

# The Linguistic Interpretation of Non-emblematic Gestures Must be agreed in Dialogue: Combining Perceptual Classifiers and Grounding/Clarification Mechanisms

Andy Lücking and Alexander Mehler and Alexander Henlein

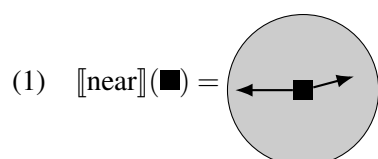
Goethe University Frankfurt, Text Technology Lab  
 {luecking,mehler,henlein}@em.uni-frankfurt.de

## 1 Introduction

Non-emblematic manual gestures pose a double challenge for semantic theories: Firstly, gestures are instances of *visual communication*, so their interpretation requires a means of perceptual classification. Secondly, according to gesture studies, non-emblematic gestures lack “standards of form” (McNeill, 1992, p. 22). In other words, there is no lexicon of such gestures (as opposed to emblematic ones). Accordingly, the linguistic interpretation of gestures – that is, the classification of a gesture occurrence by means of verbal labels from a natural language – leaves room for interpretation. If this room for interpretation is to be resolved, it must be negotiated in dialogue (“What does the speaker/gesturer mean by the gesture?”). Therefore, the linguistic meaning of non-emblematic gestures, if unclear or important for the understanding of the utterance, must be agreed in dialogue.

## 2 Perceptual Classification and CVM

Following formal semantics work on spatial language (Zwarts, 1997, 2003) and the psychophysics of biological movement (Johansson, 1973; Johansson et al., 1980), a uniform, imagistic extension of semantic models, respectively the lexical semantics of certain predicates, is accomplished in terms of vector sequences. For instance, the spatial preposition *near* has the vector denotation in (1) (Zwarts, 2003). The reference object is represented by the black rectangle, the two arrows indicate two of the vectors from the denotation (the gray area; boundaries should be fuzzy, of course).



Johansson and colleagues showed that the perception of dynamic events, that is, events that involve motion, can be modeled in terms of vectorial representations, too. The vector-based representations provide useful explications of the visual components of lexical items, dubbed *conceptual vector meaning* (CVM) (Lücking, 2013). CVMs are also candidates for explicating *what* a perceptual classifier actually has learned.

Larsson (2015, 2020) makes classification the core of meaning so that the type *Meaning* (*Mng*) of a lexical entry involves a classifier (*clfr*) from the outset. As Larsson (2020) emphasizes, classifiers provide a computational spell out of (perceptual) *judgments* in TTR: a situation *s* is of type *T*,  $s : T$ , if the classifier associated with *T* returns *T* when applied to *s* (i.e.,  $\llbracket T \rrbracket.\text{clfr}(\text{par}, s) = T$ ). As is known from human vision, people classify objects and events by comparing a visual percept with stored image (Ullman, 1996, §6). CVMs are representations of stored images, so we add them to *Meaning*:

$$(2) \quad Mng := \left[ \begin{array}{l} \text{par} : Rec \\ \text{cvm} : Type \\ \text{bg} : RecType \\ \text{fg} : \text{bg} \rightarrow RecType \\ \text{clfr} : \text{par} \rightarrow \text{bg} \rightarrow \text{cvm} \rightarrow RecType \end{array} \right]$$

The classifier in (2) now involves an additional layer of computation, namely a geometric comparison *G* of the percept (from ‘par → bg’) with the value of ‘cvm’.<sup>1</sup> Let us illustrate this with the simple example of *near*. Using *near*’s CVM from (1), the meaning of *near* can be expressed as follows, where, following Zwarts (2003),  $\text{place}(\mathbf{v}, x)$  denotes a vector emanating from object *x*:

<sup>1</sup>Ideally, there also should be a feedback loop such that each successful or unsuccessful classification updates (confirms or modifies) *Mng.cvm*.

$$(3) \llbracket \text{near} \rrbracket = \left[ \begin{array}{l} \text{par} : \text{Rec} \\ \text{cvm} = \left\{ \mathbf{v} \mid \text{place}(\mathbf{v}, \text{bg}, x) \right\} : \text{Type} \\ \text{bg} = \left[ \begin{array}{l} x : \text{Ind} \\ \mathbf{v} : \text{Vec} \end{array} \right] : \text{RecType} \\ \text{l} : \mathbb{R} \\ \text{fg} : \text{bg} \rightarrow \text{near}(\text{bg}.x) \\ \text{clfr} : \text{par} \rightarrow \text{bg} \rightarrow \text{cvm} \rightarrow \text{RecType} \end{array} \right]$$

The classifier for *near*,  $\llbracket \text{near} \rrbracket.\text{clfr}$ , applies to situations  $r$  involving an individual and a vector of a certain length ‘l’:

$$(4) r = \left[ \begin{array}{l} x : \text{Ind} \\ \mathbf{v} : \text{Vec} \\ \text{l} = \|\mathbf{u}\| : \mathbb{R} \end{array} \right]$$

$$(5) \llbracket \text{near} \rrbracket.\text{clfr}(\text{par}, \text{cvm}, r) = \begin{cases} \text{near}(r.x) & \text{if } G[(r.l \cdot \text{par}.w), \text{cvm}] > \text{par}.t \\ -\text{near}(r.x) & \text{else} \end{cases}$$

$G$  is an algorithm from computational geometry (Sack and Urrutia, 2000), which compares the weighted input of situation  $r$  with the stored CVM information. In this case,  $G$  just has to perform a distance calculation.

### 3 Speech–Gesture Monitoring in Dialogue

The default integration of speech and gesture – namely that a gesture  $g$  directly exemplifies it affiliate – can now be expressed as follows:


$$(6) \textit{Affiliation Default} \quad \llbracket \ulcorner \textit{affiliate} \urcorner \rrbracket.\text{clfr}(\text{par}, \text{cvm}, \pi_v(g)) \mapsto \ulcorner \textit{affiliate} \urcorner$$

That is, a vectorized gesture movement figures as the background situation onto which the classifier associated with the gesture’s affiliate in speech applies. This immediately gives rise to a notion of speech–gesture mismatch, or inconsistency:

$$(7) \textit{Speech–Gesture Mismatch} \quad \text{If } \llbracket \ulcorner \textit{affiliate} \urcorner \rrbracket.\text{clfr}(\text{par}, \text{cvm}, \pi_v(g)) \not\mapsto \ulcorner \textit{affiliate} \urcorner, \text{ an inconsistency between speech and gesture } g \text{ occurs.}$$

We note again that (7) is a simplification, since gestures that attach to frame elements that are associated with the surface affiliate expression are not taken into account. Apart from this simplification, a mismatch according to (7) can trigger multimodal clarification interaction (Ginzburg and Lücking, 2021).

Example (8), taken from Lücking et al. (2024), is constructed following SaGA dialogue V10, 3:19 (Lücking et al., 2010) where R talks about staircases and makes a spiral gesture (8-a). Then F poses the verbal clarification request whether the linguistic interpretation of R’s multimodal utterance is the hyponym “spiral staircase” (8-b), which can be confirmed or rejected (8-c).

- (8) a. R: Inside the hall was an imposing staircase. 
- b. F: Do you mean a spiral staircase?
- c. R: Yes/No.

The spiral gesture from example (8) does not directly match  $\llbracket \text{staircase} \rrbracket.\text{clfr}$ , but it does correspond to  $\llbracket \text{spiral-staircase} \rrbracket.\text{clfr}$ . This raises the issue  $q_0 = \text{“?Mean}(R, u_0, \text{‘spiral staircases’})\text{”}$  as F’s MaxQUD, where  $u_0$  is the multimodal sub-utterance consisting of the noun *staircases* and the wounded gesture. Parameter Identification is triggered, leading to F’s clarification question, which is co-propositional to  $q_0$ .

### 4 Conclusions

We formally defined speech–gesture congruence and mismatch, in particular the latter underlies multimodal clarification interaction. The sample analysis shows a sometimes intricate interaction of QUD accommodation and perceptual gesture classification, mechanisms which call for further exploration in future work. A couple of processing predictions of our model can already be derived, however, including the following ones.

- The ease of the linguistic interpretation of a gesture depends on the degree of conventionalization (strength) between lexemes and their associated CVMs.
- The linguistic interpretation becomes more difficult when the gesture gives rise to a vectorial model that is not lexicalized in terms of a CVM.

## Acknowledgement

Support by the *Deutsche Forschungsgemeinschaft* (DFG), grant number 502018965, is gratefully acknowledged.

Joost Zwarts. 2003. Vectors across spatial domains: From place to size, orientation, shape, and parts. In Emile van der Zee and John Slack, editors, *Representing Direction in Language and Space*, pages 39–68. Oxford University Press, Oxford, NY.

## References

- Jonathan Ginzburg and Andy Lücking. 2021. [Requesting clarifications with speech and gestures](#). In *Proceedings of the 1st Workshop on Multimodal Semantic Representations*, MMSR, pages 21–31. Association for Computational Linguistics.
- Gunnar Johansson. 1973. [Visual perception of biological motion and a model for its analysis](#). *Perception & Psychophysics*, 14(2):201–211.
- Gunnar Johansson, Claes von Hofsten, and Gunnar Jansson. 1980. Event perception. *Annual Review of Psychology*, 31:27–63.
- Staffan Larsson. 2015. [Formal semantics for perceptual classification](#). *Journal of Logic and Computation*, 25(2):335–369.
- Staffan Larsson. 2020. Discrete and probabilistic classifier-based semantics. In *Proceedings of the Probability and Meaning Conference, PaM 2020*, pages 62–68, Gothenburg. Association for Computational Linguistics.
- Andy Lücking. 2013. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. De Gruyter, Berlin and Boston.
- Andy Lücking, Kirsten Bergmann, Florian Hahn, Stefan Kopp, and Hannes Rieser. 2010. [The Bielefeld speech and gesture alignment corpus \(SaGA\)](#). In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, LREC 2010, pages 92–98, Malta. 7th International Conference for Language Resources and Evaluation.
- Andy Lücking, Alexander Henlein, and Alexander Mehler. 2024. [Iconic gesture semantics](#). *Preprint*, arXiv:2404.18708.
- David McNeill. 1992. *Hand and Mind – What Gestures Reveal about Thought*. Chicago University Press, Chicago.
- Jörg-Rüdiger Sack and Jorge Urrutia. 2000. *Handbook of computational geometry*. Elsevier, Amsterdam, The Netherlands.
- Shimon Ullman. 1996. *High-Level Vision*. A Bradford Book. MIT Press, Cambridge, MA.
- Joost Zwarts. 1997. Vectors as relative positions: A compositional semantics of modified PPs. *Journal of Semantics*, 14(1):57–86.