How Does Spoken Language Shape Iconic Gestures?

Sotaro Kita
University of Birmingham

Aslı Özyürek
Center for Language Studies, Radboud University
and
Max Planck Institute for Psycholinguistics

In this paper we review a series of studies that show empirically that the lexical and syntactic structure a speaker uses to express information influences the way iconic gestures are shaped. The findings come from cross-linguistic comparisons, second language acquisition, as well as experimental studies conducted within a single language. We argue that the findings provide direct evidence for McNeill's claim that gestures are a necessary component of thinking and speaking, and that gestures reflect an online dynamic process of a 'dialectic between imagery and language'.

1. Introduction

Our interest in gesture studies and in particular how gestures are related to the speaking process originated while we were both students in David McNeill's lab at the University of Chicago, and has continued in our collaborations at the Max Planck Institute (the 'Gesture Project'), and thereafter. In our graduate and postdoctoral work, we have been fascinated by and pursued McNeill's idea that gestures are a necessary component of thinking and speaking, and that gestures reflect an online dynamic process of the 'dialectic between imagery and language' (McNeill, 1985; 1992; 2005), currently called the Growth Point Theory (McNeill & Duncan, 2000; McNeill, 2005). In our ongoing work we attempt to provide empirical evidence for this idea, in particular for one side of the dialectic; the influence of language structure on how imagery is shaped for speaking and iconic gestures.

As students coming from Japan and Turkey, and speaking native languages that differ typologically from English, we were both interested in testing the idea McNeill put forward in *Hand and Mind* (1992) that gestures, as reflections of imagery, should be similar across different languages. However, counter to our initial intuitions we have found this not to be the case (Özyürek & Kita, 1999; Kita & Özyürek, 2003). Since then we have been developing a theoretical framework, the Interface Hypothesis, to account for the fact that iconic gestures depict the same event in different ways, depending on the specific lexical item or syntactic structure used to describe the event in the concurrent speech.

In this paper we first outline competing accounts of how iconic gestures are generated during speaking and their relation to imagery. Then we describe the Interface Hypothesis of online relations between speaking and gesturing, as well as empirical evidence that supports this hypothesis. We end by discussing the links between Growth Point Theory and the Interface Hypothesis.

2. Three Hypotheses of how Iconic Gestures are Generated

According to the first hypothesis of how iconic gestures are generated in relation to speech, the Free Imagery Hypothesis, gestures are generated from imagery that is formed 'prelinguistically', that is, independent of linguistic formulations. Krauss and his colleagues (1996; 2000), for example, suggested that gestures are generated from spatial imagery in working memory, activated at the moment of speaking. More specifically, the 'spatial-dynamic feature selector' picks up spatial features in the spatial working memory that are a part of the idea to be conveyed, and these features define the content of a gesture. A subset of these features may contribute to lexical retrieval in speech by cross-modal priming. Other than that, however, the 'spatial-dynamic feature selector' has no access to the computation involved in the speech-production process, and thus the gestures cannot be shaped by the linguistic formulation. Unlike Krauss and his colleagues, de Ruiter (2000) proposed that representational gestures are generated by a part of the speech production process proper, the 'conceptualizer' (in the sense of Levelt, 1989), which produces a pre-verbal message to be fed into the 'formulator' (i.e., the linguistic formulation module). Despite this difference, in both Krauss' and de Ruiter's models, gestures are generated before and without access to processes of linguistic formulation. Consequently, both models support the Free Imagery Hypothesis, which predicts that the encoding of a concept in gesture is *not* influenced by the details of how that same concept is linguistically encoded in speech.

The second view is the Lexical Semantics Hypothesis, according to which gestures are generated from the semantics of lexical items in their accompanying speech. Butterworth and Hadar (1989) proposed that a lexical item generates iconic gestures from one or more of its semantic features that can be interpreted spatially. Unlike the Free Imagery Hypothesis, the prediction of the Lexical Semantics Hypothesis is that representational gestures do not encode what is not encoded in the concurrent speech. The Lexical Semantics Hypothesis further states that the source of gestures lies strictly at the lexical level rather than at the levels of syntax and discourse.

The third view is the Interface Hypothesis (Kita & Özyürek, 2003), according to which gestures originate from an interface representation, which is spatio-motoric, and organized for the purpose of speaking. In other words, gestures are produced from a type of 'thinking for speaking' in the sense of Slobin (1987;

1996) (Kita, 1993; 2000; Kita & Özyürek, 2003; McNeill, 2000; McNeill & Duncan, 2000). In this view, speaking imposes constraints on how information should be organized. The organization has to conform to the existing lexical and constructional resources of the language (Slobin, 1996) and the linear nature of speech (Levelt, 1989). Furthermore, the limited capacity of the speech production system imposes constraints as well. Rich and complicated information is organized into smaller packages so that each package has the appropriate informational complexity for verbalization within a processing unit. This may involve feedback from speech 'formulation' processes to 'conceptualization' processes (Kita, 1993; Vigliocco & Kita, 2006). In sum, the optimal informational organization for speech production is determined by an interaction between the representational resources of the language and the processing requirements of the speech production system.

In line with this view of speaking, the Interface Hypothesis proposes that gestures are generated during the conceptual process that organizes spatio-motoric imagery into a suitable form for speaking. Thus, it predicts that the spatio-motoric imagery underlying a gesture is shaped *simultaneously* by 1) how information is organized in a readily accessible linguistic expression that is concise enough to fit within a processing unit for speech production, and 2) the spatio-motoric properties of the referent (which may or may not be verbally expressed).

3. Evidence for the Interface Hypothesis

The initial evidence for the Interface Hypothesis was based on cross-linguistic differences in how motion events are linguistically expressed, and how they are reflected in cross-linguistic differences in gestures. The cross-linguistic variation in gestural representation was first demonstrated in a comparison of gestures produced by Japanese, Turkish, and English speakers in narratives that were elicited by an animated cartoon (*Canary Row* with Sylvester and Tweety, used often by David McNeill, his students and colleagues). There were two events in the cartoon for which the linguistic packaging of information differed between English on one hand and Japanese and Turkish on the other. In the first event, a protagonist swung on a rope, like Tarzan from one building to another. It was found that English speakers all used the verb "swing," which encoded the arc shape of the trajectory, and Japanese and Turkish speakers used verbs such as "go," which did not encode the arc trajectory. This is because Japanese and Turkish do not have an agentive intransitive verb that is equivalent to English 'swing', nor is there a straightforward paraphrase. Presumably, in the conceptual planning phase of the utterance describing this event, Japanese and Turkish speakers get feedback from speech formulation processes and create a mental representation of the event that does not include the trajectory shape. If gestures reflect this planning process, the gestural contents should differ cross-linguistically in a way analogous to the difference in speech. It was indeed found that Japanese and Turkish speakers were more likely to produce a straight gesture, which does not encode the trajectory shape, and most English speakers produced only gestures with an arc trajectory (Kita, 1993; 2000; Kita & Özyürek, 2003).

In the second event, in which one of the protagonists rolled down a hill, the description differed cross-linguistically along the lines discussed by Talmy (1985). English speakers used a verb and a particle or preposition to express the 'manner' (rolling) and 'path' (descending) of the event within a single clause (e.g., "he rolled down the hill"). In contrast, Japanese and Turkish speakers separated manner and path expressions into two clauses (e.g., "he descended as he rolled"). This difference in the clausal packaging of information should have processing consequences, because a clause approximates a unit of processing in speech production (Garrett, 1982; Levelt, 1989). It is plausible that manner and path are processed within a single processing unit for English speakers, but in two separate units for Japanese and Turkish speakers. Thus, as compared to English speakers, Japanese and Turkish speakers should separate the images of manner and path more often so as to process them in turn. The gesture data are consistent with this prediction (Özyürek & Kita, 1999; Kita & Özyürek, 2003). Japanese and Turkish speakers were more likely than English speakers to produce separate gestures for manner and path.

Furthermore, in the description of both the swing and the roll events, the same gestures that showed the linguistic effects described above also simultaneously expressed spatial information never encoded in speech. More specifically, these gestures systematically expressed the left-right direction of the observed motion. For example, if the cat swung from the right side of the TV screen to the left, the gestures always consistently represented the motion this way, even though this information was never expressed in speech. This phenomenon was first reported for English by McCullough (1993) and was found to be the same in all languages in our study. Further evidence for gestures encoding information not encoded by speech has also been reported in children's explanations of scientific reasoning (e.g., Goldin-Meadow, 2003).

Thus, the above studies showed, first, that the gestural packaging of information parallels linguistic packaging of the same information in speech. This finding contradicts the prediction made by the Free Imagery Hypothesis, according to which the representations that underlie gestures are determined at a level of the speech production process that has no access to the details of linguistic formulation. Second, the above studies showed that gestures encoded spatial details that were never verbalized. This poses a critical problem for the Lexical Semantics Hypothesis, as does the finding of relationships between the syntactic and gestural packaging of manner and path information. According to this hypothesis, the content of gestures should be determined only at the lexical level. Japanese, Turkish, and English speakers all used one word referring to manner and a second word referring to path, but nevertheless the gestures differed in accordance with differences at the syntactic level. On the basis of these results, it was concluded that the content of gestures is shaped *simultaneously* by how speech organizes information about an event and by spatial details of the event, which may or may not have been expressed in speech.

One problem with the studies described so far was that the differences in iconic gestures used by English speakers on the one hand and Turkish and Japanese speakers on the other might have causes other than the syntactic packaging of information, such as broader cultural differences. In order to eliminate this possibility, we looked at a new set of manner and path descriptions to see how gestures are shaped when Turkish and English speakers use similar versus different syntactic constructions (Özyürek, Kita, Allen, Brown & Furman, 2005). We found that the cross-linguistic difference in gestures emerges *only* when speakers of Turkish and English used different syntactic means (i.e., one versus two clause constructions) but *not* when the speakers of the two languages used comparable syntactic packaging of information. That is, when only the manner or only the path of a given event was expressed in speech, similar content was expressed in gesture regardless of the language (e.g., English speakers produced manner-only gestures when they expressed only manner in speech just as Turkish speakers did). Thus, it is not the case that English speakers always preferred gestural representations in which manner and path were expressed simultaneously in one gesture. Rather, the cross-linguistic difference in gesture was observed only when the syntactic packaging of manner and path differed. This provides further support for the view that gestural representation is shaped in the process of packaging information into readily verbalizable units for speaking, as proposed by the Interface Hypothesis.

Further evidence came from a study of how native speakers of Turkish gestured when they described the manner and path of motion events in their second language, English. Özyürek (2002a) compared Turkish speakers at different proficiency levels of English (beginners, intermediate and advanced). The most proficient group typically used one-clause expressions of manner and path in speech, and produced gestures that expressed manner and path simultaneously, similar to native speakers of English (Kita & Özyürek, 2003; Özyürek et al., 2005). In contrast, the groups with lower proficiency typically used two-clause expressions (i.e., they transferred their preferred L1 structure into their L2), and produced separate gestures for manner and path, just as they would do when speaking in Turkish. This result indicates that how one syntactically packages information in speech shapes how one gesturally packages information both in L1 and L2.

Finally, a more direct demonstration of how spoken language shapes iconic gestures in an online and dynamic way comes from a study of English speakers (Kita, Özyürek, Allen, Brown, Furman, & Ishizuka, submitted). This study manipulated stimulus events in such a way that English speakers produced both one-clause (e.g., "he rolled down") as well as two-clause descriptions (e.g., "he went down as he rolled") of manner and path. The experimental manipulation used Goldberg's (1997) insight that when the causal link between manner and path is stronger, the preference for one-clause construction is stronger. It was found that gestures expressing manner and path simultaneously were more common in one-clause than two-clause descriptions. In contrast, manner only and path only gestures were more common in two-clause than one-clause descriptions. In other words, the gestural representation of manner and path changed depending on what type of syntactic packaging of manner and path the speaker decided to use for a given utterance. This further substantiates the contention of the Interface Hypothesis that iconic gestures are products of online, utterance-by-utterance conceptualization for speaking. The typological preferences of a given language and how it can shape iconic gestures can change if the speaker chooses a non-typological construction.

In summary, even when speakers describe the same event, gestural depiction of the event varies, depending on how the concurrent speech packages the information about the event at the lexical (Kita, 1993; Kita & Özyürek, 2003) and the syntactic level. The syntactic effect on iconic gestures has been demonstrated in both crosslinguistic (Kita & Özyürek, 2003; Özyürek & Kita, 1999; Özyürek et al., 2005) and single-language studies (for English: Kita et al., submitted; for English as a second language: Özyürek, 2002a). The linguistic effect on iconic gestures provides evidence against the Free Imagery Hypothesis, which posits that iconic gestures are generated without input from linguistic formulation processes. The content of iconic gestures, however, is not completely determined by the content of concurrent speech (Goldin-Meadow, 2003; McCullough, 1993; Kita & Özyürek, 2003). That is, gestures systematically encode spatio-motoric information that is not encoded in concurrent speech, which contradicts the Lexical Semantics Hypothesis. Taken together, the evidence supports the Interface Hypothesis. That is, iconic gestures are generated at the interface between spatio-motoric thinking and language production processes, where spatio-motoric imagery is organized into readily verbalizable units.

4. Conclusion

The evidence we have provided so far for the Interface Hypothesis is also direct evidence for David McNeill's claim that gestures are the product of an online 'dialectic between imagery and language'. More specifically, crosslinguistic investigations show convincingly that the linguistic packaging of information shapes iconic gestures online.

This does not mean, however, that this is the only comprehensive mechanism that underlies the 'dialectic between imagery and language' during speaking or that governs how iconic gestures are shaped in general. In the line of work presented here we have not discussed the roles of the communicative-social context (Özyürek, 2002b; Bavelas, this volume), the discourse context (i.e., the "field of oppositions" in discourse (McNeill, 1992)), nor the tight temporal relations between speech and gesture that play a role in this dialectic. These are elaborated in the most recent version of the Growth Point Theory (McNeill, 2005). We believe that more cross-linguistic work on these aspects of the relations between speech and gesture is needed in the future. For example, there are initial findings that speech-gesture synchrony differs across different languages (Özyürek, 2005; also discussed in McNeill, 2005). Future research in these domains will shed light on the dynamic intertwining of speech and gesture.

References

Butterworth, B., & Hadar, U. (1989). Gesture, speech, and computational stages: A reply to McNeill. *Psychological Review, 96,* 168-174.

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.

Garrett, M. F. (1982). Production of speech: Observations from normal and pathological language use. In A. W. Ellis (Ed.), *Normality and pathology in cognitive functions* (pp.19-76). London: Academic Press.

Goldberg, A.E. (1997). The relationships between verbs and constructions. In M. Verspoor, K. D. Lee, & E. Sweetser (Eds.), *Lexical and syntactical constructions and the construction of meaning* (pp.383-398). Amsterdam: John Benjamins.

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

Kita, S. (1993). *Language and thought interface: A study of spontaneous gestures and Japanese mimetics.* Unpublished doctoral dissertation. University of Chicago.

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

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.

Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., & Ishizuka, T. (submitted). Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Manuscript submitted to *Language and Cognitive Processes.*

Krauss, R.M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In M. Zanna (Ed.), *Advances in experimental social psychology, 28* (pp.389-450). Tampa: Academic Press.

Krauss, R.M., Chen, Y., & Gottesman, R.F. (2000). Lexical gestures and lexical access: A process model. In D. McNeill (Ed.), *Language and gesture* (pp.261-283). Cambridge: Cambridge University Press.

Levelt, W.J.M. (1989). *Speaking.* Cambridge, MA: MIT Press.

McCullough, K.E. (1993). Spatial information and cohesion in the gesticulation of English and Chinese speakers. A paper presented at the annual meeting of American Psychological Society. Chicago.

McNeill, D. (1985). So you think gestures are non-verbal? *Psychological Review, 92,* 350-371.

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

McNeill, D. (2000). Analogic/analytic representations and cross-linguistic differences in thinking for speaking. *Cognitive Linguistics, 11,* 43-60.

McNeill, D., & Duncan, S. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and gesture* (pp.141-161). Cambridge: Cambridge University Press.

McNeill, D. (2005). *Gesture and thought.* Chicago: University of Chicago Press.

Özyürek, A. (2002a). Speech-gesture synchrony in typologically different languages and second language acquisition. In B. Skarabela, S. Fish & A. H. J. Do (Eds.), *Proceedings of the 26th annual Boston University conference on language development* (pp.500-509). Somerville, MA: Cascadilla Press.

Özyürek, A. (2002b). 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.

Özyürek, A., & Kita, S. (1999). Expressing manner and path in English and Turkish: Differences in speech, gesture, and conceptualization. In M. Hahn & S. C. Stoness (Eds.), *Proceedings of the twenty-first annual conference of the Cognitive Science Society* (pp.507-512). Mahwah, NJ: Lawrence Erlbaum.

Özyürek, A. (2005). What do speech-gesture mismatches reveal about speech and gesture integration?: A comparison of Turkish and English. In C. Cang, M. Houser, Y. Kim, D. Mortensen, M. Park-Doob, M. Toosarvandani (Eds). *Proceedings of the 27th meeting of the Berkeley Linguistics Society* (pp.449-456) Berkeley, CA: Berkeley Linguistics Society.

Özyürek, A., Kita, S., Allen, S., Furman, R., & Brown, A. (2005). How does linguistic framing of events influence co-speech gestures? Insights from cross-linguistic variations and similarities. *Gesture, 5(1), 215-237.*

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 meeting* (pp. 435-445), Berkeley, CA: Berkeley Linguistics Society.

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.

Talmy, L. (1985). Semantics and syntax of motion. In T. Shopen (Ed.), *Language typology and syntactic description, Vol.3, Grammatical categories and the lexicon* (pp.57-149). Cambridge: Cambridge University Press.

Vigliocco, G. & Kita, S (2006). Language-specific properties of the lexicon: Implications for learning and processing. *Language and Cognitive Processes, 21*(7-8), 790-816.