach of Goldberg (1995), according to which linguistic structures containing both semantic and syntactic components are central to language processing. Combining these two approaches, we hypothesize that people select a construction before retrieving words and gestures. The construction provides an abstract plan for speaking that includes the semantic-syntactic information used in the retrieval of both words and spatial representations at appropriate places in the utterance. The spatial representations are the basis of symbolic gestures and so this approach helps to identify where specific gestures will occur. Following Jackendoff, we hypothesize that the spatial representations are abstract in nature (Avraamides et al.,2004). We propose that these abstract representations may be instantiated either as internal mental images or externally as gestures.

This account predicts that the information conveyed in linguistic gesture will be tightly tied to the accompanying language, since both language and gesture derive from the same construction. This helps to explain why linguistic gestures provide little information not included in the accompanying language. What little extra information is included in gesture is information required for the instantiation of an abstract spatial schema (a spatial element of a linguistic construction) in a particular situation. For instance, a gesture representing an observed leftward movement is usually performed in a leftward direction (Kita & Ozyurek, 2003), since a linear gesture must have *some* direction. But the gestural reproduction of the stimulus is limited to what is necessary to instantiate an abstract spatial schema as a physical hand movement. Thus, this account provides a mechanism for selecting perceptual features for inclusion in gestural representation, in contrast to unconstrained perceptual accounts (de Ruiter, 2007; Hostetter & Alibali, 2008). This account also helps to explain the observed temporal synchrony between gestures and utterances of similar meaning (McNeill, 1998).

Modeling Language

We evaluate our conception of linguistic spatial gesture by modeling the findings reported by Kita and Ozyurek (Kita & Ozyurek, 2003). Native speakers of English, Japanese, and Turkish were shown a cartoon and asked to describe it to another person. In one scene, a cat (Sylvester) jumps out the window of an apartment building, grabs onto a hanging rope and swings across the street to another building. In another scene, the cat, after swallowing a bowling ball, rolls down the street. English speakers described both path (down/across the street) and manner of locomotion (swing or roll) in a single clause, such as (with clauses marked by square brackets):

English-Swing: [The cat swings across the street.]

English-Roll: [The cat rolls down the street.]

In contrast, speakers of Japanese and Turkish (hereafter J/T) described path and manner in two separate clauses, paraphrased roughly as:

J/T-Swing: [[The cat goes across the street], [ ]]
J/T-Roll: [[The cat goes down the street], [as he rolls]]

Note that J/T lack an appropriate equivalent to "swings" in this context, an unusual lacuna in both these languages, and so the manner is not described verbally, but is often depicted by a gesture following the spoken clause, the position where a dependent clause describing manner normally occurs, as in the J/T-Roll sentence.

The clausal structure of the four sentences, above, corresponds to linguistic *constructions* as characterized by Goldberg (1995). A linguistic construction is a semantic-syntax pair that also specifies the mapping between semantics and syntax. While her theory focuses primarily on clausal constructions, Goldberg considers the construction framework to be applicable to all levels of the language down to words. Thus, the J/T description of the roll event consists of two constructions nested within a larger construction, as shown in J/T-Roll, above.

Table 1 outlines a simplified English intransitive motion construction, characterizing the semantic and syntactic components of the clause in English-Roll, adapted from Goldberg (1995).

Table 1: A simplified intransitive motion construction.

Semantics	THEME	MOVE	GOAL
Lexical items	"He"	"rolls"	"down the street"
Syntax	subject	verb	oblique prep. phrase

We omit many details. A construction has semantic content beyond that indicated by standard semantic categories, such as those shown here; for example, this construction denotes movement along a path. For Goldberg, the verb has a centrality not depicted here and constructions include rules for mapping from semantics to syntax that we omit. Note that the lexical items are not part of the construction, but instead are added to the construction in the course of its application.

We adopt a simplified process model for language production based on constructions, consisting of the following sequence:

1. **Construction retrieval/instantiation.** A construction is selected based on the match of its semantic components to the situation, in the process of which those semantic components are instantiated.
2. **Lexical retrieval.** Lexical items (e.g., words) are retrieved for each semantic component in turn (from

left to right in Table 1) based on the semantics-syntax mapping specified by the construction as well as by its semantic content.

In the course of the first, construction retrieval, step, semantic components in the construction are instantiated with concepts and/or spatial representations. Following Jackendoff (2002), we hypothesize that some semantic categories are instantiated with irreducibly spatial representations. In fact, Jackendoff argues that the semantics of the MOVE component in an intransitive motion construction is exclusively spatial in nature. Note in English, the MOVE component represents the *manner* of movement (e.g., swinging, rolling). In contrast, this manner of movement component is absent from the intransitive motion constructions in J/T; instead, the manner of movement is represented by a separate dependent clause following the intransitive motion clause.

We hypothesize that the instantiated spatial components of constructions at all levels (multi-clausal, clausal, lexemes), resulting from step 1, constitute the basis for gesturing during speech.

Modeling Gesture

Kita and Ozyurek (2003) categorize the manual gestures found to accompany utterances English/J-T-Swing/Roll, above, into one of three types:

1. *Manner only*: e.g., a circular motion for rolling.
2. *Trajectory only*: e.g., a straight-line motion from left to right.
3. *Conflated*: depicting both manner and trajectory, e.g., a looping left-to-right movement for rolling.

As a manner-only gesture is not possible for denoting swinging, only trajectory and conflated gestures accompanied the swing utterances. In general, the authors found that the language groups differed in their gestures in a manner corresponding to the structure of their utterances: English speakers often made conflated gestures only, whereas J/T speakers more often made manner only and trajectory only gestures. Note that the language groups did not differ in their overall production of conflated gestures, but in the tendency to produce *only* conflated gestures, which was more common in English. Based on these findings, the authors proposed that the production of gestures is influenced by the structure of language in the planning stage of speech production.

Kita and Ozyurek also noted that among all language speakers, the direction of gestures (e.g., left to right) generally corresponded to the direction observed in the cartoon, but was never mentioned in the utterances. On this basis, they posited a separate line of influence of perception on gesture, unrelated to language. In contrast to this, we propose a unified account of gesture and language production.

We hypothesize that the spatial components of constructions at all levels (discourse, multi-clausal, clausal, lexemes) constitute the basis for gesturing during speech. We explain the selection of spatial features of an event for gestural representation in terms of the requirement to instantiate an abstract spatial representation to produce both speech and gesture. Since a translation gesture must have *some* direction, the reproduction of the observed direction is simply part of this instantiation process.

Although Kita and Ozyurek did not report the correspondence between gesture and language in a fine-grained manner, we have inferred from their reported data the correspondence outlined in Tables 2 and 3. Note that there is no manner clause for Swing descriptions available to J/T speakers. We make certain assumptions based on the common observation that symbolic gestures co-occur with like-meaning language (McNeill, 1998). Thus, manner-only and conflated (manner+trajectory) gestures accompany manner language (the verb in English, the adverbial post-clause in J/T), while trajectory-only and conflated gestures accompany path language (the prepositional phrase in English, the intransitive motion clause in J/T). The relative frequency of the two possible gestures for the two respective language segments of interest (path vs. manner language) is the focus of our model.

Table 2. Language and accompanying gestures during Roll description observed and predicted by model.

		% Ss observed	**Model**
English			
Language	**Gesture**		
Manner verb	Conflated	66	51
Manner verb	Manner only	13	18
Path phrase	Conflated	53	68
Path phrase	Trajectory only	39	28
Japanese / Turkish			
Language	**Gesture**		
Manner clause	Conflated	59	76
Manner clause	Manner only	40	16
Path clause	Conflated	25	31
Path clause	Trajectory only	67	66

Table 3. Language and accompanying gestures during Swing description observed and predicted by model.

		% Ss observed	Model
English			
Language	**Gesture**		
Manner verb	Conflated	88	93
Path phrase	Conflated	81	88
Path phrase	Trajectory only	7	3
Japanese/Turkish			
Language	**Gesture**		
[Manner clause position]	Conflated	75	88
Path clause	Conflated	37	30
Path clause	Trajectory only	63	70

The construction approach provides a useful framework for understanding both planning and online production of speech. In the present context, it offers an explanation of how ongoing speech can be influenced by elements of the speech plan that are executed before and after the currently executed (spoken) element, an explanation that can be extended to gesture. Specifically, we hypothesize that spatial semantic components within the same construction will have a greater influence on one another (via priming, etc.) than those in separate constructions. Further, this influence will be greater the lower the shared construction is in the construction hierarchy, since spatial representations are more concrete and less abstract lower in the hierarchy. Thus, conflated gestures, representing both trajectory and manner, are proportionally more common during the path language in English than in J/T because the path language in English shares the same construction as the manner language, in contrast to J/T where manner language is in a separate low-level construction. However, conflated gestures do occur sometimes in J/T because the respective clauses describing path and manner are contained within the same higher-level construction.

Similarly, clause structure can help to explain how an executed gesture influences the selection of a subsequent gesture. In English, the type of gesture selected to express manner has a great influence on the subsequent gesture selected to express path, whereas in J/T there is no apparent influence of the selection of path-describing gesture on the subsequent manner-describing gesture. This finding is explained by the occurrence of the two gestures within a single clause in English, but in separate clauses in J/T.

Model of Gesture and Language

The models of gesture and language production were developed within ACT-R (Anderson et al.,2004). ACT-R is a hybrid symbolic/sub-symbolic production-based system. ACT-R consists of a number of modules, buffers, and a central pattern matcher. Since ACT-R is not well-suited to represent structured representations, such as nested linguistic constructions, we attempt to capture the retrieval of spatial representations using ACT-R's partial matching capability. Specifically, the relative similarity of pairs of related spatial representations is modulated to reflect their proximity in the construction hierarchy, as is their capability to prime one another.

To represent space, we have developed a version of ACT-R, ACT-R/E, that utilizes a spatial theory called Specialized Egocentrically Coordinated Spaces (SECS) (Harrison & Schunn, 2003). SECS provides two egocentric spatial modules, which are responsible for the encoding and transformation of representations in service of navigation (configural) and manipulation (manipulative). Our model currently includes configural spatial representations.

Non-default ACT-R parameter settings are listed in Table 4. Manner chunk similarity refers to the associative similarity between the manner chunk in a language construction and an imaginal or gestural spatial representation. Similarly for path chunk similarity. Note that similarities are greater in English than in J/T, reflecting the increased priming by a linguistic construction lower in the construction hierarchy compared to a higher-level construction. Overall, for both language groups, there was a higher rate of conflated gestures for the swing description than for the roll description, possibly due to the smaller number of gesture types available for swing (i.e., the absence of a manner-only gesture). This may explain the need for weaker manner chunk similarities for the Roll models relative to the Swing models. The reduction of base level learning rates in English relative to J/T reflects the priming of later gesture selection by the previously-selected gesture in English, unlike in J/T.

Table 4. Non-default ACT-R parameter settings.

Parameter	English swing	J/T swing	English roll	J/T roll
Enable partial matching	true	true	true	true
Activation Noise	0.3	0.3	0.3	0.3
Retrieval threshold	-6.0	-6.0	-6.0	-6.0
Base level learning rate	0.2	0.9	0.5	0.9
Manner chunk similarity	-0.1	-1.0	-0.9	-3.0
Path chunk similarity	-0.2	-1.0	-0.1	-1.0

In describing the Roll situation, the English-language model first retrieves, and instantiates the semantics of, an intransitive motion clause construction, based on the observed event (see Table 1.) The instantiated construction forms the plan for all further retrievals, gestures, and utterances for the clause. First, the model retrieves and utters the first clause argument, the THEME (e.g., "the cat). Next it retrieves a manner verb corresponding to the MOVE argument. The verb contains a spatial representation that strongly primes a *manner* gesture representation, but the clause construction itself carries path-following meaning and so contains a spatial representation that weakly primes a *trajectory* gesture representation—weakly, because the clause construction is a higher-level construction than the verb. Although the priming of a trajectory gesture is weaker than that of a manner gesture in English, it is stronger than the priming of the corresponding "non-matching" gestures in J/T, because those gestures are primed by a still higher-level construction. As a result, English speakers more often retrieved both manner and trajectory gesture representations, fusing them into a conflated gesture. However, when only the manner gesture representation is retrieved, then a manner-only gesture will be performed. The manner verb is then uttered together with the selected gesture.

Next, path description language (spec. a prepositional phrase) is retrieved based on the instantiated GOAL. Once retrieved, this path phrase's spatial trajectory representation strongly primes a trajectory gesture. At the same time, the construction's MOVE representation *weakly* primes the manner gesture, weakly because it is at a higher level than the path language. Also, if the manner gesture was retrieved and performed earlier with the verb, that earlier retrieval makes an additional contribution to its activation, making it more likely to be retrieved again; no such priming occurs in J/T because the two successive gestures occur in separate constructions. If the manner representation is retrieved together with the trajectory representation, then the GOAL utterance is accompanied by a conflated gesture. If only a trajectory representation is retrieved, then it is accompanied by a trajectory-only gesture.

The J/T models function in a similar manner to this illustration, differing primarily in the nested structure of its constructions.

Given that individual variability is typically quite high for gesturing, the predictions of our model are rather similar to the observed pattern of behavior (Tables 2 and 3) and were all within the 95% confidence interval. r^2 was .63 for Roll and .98 for Swing.

Discussion

We have introduced the contrast between cognitive and linguistic spatial symbolic gestures in hopes of resolving apparently conflicting evidence in the literature. Cognitive gestures help us to determine *what* to say in a spatially complex domain, while linguistic gestures help us to express what we have determined to say. Obviously these two types of gesturing may be intermixed in a given situation, but certain experimental situations clearly encourage one type of gesturing over the other for a given population.

With regard to linguistic gestures, we hypothesize that gestures are generated on the basis of spatial components within linguistic representations (Jackendoff, 2002). The grammatical framework we adopt is that of constructions (Goldberg, 1995) in which lexical items, clauses, and more complex linguistic expressions may all be viewed as constructions, i.e., semantic-syntactic pairings whose semantic content, we hypothesize, includes abstract spatial components. The spatial semantic content at all levels of the construction hierarchy constitutes the basis for gesturing.

From this viewpoint, linguistic gestures are largely constrained by language generation. Specifically, perceptual information is incorporated in gesture during the course of instantiating linguistic structures. This obviates the need to hypothesize a separate, independent source of perceptual input into gesturing, together with the problems such a hypothesis entails: of explaining that mechanism and, especially, of explaining the selection of perceptual features to represent gesturally. As the information conveyed in gesture is largely limited to that conveyed in language, it would appear inappropriate to posit an unconstrained source of perceptual input into gesture production.

From our perspective, linguistic gesture and language are intimately related. Our model is an explicit computational / process account of McNeill's proposal that gesture and speech arise from a single process of utterance formation (McNeil, 1992, p. 29-30).

Although not addressed in this model, many factors modulate the rate of gesturing, such as social stimulation (Alibali et al., 2001), exposure to perceptual vs. verbal information (Hostetter & Hopkins, 2002), etc. The idea of an activation threshold governing the elicitation of gesturing, proposed by Hostetter and Alibali (2008), may be useful in explaining the expression of spatial representations externally in gesture rather than internally in imagery.

Acknowledgments

This work was supported by the Office of Naval Research under job order number N0001409WX20173. The views and conclusions contained in this document should not be interpreted as necessarily representing official policies, either expressed or implied, of the U.S. Navy.

References

Alibali, M. W. (2005). Gesture in Spatial Cognition: Expressing, Communicating, and Thinking About Spatial Information. *Spatial Cognition and Computation, 5*(4), 307–331.

Alibali, M. W., & Goldin-Meadow, S. (1993). Transitions in learning: What the hands reveal about a child's state of mind. *Cognitive Psychology, 25*, 468-523.

Alibali, M. W., Heath, D. C., & Myers, H. J. (2001). Effects of visibility between speaker and listener on gesture production: Some gestures are meant to be seen. *Journal of Memory and Language, 44*, 169-188.

Anderson, J. R., Bothell, D., Byrne, M. D., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111*(4), 1036-1060.

Avraamides, M. N., Loomis, J. M., Klatzky, R. L., & Golledge, R. G. (2004). Functional Equivalence of Spatial Representations Derived From Vision and Language: Evidence From Allocentric Judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(4), 801-814.

Beattie, G., & Coughlan, J. (1999). An experimental investigation of the role of iconic gestures in lexical access using the tip-of-the-tongue phenomenon. *British Journal of Psychology, 90*, 35-56.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. . *Semiotica, 123*, 1-30.

Bernardis, P., & Caramelli, N. (2007). Semantic priming between words and iconic gestures. In S. Vosniadou, D. Kayser & A. Protopapas (Eds.), *Proceedings of the Second European Cognitive Science Conference* (pp. 614-619). Delphi, Greece: Lawrence Erlbaum Associates.

Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysio-logical evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology, 25*(7-9), 1114-1128.

Chu, M., & Kita, S. (2008). Spontaneous Gestures During Mental Rotation Tasks: Insights Into the Microdevelopment of the Motor Strategy. *Journal of Experimental Psychology: General 137* (4), 706–723.

de Ruiter, J. P. (1995). Why do people gesture at the telephone? In M. Biemans & M. Woutersen (Eds.), *Proceedings of the Center for Language Studies: Opening, Academic Year 95-96* (pp. 49-55). Nijmegen: University of Nijmegen.

de Ruiter, J. P. (2007). Postcards from the mind: The relationship between speech, imagistic gesture, and thought. *Gesture, 7*(1), 21-38.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

Harrison, A. M., & Schunn, C. D. (2003). *ACT-R/S: Look Ma, No "Cognitive-map"!* Paper presented at the International Conference on Cognitive Modeling.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review, 15*(3), 495-.

Hostetter, A. B., & Hopkins, W. D. (2002). The effect of thought structure on the production of lexical movements. *Brain and Language 82*, 22–29.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

Kita, S., & Davies, T. S. (2009). Competing conceptual representations trigger co-speech representational gestures. *Language and Cognitive Processes, 24*(5), 761-775.

Kita, S., & Ozyurek, 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.

Lozano, S. C., & Tversky, B. (2006). Communicative gestures facilitate problem solving for both communicators and recipients. *Journal of Memory and Language, 55*, 47–63.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D. (1998). Models of Speaking (To Their Amazement) Meet Speech-Synchronized Gestures. In D. McNeill (Ed.), *Language and gesture: window into thought and action.* Hillsdale, NJ: Lawrence Erlbaum Assoc.

Melinger, A., & Kita, S. (2007). Conceptualisation load triggers gesture production. *Language and Cognitive Processes, 22*(4), 473-500.

So, W. C., Kita, S., & Goldin-Meadow, S. (2009). Using the Hands to Identify Who Does What to Whom: Gesture and Speech Go Hand-in-Hand. *Cognitive Science, 33*, 115-125.

Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L., Schunn, C. D., & Kirschenbaum, S. S. (2006). The relationship between spatial transformations and iconic gestures. *Spatial Cognition and Computation, 6*(1), 1-29.