The Role of Iconic Gestures in Production and Comprehension of Language: Evidence from Brain and Behavior

Asli Özyürek

Radboud University Nijmegen, Center for Language Studies &
Max Planck Institute for Psycholinguistics

Abstract. Speakers in all cultures and ages use gestures as they speak (i.e., cospeech gestures). There have been different views in the literature with regard to whether and how a specific type of gestures speakers use, i.e., iconic gestures, interacts with language processing. Here I review evidence showing that iconic gestures are not produced merely from the spatial and/or motoric imagery but from an in interface representation of imagistic and linguistic representation during online speaking Similarly, for comprehension, neuroimaging and behavioral studies indicate that speech and gesture influences semantic processing of each other during online comprehension. These findings show overall that processing of information in both modalities interacts during both comprehension and production of language arguing against models that propose independent processing of each modality. They also have implications for AI models that aim to simulate cospeech gesture use in conversational agents.

Keywords: iconic, cospeech gesture, interface, production, comprehension, brain, behavior.

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, [1]). However, during everyday face-to-face communication, we almost always 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 [2]. 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

S. Kopp and I. Wachsmuth (Eds.): GW 2009, LNAI 5934, pp. 1–10, 2010.

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.

Speakers use gestures at all ages (starting from around 9 months) and cultures. The use of gesture is so robust in human communication that it is visible in people blind from birth, when people talk on the phone –albeit less than during face-to face communication [3]- and can be found in sign languages where the same modality is used for both sign and gesture (see for a review [4]).

Research on gestures that people produce while speaking has identified different types of gestures [2],[5]. 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, represent meaning by their resemblance to different aspects of the events they depict (e.g., wiggling fingers crossing space to represent someone walking) and rely more on speech for their meaning. 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 paper.

It is important to note here that previous research has shown evidence both of speaker-oriented (cognition centered) as well as addressee-oriented (context centered) factors in shaping gestures and their relation to speech. Here I will review speaker-oriented evidence to explain the interactions between speech and gesture-without denying that social context or communicative intention to convey a message designed for the addressee are also additional factors that shape iconic gesture production (e.g., [3], [5], [14]) and are needed for a full account of speech and gesture production and comprehension.

2 Previous Studies on Relations between Gesture and Speech

Previous work by McNeill ([6],[2]) has shown that iconic 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 representational format than speech, the two modalities are systematically related to each other and convey the speaker's meaning together as a "composite signal" [7]. This unified meaning representation is achieved by semantic relatedness and temporal congruity

between speech and gesture [2]. 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 [8].

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 [2]. All three phases together constitute a gesture phrase. McNeill [2] has also 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. The question I address here it whether two streams of communication interact and are integrated during the language production and comprehension process or alternatively can be conceived as two independent but parallel streams of communication. Most studies and models of gesture processing have been designed for production but less is known about the interaction processes between the two for comprehension. The purpose of this paper is then to review recent evidence showing that speech and gesture interact during both production (section 4.1) and comprehension (section 4.2) of language. Before that I briefly outline some competing views proposed about the relations between speech and gesture during processing in section 3.

3 Models of Speech and Gesture Processing: Competing Views

Even though the speech and gesture seem tightly coordinated according to behavioral measures, there is controversy in their literature with regard to their underlying interaction during the production and comprehension processes.

According to some views ([9],[10],[11]) speech and gesture are processed independently and in a parallel fashion (i.e., that explains their overt coordination at the behavior level). According to these view gestures are generated and processed directly and solely from the spatial and motoric representations, whereas speech is generated from propositional representations and without interactions between the two during the production process. For example according to Krauss [10], gestures are generated from spatial 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 (from gesture to speech). However how information is semantically or grammatically encoded for example would not change the representational format of such gestures. Also according to a new framework, Gesture as Simulated Action (GSA) [9], gestures arise simply out of simulations of actions and do not interact with the language production process.

However, according to other views ([12],[13],[2],[8],[14]) 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.

Even though most models have been proposed for production but not comprehension, the existing production models also have different views for how listeners/viewers might comprehend information from both modalities. The independence models claim that gesture is used–if ever–as "add-on" information during comprehension and only after speech has been processed [15]). However, interaction models [16] claim that there are mutual, simultaneous and even obligatory interactions between processing of speech and gesture during comprehension.

Below I review studies from my own collaborative work that provide evidence for the fact that speech and gesture processing interact during *both* in production and comprehension of utterances, arguing against the independent and sequential models of processing.

4 Evidence for Interactions between Speech and Gesture

4.1 Production

As a first step to test whether speech and gesture processing interacts during production we investigated whether gestures of the same motion event would differ according the language- specific semantic and grammatical encoding of spatial information in different languages 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 generated from and shaped solely by spatial representations (i.e., which would be similar across speakers of different languages). However according to interaction models (i.e., specifically the Interface Model [8], the linguistic encoding of the event would change the shape of gestures, due to an interaction between linguistically formulating the message (i.e., specific for requirements of each language) and the formation of the gesture during online production.

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 ([8], [17]). 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 ([18], [8]).

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 [19]. 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 ([20], [21]), 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 English speakers. The gesture data confirmed this prediction ([17], [8]). 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) [22].

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 [8].

These findings are line with the view (i.e., Interface Hypothesis, [8]) 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.

However one concern regarding the above studies was that different gestures produced by speakers of different languages could have still originated from spatial representations that are shaped differently due to difference cultural ways of thinking or habitually using language in a certain way-i.e., in line with Whorfian Hypothesis [23]. If this were the case the difference in gestures across speakers of different languages would not be evidence for the online interaction between gesture and language production processes but rather gestures could still be considered to be originated and shaped solely by the spatial representations (i.e., shaped in language-specific ways a priori to the encoding of each message). To clear out which of these processes could be responsible for our initial findings about gestural differences across languages, we asked English speakers to describe motion events using different syntactic frames –one in which manner and path expressed in one verbal clause (i.e., roll down) and one where manner and path are in separate clauses (i.e., went down the hill rolling) (less preferred but not ungrammatical for English speakers). We found that English speakers gestures changed with the syntactic frames they chose reflecting differences in the same way we found between English and Turkish speakers' gestures [24]. These findings rule out the possibility that spatial gestures are generated from language- or culture-specific spatial representations prior to the online linguistic formulation of the event. If the former were the case, we would have expected English speakers to use also conflated gestures when they used the less preferred syntactic frame –but instead they used segmented gestures as Turkish speakers. This finding provided further evidence that iconic gestures are shaped by speaker's online syntactic choices rather than a priori by habitual language-specific representations.

4.2 Comprehension

Neural Evidence: If speech and gestures are two interacting systems of communication in comprehension as well as in production then we expect speech and gesture processing to use similar neural resources during comprehension. Even though previous research has shown that listeners/ viewers pay attention to gestures and pick up information from gestures [25], only recently researches have begun to investigate the interactions between speech and gesture during language comprehension. In two studies we investigated the neural correlates of speech and gesture comprehension.

One of these studies used an ERP (event related potentials) technique, which measured electrophysiological responses to events by electrodes attached to the scalp as listeners/viewers listened sentences and saw accompanying gestures. In the sentence-gesture pairs we manipulated the semantic fit of a verb or of a temporally overlapping iconic gesture to the preceding sentence context (see Table 1). In the control condition both a critical verb and accompanying gesture fitted semantically to the previous sentence context. In the experimental conditions either speech or gesture or both did not fit semantically to the previous context. Recordings were measured, time-locked to the beginning of the critical verb and stroke of gesture, which were presented simultaneously. The results showed similar N400 effects (showing

Table 1. Examples from speech gesture pairs used in [26]

Control condition (speech and gesture match to previous context)
(1) He slips on the roof and [rolls down]
G: ROLL DOWN
Experimental conditions (speech and/or gesture (in bold) mismatch to previous context)
(2) He slips on the roof and **[writes]** a note (speech mismatch only)
G: ROLL DOWN
(3) He slips on the roof and [rolls down] (gesture mismatch only)
G: WRITE
(4) He slips on the roof and **[writes]** a note (speech and gesture mismatch)
G: WRITE

detection of semantic unfit) for sentences where either language or gesture did not fit semantically to the previous context. These results show that the information form both speech and gesture are integrated to previous context of the utterance at the same time providing evidence against independent and sequential models of speech and gesture comprehension processes [26]. Note that if gesture was processed after the verb or vice versa we would have expected either speech or gesture anomaly to be detected later than 400 ms but we did not.

In the second study we used fMRI technique to identify brain regions activated during understanding iconic gestures versus verbs in a sentence context using the same stimuli (Table 1) in the ERP study above. Integration load was expected to vary with this manipulation due to the increased load of semantic integration, 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, (LIPC) corresponding to Brodmann area (BA) 45. That is, gesture mismatches as well as speech mismatches recruited LIPC showing common areas of processing of semantic information from both modalities. Intraparietal and superior temporal regions also showed gesture and language specific responses respectively for mismatches than matches [27].

Gesture-mismatch activating similar areas as those of language mismatch are in line within a neurobiological theory of language, '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 ([28], [29]). 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.

Behavioral Evidence: Thus both the ERP and the fMRI measurements show that the brain comprehends speech and gesture in relation to a previous sentence context in similar ways; both are processed as semantically, using similar time course and neural correlates. However these studies do not directly show whether the semantic processing of each modality interacts with the other. Thus in a third study we investigated this possibility in a behavioral experiment [16]. We asked whether listeners/viewers do process the meaning of speech and gesture separately or whether the meaning of one interacts with processing the meaning of the other during comprehension. We presented participants with action primes (someone chopping vegetables) and bi-modal speech and gesture targets. Participants were faster and more accurate to relate primes to targets that contained congruent (Speech: "CHOP"; gesture: CHOP) versus incongruent information (Speech: "CHOP"; gesture: TWIST). Moreover, the strength of the incongruence affected processing, with fewer errors for weak (Speech: "CHOP"; gesture: CUT) versus strong incongruities (Speech: "CHOP"; gesture: TWIST). Furthermore, this influence was bi-directional. A follow up study demonstrated that gesture's influence on speech was obligatory. That is even though subjects were asked only to decide whether the verb followed an action prime matched to the prime, whether gesture was congruent or incongruent to the accompanying verb influenced subjects responses. These results show that listeners/viewers process the meaning of one modality in relation to the meaning of the other rather than processing each in an independent manner.

5 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 -at both semantic and syntactic levels for production and semantic for comprehension- and recruiting similar neural correlates brain rather than being processed in a distinct modular fashion. It is important to note here that the model proposed by Interface Hypothesis [8] for production is also successfully implemented in AI models that try to simulate iconic gesture production in conversational agents [30]. In the future it would be useful to see whether AI models can be also extended to comprehension which simulates the interaction between the two modalities as proposed in Integrated Systems Hypothesis [16].

Further research is necessary to delineate the exact level where these cross modal semantic interaction processes take place during processing as well as the role of communicative intentions of the speakers in gesture processing and to situate gesture production and comprehension in a larger interactional-situational context than we have done so far.

Acknowledgements. The research reviewed here was supported by Netherlands Organization for Scientific Research (NWO), 051.02.040; US National Science Foundation (NSF) and the Max Planck Institute for Psycholinguistics, Netherlands.

References

1. Calvert, A.: Crossmodal processing in the human brain: Insights from functional neuroimaging studies. Cerebral Cortex 11, 1110–1123 (2001)
2. McNeill, D.: Hand and mind. University of Chicago Press, Chicago (1992)
3. Bavelas, J.B., Gerwing, J., Sutton, C., Prevost, D.: Gesturing on the telephone: Independent effects of dialogue and visibility. Journal of Memory and Language 58, 495–520 (2008)
4. Özyürek, A.: Gesture in sign and spoken language. In: Pfau, R., Steinbach, M., Woll, B. (eds.) Sign language: An international handbook. Mouton, Berlin (in press)
5. Kendon, A.: Gesture. Cambridge University Press, Cambridge (2004)
6. McNeill, D.: So you think gestures are nonverbal? Psychological Review 92, 350–371 (1985)
7. Clark, H.: Using language. Cambridge University Press, Cambridge (1996)
8. Kita, S., Özyürek, A.: 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 (2003)
9. Hostetter, A.B., Alibali, M.W.: Visible embodiment: Gestures as simulated action. Psychonomic Bulletin and Review (2008)
10. Krauss, R.M., Chen, Y., Gottesman, R.: Lexical gestures and lexical access: A process model. In: McNeill, D. (ed.) Language and Gesture, pp. 261–284. Cambridge University Press, Cambridge (2000)
11. Feyreisen, P., Lanoy, J.D.: Gestures and speech: Psychological investigations. Cambridge University Press, Cambridge (1991)
12. De Ruiter, J.P.: The production of gesture and speech. In: McNeill, D. (ed.) Language and Gesture, pp. 284–312. Cambridge University Press, Cambridge (2000)
13. Mayberry, R., Jaques, J.: Gesture productino during stuttered speech: insights into the nature of speech-gesture integration. In: McNeill, D. (ed.) Language and Gesture, pp. 199–215. Cambridge University Press, Cambridge (2000)
14. Özyürek, A.: 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 (2002)
15. Krauss, R.M., Morrel-Samuels, P., Colasante, C.: Do conversational hand gestures communicate? Journal of Personality and Social Psychology 61, 743–754 (1991)
16. Kelly, S., Özyürek, A., Maris, E.: Two sides of the same coin: Speech and gesture interact to enhance comprehension. Psychological Science (in press)
17. Özyürek, A., Kita, S.: Expressing manner and path in English and Turkish: Differences in speech, gesture, and conceptualization. In: Hahn, M., Stoness, S.C. (eds.) Proceedings of the twenty first annual conference of the Cognitive Science Society, pp. 507–512. Lawrence Erlbaum, Mahwah (1999)
18. Kita, S.: How representational gestures help speaking. In: McNeill, D. (ed.) Language and Gesture, pp. 261–284. Cambridge University Press, Cambridge (2000)
19. Talmy, L.: Semantics and syntax of motion. In: Shopen, T. (ed.) Language typology and syntactic description. Grammatical categories and the lexicon, vol. 3, pp. 57–149. Cambridge University Press, Cambridge (1985)
20. Bock, K.: Towards a cognitive psychology of syntax: Information processing contributions to sentence formulation. Psychological Review 89, 1–47 (1982)
21. Levelt, P.: Speaking. MIT Press, Cambridge (1989)

22. Özyürek, A., Kita, S., Allen, S., Furman, R., Brown, A.: How does linguistic framing influence co-speech gestures? Insights from cross-linguistic differences and similarities. Gesture 5, 216–241 (2005)
23. Pederson, E., Danziger, E., Wilkins, D., Levinson, S.C., Kita, S., Senft, G.: Semantic typology and spatial conceptualization. Language 74, 557–589 (1998)
24. Kita, S., Özyürek, A., Allen, S., Brown, A., Furman, R., Ishizuka, T.: Relations between syntactic encoding and co-speech gestures: Implications for a model of speech and gesture production. Language and Cognitive Processes 22(8), 1212–1236 (2007)
25. Beattie, G., Shovelton, H.: Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. Semiotica 123, 1 (1999)
26. Özyürek, A., Willems, R., Kita, S., Hagoort, P.: On-line integration of information from speech and gesture: Insight from event-related potentials. Journal of Cognitive Neuroscience 19(4), 605–616 (2007)
27. Willems, R., Özyürek, A., Hagoort, P.: When language meets action. The neural integration of speech and gesture. Cerebral Cortex 17, 2322–2333 (2007)
28. Hagoort, P.: How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. Neuroimage 20, S18–S29 (2003)
29. Hagoort, P., Hald, L., Bastiaansen, M., Petersson, K.M.: Integration of word meaning and world knowledge in language comprehension. Science 304, 438–441 (2004)
30. Kopp, S., Bergmann, K., Wachsmuth, I.: Multimodal communication from multimodal thinking – Towards an integrated model of speech and gesture production. Semantic Computing 2(1), 115–136 (2008)