# 11

# Processing of Multi Modal Semantic Information: Insights from Crosslinguistic Comparisons and Neurophysiological Recordings

Aslı Özyürek

## 1. Introduction

Face-to-face communication involves continuous coordination and processing of information across modalities such as from speech, lips, facial expressions, eye gaze, hand gestures etc. Previous studies investigating multi modal processing during communication have focused mostly on the relationship between lip movements and speech (e.g., McGurk effect, Calvert, 2001). However, during everyday face-to-face communication, we almost also use and view meaningful hand movements, i.e., *gestures*, along with speech. Although both gestures and lip movements are examples of the natural co-occurrence of auditory and visual information during communication, they are fundamentally different with respect to their relationship to the speech they accompany. Whereas there is a clear one-to-one overlap of speech sounds and lip movements in terms of their form, the mapping between the forms of gesture and speech is different (see McNeill, 1992). Consider for example an upward hand movement in a climbing manner when a speaker says: "He climbed up the ladder". Here, the gesture might depict the event as a whole, describing the figure (crawled hands representing as that of the agent, 'he'), manner ('climb') and direction ('up') simultaneously. In speech, however, the message unfolds over time, broken up into smaller meaningful segments (i.e. words).

Because of such differences, the mapping of speech and gesture information has to happen at a higher, semantic level.

In this paper I will address the question of what are the mechanisms that underlie processing of such high level multi modal semantic information, specifically conveyed through speech and hand gestures both during production and comprehension of utterances.

Research on gestures that people produce while speaking has identified different types of gestures (see McNeill, 1992; Kendon, 2004 for a review). Some of the hand gestures that speakers use, such as emblems, are highly conventionalized and meaningful even in the absence of speech (e.g., a thumbs up gesture for O.K.). Some others, such as pointing gestures are meaningful in the context of both the speech and the extra linguistic context of the utterance that the point is directed to (e.g., pointing to a lamp and say " turn on that lamp"). However, others are less conventionalized, do not necessarily need extra linguistic context for their use and represent meaning by their resemblance to different aspects of the events they depict (e.g., wiggling fingers crossing space to represent someone walking). The latter have been called *iconic* or *representational* gestures in the literature and how they are processed in relation to speech both during production and comprehension of utterances is the topic of this talk.

## 2. Previous studies on production and comprehension of iconic gestures in relation to speech

Previous work by McNeill (1985, 1992) has shown that gestures reveal speakers' imagistic representations during speaking. For example, a circular hand gesture representing the shape of a table, which accompanies the speech referring to the table, provides information about the speaker's mental image of the table at the moment of speaking. Due to differences in modality, iconic gestures reveal information in a different schema than verbal expressions. Gestures represent meaning as a whole, not as a construction made out of separate meaningful components as in speech.

However, although gestures reveal the information in a different representa-

tional format than speech, the two modalities are systematically related to each other and convey the speaker's meaning together as a "composite signal" (Clark, 1996). This unified meaning representation is achieved by semantic relatedness and temporal congruity between speech and gesture (McNeill, 1992). First of all, there is *semantic overlap* between the representation in gesture and the meaning expressed in the concurrent speech, although gesture usually also encodes additional information that is not expressed in speech. Consider the example of a narrator telling an animated cartoon story. In the relevant scene, a cat that has swallowed a bowling ball rolls down the street into a bowling alley from left to the right on the TV screen. The narrator describes this scene with the sentence "the cat rolls down the street," accompanied by a hand gesture consisting of the hand moving from left to right while the fingers wiggle repetitively. In this example, a single gesture exhibits simultaneously the manner, the change of location, and the direction of the movement to the right. Speech expresses the manner and the path of the movement, but not the direction. Thus there is informational overlap between speech and gesture, but also additional information in the gesture (McCullough, 1993; Kita & Özyürek, 2003).

The second systematic relationship between speech and gestures is *temporal*. A gesture phrase has three phases: preparation, stroke (semantically the most meaningful part of the gesture), and retraction or hold (McNeill, 1992). All three phases together constitute a gesture phrase. McNeill (1992) has shown that in 90% of speech-gesture pairs, the stroke coincides with the relevant speech segment, which might be a single lexical item or a phrase. For example the stroke phase of the *climb up* gesture exemplified above is very likely to occur during the bracketed part of the following utterance "he [climbed up] the ladder".

Thus research has shown that, at least at the surface level, there is semantic and temporal coordination in the production of semantic information in the two channels during communication. What about comprehension of such gestures during in relation to speech? Is there evidence that listeners pay attention to and comprehend the information coming from gestures?

A line of studies has investigated whether listeners use information from gestures in their comprehension of speaker's overall message. For example Graham and Argyle (1975) had speakers describe abstract line drawings with and without

gestures, and required listeners to make drawings on the basis of the speakers' input. Listeners were more accurate in their drawings in the speech-and-gesture condition than in the speech-alone condition. In another study, Beattie and Shovelton (1999) have shown that questions about the size and relative position of objects in a speaker's message are better answered by listeners when gestures are part of the description than when gestures are absent.

Another set of studies has investigated whether listeners pick up the information in gesture when gesture conveys different information than speech. McNeill and colleagues (1999) presented listeners with a videotaped narrative in which the semantic relationship between speech and gesture was manipulated, and then asked them to retell the narrative they heard. In the videotaped narrative, an actor told a story about an animated cartoon, and the story contained utterances with three types of speech and gesture relationship. In one type, gesture conveyed contradictory information in a way that never occurs in natural communication. For example, the narrator said, "he is running ahead of it," while making a climbing gesture with the hands. In the second type of utterance, gestures conveyed complementary information to speech. For instance, the speaker said, "he went across the street," while making a running gesture along a horizontal path. Finally, in the third type of utterance, gesture conveyed information compatible with speech. The results showed that listeners do attend to the information conveyed in gesture when that information complements or even contradicts the information conveyed in speech, since they incorporated information from the gestures in their retellings of the narratives. Additional support for this claim comes from research that found that adults assess children's answers to a Piagetian conservation task by taking into account children's both speech and gestures, when gestures convey different information than their speech (e.g., Goldin-Meadow & Sanhofer, 1999).

While the above studies have investigated whether listeners pick up information from gestures in the presence of both speech and gesture, other studies have investigated how much information can be gleaned from gestures in the *absence* of speech. These studies have shown that some meaning is extracted from gesture when viewed without speech, but this is limited, and that there is variability in the meaning attributed to gesture. For example, Feyereisen et al. (1988) showed subjects videotaped gestures (without sound) excerpted from classroom lectures and

asked them to choose one of the three possible interpretations; a) the words used in the accompanying speech (i.e., lexical affiliate), b) the meaning most frequently attributed to the gesture by an independent group of judges, and c) a random meaning. The results showed that viewers preferred the meaning most frequently attributed to the gesture by an independent group of judges, rather than the original words used in the accompanying speech. Feyereisen et al. concluded that iconic gestures, in the absence of speech, convey imprecise meaning to listeners/viewers and not necessarily what the speaker intended together with his speech. Thus, the true "iconicity" of such gestures is debatable and that they seem to convey their true meaning *only* viewed with speech.

In summary, the comprehension studies show that there is evidence that listeners pay attention to gestures, glean information from them, use this information to understand speaker's intended message, and the comprehension of such gestures is integrated together with speech.

## 3. Two opposing models of speech and gesture processing

Even though the speech and gesture seems tightly coordinated according to behavioral measures, there is controversy in their literature with regard to their underlying interaction during the production and comprehension processes. Many researchers accept that such gestures are generated from spatial and motoric representations of events, objects etc. However, there are two opposing views with regard to their further processing in relation to speech. According to some views (Krauss, 2000; Feyreisen & Lanoy, 2001) speech and gesture are processed *independently* but in a *parallel* fashion (i.e., that explains their overt coordination at the behavior level). For example according to Krauss, gestures are generated only from the spatial and representations, "prelinguistically", and independent from how certain information is linguistically formulated. One of the functions of gestures is to keep memories of such representations active and facilitate lexical retrieval through cross-modal priming. However how information is semantically or grammatically encoded for example would not change the representational format of such gestures.

However, according to other views (de Ruiter , 2000; Mayberry, 2000; McNeill, 1992; Kita &, Özyürek 2003, Özyürek, 2002) there is *interaction* between the production of two systems either at the conceptual, or grammatical encoding level of speech production process–even though there is further controversy with regard to which level the interaction occurs and to what extent among the latter set of researchers. According to such views how elements of events are encoded semantically or grammatically during online speaking will change the representational format of gestures.

Even though most models have been proposed for production but not comprehension, the two models also have different views about how listeners/viewers process information from both modalities. The *independent* models claim that gesture is used–if ever–as "add-on" information during comprehension and only after speech has been processed. However, *interaction* models claim that the integration of speech and gesture is early and done simultaneously as part of one process during comprehension, rather than as two separate sequential processes.

Below I report two studies that provide evidence for the fact that speech and gesture processing interact during both production and comprehension of utterances, arguing against the independent, sequential models of processing.

## 4. Study 1: Speech and gesture interaction during production process

In this study the hypothesis that whether gestures of the same motion event would differ according the language specific encoding of semantic and grammatical encoding of spatial information depicted was tested. The *independence* models would predict that the way certain elements of an event are encoded linguistically will not change the form of gestures, since gestures are merely generated from spatial representations. However according to *interaction* models, the linguistic encoding of the event would change the shape of gestures, due to an interaction at the conceptual level of encoding for speaking.

The cross-linguistic variation in gestural representation was demonstrated by comparing how Japanese, Turkish, and English speakers verbally and gesturally

express motion events, which were presented as a part of an animated cartoon (Kita & Özyürek, 2003; Özyürek & Kita, 1999). Japanese and Turkish differed from English typologically which allowed us to look whether and how gestures of the same event differed due to linguistic encoding possibilities among the speakers of these languages.

Two analyses were carried out. The first analysis concerned an event in which 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 encodes the arc shape of the trajectory, and Japanese and Turkish speakers used verbs such as *go*, which does not encode the arc trajectory. In their conceptual planning phase of the utterance describing this event, Japanese and Turkish speakers presumably got feedback from speech formulation processes and created 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 just gestures with an arc trajectory (Kita, 1993, 2000; 2002; Kita & Özyürek, 2003).

The second analysis concerned how speech and gesture express the Manner and Path of an event in which the protagonist rolled down a hill. It was found that verbal descriptions differed cross-linguistically in terms of how manner and path information is lexicalized along the lines discussed by Talmy (1985). English speakers used a Manner verb and a Path particle or preposition (e.g., *he rolled down the hill*) to express the two pieces information within one clause. In contrast, Japanese and Turkish speakers separated Manner and Path expressions over two clauses, path as in the main clause and manner as in the subordinated clause (e.g., *he descended as he rolled*). Given the assumption that a clause approximates a unit of processing in speech production (Bock, 1992; Garrett, 1982; Levelt, 1989), presumably English speakers were likely to process both Manner and Path within a single processing unit, whereas Japanese and Turkish speakers were likely to need two processing units. Consequently, Japanese and Turkish speakers should be more likely to separate the images of Manner and Path in preparation for speaking so that two pieces of information could be dealt with in turn, as compared to Eng-

lish speakers. The gesture data confirmed this prediction (Özyürek & Kita, 1999; Kita & Özyürek, 2003). In depicting how an animated figure rolled down a hill having swallowed a bowling ball in the cartoon, Japanese and Turkish speakers were more likely to use separate gestures, one for manner and one for path and English speakers were more likely to use just one gesture to express both manner and path.

These findings were further replicated in a recent study where Turkish and English speakers were asked to talk about 10 different motion events that involved different types of manner (jump, roll, spin, rotate) and path (descend, ascend, go around ). Furthermore in cases where only manner or only path was expressed in an utterance, speakers of both languages were more likely to express congruent information in gesture to what is expressed with speech (e.g., he went down the slope: Gesture: index finger moving down expressing *just* the path information) (Özyürek et al, 2005).

In addition to the cross-linguistic variation in gestural representation, it was found that gestures encoded certain spatial details of motion events that were never verbalized due to modality. For example, in the description of the above two motion events, none of the participants in any of the languages verbally encoded whether the motion was to the right or to the left, but this information was reflected in the direction of the gestures very accurately (Kita & Özyürek, 2003). This evidence shows that there are also aspects of representation in gesture that may not have interacted with the linguistic conceptualization of the event but are derived from the spatial representation of the event.

These findings are line with the view that the representations underlying a gesture is shaped *simultaneously* by 1) how information is organized according to easily accessible linguistic expression in a given language and at the moment of speaking and 2) the spatio-motoric properties of the referent which may or may not be verbally expressed. These findings are counter evidence for the models that argue that the only source that shapes gestural information is spatial representations independent of linguistic conceptualization for speaking.

## 5. Study 2: Neural correlates as evidence for speech and gesture integration during comprehension

The second study investigates the neural locus (by using fMRI) of integration of information from speech and gesture during comprehension as well as the time course of this process using ERP measures.

*fMRI study*: fMRI was used (N=12) to identify brain regions activated during observation of iconic gestures which naturally accompany speech (Willems, Özyürek, Hagoort, 2005; in press).

We manipulated the semantic fit of a verb (language) or of an iconic gesture (action) to the preceeding sentence context. Integration load was expected to vary with this manipulation, thereby showing regions specific for speech and gesture processing as well as areas common to the integration of both information types into the prior sentence context.

Analysis of both gesture and speech mismatch versus correct conditions showed overlapping areas for both comparisons in the left inferior frontal gyrus (LIFG), corresponding to Brodmann area (BA) 45. That is, gesture mismatches as well as language mismatches recruited LIFG showing common areas of processing of semantic information from both modalities. Intraparietal and superior temporal regions also showed gesture and language specific responses.

Gesture-mismatch activating similar areas as those of language mismatch are in line within a neurobiological theory of language, according to which 'Broca's complex' (including BA 47, 45, 44 and the ventral part of BA 6) in the left frontal cortex, serves as a unification space for language comprehension, in which lexical information retrieved from memory (i. e. from the mental lexicon) is integrated into a unified representation of a multi-word utterance, such as a sentence (Hagoort, 2003, Hagoort, in press; Hagoort et al., 2004). The current findings further suggest that integration of semantic information from linguistic elements as well as from both language and gesture share similar processes during comprehension.

*ERP study*: To investigate further the time course of integration of semantic information from gesture versus language, we gave subjects similar pairings as in the fMRI study. That is, either gesture or speech mismatched to previous sentence context. Electrophysiological recordings were measured, time-locked to the begin-

ning of the critical verb and stroke of gesture which were presented simultaneously. The results showed similar N400 effects for both language as well as gesture mismatches. These results further show that the integration of speech and gesture into previous context of the utterance is early and similtaneous and at the same time providing evidence against *independent* and sequential models of speech and gesture comprehension processes (Özyürek et al, in press).

## 6. Conclusion

Both the results of the production and the comprehension studies reported above suggest that multi modal semantic information, specifically from speech and gesture, is processed in an interactive way, simultaneously and in similar parts of the brain rather than being processed in a distinct modular fashion. Further research is necessary to delineate the exact level where these cross modal semantic interaction processes take place both during the production and comprehension of utterances and their gestural accompaniments.

### References

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.

Beattie, G., & Shovelton, H. (2002). An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British Journal of Psychology*, 93, 179–192.

Butterworth. B., & and Beattie, G. (1978). Gesture and silence as indicators of planning in speech. In: N. Campbell & P. T. Smith (Eds.). *Recent advances in the psychology of language: formal and experimental approaches* (347–360). NewYork: Plenum.

Calvert, A. (2001). Crossmodal processing in the human brain: Insights from functional neuroimaging studies. *Cerebral Cortex*, 11; 1110–1123.

Clark, H. (1996). *Using language*. Cambridge: CUP

Feyereisen, P., van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: What can be understood without speech? *Cahiers de Psychologie Cognitive*, 8, 3–25.

Feyreisen, P. & Lanoy, J. D. (1991). *Gestures and speech: Psychological investigations.* Cambridge: CUP

Furuyama, N. (2000). Gestural interaction between the instructor and the learner in origami instruction. In McNeill (ed.). *Language and Gesture.* 99–118. Cambridge: CUP

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, 10, 57–67.

Goldin-Meadow, S., & Sanhofer, C. M. (1999). Gesture conveys substantive information about a childs' thoughts to ordinary listeners. *Developmental Science*, 2, 64–74.

Hagoort, P. (2003). How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage*, 20, S18–S29.

Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K. M. (2004), Integration of word meaning and world knowledge in language comprehension. *Science*, 304, 438–441.

Kendon, A. (2004). *Gesture.* Cambridge: University of Cambridge Press

Kita, S. (2000). How representational gestures help speaking. In McNeill (ed.). *Language and Gesture.* pp. 162–186. Cambridge: CUP

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

Krauss, R. M. , Chen, Y. & Gottesman, R. (2000). Lexical gestures and lexical access: A process model. In McNeill (ed.). *Language and Gesture.* 261–284. Cambridge: CUP

Levelt, P. (1989). *Speaking.* MIT Press.

Mayberry, R. & Jaques, J. (2000). Gesture productino during stuttered speech: insights into the nature of speech-gesture integration. In McNeill (ed.). *Language and Gesture.* 199–215. Cambridge: CUP

McNeill, D. (1985). So you think gestures are nonverbal? *Psychological Review*, 92, 350–371.

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

McNeill, D., Cassell, & McCullough, K-E. (1999). Communicative effects of speech- mismatched gestures. *Research on Language and Social Interaction*, 27, 223–238.

De Ruiter, J. P. (2000). The production of gesture and speech. In McNeill (ed.). *Language and Gesture.* 284–312. Cambridge: CUP

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

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* (507–512). Mahwah, NJ: Lawrence Erlbaum.

Özyürek, A., Kita, S., Allen S., Furman, R., Brown, A. (2005). How does linguistic framing influence co-speech gestures? Insights from cross-linguistic differences and similarities. *Gesture* 5, 216–241

Özyürek, A., Willems, R., Kita. S,. & Hagoort, P. (in press). On-line integration of information from speech and gesture: Insight from event-related potentials. *Journal of Cognitive Neuroscience.*

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

Willems, R., Özyürek, A, Hagoort, P. (2005). When language meets action: A first fMRI study on comprehension of gesture and speech. *Poster for 2005 Cognitive Neuroscience Conference, NY*

Willems, R., Özyürek, A., Hagoort, P. (in press). When language meets action. The neural integration of speech and gesture. *Cerebral Cortex.*