

# Chapter 1

## Gesture

Andy Lücking

Goethe University Frankfurt

The received view in (psycho-)linguistics, dialogue theory and gesture studies is that co-verbal gestures, i.e. hand and arm movement, are part of the utterance and contributes to meaning. The relationships between gesture and speech obeys regularities which are consequently captured in grammar by means of a gesture-grammar interface. This chapter reviews constraints on speech-gesture integration and summarises their implementations into HPSG frameworks. Pointers to future developments conclude the exposition.

### 1 Why gestures?

In seminal works, McNeill (1985; 1992) and Kendon (1980; 2004) argue that co-verbal gestures, i.e. hand and arm movements, can be likened to words and are part of a speaker's utterance. Accordingly, integrated speech-gesture production models have been devised (Kita & Özyürek 2003; de Ruiter 2000).

This section highlights some phenomena particularly important for grammar, including *mixed syntax* (Slama-Cazacu 1976):

- (1) He is a bit [*circular movement of index finger in front of temple*].

In 1, a gesture replaces a position that is usually filled by syntactic constituent. With regard to deictic gestures Fricke (2012) argues that deictic words within noun phrases – her prime example is German *so* –, provide a *structural*, that is, *language-systematic* integration point between the vocal plane of conventionalized words and the non-vocal plane of body movement. Therefore, on this conception, grammar is inherently multimodal.



## 2 Kinds of gestures

The previous section talks about iconic and deictic gestures. What is meant by this distinction?

This section provides a brief taxonomy, probably in terms of




- iconic (representational)
- deictic (pointing)
- interactive (dialogue-oriented)
- beat (rhythmic)
- emblem (lexicalized)


Besides categorially distinguished gesture classes, gestures can also be specified according to multi-dimensional characteristics.

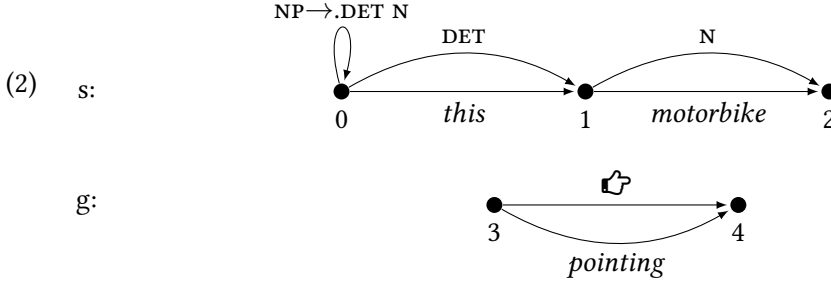
Gestures have to be partitioned into at least three phases: preparation, stroke, and retraction. Unless stated otherwise, when talking about gestures in the following, the stroke phase, which is the “gesture proper” or the “semantic interpretable” phase, is referred to.

In any case, the spontaneous, usually co-verbal hand and arm movement considered in this chapter are different from the signed signs of sign languages (Markus Steinbach and Anke Holler, this volume) and pantomime (not spontaneous and co-verbal).

## 3 Precursors


Using typed feature structures in order to represent form and meaning of gestures goes back to computer science approaches to human-computer interaction (HCI) (cf. Lücking & Pfeiffer 2012). The *QuickSet* system (Cohen et al. 1997) allows users to operate on a map and move objects or create barb wires (the project was funded by a military fund) by giving verbal commands and manually indicating coordinates. The system processes voice and pen (gesture) input by assigning signals from both media representations in the form of attribute-value matrices (AVMs) (Johnston 1998; Johnston et al. 1997). The common representation format is a precondition for implementing multimodal integration of speech and gestures in terms of unification (cf. Lücking et al. 2006). For instance, *QuickSet* will move a vehicle to a certain location on the map when asked to *Move this*[] *motorbike to here*[, where ‘’ represents a touch gesture.

Since a linguistic unification-based grammar rests on a conventional, “uni-modal” parser, Johnston (1998) and Johnston et al. (1997) developed a *multimodal chart parser*. A multimodal chart parser consists of two or more layers and allows for layer-crossing charts. The multimodal NP *this* *motorbike*, for instance, is processed in terms of a multimodal chart parser covering of a speech (s) and a gesture (g) layer:



A multimodal chart or *multichart* is defined in terms of sets of identifiers from both layer. Possible multicharts from (2) include the following ones:

- (3) multichart 1: {[s,0,1], [g,3,4]}  
 multichart 2: {[s,1,2], [g,3,4]}  
 ...

The meaning (*content*) of a gesture of category (*cat*) *spatial\_gesture* like ‘’ is represented as a *latitude-longitude* coordinate pair (Johnston 1998):

- (4) 
$$\left[ \begin{array}{ll} \text{cat:} & \text{spatial\_gesture} \\ \text{content:} & \left[ \begin{array}{ll} \text{fsType:} & \text{point} \\ \text{coord:} & \text{latlong}(x, y) \end{array} \right] \end{array} \right]$$

The basic rule for integrating spatial gestures with speech commands is the *basic integration scheme* (Johnston 1998; Johnston et al. 1997), reproduced in (5):

$$(5) \left[ \begin{array}{l} \text{lhs :} \\ \text{rhs :} \\ \text{constraints :} \end{array} \left[ \begin{array}{l} \left[ \begin{array}{l} \text{cat :} \quad \textit{command} \\ \text{modality :} \quad \boxed{2} \\ \text{content :} \quad \boxed{1} \\ \text{time :} \quad \boxed{3} \end{array} \right] \\ \left[ \begin{array}{l} \text{dtr1 :} \\ \text{dtr2 :} \end{array} \left[ \begin{array}{l} \left[ \begin{array}{l} \text{cat :} \quad \textit{located\_command} \\ \text{modality :} \quad \boxed{6} \\ \text{content :} \quad \boxed{1}[\text{location } \boxed{5}] \\ \text{time :} \quad \boxed{7} \end{array} \right] \\ \left[ \begin{array}{l} \text{cat :} \quad \textit{spatial\_gesture} \\ \text{content :} \quad \boxed{5} \\ \text{modality :} \quad \boxed{9} \\ \text{time :} \quad \boxed{10} \end{array} \right] \end{array} \right] \end{array} \right] \left\{ \begin{array}{l} \text{overlap}(\boxed{7}, \boxed{10}) \vee \text{follow}(\boxed{7}, \boxed{10}, 4) \\ \text{total-time}(\boxed{7}, \boxed{10}, \boxed{3}) \\ \text{assign-modality}(\boxed{6}, \boxed{9}, \boxed{2}) \end{array} \right\}$$

The AVM in (5) implements a mother-daughter structure along the lines of a context free grammar rule, where a left-hand side (*lhs*) expands to a right-hand side (*rhs*). The right-hand side consists of two constituents (daughters *dtr1* and *dtr2*), a verbal expression (*located\_command*) and a gesture. The semantic integration between both modalities is achieved in terms of feature-structure sharing, see tag  $\boxed{5}$ : the spatial gesture provides the location coordinate for the verbal command.

The bimodal integration is constrained by a set of restrictions, mainly regulating the temporal relationship between speech and gesture (see tags  $\boxed{7}$  and  $\boxed{10}$  in the ‘constraints’ field): the gesture may overlap with its affiliated word in time, or follow it in at most four seconds.

## 4 Gestures in HPSG

With regard to grammar-gesture integration, three main phenomena have to be dealt with:

- What is the meaning of a gesture?
- What is the affiliate of a gesture, that is, its verbal attachment site?

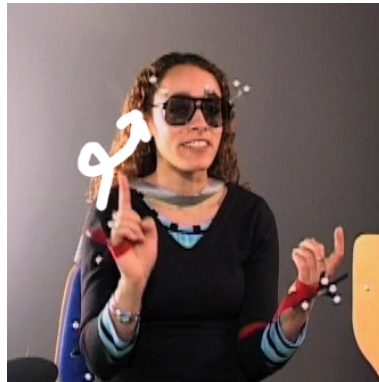


Figure 1: “I think that shall be staircases.”

- What is the result of multimodal integration?

The following example (taken from Lücking (2013a: 189)) illustrates these issues:

- (6) Ich g[laube das sollen TREP]pen sein.  
 I think that shall staircases be.  
 “I think that shall be staircases.”

The first syllable of the German noun *Treppen* (staircases) carries main stress, indicated by capitalization. The square brackets indicate the temporal overlap between speech and gesture, which is shown in Figure 1. Obviously, mere temporal synchrony is too weak an indicator of affiliation, but prosody seems point the right way.

Note that the gesture attributes to the noun it attaches to: from the multimodal utterance the observer retrieves the information that the speaker talks about spiral staircases.

This interpretation changes with a different affiliation of the gesture. Suppose the gestures is produced in company to (presumably stressed) *glaube* (think). Now the spiral movement is interpreted as a metaphorical depiction of psychological process. Thus, the interpretation of a gesture depends on the integration point (affiliation), which in turn seems to be marked by prosody. The resulting multimodal utterance may express a richer content than speech alone.

On accounts that try to deal with the three phenomena, see Alahverdzhieva (2013) diss Alahverdzhieva & Lascarides (2010) Lücking (2013a) diss Lücking (2013b) Lücking (2016)

## 4.1 Pointing Gestures

Synchrony of speech and co-verbal pointing gestures is licensed by a prosodic constraint on the sign type *deictic-word*. The semantic attachment site is underspecified using underspecification techniques of MRS (Copestake et al. 2005). (Alahverdzhieva & Lascarides 2011)

## 4.2 Iconic Gestures

Synchrony of speech and co-verbal pointing gestures is licensed by a prosodic constraint on the sign type *depict-word*. The semantic attachment site is underspecified using underspecification techniques of MRS (Copestake et al. 2005). (Alahverdzhieva & Lascarides 2010)

The prosodic constraint is similar to the (**phonetic-kinematic**) interface relating speech and gestures into a *multimodal ensemble* (Kendon 2004), defined in the unification-based account of Lücking (2013a).

# 5 Outlook

What are (still) challenging issues with respect to grammar-gesture integration?

- gestalt phenomena: the forms depicted by gesture are often incomplete and have to be completed by drawing on gestalt principles or everyday knowledge
- meaning-carrying gestures feature vs. insignificant gestures features
- “Portmanteau gestures”, like a spiral upward movement, which cannot simply be decomposed into a circular and a straight movement aspect (Lücking 2013a)
- building multimodal parsers (Alahverdzhieva et al. 2012), but see also Section 3 above

## References

Alahverdzhieva, Katya. 2013. *Alignment of speech and co-speech gesture in a constraint-based grammar*. School of Informatics, University of Edinburgh dissertation.

- Alahverdzhieva, Katya, Dan Flickinger & Alex Lascarides. 2012. Multimodal grammar implementation. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: human language technologies* (NAACL-HLT 2012), 582–586. Montreal, Canada.
- Alahverdzhieva, Katya & Alex Lascarides. 2010. Analysing language and co-verbal gesture in constraint-based grammars. In Stefan Müller (ed.), *Proceedings of the 17th international conference on head-driven phase structure grammar (hpsg)*, 5–25. Paris.
- Alahverdzhieva, Katya & Alex Lascarides. 2011. An HPSG approach to synchronous speech and deixis. In Stefan Müller (ed.), *Proceedings of the 18th international conference on head-driven phrase structure grammar*, 6–24. University of Washington.
- Cohen, Philip R., Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen & Josh Clow. 1997. QuickSet: multimodal interaction for distributed applications. In *Proceedings of the fifth acm international conference on multimedia* (MULTIMEDIA '97), 31–40. Seattle, Washington, USA. DOI:10.1145/266180.266328
- Copestake, Ann, Dan Flickinger, Carl Pollard & Ivan A. Sag. 2005. Minimal recursion semantics: an introduction. *Research on Language and Computation* 3(4). 281–332.
- Fricke, Ellen. 2012. *Grammatik multimodal. wie Wörter und Gesten zusammenwirken*. Vol. 40 (Linguistik – Impulse und Tendenzen). Berlin & Boston: De Gruyter.
- Johnston, Michael. 1998. Unification-based multimodal parsing. In *Proceedings of the 36th annual meeting on association for computational linguistics – volume i*, 624–630. Montreal, Quebec, Canada: Association for Computational Linguistics.
- Johnston, Michael, Philip R. Cohen, David McGee, Sharon L. Oviatt, James A. Pittman & Ira Smith. 1997. Unification-based multimodal integration. In *Proceedings of the eighth conference on european chapter of the association for computational linguistics*, 281–288. Madrid, Spain: Association for Computational Linguistics.
- Kendon, Adam. 1980. Gesticulation and speech: two aspects of the process of utterance. In Mary Ritchie Key (ed.), *The relationship of verbal and nonverbal communication*, vol. 25 (Contributions to the Sociology of Language), 207–227. The Hague: Mouton.
- Kendon, Adam. 2004. *Gesture: visible action as utterance*. Cambridge, MA: Cambridge University Press.

- Kita, Sotaro & Aslı Özyürek. 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(1). 16–32. DOI:10.1016/S0749-596X(02)00505-3
- Lücking, Andy. 2013a. *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. Berlin & Boston: De Gruyter. Zugl. Diss. Univ. Bielefeld (2011).
- Lücking, Andy. 2013b. Interfacing speech and co-verbal gesture: exemplification. In *Proceedings of the 35th annual conference of the german linguistic society* (DGfS 2013), 284–286. Potsdam, Germany.
- Lücking, Andy. 2016. Modeling co-verbal gesture perception in type theory with records. In M. Ganzha, L. Maciaszek & M. Paprzycki (eds.), *Proceedings of the 2016 federated conference on computer science and information systems*, vol. 8 (Annals of Computer Science and Information Systems), 383–392. IEEE. **Winner best paper award**. DOI:10.15439/2016F83
- Lücking, Andy & Thies Pfeiffer. 2012. Framing multimodal technical communication. with focal points in speech-gesture-integration and gaze recognition. In Alexander Mehler & Laurent Romary (eds.), In collab. with Dafydd Gibbon, *Handbook of technical communication*, vol. 8 (Handbooks of Applied Linguistics), chap. 18, 591–644. Berlin & Boston: De Gruyter Mouton.
- Lücking, Andy, Hannes Rieser & Marc Staudacher. 2006. Multi-modal integration for gesture and speech. In David Schlangen & Raquel Fernández (eds.), *Proceedings of the 10th workshop on the semantics and pragmatics of dialogue* (Brandial'06), 106–113. Potsdam: Universitätsverlag Potsdam.
- McNeill, David. 1985. So you think gestures are nonverbal? *Psychological Review* 92(3). 350–371.
- McNeill, David. 1992. *Hand and mind – what gestures reveal about thought*. Chicago: Chicago University Press.
- de Ruiter, Jan Peter. 2000. The production of gesture and speech. In David McNeill (ed.), *Language and gesture*, chap. 14, 284–311. Cambridge, UK: Cambridge University Press.
- Slama-Cazacu, Tatiana. 1976. Nonverbal components in message sequence: “Mixed Syntax”. In William C. McCormick & Stephan A. Wurm (eds.), *Language and man. anthropological issues* (World Anthropology), 217–227. The Hague & Paris: Mouton.