Integrating Gesture Meaning and Verbal Meaning for German Verbs of Motion: Theory and Simulation

Kirsten Bergmann*,, Florian Hahn*, Stefan Kopp*,**, Hannes Rieser*, Insa Röpke***[1]
*Collaborative Research Center "Alignment in Communication" (CRC 673)
**Center of Excellence "Cognitive Interaction Technology" (CITEC)
P.O. Box 100 131, 33501 Bielefeld, Germany

Abstract

When verbs of motion are accompanied by gestures, this comes along with a relatively complex relation between the two modalities. In this paper, we investigate the semantic coordination of speech and event-related gestures in an interdisciplinary way. First, we explain how to efficiently construct a speech-gesture-interface for a gesture which accompanies a verb phrase from a theoretical viewpoint. Resting upon this analysis, we also provide a computational simulation model which further explicates the relation between the two modalities based on activation-spreading within dynamically shaped multi-modal memories.

Keywords: Gesture semantics; Event-related gestures; Iconic gestures; Speech-gesture interface; Theoretical reconstruction; Computational simulation; Interdisciplinary methodology

Introduction

The description of actions, motions, or events often comes along with speakers' use of co-speech gestures. Event-related gestures take place, e.g., when giving route descriptions (e.g., Allen (2003)), when describing motion in space (e.g., Kita and Özyürek (2003)), or when simulating an activity pantomimically (e.g., Müller (1998)). In contrast to shape-depicting or localizing gestures referring to static objects, event-related gestures pose even greater challenges for gesture research – basically due to the more complex relation of event-related gestures with accompanying speech. As an illustration, consider the example depicted in Figure 1 from a route description dialogue in which an iconic gesture accompanies a verb of motion. The utterance occurs while one speaker describes how to walk through a park passing a pond. The speaker utters "Gehst quasi drei Viertel um den Teich herum" (engl.: "(You) roughly walk three quarters around the pond (around)"), while a round shape is depicted in temporal overlap with "drei Viertel um den Teich herum" ("three quarters around the pond around"), i.e., with most of the verb phrase (see Figure 2 for a visualization).

Such a broad scope of the gesture is characteristic of event-related gestures. While we can often identify a single word (or a relatively short phrase) as verbal affiliate for object-related gestures, the stroke of event-related gestures is often spread over the entire verb phrase. Since in the prototypical case the number of arguments controlled by a verbal predicate is larger than it is for noun phrases, an interface between speech and gesture has to be more complex in order to link the two modalities adequately.

Another issue is the fact that verb phrases can feature so-called "sentence brackets", as in our example. Here, due to a sentence bracket, the finite verb stem is separated from its prefix. Together they embrace the German "Mittelfeld". In our example, the stem ("gehst") and its prefix ("herum") embrace the noun phrase "quasi drei Viertel um den Teich". It is important to note that both the prefix and the finite verb stem cannot be fully interpreted on their own even if they are separated on the surface. For that reason, we mark them in the syntax representation with an asterisk (see Figure 2).

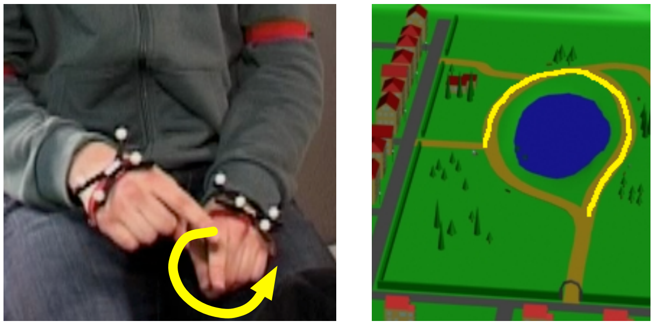

Figure 1: An event-related gesture (left) depicting the way to be walked around a pond (right)

The objective of this paper is to contribute to the understanding of how event-related gestures are connected with speech, whereby we aim to explore the phenomenon from an interdisciplinary viewpoint in which theoretical reconstructions as well as computational and cognitive generation models are developed in tandem. Taken together, these two perspectives provide new insights and a more comprehensive understanding of how event-related gestures are used. In both research lines we already have plenty of experience with NP-related gestures on which we build our current modeling attempts. These include a gesture typology and a partial ontology for noun phrase-aligned gestures (Rieser, 2010), as well as a generation network for iconic gestures (Bergmann & Kopp, 2009) which is integrated into an overall production framework to generate speech-gesture utterances based on an activation-spreading account within dynamically shaped multi-modal memory (Kopp, Bergmann, & Kahl, 2013).

Both research lines are based on empirical data from a systematically annotated corpus, the Bielefeld Speech-and-Gesture Alignment-corpus (SaGA; cf. Lücking, Bergmann, Hahn, Kopp, and Rieser (2012)) which consists of 25 dialogues of dyads engaged in route descriptions. The primary purpose of such route descriptions is to request actions. Therefore this kind of discourse contains a large number

[1] Authors in alphabetical order.

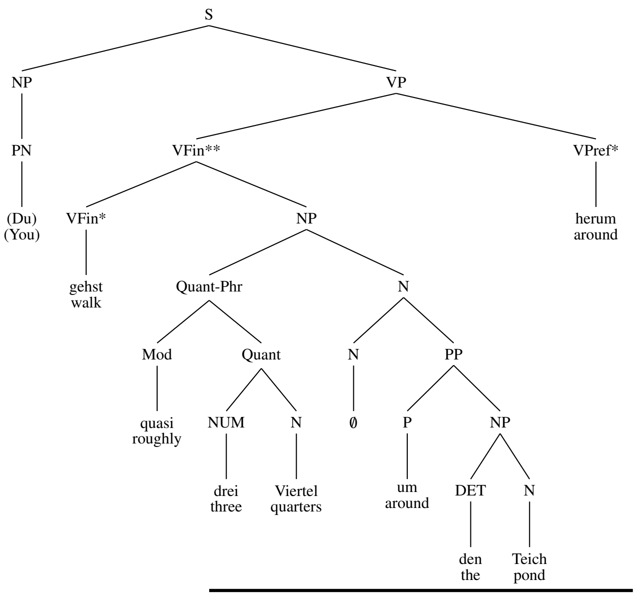

Figure 2: Syntax representation (stroke indicated by line)

of instructions, the majority of them referring to landmarks (Daniel and Denis (1998), e.g., "turn right at X", "walk along X"), as in our example cited above. This paper illustrates our joint work on event-related gestures with respect to the example mentioned above originating from the SaGA-corpus.

Interface Constructions for Gestures Accompanying Verb Phrases

By intuition, five readings of our example gesture are possible. The function of the gesture could be to be related to (i) "herum" (prefix reading), (ii) "um den Teich" (PP reading), (iii) "herumgehen" (finite verb reading), (iv) "drei Viertel um den Teich" (NP reading) or (v) "gehst drei Viertel um den Teich herum" (VP reading). We select (v) here, since the stroke overlaps with both the prefix and the object noun phrase.

Due to our work with the SaGA-corpus, we are aware of the fact that gesture use is bound to speech acts. However, at present we abstract from direct and indirect speech acts and only treat the embedded proposition. We have developed a general methodology for interfacing speech meaning and gesture meaning. In this framework we can treat most of the interpretation problems we systematically discussed related to the corpus. Using an interface methodology, we concentrate on the static semantics of speech-gesture occurrences. The compositionality issues we take up are similar to those found in proposition-related speech-gesture interfaces by Giorgolo, Lücking or Lascarides and Alahverdzhieva. They differ from Lascarides and Stones' proposals, simply, because we do not reach the discourse level here. See Rieser (2013) for an overview of the options available.

Despite gestures having an independent meaning, we assume that gesture meaning constrains verbal meaning. Given this assumption, we aim at constructing a *multi-modal proposition*. We first provide an independent semantics for the speech part and the gesture part, couched in two different logics. Both are derived from more fundamental information: on the one hand lexicon entries and on the other hand gesture features. Both are extended for interfacing and subsequently fused into the interface proper generating a unified semantics. Here, verbal meaning is considered as functor for the gesture meaning as argument. The constraints are modeled using typed lambda calculus. In the end, the "gesture logic" is embedded into the "speech logic".

The semantic representation for the verb phrase in our example (ignoring "quasi") is provided using a Montague-Parsons-Reichenbach inspired event ontology. It roughly represents a yet undetermined agent x who is engaged in a "herumgehen"-event. The theme of this event is the not otherwise specified path around the pond:

$$\begin{aligned}&\lambda x.\ \exists e z\ 3/4x_1\ \exists \mathrm{F}\\&(\text{WALK-AROUND}(e) \wedge \mathrm{AGENT}(e,x) \wedge \mathrm{THEME}(e,x_1) \wedge\\&\mathrm{F}(x_1,z) \wedge \text{AROUND}(x_1, \iota y(\text{POND}(y))))\end{aligned}$$

The semantic representation of the gesture is got using the annotated gesture features. The combination of the gesture's features and their values maps to a simple first order predicate logic formula. However, this mapping is only partial since not all gesture predicates are interpreted. With respect to our example, the mapping result is a semantic representation of a gesticulated circular trajectory, as shown in the attribute value matrix in Figure 3.[2]

$$\begin{bmatrix}\text{Path of Wrist–ARC<ARC<ARC<ARC} & \exists x_2(\text{CIRCULAR-TRAJECTORY}(x_2))\\ \text{Representation Technique–Drawing} & \end{bmatrix}$$

Figure 3: Partial ontology for the gesture under consideration

The interface is built adding parameters to both the gesture and the speech representation. In order to calculate the multimodal proposition, the speech representation is extended by a lambda abstracted predicate:

$$\begin{aligned}&\lambda Y.\ \lambda x.\ \exists e z\ 3/4x_1\ \exists \mathrm{F}\\&(\text{WALK-AROUND}(e) \wedge \mathrm{AGENT}(e,x) \wedge \mathrm{THEME}(e,x_1) \wedge\\&\mathrm{F}(x_1,z) \wedge \text{AROUND}(x_1, \iota y(\text{POND}(y))) \wedge Y(z))\end{aligned}$$

In addition, an identity relation with two relata is added to the gesture representation: the variable for the circular trajectory and a lambda abstracted variable. In the interface the relation provides for the identification of the trajectory variable with the second variable of the F-predicate which represents the not otherwise specified path around the pond. Thereby, the gesture semantics constrains the speech semantics in the following way: The set of models in which the final multimodal proposition is true is restricted to those including a circular trajectory which stands in a relation to the path. The resulting extended gesture representation is as follows:

$$\lambda z_2.\exists x_2(\text{CIRCULAR-TRAJECTORY}(x_2)\ \wedge\ =(x_2,z_2))$$

[2] At this point, we abstract from the details of this trajectory which, *inter alia*, consists of four bent segments.

Using the gesture representation as argument for the speech functor, the simplified calculated interface formula is as follows:

$$\begin{aligned}&\lambda x.\ \exists ez\ 3/4x_1\ \exists \mathrm{F}\\&(\textsc{walk-around}(e) \wedge \mathrm{AGENT}(e,x) \wedge \mathrm{THEME}(e,x_1) \wedge\\&\mathrm{F}(x_1,z) \wedge \textsc{around}(x_1, \iota y(\textsc{pond}(y))) \wedge\\&\textsc{circular-trajectory}(z))\end{aligned}$$

Completing the formula by integrating the representation of the agent, results in the following:

$$\begin{aligned}&\exists ez\ 3/4x_1\ \exists \mathrm{F}\\&(\textsc{walk-around}(e) \wedge \mathrm{AGENT}(e, \mathrm{ADDRESSEE}) \wedge\\&\mathrm{THEME}(e,x_1) \wedge \mathrm{F}(x_1,z) \wedge \textsc{around}(x_1, \iota y(\textsc{pond}(y))) \wedge\\&\textsc{circular-trajectory}(z))\end{aligned}$$

As a whole, the interface provides the multi-modal meaning for the speech-gesture occurrence. It represents the "herumgehen"-event around a pond. The addressee of the utterance is the agent engaged in this event. The gesture contributes to the multi-modal meaning by adding the information that a circular trajectory stands in a relation to the not otherwise specified path around the pond.

Computational Simulation

Based on the theoretical issues discussed above, we now go further into the relation between gesture and speech from a computational simulation viewpoint. In particular, we provide an explanation of semantic coordination between the two modalities based on activation-spreading within dynamically shaped multi-modal memories, in which coordination arises from the interplay of visuo-spatial and linguistically shaped representations.

In previous work, we developed a production model (Kopp et al., 2013) that comprises three stages: conceptualization, where a *message generator* and an *image generator* work together to select and organize information to be encoded in speech and gesture, respectively; formulation, where a *speech formulator* and a *gesture formulator* determine appropriate verbal and gestural forms for this; *motor control* and *articulation* to finally execute the behaviors. The production architecture is based on a multi-modal memory model that comprises visuo-spatial knowledge representations (e.g., mental images), symbolic-propositional representations, and supra-modal associations for concepts like 'round' or 'left-of', which are assumed to link the respective visuo-spatial properties to corresponding denotations in propositions. The *message generator* works on the propositional representations to compose preverbal messages that the LTAG-based *speech formulator* can process. The *image generator* works on visuo-spatial informations about the object or event to be described. The result is passed down to the *gesture formulator* which derives a gesture form specification using a Bayesian decision network that was learned from the SaGA-corpus data (GNetIc; Bergmann and Kopp (2009)). These generation networks combine speaker-specific characteristics of gesture use (captured in data-based conditional probabilities) with common patterns of how meaning is mapped onto gesture form (captured in rule-based decision nodes).

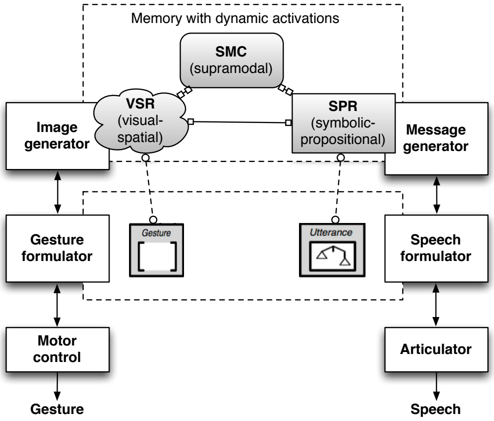

Figure 4: Model of speech-gesture production

Now, to account for the particular case of event-related speech-gesture utterances – going beyond gestures accompanying noun phrases – we are advancing the production model with respect to several issues. First, the knowledge representations are extended to deal with *dynamic* mental imagery. In particular, the set of supra-modal concepts has to implement actions specified by start and target position. Here the multi-modal information, resulting from the interface, indicates what the scope of the supra-modal concepts has to be. In addition, the formulators (LTAG grammar and GNetIc model), must be able to produce event-related sentences and gestures. For the GNetIc model, this add-on means that further representation techniques like 'drawing-posturing' (the agent is represented by the hand which is drawing the path's trajectory) come into play. Notably, for our domain of application, the physical form of these gestures does not necessarily differ from shape-depicting gestures for static objects.

In this model the production process for our example starts from a simplified communicative goal "walk-around (ADDRESSEE, POND, START-LOC, DEST-LOC)" which partially captures the event-based theoretical reconstruction as described in the previous section. Based on this communicative intention the *image generator* and the *message generator* induce activations within the respective representations. The multi-modal memory strives for coherence by invoking supra-modal association concepts. If such a concept's visuo-spatial part matches with highly activated entries in the visuo-spatial representation, a symbolic-propositional representation of the spatial concepts, bound to the specific entity, is created (e.g., for 'round', 'three quarters', or 'clockwise'). Now, a dynamic cognitive simulation runs to model the spreading of activation across the linked multi-modal memory structures. At any time, the generators can independently retrieve entries based on their activation and try to compose structured messages to be sent to the respective formulators. Depending on how much time is available for the process of meaning coordination, the multi-modal representations are more or less well coordinated when the formulators start with their generation work of turning the respective messages into verbal and gestural forms. Temporal coordination results from a synchrony constraint for onsets of co-expressive behaviors realized in our virtual agent software (Kopp & Wachsmuth, 2004).

This way, the system allows to produce different variants of our multi-modal example utterance, e.g., the words "you walk around the pond" accompanied by a gesture depicting the trajectory of the path around the pond as specified in the multi-modal proposition in the previous section. Note that *via* the constraint induced by the gesture, roundness is added to the speech meaning. So the gesture would be non-redundant to speech, supplementing the verbal utterance with information about the path's shape as well as the fact that the addressee only has to walk three quarters of the way. If more time is available, however, the contents expressed verbally and/or gesturally tend to converge. Thus, it is more likely that both generators retrieve similar contents, and accordingly, the *speech formulator* is now enabled to plan a sentence like "You walk three quarters around the circular pond" which encodes information about the path around the pond, however, this time in redundancy with speech.

Conclusion

This paper provides an interdisciplinary view on the integration of gesture meaning and verbal meaning for German verbs of motion. We discussed an example utterance of an iconic gesture accompanying a verb of motion from a theoretical perspective which resulted in the construction of a relation between the gesture representation and the event representation depicted verbally. Complementing the theoretical reconstruction with a computational modeling viewpoint enabled us to further spell out the relation between gesture and speech. Our simulation model offers a cognitive account of how meanings are coordinated across both modalities and, thus, explains how different variants of information distribution might emerge (for details see Kopp et al. (2013)).

With these modeling instruments at hand, it is now possible to explore different integration mechanisms for speech and gesture like the overall production costs, available resources etc. We are also able to investigate the *temporal* relationship between both modalities, e.g., by testing whether we can simulate empirical findings stating that temporal synchrony follows semantic synchrony (Bergmann, Aksu, & Kopp, 2011). Another focus of our future work is a scope extension, in a first step, towards other kinds of gestures accompanying verb phrases, e.g., perception verbs (e.g., 'see') or static verbs (e.g., 'stand'). In a second step, we aim to apply our methodology to gesture use in dialogues, considering speech acts and dialogue acts. We are confident that the interdisciplinary methodology we have initialized in this paper enables us to deal with these challenging issues.

Acknowledgments

This research is supported by the German Research Foundation (DFG) in the Collaborative Research Center 673 "Alignment in Communication" and the Center of Excellence 277 "Cognitive Interaction Technology" (CITEC).

References

Allen, G. (2003). Gestures accompanying verbal route directions: Do they point to a new avenue for examining spatial representations? *Spatial Cognition and Computation*, *3*(4), 259–268.

Bergmann, K., Aksu, V., & Kopp, S. (2011). The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *Proceedings of GESPIN 2011.*

Bergmann, K., & Kopp, S. (2009). GNetIc—Using Bayesian decision networks for iconic gesture generation. In *Proceedings IVA 2009* (pp. 76–89). Berlin: Springer.

Daniel, M., & Denis, M. (1998). Spatial descriptions as navigational aids: A cognitive analysis of route directions. *Kognitionswissenschaft*, *7*, 45-52.

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.

Kopp, S., Bergmann, K., & Kahl, S. (2013). A spreading-activation model of the semantic coordination of speech and gesture. In *Proceedings of CogSci 2013.*

Kopp, S., & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Computer Animation and Virtual Worlds*, *15*, 39-52.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., & Rieser, H. (2012). Data-based Analysis of Speech and Gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its Applications. *Journal on Multimodal User Interfaces*, 287–317.

Müller, C. (1998). *Redebegleitende Gesten: Kulturgeschichte–Theorie–Sprachvergleich.* Berlin: Berlin Verlag.

Parsons, T. (1990). *Events in the Semantics of English. A Study in Subatomic Semantics*. MIT Press, Cambridge, Massachusetts.

Reichenbach, H. (1947). *Elements of Symbolic Logic*. The Macmillan Company, New York.

Rieser, H. (2010). On Factoring out a Gesture Typology from the Bielefeld Speech-And-Gesture-Alignment Corpus (SaGA). In *Gesture in Embodied Communication and Human-Computer Interaction.* Berlin: Springer.

Rieser, H. (2013). Speech-gesture Interfaces. An Overview. In *Proceedings of 35th Annual Conference of DGfS* (pp. 282–283).

Thomason, R. H. (1974). *Formal Philosophy. Selected Papers of Richard Montague*. Yale University Press, New Haven and London.