Iconic Gesture Semantics

Andy Lücking, Alexander Henlein and Alexander Mehler

Goethe University Frankfurt, Text Technology Lab,
Robert-Mayer-Straße 10, Frankfurt a/M, D-60325, Germany.

*Corresponding author(s). E-mail(s): luecking@em.uni-frankfurt.de;
Contributing authors: henlein@em.uni-frankfurt.de;
mehler@em.uni-frankfurt.de;

Abstract

The "meaning" of an iconic gesture is conditioned on its informational evaluation. Only informational evaluation lifts a gesture to a quasi-linguistic level that can interact with verbal content. Interaction is either vacuous or regimented by usual lexicon-driven inferences. Informational evaluation is spelled out as extended exemplification (extemplification) in terms of perceptual classification of a gesture's visual iconic model. The iconic model is derived from Frege/Montague-like truth-functional evaluation of a gesture's form within spatially extended domains. We further argue that the perceptual classification of instances of visual communication requires a notion of meaning different from Frege/Montague frameworks. Therefore, a heuristic for gesture interpretation is provided that can guide the working semanticist. In sum, an iconic gesture semantics is introduced which covers the full range from kinematic gesture representations over model-theoretic evaluation to inferential interpretation in dynamic semantic frameworks.

Keywords: Iconic Gesture Semantics, Exemplification, Extemplification, Perceptual Classification, Speech–Gesture Integration, Semantic Theory

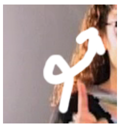
Contents

1 Introduction

> "as we see *them*, we see something *in them*."
>
> Jürgen Streeck (2008, p. 286)

"So You Think Gestures are Nonverbal?" was asked by McNeill in 1985, and the close interweaving of gesture and speech in utterance production has been emphasized since then (see, e.g., Özyürek, 2014). And indeed, speech and gesture also seem to interact semantically, as has been pointed out frequently (see McNeill, 1992; Kendon, 2004, for two seminal sources), and is, for instance, most clearly shown by examples where the gesture can be construed in such a way that it adds to speech meaning, as in (1), where the gesture can be understood to provide shape information specifying the type of the staircases talked about:[1]

[1] The example is taken from dialogue V10 from the SaGA corpus (Lücking et al, 2010), starting at minute 3:19.

(1) “Ich g[laube das  sollen TREP]pen sein” (*I think that should be staircases*; capitalization indicates main stress of the first syllable of the noun *Treppen* ‘staircases’, square brackets indicate the temporal alignment of speech and gesture)

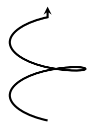
If the gesture is construed in such a way that it makes the concept *wounded* or *spiral* salient, then there is a strong inclination to interpret the noun *staircases* in terms of its hyponym *spiral staircases*. The modest conditional formulation – *if the gesture is construed in such a way that ...* – is indeed part of the iconic gesture semantics developed in the following. We argue that the *informational evaluation* of a gesture lifts its visual properties to quasi-linguistic status, which can interact with natural language semantics and reasoning in the first place (we argue further that this reasoning is mainly driven by lexical knowledge). This is because gestures and words are quite different things: It is widely assumed that the meaning of natural languages are conventionalized, largely arbitrary systems, while iconic gestures are driven by non-arbitrary, visuo-spatial properties.[2] This ambivalence – the verbal and nonverbal characteristics of iconic gestures – has also beset iconic gesture semantics: On the one hand, there are gesture semantics that assume a shared semantic representation of speech and gesture, namely the predicate constants of the formal semantic representation language (Lascarides and Stone, 2009). This is obviously a word-like construal of iconic gesture meaning. On the other hand, Giorgolo (2010) developed a mereotopological model of the interpretation of iconic gestures, accounting for the nonverbal, visual nature of iconic gestures.[3]

[2] See, for instance, Lewis (1969) respectively Alibali (2005); Duncan (2002); Bavelas and Chovil (2000).

[3] The status of other works is less clear. Rieser and Poesio (2009), for instance, pursue a labelling approach but restrict the gestural predicates to spatial ones, which give rise to partial ontology descriptions (see also Hahn and Rieser, 2010). Hence, they use labels to describe spatial configurations. Lücking (2013a) employs a spatial, vector-based model, but only to the effect of deriving verbal labels therefrom. This shows that both, labelling and spatial, views should be kept apart and combined in a unified approach, which is one of the goals of this paper.

The difference between both approaches can be illustrated by means of the gesture from (1): The first approach – *labelling theories* – assigns semantic predicates to the gesture which express its meaning(s)[4] (2a), the second one – *spatial theories* – models the meaning of the gesture in terms of a spatial structure, or iconic model (2b).

(2) a. labelling theories: $\mathit{spiral}(x) \vee \mathit{curled}(x) \vee \mathit{twined}(x) \vee \ldots$

b. spatial theories: 

We conjecture that this ambivalence is due to the "twofoldedness" of gestures diagnosed in gesture research: "as we see *them*, we see something *in them*." (Streeck, 2008, p. 286; original emphsis). We aim at an explanation and a reconciliation of this ambiguity, since from a theoretical perspective, this situation is unsatisfactory for several reasons:

Firstly, labelling theories lack a principled way of assigning predicates to iconic gestures; semantic representations are brought about "by hand", merely following (not necessarily justified, see section 4.4) interpretive assumptions.

Secondly, spatial theories fail to deliver a computational procedure for deriving iconic models from kinematic gesture representations.[5]

Thirdly, the meaning representations in (2a) and (2b) are obviously related in a non-arbitrary way. However, the relation between labelling and spatial theories has not been addressed in gesture semantics so far. We provide the basis of a unified iconic gesture semantics that takes these desiderata into account. Hence, gesture semantics is orthogonal (and intuitively and logically prior, cf. sections 4.3 and 4.4) to recent

[4] Since gestures are considered vague or underspecified, a single gesture usually is compatible to several predicates (Lascarides and Stone, 2009).

[5] The gesture interpretation function ϕ assumed by Giorgolo (2010), for instance, is not spelled out.

investigations of the inferential interplay of speech and gesture (e.g., Ebert and Ebert, 2014; Schlenker, 2019; Esipova, 2019).[6]

Our starting point is a spatial theory, namely a vector space model, following work by Zwarts (1997); Zwarts and Winter (2000); Zwarts (2017) (section 3). Our first contribution is a formal notion of *gesture space* modelled as an oriented vector space in terms of an adaptation of vector space semantics. We then show how kinematic gesture representations can be interpreted in terms of vector sequences in gesture space (addressing the second issue raised above).

Our second contribution is a shift in semantic architecture: we propose to employ vector spaces as models of *intensions* of certain lexical items (section 4). This move paves the way for explaining the relation between labelling and spatial theories (third issue): iconic models and semantic predicates are linked *via* a semiotic relation reminiscent of *exemplification* (Goodman, 1976). We will argue that in order to maintain the semiotic nature of iconic gestures (i.e., avoiding treating them as "mere" objects or events), a modified notion of exemplification is needed. This is perhaps the most consequential contribution (why we have said that it is a shift in semantic architecture): The notion of meaning in mainstream possible worlds semantics, $[\![\cdot]\!]$, is such that the meaning $[\![\alpha]\!]$ of a word α when applied to a world w returns α's extension in w.[7] The notion of meaning emerging from extended exemplification ("extemplification", which is needed for assigning "meaning" to gestures), to the contrary, has to be such that when applied to an object (in some world) it returns a linguistic label for that object. A consequence of this view is that iconic gesture semantics cannot be fully spelled out within a Frege/Montague framework (and this is most likely the reason that there is still no formal iconic gesture semantics). We therefore provide a heuristic for interpreting gestures within possible worlds semantics (section 4.2).

[6] "It should be emphasized that we will not seek to explain how a gesture [. . .] comes to have the content that it does" (Schlenker, 2018, p. 296).

[7] We are simplifying somewhat, of course, ignoring intensional expressions and contexts. This does not touch on the general point we want to make, however.

The general architecture of the iconic gesture semantics is shown in Figure 1, extending an image used by Giorgolo (2010). Speech is interpreted as usual by a denotation function within frames of reference. The kinematic representation of a gesture, however, is translated via function ϕ into an iconic model. This model intersectively constrains the spatial projection of the entities talked about within the spatial frame of reference. Hence, iconicity is semantically differentiated in two aspects: a direct interpretation of a gesture's form (ϕ), and a spatial model interacting with verbal meaning. This part of gestures semantics can be integrated into spatially extended but otherwise traditional Frege/Montague models. The spatial iconic models are then *informationally evaluated* in terms of the predicate constants of the semantic language (extended exemplification, or classification). This step brings about quasi-linguistic gesture meaning that can interact with speech meaning (e.g., lexical meaning; section 4.3). We discuss repercussions of iconic gesture semantics with regard to semantic theorizing in general in section 5.

A word of warning. We consider iconic gesture semantics a theoretical contribution. As such, the following work is largely conceptual: the starting point is a standard semantic framework in the tradition of Frege/Montague. It is extended by a spatial framework to capture the semantic, that is, truth-functional contribution of iconic co-speech gestures. Furthermore, and in line with the above-said, we take seriously the spatial (as opposed to linguistic[8]) nature of iconic gestures. As a consequence, many spatial, vectorial representations are used, which may be unusual for textbook semantics. These are needed to spell out a fully compositional procedure for iconic gesture semantics, and an algorithm for deriving gesture interpretations. We only come back to more concrete examples in section 4. The theory and framework we are proposing is nonetheless rooted in much previous work and empirical findings, some

[8]We use the term 'linguistic' and its morphosyntactic variants to denote constituents of speech (i.e., words phrases, clauses) in contrast to gestures. Given a bulk of gesture research that tries to collect evidence that gestures are linguistic, too, this distinction can be questioned, and we won't argue that. It is nonetheless a handy distinction.

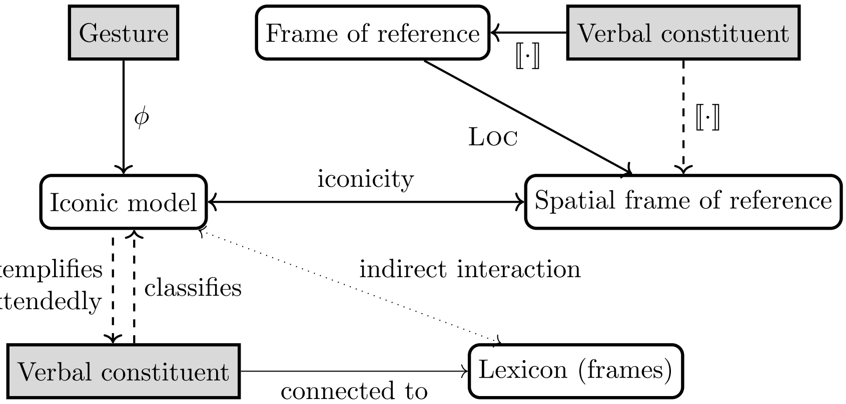


Fig. 1: Architecture of iconic gesture semantics. See main text for details and pointers to pertinent sections.

of which are discussed in the following section to derive some "guidelines" which a decent iconic gesture semantics has to live up to.

2 A brief gesture primer

By 'gesture', we refer to hand and arm movements that accompany speech and are related to the narrative. In early taxonomies, several kinds of manual gestures have been distinguished, namely *metaphorics*, *deictics*, *iconics*, *beat gestures* (McNeill, 1992).[9] These gesture classes are, however, not mutually exclusive: iconic gestures may exhibit rhythmic patterns, and iconic features can overlay deictic gestures. For instance, while pointing at a wooden disc, the index finger can rotate, tracing the outline of the referent (Kranstedt et al, 2006a). Hence, gesture "classes" are better conceived as *dimensions* of gesture meaning (McNeill, 2005).

Deictic gestures can be subdivided according to targeting an object, indicating a direction or region (Rieser, 2004), or being abstract (e.g., pointing at a location in gestures space; McNeill et al 1993). Based on gesture research, a more specific classification of iconic gestures can be given, too: in iconic gesturing, the hands virtually

[9]The original taxonomy includes a fifth class, namely *cohesives*, which is usually not adopted in later taxonomies. In any case, manual gestures are distinguished from signs of a sign language and from emblems.

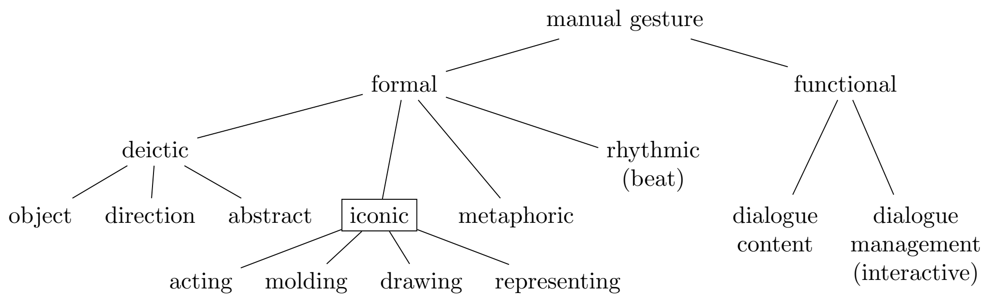


Fig. 2: Dimensions of classifying gestures. The focus is on iconic gestures.

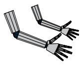
do what the hands do in everyday life and artwork, namely *acting*, *molding*, *drawing*, and *representing* (Streeck, 2008; Müller, 2014, we adopt Müller's quadrinomial nomenclature; Streeck distinguishes even 12 *practices*). *Acting* classifies a miming action. *Molding* describes three-dimensional sculpturing, *drawing* two-dimensional tracing. In *representing*, the hand/arm is a proxy for an object. Such gestural modes of representation will play a prominent role for capturing the meaning of iconic gestures, as will become clear in the following.

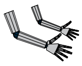
As has been pointed out by Bavelas et al (1992), there are frequent gesture occurrences that are functionally bound up with coordination or dialogue management, rather than dialogue content. Such gestures have also been termed *interactive gestures*, emphasizing that they serve pragmatic functions in dialogue.[10] A summary is provided in Figure 2. We are concerned with the iconic dimension of co-speech gestures (and the corresponding modes of representation) in the following.

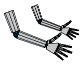
The semantic problem here is that iconic gestures are not regimented by fixed form–meaning associations (i.e., a lexicon; cf. McNeill 1992). Hence, iconic gestures do not constitute a fixed class which can simply be interpreted by the interpretation function, $[\![\cdot]\!]$, for language. In fact, the very same gesture receives quite different interpretations

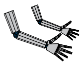
[10] This includes some uses of pointing gestures, which, while directed at the addressee, are embedded in information-state processing (Ginzburg and Lücking, 2021a): they can be assigned to the object-deictic and dialogue-management dimensions.

when produced in combination with different words and phrases, as exemplified in (3), varying a constructed example used by Esipova (2019).[11]

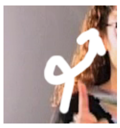
(3) a. S. might bring her  dog. ⇒ "holding/touching dog"

b. S. might bring her  dog in a box. ⇒ "holding/touching box"

c. S. might bring her  small dog. ⇒ "sizing dog"

d. S. might bring her  dog by carrying it. ⇒ "carrying dog"

Gesture interpretation also varies with respect to temporal alignment with speech. Consider again example (1), repeated in (4). The part of speech which roughly co-occurs with the gesture is indicated by brackets. Since the first syllable from the noun *Treppen* is not only part of the portion of speech that co-occurs with gesture but also has primary stress (indicated by capital letters), it is the first candidate for providing an integration point for gesture information (Loehr, 2007; Lücking, 2013a; Alahverdzhieva et al, 2017). Hence, the noun is the integration *locus* of the spiral-movement gesture.

(4) *Ich g[laube das*  *sollen TREP]pen sein*

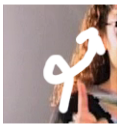
I think that should staircases be

'I think that should be staircases' + gesture overlapping with bracketed speech

The multimodal expression composed of the noun *staircases* and the spiral gesture *can* (but need not) be interpreted in such a way that the gesture depicts the shape of staircases, indicating the hyponym *spiral staircases*.

[11] The gesture is drawn by using Ressler's *Sketch* system (http://www.frontiernet.net/~eugene.ressler/).

Changing the affiliated word also changes gesture interpretation. Suppose the gesture is produced a bit earlier and the main stress is carried by the main verb, as indicated in (5), taken from Lücking (2021, p. 1206):

(5) *I[ch*  *GLAUbe das] sollen Treppen sein*
I think that should staircases be

'I think that should be staircases' + gesture overlapping with bracketed speech

Now the gesture is associated with *think* and interpreted as a metaphorical depiction of a cognitive process. The gestures' co-speech dependence in (5) is therefore bound up with *affiliation*: A gesture occurrence has a "docking point" in speech, its *lexical affiliate* (Schegloff, 1984). The affiliate guides the interpretation of the gesture (Hadar and Krauss, 1999). Note that while the affiliate was initially assumed to be a lexical item (i.e., a "single word"), it is lexical in only about 80 % of occurrences, the remainder exhibit syntactically more complex verbal attachment sites (Mehler and Lücking, 2012).

In many instances the gesture remains "informationally vacuous", however, depicting its affiliate, as in (6a,b). Such types of examples have not received much attention in the semantic literature, but need to be covered, too. Of more interest are examples such as the spiral staircase one or (6c), where there is no direct depiction relation between the gesture and its affiliate. We are concerned with a semantic framework that allows deriving the interpretations given after the arrow ('⇒'). These do not cover the full range of iconic gestures, though. We will not discuss examples such as (6d), which arguably involve question-under-discussion management (Laparle, 2021), but we briefly come back to it in the discussion in section 5.

(6) a. Stewart Robson on ESPN FC Extra Time:

https://www.youtube.com/watch?v=CiVS5_HKFY8&t=0h1m18s

You know when they go on that wheel [*circle gesture*] and throw [*throwing movement*] the dagger would you ever like to see that go wrong?

⇒ circle gesture represents the wheel, and the throwing movement pantomimically mimics *throw*

b. Toni Kroos (in Spanish)

https://www.youtube.com/watch?v=XwGkfrvXPpI&t=0h6m12s

dart? dart? Sabes que es dart? [*throws virtual dart arrow*]

⇒ the throwing movement pantomimically mimics darting

c. Daniel Levitin, Ted talk:

https://www.youtube.com/watch?v=8jPQjjsBbIc&t=0h1m30s

so I [*balancing/weighing movement*] figured, under the circumstances, I was coming out even.

⇒ figure out by weighing

d. Daniel Levitin, Ted talk 2:

https://www.youtube.com/watch?v=8jPQjjsBbIc&t=0h3m55s

some of them are obvious [locate right], some of them are not so obvious [locate left]

The co-text dependence of gestures obvious from (6) is, however, not arbitrary. It is restricted by semantic speech–gesture congruence, or semantic compatibility, as evinced by speech–gesture mismatches. An example from a study of Cassell et al (1999) is given in (7), where the brackets indicate the segment of the utterance that overlaps with the offering gesture (i.e., an action that simulates a delivery event):[12]

(7) Granny sees him and says "oh what a nice little monkey". And then she [offers him a penny].

[12]The stimuli are two-time videotaped retellings of Sylvester and Tweety Bird cartoons, where the second recording involve a modified (mismatch) gesture.

a. normal: left hand proffers penny in the direction of listener.

b.#mismatched: left hand offers penny to self.

The speaker in (7) adopts the viewpoint of the granny, and offering is an action that is directed towards an addressee, hence the incongruity. Notice that incongruity is a stronger notion than denotational vacuity: the mismatch in (7) seems to rest on a *conflict* between speech meaning and iconic model, not just in the fact that there is no model that satisfies both. A gesture semantics should be part of the explanation of such mismatches.

To summarize, an iconic gestures semantics needs to address (at least) three basic desiderata:

D1 Gesture interpretation is strongly dependent on the accompanying speech (affiliate dependence).

D2 Speech–gesture affiliation is regimented by (in-)congruency.

D3 Gesture interpretation is an instance of visual communication, which rests on a perceptual interpretation of gestural forms.

3 Vector space semantics

To capture the visual content of an iconic gesture in line with spatial theories (briefly introduced in section 3.2), we adopt a vector space model (section 3.4). One might also try to apply projective semantics to iconic gestures, which is discussed in section 3.3).[13] In any case, some kind of "preprocessing" is required that works on the kinematic gesture representation and brings about visuo-spatial constraints – this is the heart of iconicity. To this end, a kinematic, form-based gesture representation is introduced in section 3.1. Its interpretation in vector space is discussed in section 3.5.

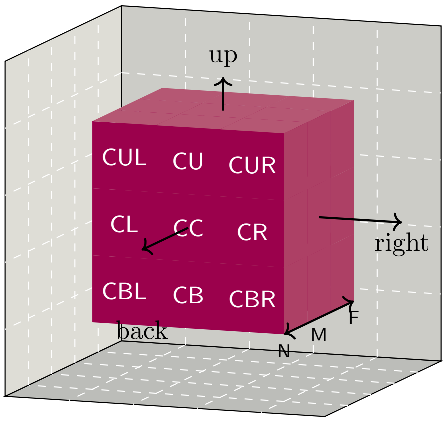
[13]This has, to our knowledge, not been seriously proposed so far (and we will see that it is a non-starter for gestures), but is alluded to by Schlenker (2019, p. 744).

3.1 Kinematic gesture representation

> "Bodily movements cannot be transcribed directly, they have to be described. Descriptions in their turn are to be read like stage directions that do not have to be pronounced but enacted."
>
> Bohle (2013, p. 1003)

The alphabet provides a ready-made transcription system for written text, and phonetic transcription systems for spoken language. How to represent iconic gestures? Given that manual gestures are visible, bodily movements, a form-based, kinematic representation is appropriate. There are several such transcription systems around, we illustrate gesture annotation according to the format used by Kopp et al (2004); Lücking et al (2010). A brief introduction to kinematic gesture representation is needed since it is the basis for the construction of iconic models in vector space (section 3.5). In fact, "annotation playing for gestures the same role as syntax representation plays for linguistic utterances" (Rieser, 2015, p. 123).

The formal description of a gestural movement is given for each hand in terms of the *handshape*, the orientations of the *palm* and the *back of the hand* (BOH), the movement trajectory (if any) of the *wrist*, and the relation between both hands (synchronicity, SYNC). The handshape can be transcribed according to a handshape nomenclature, for instance, the fingerspelling alphabet of American Sign Language, where "G" labels an extended index finger. The orientations (ORIENT) of the palm and back of the hand are specified with reference to the speaker's body (e.g., *PAB* encodes "palm away from body" and *BUP* encodes "back of hand upwards"). Movement trajectories of the whole hand are specified with respect to the wrist in terms of the described path, its direction (DIR), and the extent of the movement. Position and extent are given with reference to the *gesture space* (McNeill, 1992, p. 86–89). Originally, gesture space was a two-dimensional descriptive "map" for kinematic gesture descriptions, where the third dimension is given only as the annotation of the distance of the hand from the speaker's body (feature DIST). Hence, it seems reasonable to construe gesture space as a three-dimensional space that is oriented along the speaker's/gesturer's anatomical planes, as displayed in Figure 3. Since this representational system of gesture "phonology" is divided according to the anatomical joints of the arm it arguably implements complete coverage of gestural movements. Directional information is given in terms of 45° sections (e.g., MF 'forward', ML 'left', and MF/ML 'between forward and left', i.e., "diagonally ahead"), which was sufficient for annotation projects (see Lücking et al, 2013), but could be made more fine-grained.



CBL:	center below left
CL:	center left
CUL:	center upper left
CB	center below
CC:	center center
. . .	. . .
N:	near
M:	middle
F:	far

Fig. 3: Gesture space model. Three-dimensional extension of the two-dimensional model of McNeill (1992) (taken from Lücking (2016)).

As an example, consider the representation of the spiral gesture from our first examples (1), (4) and (5). The right hand with the index finger extended is moved upwards, whereby the back of the hand is rotated – a combined movement that generates the visual impression of a spiral. The corresponding annotation is shown in Figure 4. The feature value *LHH* ("left hand held") tells us that the left hand is not in rest position, but is actually not performing a gesture. Such information is captured

$$
\begin{bmatrix}
\textit{right hand} & \\
\text{handshape} & \begin{bmatrix} \text{shape} & \text{G} \\ \text{path} & 0 \\ \text{dir} & 0 \end{bmatrix} \\
\text{palm} & \begin{bmatrix} \text{orient} & \text{PAB>PAB/PUP>PAB} \\ \text{path} & 0 \\ \text{dir} & 0 \end{bmatrix} \\
\text{boh} & \begin{bmatrix} \text{orient} & \text{BUP>BTB/BUP>BUP} \\ \text{path} & \text{arc>arc>arc} \\ \text{dir} & \text{MR>MF>ML} \end{bmatrix} \\
\text{wrist} & \begin{bmatrix} \text{path} & \text{line} \\ \text{dir} & \text{MU} \\ \text{dist} & \text{D-EK} \\ \text{extent} & \text{small} \end{bmatrix} \\
\text{sync} & \begin{bmatrix} \text{config} & \text{BHA} \\ \text{rel.mov} & \text{LHH} \\ \text{s-loc} & \text{P-R} \\ \text{e-loc} & \text{P-UR} \end{bmatrix}
\end{bmatrix}
$$

Fig. 4: Kinematic representation of the spiral gesture from examples (1), (4) and (5).

as synchronization information (SYNC), which also hosts the starting and ending slot of a gesture movement in gesture space (SLOC respectively ELOC).

In the following, it is shown that kinematic gesture representations can be mapped onto vector sequences within vector space models. Vector space models are instances of spatial theories, that extended standard models with a spatial frame of reference.

3.2 Spatial frames of reference

The basic idea of spatial theories is that linguistic expressions are interpreted in terms of a spatial frame of reference in addition to the standard entity frame of reference. The spatial frame of reference can be understood as an abstract configuration of the space inhabited by the discourse entities (the objects and situations talked about). The spatial theory developed by Giorgolo (2010), for instance, makes use of a rather powerful mereotopological language for modeling the spatial frame. On this account, an iconic gesture is mapped via a procedure ϕ onto an iconic space, that is, a mereotopological rendering of the gesture's kinematic representation. The iconic space bears an equivalence relation (under some fixed perspective) to the spatial frame of reference –

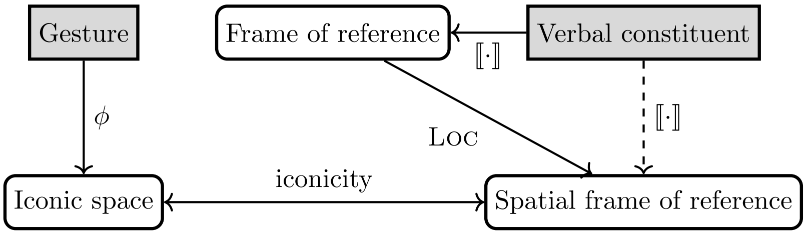


Fig. 5: The process of iconic gesture interpretation within a spatial-frame-of-reference model (reproduced after Giorgolo 2010, p. 53).

capturing the iconic gestures' visual contribution. The integration of speech and gesture is intersective: a gesture assigns the arguments of its verbal affiliate (which can be an n-ary predicate or an n-ary predicate modifier; a grammar of affiliation is presupposed on this account) a common subspace in the spatial frame of reference and requires this subspace to be equivalent to the iconic space, that is, the output of ϕ. An illustration of the architecture of the spatial model of Giorgolo (2010) is given in Figure 5, which provides a useful summary for spatial theories in general.

The approach conservatively extends the truth conditions of multimodal utterances: a multimodal utterance (an utterance consisting of a sentence and an accompanying gesture[14]) is true just in case the sentence is true in the Montagovian frame of reference, and the spatial configuration of the referents is such that it complies to the iconic space (via iconic equivalence between the spatial frame of reference and the output of ϕ). It is dynamic in the sense that a gesture is not lexicalized, in line with insights from gesture studies (see section 1). A gesture is considered a physical act. The biggest downside of Giorgolo's theory, however, is that a spell-out for ϕ is not yet provided. It therefore provides an impressive conceptual and analytical model but falls short computationally.

[14]Since integration in Giorgolo's model happens on the level of predicates or modifiers, it seems to be possible that a sentence may be coupled with several gestures.

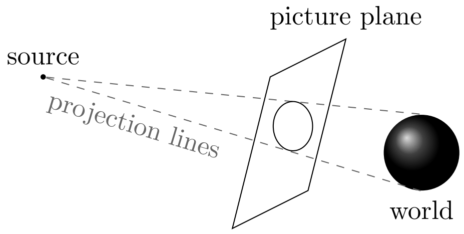


Fig. 6: Geometric projection: The pair of projection source (determining perspective) and picture plane (determining orientation) is a viewpoint. A viewpoint and a world define a scene. The set of all such scenes is the pictorial space, the content of a picture.

3.3 Applying projection semantics?

A truth-functional approach to the meaning of figurative drawings[15] has been developed by Greenberg (2011, 2021) in terms of projection semantics. A projection is a geometric projection in the sense of the artist's technique to construct perspectival paintings.[16] Projection semantics, however, uses geometric projections to *interpret* pictures. A picture expresses a content (a pictorial space), if the picture is a projection of that content (relative to a viewpoint, that is, a pair of projection source and picture plane) – see Figure 6 for a simple example.[17] Projection semantics is semantic because it offers a notion of accuracy: a picture is accurate iff (abbreviates if and only if) there is a viewpoint that provides a geometric projection from the pictorial space to the picture. Viewpoints can straightforwardly be embedded in standard possible world semantics: they provide a geometric system on top of the worlds already part of the model's domains. Further refinements can, or need to, be given in terms of different systems of depiction (in particular "impurely projective" ones like caricature; Figure 6 only shows simple line drawing) and depth constraints, among others (cf. Greenberg, 2021).

[15] "Canonical examples of pictures include architectural and engineering drawings, figurative paintings, functional illustrations sketches and illustrations, photographs, as well as many kinds of maps." (Greenberg, 2021, p. 849)

[16] See, for instance, Albrecht Dürer's *Underweysung der Messung*, in particular illustrated in the final image of the fourth volume (available at https://de.wikisource.org/wiki/Underweysung_der_Messung,_mit_dem_Zirckel_und_Richtscheyt,_in_Linien,_Ebenen_unnd_gantzen_corporen/Viertes_Buch).

[17] Formally: $[\![\mathrm{P}]\!]_{S,c} \subseteq \{\langle w, v\rangle \mid proj_S(w, v) = P\}$ (the denotation of a picture P in context c is a subset of the world–viewpoint pairs $\langle w, v\rangle$ (scenes) such that there is a projection from the world–viewpoint pairs onto P relative to a system of depiction S).

Does projection semantics also provide an adequate framework for analyzing iconic gestures? A mapping between paintings and gestures is obtained straightforwardly: the gesturer and his static pose or dynamic hand and arm movement plays the role of the picture plane which displays the projection source. The content of the gesture, as with pictures, depends on the depiction system to be at work. Given the basic figurative projection function of projection semantics, this means, that the content of a gesturer's gesture is the set of world–viewpoint pairs where a (possibly different, let us assume) gesturer performs that gesture, seen from the viewpoint in question. Such a verbatim interpretation obviously misses the semiotic point of a gesture: the content of a gesture surely is not a look-alike gesturing situation. In other words, projection semantics, when applied to iconic gestures, fails to distinguish between the gesture and the content of the gesture.

Besides this semiotic issue, geometric projections of oriented worlds face perceptual challenges. Consider an iconic gesture that can be interpreted as *rolling*: the index finger rotates in circles while the wrist is moved rightwards. The decisive feature is that there is a part of the rotation movement of the index finger that runs backwards (i.e., in the opposing direction of the wrist movement), as illustrated in Figure 7(c) (see also Bressem, 2013, p. 1088). This configuration, however, is a purely perceptual one; it can never be projected onto a physical movement (Figures 7(a) and (b)), where "going back" is simply impossible. Thus, geometric projections fall short of capturing the semiotic potential of gestures (they only account for the gesturer's movement), and they make wrong predictions with respect to gestures where the perceptual image diverges from its physical origin.

3.4 Basic vector space semantics

If mereotopological models and geometric projections do not live up to the representational and computational requirements for an iconic gesture semantics (as discussed

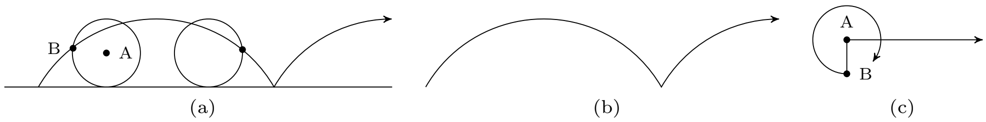


Fig. 7: Perceiving a rolling disc as simultaneous movements of points A and B (a). The movement trajectory of point B in isolation corresponds to a cycloidal trace (b). The shared movement of A and B – a horizontal rightward movement – is the perceptual reference frame for both individual movements. The resulting *percept* is that B circles around A (c). (Reproduced after Johansson, 1973, p. 207)

in sections 3 and 3.3), what kind of model can be employed to aptly interpret visual communication? In various areas of semantics and cognitive science, vector representations have been developed to this end (e.g., Talmy, 1988; Krifka, 1998; Jackendoff, 1991; O'Keefe, 1996). Within formal semantics, a vector space semantics has been introduced by Joost Zwaarts (Zwarts, 1997, 2003; Zwarts and Winter, 2000). For an overview of spatial semantics see Zwarts (2017).

Following Zwarts and Winter (2000), we assume that vectors are primitive spatial entities in natural language models. Vectors are given within a vector space V over the real numbers $\mathbb{R}$.[18] Vector spaces are closed under vector addition and scalar multiplication. Two domains are defined from this ontology:

- The domain of points: $D_p = V$ (each point is defined by a vector's endpoint)
- The domain of vectors: $D_v = V \times V$ (the Cartesian product of V)
- For each element ("point") $w \in D_p$ there is a vector space $V_w \in D_v$ (w is the zero-vector, or centre of V_w).

D_p and D_v are added to the model along with the domain of entities E. Elements from D_p are referred to as points, elements from D_v as vectors. The former are noted as $\mathbf{p}$, $\mathbf{q}$, the latter as $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$.

[18] We assume a standard Euclidean space, but this is not a necessity: spaces defined in terms of polar instead of Cartesian coordinates are formally as well as conceptually potent alternatives (Zwarts and Gärdenfors, 2016).

E (the domain of entities) and D_v are related by a couple of functions, some of which are briefly introduced in the following (we basically follow the vector space model introduced by Zwarts 1997, 2003):

- The vector space located at a concrete object denoted by an NP α is given by 'space($[\![\alpha]\!]_M$)'.[19]
- Place and axis vectors determine spatial relationships:
  - place($x, \mathbf{v}$): x is placed at the end of $\mathbf{v}$; place($\mathbf{v}, x$): the beginning point of $\mathbf{v}$ is placed at x; place($\mathbf{u}, \mathbf{v}$): the beginning point of $\mathbf{u}$ is placed at the end of $\mathbf{v}$ – see Figure 8.
  - axis($x, \mathbf{v}$) object x has an axis $\mathbf{v}$ – see Figure 9.
- Paths: sequences of axis or place vectors, see Figure 10. Paths are defined as a mapping from an interval $[0, 1]$ to vectors. Paths are notated as $\mathbf{v}[k]$. The beginning of a path is indexed as $\mathbf{v}[0]$, its end point as $\mathbf{v}[1]$.[20] Note that paths are non-temporal entities. They receive a temporal interpretation only if index k is mapped to points or intervals in time.

Vectors pointing in the same direction constitute an *axis*. The set of vectors varying only in the direction of one of their three dimensions make up a *plane*.

- The *inverse* $-\mathbf{v}$ of a vector $\mathbf{v}$ points in the opposite direction of $\mathbf{v}$, see Figure 11. The inverse of an axis A is the set of vectors $-A$ pointing in the opposite direction of A.
- The *orthogonal complement* $A^{\perp}$ of an axis or plane A is the set of vectors orthogonal to those in A ($A^{\perp} = \{\mathbf{v} \in D_v \mid \forall \mathbf{w} \in A : \mathbf{v} \perp \mathbf{w}\}$), see Figure 12.

[19] In addition to located vector spaces, also the "bounding box" of objects will be needed (Weisgerber, 2006).

[20] Hence, place($\mathbf{w}, \mathbf{v}$) $\Leftrightarrow$ $\mathbf{w}[0] = \mathbf{v}[1]$. Note that defining paths to the normalized interval $[0, 1]$ diverges from the original model of Zwarts (2003), where real or natural numbers have been used. Normalization is a simplification that allows to address a path's beginning and end points without additional mapping. Note further that for the vectorization of kinematic gesture representations introduced in section 3.5 vector sequences built by concatenating vectors head to tail are more useful. However, we stipulate that any head-to-tail vector sequence approximates a continuous vector sequence, or, conversely, that there is an interpolation for a head-to-tail sequence that equals a continuous vector sequence.

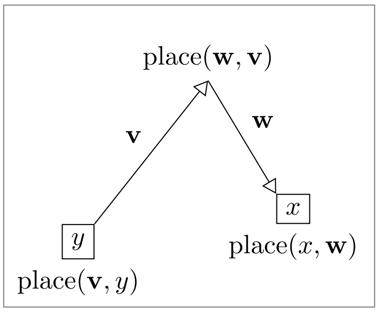


Fig. 8: Place vector

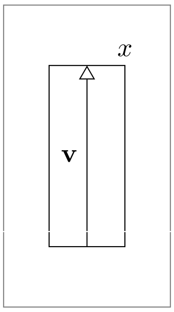


Fig. 9: Axis vector

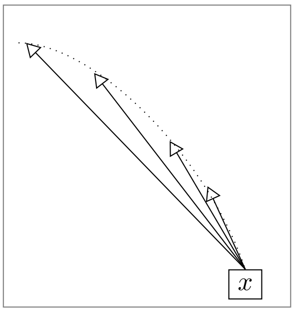


Fig. 10: Place path

- Each vector space V provides three, mutually perpendicular *orienting half-axes.* Intuitively, these axes correspond to the directions *up* (UP), *forward* or *front* (FT), and *right* (RT). These half-axes in addition to their corresponding inverses, give rise to an *oriented vector space*, see Figure 13. The orienting half-axes are determined by external reference frames, or by intrinsic or functional properties of reference objects. Except for the anatomical planes (see section 3.5 below), we will leave open further specifications in this regard.
- Given orienting axes A and A', a vector $\mathbf{v}$ can be decomposed into its *projections* onto the axes, $\mathbf{v}_A$ and $\mathbf{v}'_A$. Figure 14 shows the orthogonal components of $\mathbf{v}$ on the UP and the RT axes.

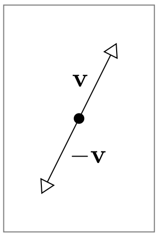


Fig. 11: Inverse of a vector

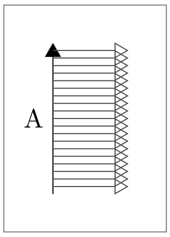


Fig. 12: Orthogonal complements of an axis

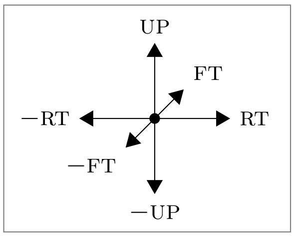


Fig. 13: Orienting half-axes and their inverses

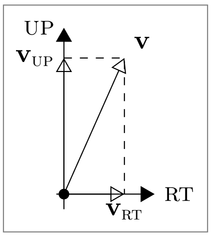


Fig. 14: Projections of a vector

Why do we need vector denotations? Vectors provide a mathematical model for spatial prepositions such as *above*. The differently shaded areas in Figure 15 show three readings of *above* (of a glass), which can all be defined in terms of sets of vectors from the vector space located at the object in question (Zwarts, 1997).[21]

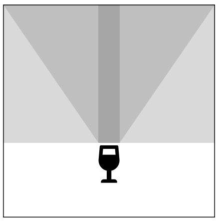


Fig. 15: $above(\mathrm{NP}) = \{\mathbf{v} \in \mathrm{space}([\![\mathrm{NP}]\!]) \mid \ldots\}$.

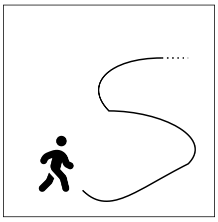


Fig. 16: $\lambda x.\lambda e[\mathrm{walk}(e, x) \wedge \exists \mathbf{v}[\mathrm{path}(e, \mathbf{v})]]$.

Vectors (that is, spatial entities) are also part of the truth conditions for movements in space. For instance, a walking event is a walking event only if there is a path of motion, be it undirected, as in Figure 16, or directed (Krifka, 1998):[22]

(8) a. Mary walked from the university to the capitol.

b. $\lambda e \exists x[\mathrm{name}(x, \text{“Mary”}) \wedge \mathrm{walk}(e) \wedge \mathrm{agent}(e, x) \wedge \mathrm{source}(e, u) \wedge \mathrm{goal}(e, c) \wedge \exists \mathbf{v}[\mathrm{place}(\mathbf{v}, u) \wedge \mathrm{place}(c, \mathbf{v})]]$

The sentence in (8a) is true if the walking event in question took place, in which case there is a path $\mathbf{v}$.

Oriented vector spaces have also been used to spell out a spatial semantics for pointing gestures (Lücking, 2022), in line with current (that is, post-Kaplanian) pointing cone construals (Bangerter and Oppenheimer, 2006; Kranstedt et al, 2006b; Lücking et al, 2015; Lücking, 2018). An illustration is shown in Figure 17.

[21] A more elegant solution is to define a probability distribution over the space surrounding the object denoted by the NP (O'Keefe, 1996).

[22] Since we need some notion of events, we make use of a neo-Davidsonian semantics in the style of Parsons (1990). On the integration of event semantics and Montague/Frege ones see Champollion (2015).

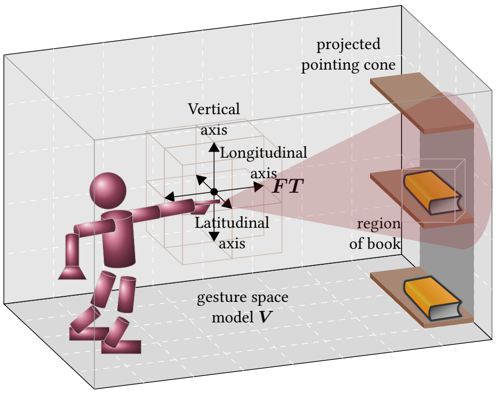


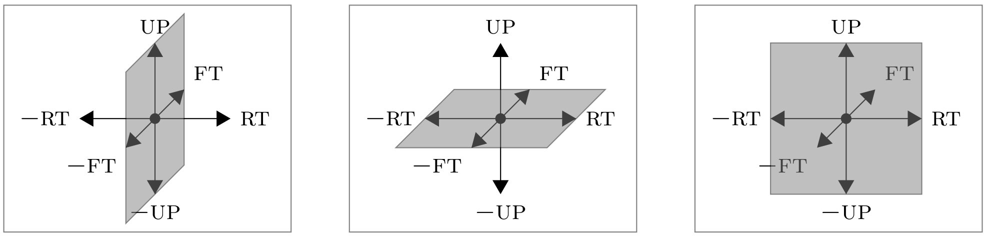
Fig. 17: Illustrating a vector space semantics for pointing gestures. The manikin points straight ahead, thereby excluding the bottom book from the set of potential referents. The space "highlighted" by the pointing gesture is defined in terms of vectors emanating from the index finger (the image is adapted from Lücking, 2022, p. 112; the book icon is taken from LaTeX's *beamer* class, Miletić et al, 2024).

In sum, vector spaces have already proven to be adequate models for spatial language as well as non-verbal demonstration acts. In the following, we show how vector spaces can be extended to provide a formal model to interpret iconic gestures, too.

3.5 Iconic models in vectorial gesture space

A main obstacle for spatial theories is to provide a mapping ϕ from kinematic gesture representations to iconic models (cf. section 3) – that is in our case to vector sequences in vector space. Such a mapping is needed for an iconic gesture semantics to get off the ground in the first place. In the following, we provide such a mapping, drawing on symbolic-computational work by Lücking (2016). The basic idea is twofold but simple:

- The notion of oriented vector space is used as a formal model of a speaker's gesture space.
- The annotation predicates that represent gesture kinematics (section 3.1) are re-interpreted in terms of vector sequences in vectorial gesture space.



(a) Sagittal plane (b) Transverse plane (c) Coronal plane

Fig. 18: Oriented planes in vector space.

This two-step process gives rise to a computational procedure to derive vector-based iconic models from the form of a gesture – the first characterizing feature of iconicity (the second one is the intersection of these iconic models with the models of the spatial configurations of the objects and events talked about, cf. section 1).

A vector space space(s) is anchored at each speaker s (that is, s is placed at the origin of V, place(s, V)). This vector space is oriented along the anatomical planes which hypothetically transect the speaker's body, see Figure 18:

(9) Anatomical planes

a. *Sagittal plane* $:= \{\mathbf{v} \in D_v \mid \mathbf{v} \perp \text{UP} \cup \text{RT} \vee \mathbf{v} \perp \text{-UP} \cup \text{-RT}\}$

b. *Transverse plane* $:= \{\mathbf{v} \in D_v \mid \mathbf{v} \perp \text{FT} \cup \text{UP} \vee \mathbf{v} \perp \text{-FT} \cup \text{-UP}\}$

c. *Coronal* (or *Frontal*) *plane* $:= \{\mathbf{v} \in D_v \mid \mathbf{v} \perp \text{RT} \cup \text{FT} \vee \mathbf{v} \perp \text{-RT} \cup \text{-FT}\}$

Speaker-centered vector spaces provide a *semantic model for gesture spaces*, that is, the space roughly in front of a speaker's chest where hand and arm movements are carried out.

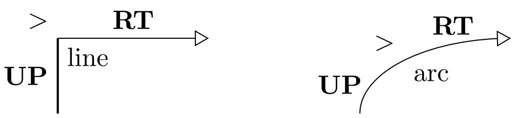
With the formal notion of gesture space as vector space, a mapping from kinematic gesture representation of simple, uni-directional movements to vectors is straightforward: the directions encoded in gesture annotation predicates are translated to orienting vectors, as spelled out in (10).

(10)

annotation	mnemonic	vector
towards TL	towards left	$-\mathbf{RT}$
TR	towards right	$\mathbf{RT}$
UP	upwards	$\mathbf{UP}$
DN	downwards	$-\mathbf{UP}$
TB	towards body	$-\mathbf{FT}$
AB	away from body	$\mathbf{FT}$

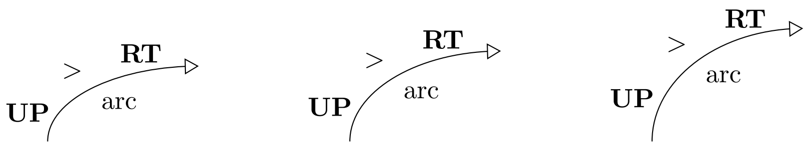
The mapping from (10) allows deriving an iconic model from a simple movement of the hand, for example, from left to right: **RT**. Moving on to vector sequences, vectors are concatenated head-to-tail, while the concatenation obeys the trajectory annotation *line* vs. *arc*, which distinguishes roundish from angular paths (see section 3.1). A minimal example is shown in (11), where the iconic models emerging from vector sequence $\mathbf{UP} >_{\text{line}} \mathbf{RT}$ respectively $\mathbf{UP} >_{\text{arc}} \mathbf{RT}$ are given.

(11)



The resulting vector sequences, $\mathbf{Z} = \mathbf{UP} >_{\text{line}} \mathbf{RT}$ and $\mathbf{Z}' = \mathbf{UP} >_{\text{arc}} \mathbf{RT}$, are both open, that is $\mathbf{Z}[0] \neq \mathbf{Z}[1]$ and $\mathbf{Z}'[0] \neq \mathbf{Z}'[1]$. Distinguishing open and closed trajectories is important and is brought about by comparing the starting and ending positions of vector trajectories in vectorial gesture space: both are part of kinematic gesture representations (cf section 3.1). This procedure – the annotation of gesture and form and its mapping onto vector sequences – involves some degree of abstraction. For instance, the sequence $\mathbf{Z} = \mathbf{UP} >_{\text{line}} \mathbf{RT}$ is compatible with numerous iconic models and does not distinguish the following ones:

(12)



The procedure, however, constrains the iconic model along the lines of (12) to be a bent trajectory (with a certain orientation, namely starting upwards, going to the right). This degree of abstraction takes into account the "sloppiness" of gesture performance (e.g., Rieser, 2011).

We notate $>_{\text{line}}$ as $\perp$ and $>_{\text{arc}}$ as $\circ$. Together with the vectorization of basic predicates in (10), there is a complete vectorization of kinematic gesture representation, that is, a spell out of function ϕ from spatial theories. The algebraic gesture derivation 'vec(γ)' is given in (13).

(13) Gesture vectorization function vec(γ) as model for ϕ.

a. $\text{vec}(\mathbf{u} >_{\text{line}} \mathbf{v}) = \begin{bmatrix}\text{traj} = \mathbf{u} \perp \mathbf{v}\end{bmatrix}$

b. $\text{vec}(\mathbf{u} >_{\text{arc}} \mathbf{v}) = \begin{bmatrix}\text{traj} = \mathbf{u} \circ \mathbf{v}\end{bmatrix}$

c. $\text{vec}(\text{s-loc,e-loc}) = \begin{cases}\text{sync : traj[0]} = \text{traj[1]} & \text{if s-loc} = \text{e-loc} \\ \text{sync: traj[0]} \neq \text{traj[1]} & \text{else}\end{cases}$

The input of 'vec' is a kinematic gesture representation γ as introduced in section 3.1.[23] Vectorization applies progressively over movement annotations (13a,b). Condition (13c) is the closure condition, which checks whether a given movement trajectory brings about a closed or an open path. Such information is necessary, for instance, for shape interpretations such as distinguishing a full circle from a four-fifths one.

Obviously, more interpretational complexity can be built on top of vector interpretations. For current purposes, the system developed so far is sufficient: we have

[23]Obviously, 'vec' takes arguments of different annotation types. This move is not uncommon in programming languages, and we stick to it for brevity's sake.

introduced a basic computational procedure for spelling out ϕ, the so far missing function that produces iconic models from kinematic gesture representations. Before detailing the semantics of gesture in relation to speech, let us briefly exemplify the interpretation of the spiral gesture shown in (1), (4) and (5). The new property here is that it actually involves two motion aspects: an internal (rotation) and a translational (going upwards) one. The corresponding annotation of the right hand is shown in (14).

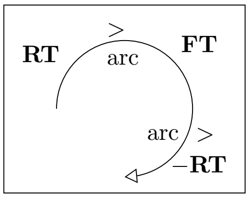
(14)
$$\begin{bmatrix} \textit{right hand} & \\ \text{handshape} & \begin{bmatrix} \text{shape G} \end{bmatrix} \\ \text{boh} & \begin{bmatrix} \text{path} & \text{arc>arc>arc} \\ \text{dir} & \text{MR>MF>ML} \end{bmatrix} \\ \text{wrist} & \begin{bmatrix} \text{path} & \text{line} \\ \text{dir} & \text{MU} \end{bmatrix} \\ \text{sync} & \begin{bmatrix} \text{s-loc} & \text{P-R} \\ \text{e-loc} & \text{P-UR} \end{bmatrix} \end{bmatrix}$$

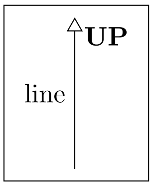
The two movements involved in the spiral gesture introduce two vector sequences. The translational one determines an open path since the starting and the end position differ. The output of the vectorization function (13) is given in (15):

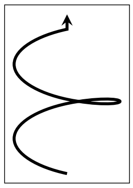
(15)
$$\begin{bmatrix} \text{traj1} = \mathbf{RT} \circ \mathbf{FT} \circ -\mathbf{RT} \\ \text{traj2} = \mathbf{UP} \\ \text{sync} : \ \text{traj2}[0] \neq \text{traj2}[1] \end{bmatrix}$$

The "vectorization" of the back of the hand (BoH) rotation gives rise to the iconic model to the left in (16), the wrist movement introduces a straight vector pointing upwards in gesture space. The *combined* movement of BoH and wrist results in a wound, spiral model.[24]

[24] Since the iconic model, in this case, is the result of a blend of two movements, one might think of such gestures as portmanteau gestures (Lücking, 2013a, 193 sqq.).

(16) BoH: 

Wrist: 

Combined: 

Let us take stock: following spatial theories of iconic gesture meaning, we formally modelled the visual content of an iconic gesture in terms of a (possibly dynamic) spatial structure, the iconic model. A crucial contribution is a spell-out for mapping ϕ, a function that brings about a systematic construction of iconic models from kinematic gesture representations. How are iconic models related to speech (the second aspect of iconicity)? Following ideas developed by Giorgolo (2010), speech and gesture are basically related in terms of intersectivity: the spatial configuration of the objects and events talked about is such that it intersects with the iconic model derived from the gesture. To complete our vector space semantics of iconic gestures, we assume a linear map between iconic models and real-world spatial configuration. A linear map is a mapping $V \mapsto W$ of vector spaces V and W that preserves the operations of vector addition and scalar multiplication. Hence, iconicity in vector spaces allows for some linear "distortion" between iconic models and spatial configurations. The most obvious one is the linear transformation of scaling: a gestural movement does not necessarily depict in life-size, so the linear map allows scaled mappings, both larger and smaller. But also rotation and reflection are possible, both are bound up with perspective.

We are now in a position to spell out a compositional semantics for speech–gesture integration. As the compositional "backbone" we assume a multimodal grammar along the lines of Alahverdzhieva et al (2017); Lücking (2013a). Such grammar extensions combine an iconic gesture with its lexical or phrasal affiliate. Schematically, a multimodal utterance $\alpha[\beta/\gamma]$ consisting of a sentence α, a co-speech gesture γ and its affiliate β is true, iff α is true and there is an intersection of the iconic model of γ and the spatial configuration ‘space(β)’ projected from β.[25] Given a standard Montagovian model, which is extended by vector spaces as outlined in section 3.4, the semantics of multimodal utterances is compositionally derived straightforwardly: α, including constituent β, is recursively interpreted as usual (cf. Heim and Kratzer, 1998; Chierchia and McConnell-Ginet, 2000). Additionally, the vector space projected from the verbal affiliate is evaluated against the linear maps from the vector space obtained from the vectorization of the iconic gesture. The recursive, truth-functional interpretation of a multimodal utterance within a simple quantifier-free, first-order fragment involving transitive verbs is spelled out in (17). The only new, multimodal rule is given in (17d).

(17) *Semantic fragment for multimodal utterances.*

a. $\llbracket[_{\mathrm{A}}\mathrm{B}]\rrbracket = \llbracket\mathrm{B}\rrbracket$, for any lexical category A and lexical entry B.

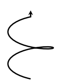
b. $\llbracket[\mathrm{NP}\ \mathrm{VP}]\rrbracket = 1$ iff $\llbracket\mathrm{NP}\rrbracket \in \llbracket\mathrm{VP}\rrbracket$.

c. $\llbracket[\mathrm{V}\ \mathrm{NP}]\rrbracket = \{x \mid \langle x, \llbracket\mathrm{NP}\rrbracket\rangle \in \llbracket\mathrm{V}\rrbracket\}$

d. $\llbracket[_{\mathrm{MM}}\mathrm{A}\ \gamma]\rrbracket = \{f \mid f : \mathrm{vec}\,\gamma \mapsto \mathrm{space}(\llbracket\mathrm{A}\rrbracket)\}$, for a multimodal category MM,[26] an iconic gesture γ, and a linear map f.

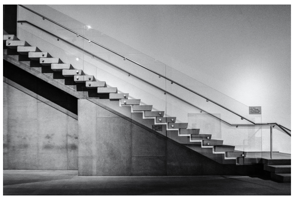
A multimodal utterance $\alpha[\beta/\gamma]$ including an iconic gesture γ and its affiliate β is true iff $\llbracket[\alpha]\rrbracket = 1 \wedge \exists f.f : \mathrm{vec}(\gamma) \mapsto \mathrm{space}(\llbracket[\beta]\rrbracket)$.

Consider, for example, a sentence that includes the noun *staircase* and an accompanying spiral gesture γ. The denotation of *staircase* in some worlds is like in Figure 19a and like Figure 19b in other worlds.

The axis path of the staircase in Figure 19a within its vector space ‘space($\llbracket$staircase$\rrbracket$)’ is , shown in Figure 20a; the one of the staircase in Figure 19b is , shown in Figure 20b. A linear map with the domain of the spiral gesture,

[25]For the sake of simplicity, we assume that a multimodal utterance involves just one gesture, but the set-up scales up straightforwardly.

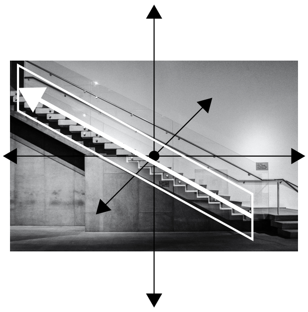
[26]From multimodal grammars, that is, a situated word (Alahverdzhieva et al, 2017) or a multimodal ensemble (Lücking, 2013a).



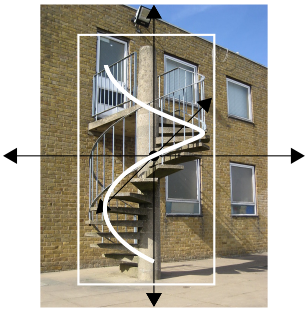
(a) *Grand Staircase* (Robert N. Clinton at www.cybershutterbug.com, CC BY-NC-ND 4.0 DEED).

(b) *Concrete Spiral Staircase* (Pat Joyce at www.flickr.com, CC BY-NC 2.0 DEED).

Fig. 19: Two kinds of staircase denotations.



(a) Axis path of *Grand Staircase* within its oriented vector space.



(b) Axis path of *Concrete Spiral Staircase* within its oriented vector space.

Fig. 20: Axis vectors within 'space($[\![$staircase$]\!]$)'.

$$\mathrm{vec}\left(\begin{bmatrix} \textit{right hand} & \\ \text{handshape} & \begin{bmatrix}\text{shape G}\end{bmatrix} \\ \text{boh} & \begin{bmatrix}\text{path} & \text{arc>arc>arc} \\ \text{dir} & \text{MR>MF>ML}\end{bmatrix} \\ \text{wrist} & \begin{bmatrix}\text{path} & \text{line} \\ \text{dir} & \text{MU}\end{bmatrix} \\ \text{sync} & \begin{bmatrix}\text{s-loc} & \text{P-R} \\ \text{e-loc} & \text{P-UR}\end{bmatrix}\end{bmatrix}\right) = \begin{bmatrix}\text{traj1} = \mathbf{RT} \circ \mathbf{FT} \circ -\mathbf{RT} \\ \text{traj2} = \mathbf{UP} \\ \text{sync}: \ \text{traj2}[0] \neq \text{traj2}[1]\end{bmatrix},$$ only exists for the latter, but not the former staircase's axis path. Hence, the spiral gesture is compatible

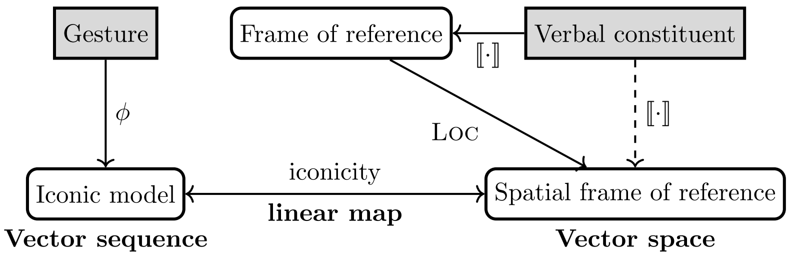


Fig. 21: Vector space semantics as spatial theory

with worlds where ⟦staircase⟧ is a singleton or a proper set of objects of the kind displayed in Figures 19b and 20b, respectively, but not of the kind shown in Figures 19a and 20a, respectively. Hence, spatial theories give rise to a well-behaved, truth-functional semantics of iconic gestures as a conservative extension of a Montagovian semantic framework. Note that (17d) provides us with a first formal notion of *multimodal well-formedness*, requested by desideratum D2 in section 2: a speech–gesture mismatch occurs if there is no linear map f which embeds the iconic model into the spatial configuration projected from the verbal constituents. The outline of a vector space-based iconic gesture semantics is shown in Figure 21, where the instances of the bottom row model components are given as boldface labels. The semantics captures the visuo-spatial contribution of iconic gestures and is semantic since it provides truth conditions. Given the spell-out of function ϕ, the semantics is also compositional. Having established a solid semantic backbone, we turn to how it interacts with labelling theories.

4 Perceptual classification and extemplification

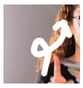
> "On the one hand, the Frege/Montague research program, based on the idea that truth-conditions are the core ingredient of clause meaning and that meanings of complex expressions are computed from the meanings of the parts, has been extremely successful. On the other, it did not really address the central question: What, precisely, are the meanings of the smallest parts, the meanings of words, or rather, lexemes?"
>
> Manfred Krifka (2012, p. 223)

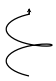
The spatial, vectorial theory of iconic gesture meaning harmonizes iconic gestures with truth-conditional semantics. However, just as semantics in the Frege/Montague tradition cannot provide an answer to the question "What *is* the meaning of ⌜a certain word⌝?", it also lacks an answer to the question "What *is* the meaning of ⌜a certain gesture⌝?" We will see that an answer to the first question also provides a solution to the second one.

But when is meaning at stake in communicating with gestures at all? There are at least two kinds of situations: (i) gesture uptake (Gullberg and Kita, 2009) and clarification interaction (Ginzburg and Lücking, 2021b); (ii) the analysis of gestures within gesture studies (e.g., Hadar, 2013). Note that both kinds of situations involve quite different perspectives: the first one is the perspective of the interlocutors in interaction, and the second one is that of the researcher studying this interaction. Let us briefly turn to each of them.

Somewhat surprisingly, addressees do not always informationally evaluate the speaker's gestures. This has been tested, for instance, by Gullberg and Kita (2009) in a drawing response study, where participants had to draw a situation that they saw described in a video of a speaker using speech and gesture. The speaker's gesture included a target gesture, that is, a gesture that displayed information not verbalized in speech (e.g., the direction of a movement). The authors found that the drawings included the information exclusively gestured more often if the speaker gazed at the target gesture. Hence, interlocutors themselves make a distinction between (mostly peripheral) *seeing* a gesture and *interpreting* a gesture (gesture uptake, or informational evaluation, to which we turn shortly).[27]

Iconic gestures can also be part of clarification interaction, which is of interest from a dialogue semantics point of view (Healey et al, 2015). Consider again the spiral gesture example, re-given in its English translation in (18). Following A's utterance, addressee B can inquire about the gesture by either reproducing it (1), or by informationally evaluating it (2).

(18) A: I think that should be  staircases

B: (1)  ? (2) Do you mean spiral staircases?

B's two kinds of responses correspond to two different clarification strategies: confirmation questions and intended meaning requests.[28] In (1), B asks to confirm the perceived gestural movement, or the intended iconic model, respectively. Thus, (1) is a nonverbal variant of verbal "Have I heard correctly? Did you say u", or "Do you mean u?", for some verbal constituent u. This reading does not seem to be available for (2), however, which addresses the intended meaning of the gesture: "Do you mean z as the content of γ?"[29]

In gesture studies and Conversational Analysis, *describing* what the hands do is a longstanding and useful way of characterizing a gesture's meaning. For instance, "Palm-Up-Open-Hand is recurrently used to communicatively present, give, offer, show

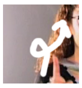
[27] This is reminiscent of the twofoldedness of gestures pointed out in section 1: seeing gestures (iconic model) and seeing something in gesture (informational evaluation).

[28] On different types of clarification requests see Ginzburg (2012, §6.2).

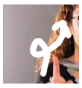
[29] This confirms, to our minds, the diagnosis of Giorgolo (2010, p. 91 sq.), namely that the "meaning" (i.e., spatial model) of iconic gestures has to be kept apart from communicating with these gestures.

a discursive object" (Müller, 2017, p. 279). A certain gesture form is labelled to perform a certain (range of potential) action(s), and based on this description, the informal analysis of the gesture and its relation to speech can be given.

This points to an important, albeit trivial-sounding fact, namely that *the interpretation of a multimodal utterance is conditioned on the informational evaluation of the gesture involved.* If this is true, then a gesture fails to contribute content, *unless* interlocutors agree on an informational evaluation of the gesture. Using the nondeniability test for gestures (Ebert, 2014; Schlenker, 2019), (19) and (20) show that this is indeed the case.[30]

(19) a. A: Inside the hall was an imposing  staircase.

b.# B: No, that's not true. The staircase was actually straight.

(20) a. A: Inside the hall was an imposing  staircase.

b. B: Do you mean a spiral staircase?

c. A: Yes.

d. B: No, that's not true. The staircase was actually straight.

Non-deniability is a feature of non-at-issue contributions. However, other contexts induce non-deniable contents, too, in particular the antecedents of indicative conditional sentences. Their consequences cannot be picked out by negation: The negation of a sentence of the form "If A then C" is either the conjunction "A and not C" or the conditional "If A then not C" (cf. Egré and Politzer, 2013).[31] A denying continuation of an indicative conditional targets the asserted implication:

[30]The quoted authors conclude from these tests that gestures contribute non-at-issue content. However, this follows only if failing to contribute at-issue content is the same as contributing non-at-issue content, as has been pointed out by Hunter (2019, p. 326 sq.). We do not subscribe to this implicit assumption, since according to spatial gesture semantics the gesture in this case simply remains sub-linguistic, visual – semantically inert.

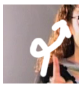
[31]A weak version of denial of indicative conditionals has "[. . .] then possibly not C" as consequence.

(21) A: If the staircase is spiral, it is an imposing one.

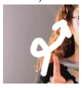
a.# B: No, that's not true. The staircase is imposing.

b. B: No, that's not true. The staircase is imposing even without being spiral.

Hence, we would expect contexts of conditioned, but not explicitly agreed, gesture interpretation to involve nondeniable consequences. Accordingly, (22) is in line with the predictions of our theory:

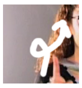
(22) *If*  *is interpreted as "spiral", then* in the hall was an imposing spiral staircase.

a.# No, that's not true. The staircase was actually straight.

b. No, that's not true. The staircase was actually straight, even if you interpret  as *spiral*.

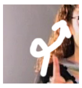
Things are different if the antecedent condition is fulfilled, which happens if the gesture is lifted to a quasi-linguistic status due to explicitly agreed informational evaluation, as in (20). From "If A then C" and "A", "C" follows and be negated.

Ebert (2024) distinguishes two additional non-at-issue tests for co-speech gestures, projection, and ellipsis. Let us turn to each in turn. The first one tests the fact that non-at-issue contents project across sentential operators:

(23) a. It is not the case that in the hall was an imposing  staircase.

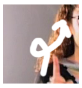
b.# No, the staircase was actually straight.

Again, this behaviour is explained by conditioned interpretation:

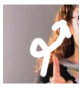
(24) a. *If*  *is interpreted as "spiral", then* it is not the case that in the hall was an imposing spiral staircase.

b.#No, the staircase was actually straight.

The second test says that the co-speech gesture contribution is ignored in ellipsis constructions:

(25) In the hall was an imposing  staircase, and a window, too.

Ignoring for now that resolving elliptical constructions is a complex process in itself (see, e.g., Ginzburg and Cooper, 2004), then a conditional context produces a straightforward interpretation of the multimodal utterance, where *spiral* does not need to take scope over *window*:

(26) *If*  *is interpreted as "spiral", then* in the hall was an imposing spiral staircase, and a window, too.

Hence, iconic gesture semantics can *explain* observations concerning the information status of iconic gestures, namely in terms of conditioned interpretation.[32]

Schematically, using again the template $\alpha[\gamma/\beta]$ for a multimodal utterance α which includes an affiliate β and the affiliated gesture γ:

(27) *Information-evaluation conditioned interpretation*:

If a gesture is interpreted to mean m, then the utterance is interpreted as $\alpha[R(m, \beta)]$.

'$R(m, \beta)$' in (27) means that m, the supposed gestural meaning, is applied to the meaning of its affiliate β via a relation R. In the simplest case, R is identity, namely when m extendedly exemplifies β (in a sense explained shortly in section 4.2) and is informationally evaluated as β. This is the standard case of affiliation, in which

[32] Some authors had the intuition that the ellipsis test brings about a different result for pro-speech gestures (Schlenker and Chemla, 2018). As will be discussed in section 5 by the example of non-lexicalized iconic models, we show that there is no clear difference in the information status of co- and pro-speech gestures. The latter, when taken to be produced in purpose for the sake of communication (i.e., as *foreground* not as *background* gesture; Cooperrider, 2017) may invoke a stronger obligation for informational evaluation, potentially explaining the gradience of information status (Barnes and Ebert, 2023).

the gesture intuitively "reduplicates" its affiliate, see examples (6a,b) in section 2 – and this is the reason why gestures, even if informationally evaluated, mostly remain semantically vacuous. Only if $m \neq \beta$, as in the spiral staircase example, an inference step is needed to combine gesture interpretation and affiliate meaning. Here, standard mechanisms from dynamic semantics resolving R apply, as is pointed out in section 4.3.

In the following, we develop a labelling approach based on visual perception theories and perceptual classification. The objects of classification will be iconic models, thereby reconciling spatial and labelling theories. To make this work, we effect a non-trivial modification of the architecture of the semantic framework: we employ vector models not only as visual denotations, but also as representations of *the meaning of lexical items*. Having the same kinds of representations available for denotations, iconic models, and meanings paves the way for spelling out a computational approach to extended exemplification/classification and the (in)congruence between words and gesture interpretations beyond denotational vacuity (see section 3). The architecture of the lexical, classifier-based extension of spatial theories is shown in Figure 1, re-given as Figure 22.[33]

We motivate the use of vector intensions from biological motion perception in section 4.1, which provides the conceptual interface for *computing* exemplification relations.[34] We then develop an exemplification heuristic as a "linguistic toolkit" for semanticists working on gestures (section 4.2). In section 4.3, we briefly show how

[33]The iconic gesture semantics theory is, of course, not a theory of speech–gesture production. However, it should be emphasized that it is compatible with most multimodal production models (apart from the fact that it incorporates empirical findings on (lack of) informational evaluation, see Gullberg and Kita (2009) and the above discussion). For instance, the *Sketch Model* (de Ruiter, 2000) assumes an abstract spatio-temporal representation alongside verbal ones. This clearly corresponds to our separation of iconic models and spatial frames of reference, respectively the lexicon. The *Lexical Access Model* (Krauss et al, 2000) emphasizes a gesture's facilitation of word retrieval (see also Hadar, 1989); the underlying connection between iconic models and verbal items is precisely captured by the extemplification relation. The *Interface Model* rests on a continuous "negotiation process" between verbal and gestural production channels (Kita and Özyürek, 2003). The iconicity relation mediates between these two generation streams, but as it stands is not a temporal, incremental process. It is less clear to our minds, however, how iconic gesture semantics relates to the heterogeneous "multimodal idea units" postulated by the *Growth Point Model* (McNeill and Duncan, 2000). In any case, we believe that it is an advantage that our iconic gesture semantics is compatible with a wide range of empirical findings and cognitive, psycholinguistic models.

[34]Hence, we relate semantics to non-linguistic cognitive activities such as perception, but contra Lewis (1970, p. 19), we do not think that "confusion comes of mixing these two topics".

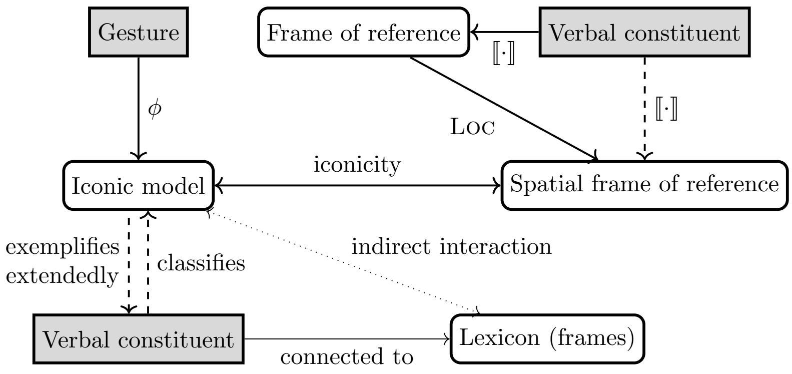


Fig. 22: Classifier-based iconic gesture semantics

gesture integrates with speech if relation R from (27) is not identity by re-using standard methods from dynamic semantics. Extemplification is applied in section 4.4 to showcase semantic analyses of some constructed, but widely discussed examples.

4.1 Excursus: Vectorial biological movement analysis and perceptual classification

Using geometric representations as representations for meanings is not new in natural language semantics and the philosophy of language (e.g., Warglien et al, 2012; Zwarts and Gärdenfors, 2016).[35] In cognitive science, it is commonplace that lexical items have both symbolic and visual meaning components (Paivio, 1986). A related dual coding approach has already been developed for the synthesis of three-dimensional iconic gestures in terms of *imagistic description trees* (Sowa, 2006). Computational linguistics developed "words-as-classifiers" approaches (Kennington and Schlangen, 2015; Larsson, 2015), where the meaning of perception-related words is perceptually grounded (Harnad, 1990; Steels and Belpaeme, 2005). Such perceptual groundings – like speaker judgments (O'Keefe, 2003) – can even be visualized (as illustrated in

[35] Chalmers (2002) makes the case for *epistemic intensions*, but does not provide a formal (e.g., vector based) model.

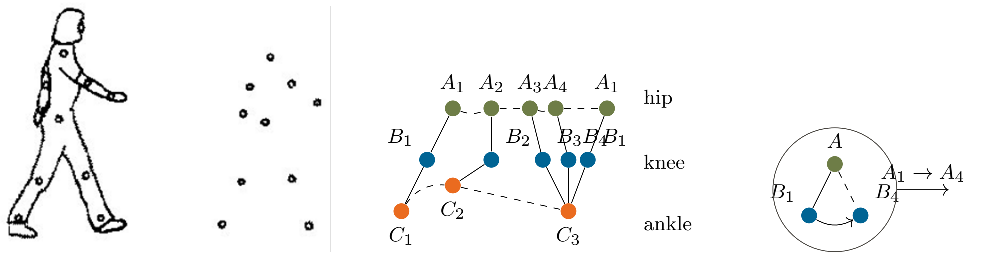
Figure 15). Bartsch (1998) developed a comparable approach in semantics and philosophy of language. Following this direction of thrust, it seems appropriate to underpin visual communication with insights from visual perception – ultimately, gestures are physical actions. Let us consider the example of biological motion and motion verbs. Motion verbs vary along two dimensions: manner and path (Engelberg, 2000). The eigenmovement distinguishes motion verbs according to manner, regardless of the distance travelled:

(28) $$\left\{ \begin{array}{c} \text{run} \\ \text{walk} \\ \text{stroll} \\ \text{saunter} \\ \dots \end{array} \right\}$$

Translational movement gives rise to a path that distinguishes motion verbs irrespective of the manner of motion:

(29) $$\left\{ \begin{array}{c} \text{run} \\ \text{detour} \\ \text{circle} \\ \text{criss-cross} \\ \dots \end{array} \right\}$$

The path component – the translational dimension of motions – is already covered by the vector denotations within the spatial model (see section 3), which truth-conditionally distinguish the verbs in (29). But what about the manner dimension? In an important series of experimental studies on the perception of biological motion, Johansson (1973, 1976); Johansson et al (1980) investigated, among others, the difference between *walking* and *running*. How are we able to consistently tell both motion manners apart across different human (and presumably some non-human) individuals? Is there an abstract perceptual commonality between running events on the one hand and walking events on the other hand? Johansson and colleagues placed little lights at the anatomical joints that actually bring about the movement (the so-called *motion* *carriers*) – see Figure 23a. The recorded light pattern (Figure 23b) is then shown to participants, who "saw" and correctly classified a walking event. "How can 10 points moving simultaneously on a screen in a rather irregular way give such a vivid and definite impression of human walking?" (Johansson, 1973, p. 204). An answer to this question was found in geometric analyses of the temporal stimulus pattern. Walking is characterized by two horizontal trajectories (due to hip and knee carriers) and an up-and-down sequence (ankle) – see Figure 23c. Factoring out common movement shares, the kernel percept of a walking event is the *abstract vector model* shown in Figure 23d. If we observe something that looks like this vector model, we can classify it as *walking*.



(a) Stimulus for *walking*

(b) Carriers of *walking* movement

(c) Model of carrier movement

(d) Abstract vector model of *walking*

Fig. 23: Motion perception: *walking*. (Taken from, respectively reproduced after Johansson, 1973, p. 202 and 208)

Now singling out walking events is exactly what the meaning of the verb *walk* is supposed to achieve, and what is "pre-compiled" in model-theoretic semantics.[36] Accordingly, the model in Figure 23d provides a representation of the intensional meaning of *walk*. We, following Lücking (2013a), refer to this visuo-spatial representation of an intension as *conceptual vector meaning*, or CVM for short. Arguably, the meaning of any visuo-spatial expression comes with a CVM (cf. *dual coding*; Paivio, 1986). As above indicated, CVMs are studied in cognitive science, and computational

[36]As is common practice in semantics, we do not distinguish sharply between possible worlds semantics and model-theoretic semantics, since it is usually of no practical significance and does not impact the present discussion, but see Zimmermann (2011, 2019) for a more careful discussion.

and lexical semantics as *perceptual classifiers*, but they are still alien to compositional semantics in the Frege/Montague tradition (we briefly return to this issue in the conclusions in section 6).[37] However, we believe and have argued at some length, that perceptual classification is key to understanding iconic meaning in the sense of the informational evaluation of vector models acted out by iconic gestures. Hence, to make iconic gesture semantics accessible to mainstream formal semantics, we therefore propose an *extended exemplification heuristic* for gesture interpretation. The heuristic will be elaborated in the following section, but the main idea is as follows. To get things started, the meaning of the lexical entry in (30a) is replaced by one involving a CVM, as illustrated in (30b).

(30) a. $[\![\text{walk}]\!] = \lambda x.\lambda e[\text{walk}(e) \wedge \text{agent}(e, x) \wedge \exists \mathbf{v}[\text{place-path}(e, \mathbf{v})]]$

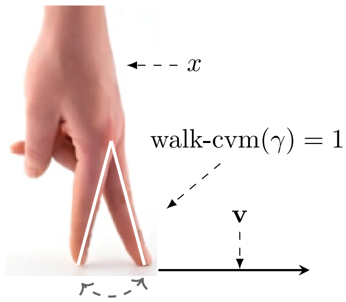
b. $[\![\text{walk}]\!] = \lambda x.\lambda e[\text{walk-cvm}(e) = 1 \wedge \text{agent}(e, x) \wedge \exists \mathbf{v}[\text{place-path}(e, \mathbf{v})]]$

(30a) is standard: the meaning of *walk*, $[\![\text{walk}]\!]$, is spelled out in terms of a function of type $\{0,1\}^{D_s^{D_e}}$. (30b) adds that the set of events E characterized by the embedded function is such that the classification of each event $e \in E$ according to the perceptual walking classifier WALK-CVM is successful.[38] The point of (30b) is that it applies to iconic gestures, too: the lexical meaning can be used to interpret a gestural movement that looks like walking as walking, as exemplified in (31).

[37] There is some related work in the framework of conceptual semantics, however, which, among others, addresses the problem of how we are able to talk about what we see (Jackendoff, 1987). On this approach, a theory of language is connected to a theory of vision, namely the 3-D models of Marr (2010). Marr's model is specialized for object recognition and identification, vector models seem to be better suited for capturing biological movement (but see Marr and Vaina, 1982). A semantic rendering of Marr's model also underlies the logic of vision of van der Does and van Lambalgen (2000), but it remains purely extensional.

[38] Computationally and conceptually, a probabilistic version of the classifier is more attractive, but for the sake of keeping things simple we assume a binary one, just returning 'true' (1) or 'false' (0).

(31)



The gesture in (31) is compatible with *walk*, but fails for, e.g., *stagger*, *crawl*, *give*, *ride*, etc. because of different, incompatible CVMs.[39] Given the successful informational evaluation of the moving fingers in terms of *walking*, this predicate figures as input to the gesture-conditioned affiliation principle (27) and provides an interpretation of the multimodal utterance of which the walking gesture is part of.

4.2 Linguistic toolkit: Extended exemplification as gesture interpretation

As a simple example, consider the toy model in (32a). The denotation of *green* is the set of three green objects. Given this, any object within the denotation can be used to *exemplify*, $\models_{\text{ex}}$, *green*, for instance, the green circle as in (32b). This is the influential reversed denotation relation (i.e., exemplification) introduced by Goodman (1976).

(32) a. $[\![\text{green}]\!] = \{$■, ●, ▶$\}$

b. ● $\models_{\text{ex}}$ green

Conversely, we can say that the predicate *green* can be used to *label* or *classify* the green circle. The exemplification of transitive and n-place predicates in general is brought about by collections of objects. Consider *throwing*, as in the throwing-a-dagger example (6a) in section 2. The lexical semantics of *throw* is given in (33).

[39] Of course, distinguishing gestures from real-world events follows from a double classification of the gesture according to a perceptual CVM and a hand classifier (a hand-cvm returns *false* for real-world events). We simply presuppose this marginal detail in the following.

(33) $[\![\text{throw}]\!] = \lambda y.\lambda x.\lambda e[\text{throw-cvm}(e) = 1 \wedge \text{agent}(e,x) \wedge \text{theme}(e,y) \wedge \exists \mathbf{v}[\text{place-path}(y,\mathbf{v})]]$

(33) gives rise to the following exemplification conditions: if there is a body movement which looks like throwing ('throw-cvm$(e) = 1$'), performed by x, and if there is something acted upon ('theme(y)') and that something is dislocated ('place-path$(y, \mathbf{v})$'; we abstract over time), we can classify this event e as a throwing event. Hence, a real-world event e exemplifies a predicate constant P if the event provides a witness for each of P's arguments, and only for the arguments. We call this *minimal exemplification.*[40] Any extended event e' which includes e will also exemplify that predicate, but not minimally.[41] Note that more specific predicates can be exemplified, such as *being thrown by Mike*, in which case Mike needs to be the agent of e, and so on.

This semiotic idea, we argue, also underlies the interpretation of iconic gestures, but with a twist: exemplification can only be applied to iconic gestures straightforwardly, if iconic gestures are taken to be objects or events in D_e and D_s, respectively. But a throwing event is different from a throwing gesture: the gesture *simulates* the real-world event. Being an event simulation rather than a real-world event itself, a gesture abstracts away over at least one object that would have been involved in the real-world event.[42] A throwing gesture (as in example 6a in section 2), in contrast to a throwing event, does not involve a theme (the object thrown) and therefore no dislocation path. In this case, the interpretation of the gesture as *throwing* involves a "virtual object", which can be thought of as a visual presupposition, since it is a precondition for interpreting the movement as throwing, regardless of linguistic context (e.g., negation). It still does involve the gesturer as an agent and a body movement which looks like throwing.

[40] Truthmaker accounts would say that e is a truthmaker for P (Mulligan et al, 1984).

[41] Being confined to minimal exemplification exempts from considerations of situational upwards persistence (Cooper, 1991).

[42] Gesture studies speak of "ad hoc abstraction" here and provide an interpretation drawing on metonymy (Mittelberg and Waugh, 2014, p. 1747).

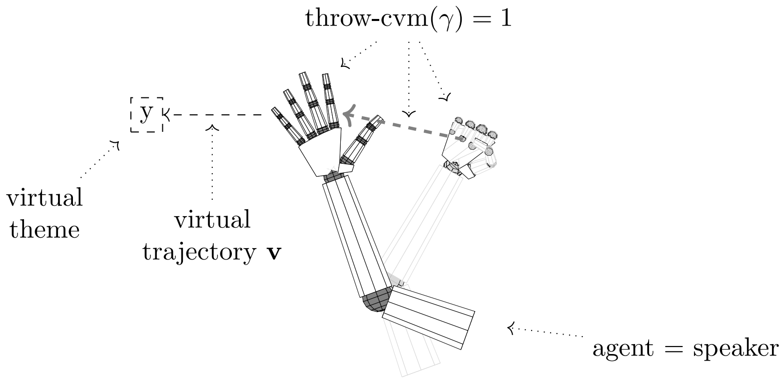


Fig. 24: Exemplifying *throw*.

Explicit representations of intensions in terms of CVMs provide the conceptual tools for expressing the required distinctions and abstractions to move from exemplification to *extended exemplification*, $\models_{\text{ext}}$, or *extemplification* as a short coinage.[43] Schematically:

(34) *Extended exemplification as gesture interpretation*

a. A gesture γ extemplifies a linguistic constituent p, $\gamma \models_{\text{ext}} p$, if p-cvm$(\gamma) = 1$ and γ is minimal wrt. p.

b. γ is minimal wrt. p iff there is a bijective mapping between (i) form features of γ, or (ii) visual, presupposed features of γ and the arguments of p.

c. If a. and b., that is, if γ exemplifies p, we can use p to informationally evaluate γ.

Step (34b) is to be brought about by the working semanticist, unless a computational classifier system is available. This is part of what makes (34) a heuristic.[44] Step (34c) delivers the input for the information-evaluation conditioned interpretation of multimodal utterances expressed in template (27) in the introduction to section 4.

[43]This is a significant improvement over Lücking (2013a), who uses Goodmanian exemplification, inheriting the above-discussed problems.
[44]In human vision, this is brought about by aligning a perceptual image with stored visual models (Ullman, 1996, in particular chapters 6 and 7).

Let us look at an example of how (34) works and offers a toolkit for semantic gesture studies, namely the throwing gesture from example (6a) in section 2. The relevant extract is given in (35):

(35) [...] throw [*throwing movement*] the dagger

Let Figure 24 be an illustration of the gesture involved in (35) – it extemplifies *throw* via the lexical meaning in (33). The movement "looks like" throwing (captured by the CVM), it is a pantomimic action (the speaker mimics the agent of the simulated action), and the imagined continuation of the stopped gestural movement evokes a virtual trajectory which triggers a virtual theme. Both the virtual trajectory and the theme are presupposed by evaluating the gesture as throwing. If this presupposition is not fulfilled, the gesture cannot "mean" throwing (it could exemplify some intransitive predicate instead, for instance, some direction instruction of a flight attendant, "Exit on the right"). Extended exemplification – in addition to a gesture's preference to directly depict its affiliate –, thus, provides some justification to interpret the gesture in (35) as *throw*. Template (27) then gives 'If the gesture is interpreted as *throw*, then [...] R(*throw*, throw) the dagger'. Obviously, R can be resolved to =, hence the multimodal utterance means '[...] throw the dagger'.

There is another interpretation available, namely the VP figuring as affiliate, that is $\lambda x.\lambda e[\text{throw-cvm}(e) = 1 \wedge \text{agent}(e, x) \wedge \text{theme}(e, \text{dagger}) \wedge \exists \mathbf{v}[\text{place-path}(y, \mathbf{v})]]$. In this case, the gesture extemplifies *throw the dagger*, which is brought about by the straightforward additional interpretive assumption that the virtual object y is identified with the dagger. Gesture interpretation then delivers "If the gesture is interpreted as *throw the dagger*, then [...] R(*throw the dagger*, throw the dagger)", which amounts to the same interpretation as the V-affiliated one. If also a subject were involved, a sentence-based extemplification can be derived, too. But recall from section 2 that the affiliate of a gesture is a lexical item in the majority of cases. Thus, the interplay of extemplification and affiliation offers some harmless leeway, which corresponds to the openness of gesture interpretation.

Throwing is an action, ‘ ’ is a shape which has been drawn by manual movement. Both, acting and drawing, correspond to two of the four modes of representation distinguished in section 2. Bridging to gesture studies, extemplification can be related to the modes of representation in the following way, where β stands for the extemplified predicate:

(36) a. *acting*: β is usually a transitive predicate whose agent is the speaker/gesturer (e.g., *throw*) or the gesturer's hand or arm (e.g., *walk*), and the theme is a virtual object

b. *molding*: β is a predicate denoting a volume

c. *drawing*: β is a predicate denoting a shape

d. *representing*: γ is identified with the discourse referent introduced by the affiliate (the gesture being a representative of a discourse referent)

From the perspective of extemplification, it is not a coincidence that – except for *representing*, which works differently – all modes of representation involve visuo-spatial predicates, featuring a CVM.

4.3 Lexicon-driven speech–gesture integration

Although the aim of this paper is a semantics of iconic gestures, this semantics is closely tight to affiliated speech. Usually, the gesture just extemplifies the affiliate, as seen in the previous section. But what has extemplification to say about examples where the extemplified predicate does not match the overtly uttered one, as in the spiral staircases example? We argue that the same kind of implicit meanings triggered by minimized contexts such as those in (37) arise if extemplified and affiliated predicate diverge.

(37) a. I can't ride my bike today. The back wheel's tire is flat.

b. The footage shows a man running on stage and stabbing Adamowicz [...]. The assailant paces back and forth, arms aloft like a victorious boxer, still holding the 15cm (six-inch) knife.[45]

The tire in (37a) is understood as the tire of the bike. The knife in (37b) is understood as the instrument of the stabbing event, and pacing back and forth the stabbing action (bridging; Clark, 1975). Accordingly, the same mechanisms used to resolve bridging anaphora apply in minimized speech–gesture affiliation pairs. An approach particularly well-suited to iconic gesture semantics is lexicon-driven and employs *frames* (Fillmore, 1968; Fillmore and Baker, 2010) to compute inferences on minimized content.[46] Corresponding lexical extensions have been developed within dynamic semantics, namely within *Discourse Representation Theory* respectively *Segmented Discourse Representation Theory* (Bos and Nissim, 2008; Irmer, 2013) and a *Type Theory with Records* (Cooper, 2010, 2023). Frames can be conceived as stereotypical situation types which are connected to lexical items ("single words"). A single word not only contributes its content, but it also *evokes* the frame it is connected to. Frame semantics is organized in a frame-base lexicon called FrameNet.[47] The lexical entry for *staircase.n*, for instance, is linked to the connecting_architecture frame. The connecting_architecture frame has a *Part* element as its core element, which is the connecting_architecture in question, and resolves to the staircase in our example. Additionally, there are nine non-core elements, namely *Connected_locations*, *Creator*, *Descriptor*, *Direction*, *Goal*, *Material*, *Orientation*, *Source*, and *Whole*. Via *frame evocation* (Irmer, 2013), the content of *staircase* in (38a) is extended by frame elements as in (38b) (using DRT's handy box notation; some frame elements are omitted for reasons of space):

[45] Taken from BBC news, https://www.bbc.com/news/world-europe-46878325, accessed 10th January 2024. Pawel Adamowicz was the mayor of Gdansk.

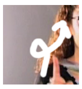
[46] To extend lexical meaning to gesture integration has been proposed early on by Rieser (2008).

[47] For a short introduction to FrameNet see Fillmore et al (2004), the FrameNet resource is available at https://framenet.icsi.berkeley.edu/.

(38) a. $\lambda x.$ $\begin{array}{|l|}\hline x \\ \hline \\ \text{staircase}(x) \\ \\ \hline \end{array}$ b. $\lambda x.$ $\begin{array}{|l|}\hline x \quad | \quad e, y_1, y_2, y_3, y_4 \\ \hline \\ e : \text{connecting\_architecture} \\ \text{Part}(e, x) \\ \text{staircase}(x) \\ \text{Creator}(e, y_1), y_1 =? \\ \text{Descriptor}(e, y_2), y_2 =? \\ \text{Direction}(e, y_3), y_3 =? \\ \text{Material}(e, y_4), y_4 =? \\ \\ \hline \end{array}$

Some remarks are in order here. The first condition instantiates the connecting_architecture frame as eventuality e. e as well as the discourse referents of the non-core arguments $y_{1...4}$ are merely implicit and are separated from the regular discourse referents of the universe (vertical bar in the top row). Implicit discourse referents were introduced by Kamp and Rossdeutscher (1994b), and a related distinction between discourse referents introduced by speech and those introduced by gesture has been argued for by Lascarides and Stone (2009). Frame evocation, thus, can be seen as a computational implementation of lexical presupposition triggering (Kamp and Rossdeutscher, 1994a). Since in this particular frame example, the part *is* the connecting_architecture talked about, all conditions that apply to e also apply to x.

If the gesture in example (19), re-given in (39):

(39) Inside the hall was an imposing  staircase.

is interpreted as *spiral*, then information-evaluation conditioned utterance interpretation from (27), re-given in (40):

(40) *Information-evaluation conditioned interpretation*:

If a gesture is interpreted to mean m, then the utterance is interpreted as $\alpha[R(m, \beta)]$.

returns '$R(\text{spiral}(z), \text{staircase}(x))$' as the affiliation argument of α. (Just as a reminder: α denotes a sentence, which is accompanied by gesture, which extemplifies m and is affiliated to expression β within α.) This information adds to the conditions of the frame-wise extended content of *staircase*:

$$(41) \quad \lambda x. \begin{array}{|l|}\hline x, z \quad | \quad e, y_1, y_2, y_3, y_4 \\ \hline \\ e : \text{connecting\_architecture} \\ \text{Part}(e, x) \\ \text{staircase}(x) \\ \text{Creator}(e, y_1), y_1 =? \\ \text{Descriptor}(e, y_2), y_2 =? \\ \text{Direction}(e, y_3), y_3 =? \\ \text{Material}(e, y_4), y_4 =? \\ \text{spiral}(z) \\ R(\text{spiral}(z), \text{staircase}(x)), R =? \\ \\ \hline \end{array}$$

The implicit discourse referents in (42) are supposed to be filled by content of different kinds: y_1 is likely to be an individual, y_4 a substance, y_3 a direction, and y_2 some property.[48] Since *spiral* has been produced by a gestural *drawing* mode of representation (or, alternatively, since it is a shape predicate), the only plausible frame element to resolve R is $R =$ *Descriptor*. We arrive at the following multimodal meaning, given that the gesture is informationally evaluated to mean *spiral* and affiliated to *staircase*:

$$(42) \quad \lambda x. \begin{array}{|l|}\hline x, z \quad | \quad e, y_1, y_2, y_3, y_4 \\ \hline \\ e : \text{connecting\_architecture} \\ \text{Part}(e, x) \\ \text{staircase}(x) \\ \text{Creator}(e, y_1), y_1 =? \\ \text{Descriptor}(e, y_2), y_2 = \text{spiral}(z), z = x \\ \text{Direction}(e, y_3), y_3 =? \\ \text{Material}(e, y_4), y_4 =? \\ \\ \hline \end{array}$$

The frame-extended predicate in (42) is processed as usual in further semantic composition.[49] Much more could (and should) be said of lexicon-based speech-gesture integration, but we believe that the above-given example is intuitively clear enough.

[48] Such constraints can nicely be captured by type constraints in a type theory (Cooper, 2011).

[49] On harmonizing DRT and Montague-style semantics, see Muskens (1996); Zeevat (1989), although the natural framework of (42) is SDRT (Asher and Lascarides, 2003).

The same mechanism derives examples discussed elsewhere, such as (43), taken from Schlenker (2018, p. 303).

(43) John [*slapping gesture*] punished his son.

If we interpret the gesture as slapping and *punished* as the lexical affiliate, then the multimodal information package '$R(\text{slapped}, \text{punished})$' is obtained. The lexical unit *punish.v* evokes the Rewards_and_punishment frame. Thus, the lexical content in (44a) is frame-wise extended to include the implicit content in (44b) (slightly abbreviated; the agent role maps to frame element 'Agent', the patient role to Evaluee; cf. Irmer, 2013):

(44) a. $\lambda y.\lambda x.\lambda e.$ $\begin{array}{|l|}\hline y, x, e \\ \hline \text{punish}(e) \\ \text{agent}(e, x) \\ \text{patient}(e, y) \\ \hline \end{array}$

b. $\lambda y.\lambda x.\lambda e.$ $\begin{array}{|l|}\hline y, x, e \quad | \quad z_1, z_2, z_3, z_4, z_5 \\ \hline e : \text{rewards\_and\_punishment} \\ \text{punish}(e) \\ \text{Agent}(e, x) \\ \text{Evaluee}(e, y) \\ \text{Reason}(e, z_1), z_1 = ? \\ \text{Degree}(e, z_2), z_2 = ? \\ \text{Instrument}(e, z_3), z_3 = ? \\ \text{Manner}(e, z_4), z_4 = ? \\ \text{Means}(e, z_5), z_5 = ? \\ \hline \end{array}$

Being an action-simulating gesture, slapping instantiates the non-core *Means* frame element: punish by slapping, see the resolved content in (45).

(45) $\lambda y.\lambda x.\lambda e.$

$$\begin{array}{|l|}\hline y, x, e, e', x', y' \quad | \quad z_1, z_2, z_3, z_4, z_5 \\ \hline \\ e : \text{rewards\_and\_punishment} \\ \text{punish}(e) \\ \text{Agent}(e, x) \\ \text{Evaluee}(e, y) \\ \text{Reason}(e, z_1), z_1 =? \\ \text{Degree}(e, z_2), z_2 =? \\ \text{Instrument}(e, z_3), z_3 =? \\ \text{Manner}(e, z_4), z_4 =? \\ \text{Means}(e, z_5), z_5 = \text{slap}(e'), e' = e \\ \text{agent}(e', x'), x' = x \\ \text{patient}(e', y'), y' = y \\ \\ \hline \end{array}$$

(45) reads as x punished y by slapping y. This captures Schlenker's intended reading for this example (Schlenker, 2018, p. 303), but is even stronger than the local context actually induced in his approach. The local context one gets without further stipulation is *If John punished his son, then slapping would be involved* (Schlenker, 2018, p. 318). But the slapping could be slapping someone else – think of John punishing his son by slapping the son's pet. In (45) this is captured by a non-maximal interpretation and dispensing condition '$y' = y$'. (The same kind of objection applies to Schlenker's helping-by-lifting example, which has an additional comitative reading.)

Furthermore, presuppositional accounts do not seem to prevent speech–gesture mismatches. Take, for instance, (46), where the spiral gesture is replaced by a slapping one:

(46) Inside the hall was an imposing [*slapping gesture*] staircase.

This gesture obviously does not extemplify its affiliate *staircase*. Furthermore, slapping, denoting an action, is not a good candidate to fill any of the frame elements evoked by *staircase* in (45). Hence, frame-based dynamic semantics algorithms would fail to integrate speech and gesture in this case and signal a mismatch. This does not seem to hold for other approaches such as those resting on local contexts and assertion-dependent presuppositions, since nothing prevents local contexts of the form "*every* *world w in which a staircase is in the hall is one in which slapping is involved*" from being computed (cf. Schlenker, 2018, p. 318 sq.).[50]

Examples such as (6c), where the verbal affiliate figuring out, is accompanied by a balancing gesture, are covered by *figure_out.v*, which evokes the Coming_to_believe frame. This frame introduces a *Means* element hosting the act performed by the agent, which enables them to figure something out. This means is extemplified by a balance scale gesture.

4.4 Applications

In this section, we apply iconic gesture semantics to some examples in order to illustrate how it is supposed to work and contribute to multimodal semantic analyses. To start with, the gesture displayed in Figure 25 evokes the visual image of a holding event. *Holding* is a two-place predicate with the meaning represented in (47a) and the extemplification mappings are explicated in (47b). Being a static action, no vector path is involved in this case.

(47) a. $[\![\text{hold}]\!] = \lambda y.\lambda x.\lambda e[\text{hold-cvm}(e) = 1 \wedge \mathit{agent}(e, x) \wedge \mathit{theme}(e, y)]$

b. Bijective iconic mappings:

– hold-cvm$(\gamma) \mapsto 1$

– speaker/gesturer $\mapsto \mathit{agent}(e)$

– space between hands $\mapsto \mathit{theme}(e)$ (i.e., the theme remains virtual)

Every content part of *hold* can be mapped onto morphological or visual features of the gesture. Hence, the gesture successfully extemplifies *hold*.

[50]Schlenker, by the way, puts a lot of weight on computability. But we would argue that the alleged "algorithm" for generating local contexts and iconic presuppositions is not computable at all. The main reason is that it operates on possible worlds. Possible worlds, however, are too much for human, cognitive processing (Partee, 1977), and since, according to philosophical arguments, they lack criteria of individuation and thereby countability (Rescher, 1999), they presumably resist any computational access. (We owe these arguments to Cooper (2023, p. 238 sqq.).) But this might be okay from a purely technical point of view. The FrameNet approach at least *is* computable (for an approach close to the present formal model see Hou et al, 2018; for a recent overview including state-of-the-art neural models see Poesio et al, 2023).

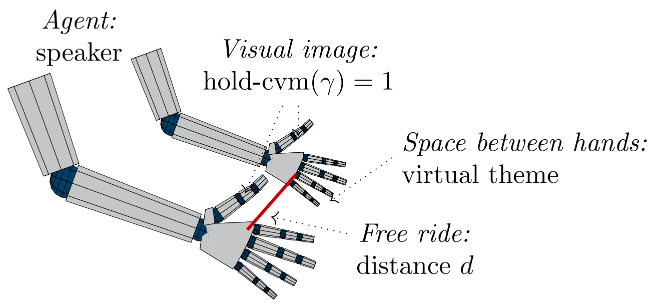


Fig. 25: Visual image of holding or touching an item. A distance (red line) comes as a *Free Ride* as studied in diagrammatic reasoning (e.g., Shimojima, 2015).

Now, this type of gesture has been claimed to mean *large* (Schlenker, 2018, p. 304, Esipova, 2019, p. 118). This interpretation does not seem to live up to iconic gesture semantics, however. There is, apart from implementational details, consensus that *large* is an adjective that is lexically associated with a measurement scale and a standard of comparison, as is expressed in (48a) (see Kennedy, 2007; Morzycki, 2009). But can this meaning be extemplified by the sample gesture? The gesture, in any case, extemplifies the property of *being this large* (*d-large* from (48a)) by means of the spatial distance d between the hands (this from a *Free Ride* type of inference from diagrammatic reasoning, Shimojima, 2015). Comparing (47) and the visual image displayed in Figure 25, however, it is apparent that the gesture does not extemplify the standard of comparison. The standard is not an *intrinsic* property of (virtual or real) sizing actions, hence *large* is not a fully visual property (as is already indicated by a lack of a CVM).

(48) a. $[\![\text{large}]\!] = \lambda x[\text{standard(large)} \leq \text{large}(x)]$,
where *large* is a measure function $\lambda x.\iota d.[x \textit{ is } d\textit{-large}]$ of type $\langle e, d\rangle$

b. Bijective iconic mappings:
– is there a LARGE-CVM?
– (from *Free Ride*) distance $d \mapsto$ *d-large*
– **?** $\mapsto$ standard (and it is unclear what to do with the agent)

Extemplification of *this large* already involves a visual inference *via* a free ride from *holding something* to *sizing*.[51] Arriving at *large* – if possible at all without further information in context[52] – needs further inferential processing.[53]

An additional trigger for pragmatic reasoning is the absence of a "lexical affiliate" (Schegloff, 1984), which is usually associated with an iconic gesture and guides its interpretation (Hadar and Krauss, 1999). The imagined affiliate in case of (3a) is the word *large* and its absence already indicates that the example is somewhat deviant: if *large* is the important property in question, then it is produced in an information structurally distinguished way, that is, being the focus expression. Omitting the focused predicate leads to a pragmatically infelicitous utterance. Pragmatically infelicitous utterances in turn trigger specific implicatures (e.g., evasion moves; Ginzburg et al, 2022). Hence, such examples are particularly ill-suited to tell apart asserted and implied contents.

Example (49) is taken from (Schlenker, 2019, p. 761).[54]

(49) This light bulb, are you going to [*speaker stretches arm overhead and rotates hand*]

Schlenker interprets the gesture as unscrewing at the ceiling. How can we derive this interpretation? *unscrew.v* gives us the following lexical information and evokes the *Closure* frame:

(50) $[\![\text{unscrew}]\!] = \lambda y.\lambda x.\lambda e[\text{unscrew-cvm}(e) = 1 \wedge \text{agent}(e, x) \wedge \text{theme}(e, y) \wedge e : \text{Closure}]$

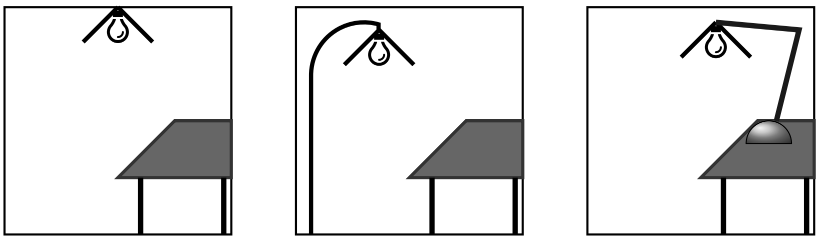
[51] The sizing mode of representation has been observed by Kendon (2004) in the Grappolo gesture family; it has also been used in iconic gesture annotation (Bergmann et al, 2014).

[52] The study described in Esipova (2019), for instance, introduces the verbal alternation *small* and *large* in experimental instructions, which is exploited by participants to reason about the gesture in terms of sizes.

[53] Note that the verbal affiliates in such examples are concrete nouns like *bottle* or *dog*. The lexical meaning of both expressions does not refer to tallness (or even size). This analysis confirms the judgments reported by Hunter (2019, p. 322), namely that a *large* reading of the gesture is unavailable for such examples.

[54] The original example has a nice picture of the gesture labelled with UNSCREW-CEILING instead of a verbal description, though.

The *Closure* frame contributes a non-core *Place* element, which the gesture exemplifies to *overhead*. This is what we get from gesture semantics. Schlenker claims further that the multimodal utterance involving the pro-speech gesture in (49) triggers the presupposition that the light bulb is at the ceiling. Since this is not part of the multimodal meaning, the location of the lamp should not be a presupposition – speaking in terms of FrameNet: the *Ground* element of the *Location_of_light* frame is not exemplified. Accordingly, we expect to find a situation where it does not hold. And indeed, stretching the arm over the head might also be necessary to unscrew light bulbs from other kinds of tall lamps, as shown in Figures 26(b) and 26(c). Hence, the impression that the light bulb is at the ceiling cannot be a presupposition of (50) – it is just an artifact of gesture interpretation (which in turn might be influenced by a defeasible abductive inference to a common location of light bulbs fixed overhead, though).



(a) Ceiling lamp (b) Floor lamp (c) Table Lamp

Fig. 26: Different lamps whose lightbulbs have to be unscrewed overhead.

This shows, we believe, the importance of a semantic toolkit for interpreting gestures in semantics, in particular, if different kinds of inferential meanings are to be kept apart.

Above all, however, iconic gesture semantics elucidates that a "linguistic lift" in terms of informational evaluation of a gesture is needed for the gesture to be able to interact with linguistic meaning in the first place. Being conditioned on a linguistic interpretation (*If the gesture is interpreted to mean "m" . . .*) explains that gestures do not contribute content, neither at-issue nor non-at-issue (cf. footnote 30); *if* they are informationally evaluated, however, they are just like additional words.

5 Discussion

5.1 "Energy spaces"

The iconic gesture semantics developed above is confined to representational gestures, that is, gestures that exhibit a rather pictorial (aka iconic) content. This is not surprising since such gestures catch the eye and have a rather clear interpretation. However, a lot of gestures occurring in natural language interactions have a strong impact on the beat dimension and/or are bound up with functional information structuring interpretations (see example 6d in section 2). A mere spatial interpretation seems to miss the point of such a gesture. Intuitively, they seem to be of a more "somatic" origin. The vector space semantics, we argue, can be adapted to such gestures, too, following ideas of Talmy (1988). He envisaged the use of *force vectors* instead of spatial place or path ones. One of the examples he discusses is the (intensional) meaning of the verb *climbing*. According to the analysis of Talmy (1988), the semantics of climb is captured in terms of two forces: one pulling downwards, one striving upwards. Mathematical vector spaces are ontologically neutral. The same formal devices can be used to model "energy spaces" consisting of force vectors. Following this idea, a domain extension is straightforward: in addition to spatial vector spaces, each individual $d \in D_e$ is assigned two force spaces, a pulling and an attracting one. Speakers occupy the respective "center of gravity" – see Figures 27 and 28. Talmy's analyses of *climbing* can be made precise in terms of an energy space spanned by the orthogonal projections of force vectors onto the downwards and upwards pulling ones in repulsion space.

Force vectors are arguably involved in verbal construction like *on the one hand ... on the other hand*: the two poles referred to are pulled apart by force vectors drawing in opposing directions. Accordingly, this verbal construction is often accompanied by two

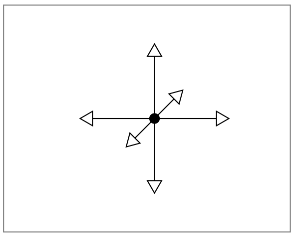


Fig. 27: Repulsion space

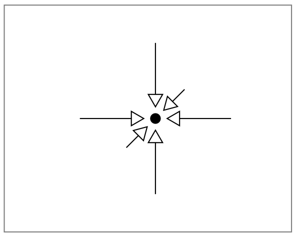
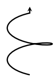


Fig. 28: Attractor space

placing gestures that locate the two poles in two different hemispheres of gesture space. Another point in this case could be gestures of uncertainty: being unsure corresponds to a lack of direction in energy space. Meandering through force vectors manifests itself in manual wiggling movements. Of course, this is speculative to a great extent, but a line of theoretical research worth to be explored in future work.

5.2 Non-lexicalized iconic models

Since a gesture's meaning contribution is conditioned on its informational evaluation, we expect no difference in at-issue contexts. The demonstrative context *like this* in (51) is supposed to shift its referent to at-issue content (Schlenker, 2019, p. 303).

(51) The staircase looked like this: 

While this works for speech, what is the at-issue contribution of the gesture in (51)? You have to informationally evaluate it first, as usual. That gestures without informational evaluation fail to contribute at-issue (or non-at-issue, for that matter) content can be seen by gestural movements that resist perceptual classification because the trajectory has no lexicalized label. An example is given in (52).[55]

(52) The inscription looked like this: 

[55] At least the authors do not know how to call this shape; it might be possible that some readers think that it resembles something they know and can name. We rely on these readers to appreciate the point of the example nonetheless.

We do think that the at-issue test still does work, and that it works for gestures. A gesture that is such that it resists perceptual classification in terms of single words just contributes its iconic model.[56]

Other gesture uses that can be subsumed under non-lexical ones are metaphoric ex(t)emplifications (such as the gesture from example 5 in section 2). We ignored them so far, but just want to mention that a frame-based approach provides a modelling clue in terms of the blending of frames (MetaNet, Petruck and Dodge, 2016).

5.3 Repercussions for semantic theories

Speech–gesture integration often leads to a reduplication in the sense that an iconic gesture extemplifies its linguistic affiliate. In other cases (i.e., if $R \neq =$), standard dynamic semantic resolutions apply. In this sense are iconic gestures not very exciting content-wise. But they are intriguing theory-wise: iconic gestures forces us to tell apart (perceptual) meaning and truth conditions. The notion of meaning needed is such that, when applied to an object, it returns a linguistic label (cf. section 1). We spell this out in terms of extemplification and perceptual classification. The resulting outline of an iconic gesture semantics according to Figure 22 involves both a standard denotational frame of reference and a classification/extemplification component. Arguably, these components cannot be reconciled with a textbook possible worlds semantics. If this is right,[57] then formal semantic thinking about iconic meaning ultimately requires looking for a different semantic theory. A suitable candidate, to our minds, is a *Type Theory with Records* (TTR, Cooper, 2023). TTR incorporates words-as-classifiers (Larsson, 2015), and can be given a probabilistic interpretation (Cooper et al, 2015), which is needed for learning and graded judgments (cf. *above*

[56] This is not puzzling from a dual coding (Paivio, 1986) or multimodal propositions (Lücking, 2013b) point of view. It is also part-and-parcel of grounding non-verbal behaviour such as motor actions (Hough et al, 2015).

[57] And there is ample supporting evidence. To name a few: the representability problem of possible worlds clashes with cognitive tractability (Lappin, 2015), which is a characterizing feature of extemplification/classification (see also footnote 50); the extensions of classifier-based semantics are indeterminate, which is not compatible with fixed universes D_e of quantification of traditional models (Larsson, 2020); the learnability problem in Montagovian models (Zimmermann, 2022) is at odds with classifier learning.

from Figure 15). TTR's modal theory does not assume possible worlds and hence fares better in terms of computability and cognitive interpretability.[58] TTR includes frames as both, situations and situation types (Cooper, 2010, 2023). Additional means of rhetorical coherence are expressed in terms of enthymemes (Breitholtz, 2020). TTR has been applied to iconic gesture (Lücking, 2016). It also provides the ontology for dialogue semantics (Ginzburg, 2012) (recall the importance of, e.g., clarification interaction discussed in section 4). After all, as is widely agreed, it is dialogue that is the "ecological niche" of multimodal interaction (Holler and Levinson, 2019; Lücking and Ginzburg, 2023). Thus, the contour not only of a compositional but also of a *computational* theory of iconic gesture semantics is emerging and will be elaborated in future work.

6 Conclusions

The theory-strategic aim of this paper was to reconcile spatial and labelling theories. This reconciliation amounts to a theory of iconic gesture semantics and has been achieved by spelling out a spatial, truth-functional account of iconic gestures, and by drawing on extended exemplification (extemplification), respectively perceptual classification. Key to this reconciliation was that gesture meaning depends on the informational evaluation of the gesture; otherwise, the gesture remains semantically inert. Iconic gesture semantics takes it seriously that gestures are not like words, and pays due attention to the characteristics of visual communication.

All the desiderata D1, D2, D3 identified in section 2 are addressed:

D1 Gesture interpretation is strongly dependent on the accompanying speech (affiliate dependence).

D2 Speech–gesture affiliation is regimented by (in-)congruency.

[58]This extends into the domain of plurality and quantification, as *Referential Transparency Theory* (RTT, Lücking and Ginzburg, 2022), which is formulated within TTR, avoids well-known processing obstacles of generalized quantifier theory.

D3 Gesture interpretation is an instance of visual communication, which rests on a perceptual interpretation of gestural forms.

D1 is captured by the affiliate being either the extemplified predicate, or by the affiliate being the frame-evoking expression which introduces an implicit discourse referent figuring as the extemplified meaning component.

D2 is captured within the spatial model in terms of an empty intersection of the gesture's iconic model and the spatial projection of its affiliate. Incongruency between speech and gesture is signalled if the gesture does neither extemplify its affiliate nor one of the frame elements evoked by the affiliate.

D3 has been the biggest obstacle to formal semantic modelling; in fact, it forces us to leave the well-trodden paths of Frege/Montague-style possible worlds semantics and to adopt a procedural, classifier-based notion of meaning. It is this repercussion of iconic gestures, we have argued, that makes visual communication a theoretically interesting object of research in formal semantics and the philosophy of language.

As a workaround for interpreting gestures in semantic analyses, an extemplification heuristic has been developed. This heuristic paves the way for systematic, empirically informed gesture studies within formal semantics and related fields. It is computationally implementable, thus enabling cross-talk between iconic gesture semantics and cognitive science.

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

Declarations

The authors have no competing interests to declare that are relevant to the content of this article.

References

Alahverdzhieva K, Lascarides A, Flickinger D (2017) Aligning speech and co-speech gesture in a constraint-based grammar. Journal of Language Modelling 5(3):421–464

Alibali MW (2005) Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information. Spatial Cognition and Computation 5:307–331

Asher N, Lascarides A (2003) Logics of Conversation. Cambridge University Press, Cambridge

Bangerter A, Oppenheimer DM (2006) Accuracy in detecting referents of pointing gestures unaccompanied by language. Gesture 6(1):85–102

Barnes K, Ebert C (2023) The information status of iconic enrichments: modelling gradient at-issueness. Theoretical Linguistics 49(3-4):167–223. https://doi.org/10.1515/tl-2023-2009

Bartsch R (1998) Dynamic Conceptual Semantics. A Logico-Philosophical Investigation Into Concept Formation and Understanding. CSLI Publications, Stanford, CA

Bavelas JB, Chovil N (2000) Visible acts of meaning: An integrated message model of language in face-to-face dialogue. Journal of Language and Social Psychology 19(2):163–194. https://doi.org/10.1177/0261927X00019002001

Bavelas JB, Chovil N, Lawrie DA, et al (1992) Interactive gestures. Discourse Processes 15(4):469–489. https://doi.org/10.1080/01638539209544823

Bergmann K, Damm O, Freigang F, et al (2014) Documentation – Sagaland. SFB 673, Project B1, Bielefeld University, https://www.phonetik.uni-muenchen.de/Bas/BasSaGAeng.html

Bohle U (2013) Approaching notation, coding, and analysis from a conversational analysis point of view. In: Müller C, Cienki A, Fricke E, et al (eds) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbooks of Linguistics and Communication Science, vol 1. De Gruyter Mouton, Berlin and Boston, chap 64, p 992–1007, https://doi.org/10.1515/9783110261318.992

Bos J, Nissim M (2008) Combining Discourse Representation Theory with FrameNet. In: Favretti RR (ed) Frames, Corpora, and Knowledge Representation. Bononia University Press, Bologna, chap 9, p 169–183

Breitholtz E (2020) Enthymemes and Topoi in Dialogue. No. 41 in Current Research in the Semantics/Pragmatics Interface, Brill, Leiden and Boston

Bressem J (2013) A linguistic perspective on the notation of form features in gestures. In: Müller C, Cienki A, Fricke E, et al (eds) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbooks of Linguistics and Communication Science, vol 1. De Gruyter Mouton, Berlin and Boston, chap 70, p 1079–1098, https://doi.org/10.1515/9783110261318.1079

Cassell J, McNeill D, McCullough KE (1999) Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. Pragmatics & Cognition 7(1):1–34. https://doi.org/10.1075/pc.7.1.03cas

Chalmers DJ (2002) On sense and intension. Philosophical Perspectives 16:135–182

Champollion L (2015) The interaction of compositional semantics and event semantics. Linguistics and Philosophy 38:31–66. https://doi.org/10.1007/s10988-014-9162-8

Chierchia G, McConnell-Ginet S (2000) Meaning and Grammar – An Introduction to Semantics, 2nd edn. MIT Press, Cambridge

Clark HH (1975) Bridging. In: Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing. Association for Computational Linguistics, TINLAP '75, pp 169–174, https://doi.org/10.3115/980190.980237

Cooper R (2010) Frames in formal semantics. In: Loftsson H, Rögnvaldsson E, Helgadóttir S (eds) Advances in Natural Language Processing. No. 6233 in Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, p 103–114, https://doi.org/10.1007/978-3-642-14770-8_13

Cooper R (2011) Copredication, quantification and frames. In: Pogodalla S, Prost JP (eds) Logical Aspects of Computational Linguistics. No. 6736 in Lecture Notes in Computer Science, Springer, Berlin and Heidelberg, p 64–79, https://doi.org/10.1007/978-3-642-22221-4_5

Cooper R (2023) From Perception to Communication. No. 16 in Oxford Studies in Semantics and Pragmatics, Oxford University Press, Oxford, UK, https://doi.org/10.1093/oso/9780192871312.001.0001

Cooper R, Dobnik S, Lappin S, et al (2015) Probabilistic type theory and natural language semantics. Linguistic Issues in Language Technology – LiLT 10(4):1–43. URL https://aclanthology.org/2015.lilt-10.4

Cooper RP (1991) Persistence and structural determination. In: Barwise J, Gawron JM, Plotkin G, et al (eds) Situation Theory and its Applications, CSLI Lecture Notes, vol 2. CSLI, Stanford, CA, chap 13, p 295–309

Cooperrider K (2017) Foreground gesture, background gesture. Gesture 16(2):176–202. https://doi.org/10.1075/gest.16.2.02coo

van der Does JM, van Lambalgen M (2000) A logic of vision. Linguistics and Philosophy 23(1):1–92. https://doi.org/10.1023/A:1005584121495

Duncan SD (2002) Gesture, verb aspect, and the nature of iconic imagery in natural discourse. Gesture 2(2):183–206

Ebert C (2014) The non-at-issue contributions of gestures. Talk given at the Workshop on Demonstration and Demonstratives, April 11-12 2014, Stuttgart

Ebert C (2024) Semantics of gesture. Annual Review of Linguistics https://doi.org/10.1146/annurev-linguistics-022421-063057

Ebert C, Ebert C (2014) Gestures, demonstratives, and the attributive/referential distinction. Handout of a talk given at Semantics and Philosophy in Europe (SPE 7), Berlin

Egré P, Politzer G (2013) On the negation of indicative conditionals. In: Aloni M, Franke M, Roelofsen F (eds) Proceedings of the 19th Amsterdam Colloquium, pp 10–18

Engelberg S (2000) Verben, Ereignisse und das Lexikon. No. 414 in Linguistische Arbeiten, Niemeyer

Esipova M (2019) Composition and projection of co-speech gestures. In: Blake K, Davis F, Lamp K, et al (eds) Proceedings of the 29th Semantics and Linguistic Theory Conference, SALT 29, pp 117–137

Fillmore CJ (1968) The case for case. In: Bach E, Harms RT (eds) Universals of linguistic theory. Holt, Rinehart & Winston, New York, NY, p 1–88

Fillmore CJ, Baker C (2010) A frames approach to semantic analysis. In: Heine B, Narrog H (eds) The Oxford Handbook of Linguistic Analysis. Oxford University Press, Oxford, p 313–340

Fillmore CJ, Baker CF, Sato H (2004) FrameNet as a “net”. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. LREC’2004, p 1091–1094

Ginzburg J (2012) The Interactive Stance: Meaning for Conversation. Oxford University Press, Oxford, UK

Ginzburg J, Cooper R (2004) Clarification, ellipsis, and the nature of contextual updates in dialogue. Linguistics and Philosophy 27(3):297–365. https://doi.org/10.1023/B:LING.0000023369.19306.90

Ginzburg J, Lücking A (2021a) I thought pointing is rude: A dialogue-semantic analysis of pointing at the addressee. In: Grosz P, Martí L, Pearson H, et al (eds) Proceedings of *Sinn und Bedeutung 25*, SuB 25, pp 276–291, https://doi.org/10.18148/sub/2021.v25i0.937, URL https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/937, special Session: Gestures and Natural Language Semantics

Ginzburg J, Lücking A (2021b) Requesting clarifications with speech and gestures. In: Proceedings of the 1st Workshop on Multimodal Semantic Representations. Association for Computational Linguistics, MMSR, pp 21–31, URL https://aclanthology.org/2021.mmsr-1.3

Ginzburg J, Yusupujiang Z, Li C, et al (2022) Characterizing the response space of questions: Data and theory. Dialogue & Discourse 13(2):79–132. https://doi.org/org/10.5210/dad.2022.203

Giorgolo G (2010) Space and time in our hands. PhD thesis, University of Utrecht

Goodman N (1976) Languages of Art. An Approach to a Theory of Symbols, 2nd edn. Hackett Publishing Company, Inc., Idianapolis

Greenberg G (2011) The semiotic spectrum. PhD thesis, Department of Philosophy, Rutgers University

Greenberg G (2021) Semantics of pictorial space. Review of Philosophy and Psychology 12:847–887. https://doi.org/10.1007/s13164-020-00513-6

Gullberg M, Kita S (2009) Attention to speech-accompanying gestures: Eye movements and information uptake. Journal of Nonverbal Behavior 33(4):251–277. https://doi.org/10.1007/s10919-009-0073-2

Hadar U (1989) Two types of gesture and their role in speech production. Journal of Language and Social Psychology 8(3-4):221–228. https://doi.org/10.1177/0261927X8983004

Hadar U (2013) Co-verbal gestures: Between communication and speech production. In: Müller C, Cienki A, Fricke E, et al (eds) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbooks of Linguistics and Communication Science, vol 1. De Gruyter Mouton, Berlin and Boston, chap 50, p 804–821, https://doi.org/https://doi.org/10.1515/9783110261318.804

Hadar U, Krauss RK (1999) Iconic gestures: The grammatical categories of lexical affiliates. Journal of Neurolinguistics 12(1):1–12. https://doi.org/10.1016/S0911-6044(99)00001-9

Hahn F, Rieser H (2010) Explaining speech gesture alignment in mm dialogue using gesture typology. In: Łupowski P, Purver M (eds) Proceedings of the 14th Workshop on the Semantics and Pragmatics of Dialogue: Aspects of Semantics and Pragmatics of Dialogue, Polish Society for Cognitive Science, Poznan, Poland, SemDial 2010, pp 99–111

Harnad S (1990) The symbol grounding problem. Physica D 42:335–346

Healey PG, Plant N, Howes C, et al (2015) When words fail: Collaborative gestures during clarification dialogues. In: Turn-Taking and Coordination in Human-Machine Interaction: Papers from the 2015 AAAI Spring Symposium, pp 23–29

Heim I, Kratzer A (1998) Semantics in Generative Grammar. No. 13 in Blackwell Textbooks in Linguistics, Blackwell, Malden, MA and Oxford, UK

Holler J, Levinson SC (2019) Multimodal language processing in human communication. Trends in Cognitive Sciences 23(8):639–652. https://doi.org/10.1016/j.tics.2019.05.006, opinion

Hou Y, Markert K, Strube M (2018) Unrestricted bridging resolution. Computational Linguistics 44(2):237–284. https://doi.org/10.1162/COLI_a_00315

Hough J, de Kok I, Schlangen D, et al (2015) Timing and grounding in motor skill coaching interaction: Consequences for the information state. In: Proceedings of SEMDIAL 2015, goDIAL 2015, pp 86–94

Hunter J (2019) Relating gesture to speech: Reflections on the role of conditional presuppositions. Linguistics and Philosophy 42:317–332. https://doi.org/10.1007/s10988-018-9244-0

Irmer M (2013) Inferring implicatures and discourse relations from frame information. Lingua 132:29–50. https://doi.org/10.1016/j.lingua.2012.11.009, special Issue: Implicature and Discourse Structure

Jackendoff R (1987) On beyond Zebra: The relation of linguistic and visual information. Cognition 26:89–114

Jackendoff R (1991) Parts and boundaries. Cognition 41:9–45

Johansson G (1973) Visual perception of biological motion and a model for its analysis. Perception & Psychophysics 14(2):201–211. https://doi.org/doi.org/10.3758/BF03212378

Johansson G (1976) Spatio-temporal differentiation and integration in visual motion perception. An experimental and theoretical analysis of calculus-like functions in visual data processing. Psychol Res 38:379–393

Johansson G, von Hofsten C, Jansson G (1980) Event perception. Annual Review of Psychology 31:27–63

Kamp H, Rossdeutscher A (1994a) DRS-construction and lexically driven inference. Theoretical Linguistics 20(2-3):165–235. https://doi.org/10.1515/thli.1994.20.2-3.165

Kamp H, Rossdeutscher A (1994b) Remarks on lexical structure and DRS construction. Theoretical Linguistics 20(2-3):97–164. https://doi.org/10.1515/thli.1994.20.2-3.97

Kendon A (2004) Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge, MA

Kennedy C (2007) Vagueness and grammar: The semantics of relative and absolute gradable adjectives. Linguistics and Philosophy 30(1):1–45. https://doi.org/10.1007/s10988-006-9008-0

Kennington C, Schlangen D (2015) Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp 292–301

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(1):16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kopp S, Tepper P, Cassell J (2004) Towards integrated microplanning of language and iconic gesture for multimodal output. In: Proceedings of the 6th International Conference on Multimodal Interfaces. ACM, New York, NY, USA, ICMI '04, pp 97–104, https://doi.org/10.1145/1027933.1027952

Kranstedt A, Lücking A, Pfeiffer T, et al (2006a) Measuring and reconstructing pointing in visual contexts. In: Schlangen D, Fernández R (eds) Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue. Universitätsverlag Potsdam, Potsdam, Brandial'06, pp 82–89

Kranstedt A, Lücking A, Pfeiffer T, et al (2006b) Deictic object reference in task-oriented dialogue. In: Rickheit G, Wachsmuth I (eds) Situated Communication. Mouton de Gruyter, Berlin, p 155–207

Krauss RM, Chen Y, Gottesmann RF (2000) Lexical gestures and lexical access: A process model. In: McNeill D (ed) Language and Gesture. Cambridge University Press, Cambridge, UK, chap 13, p 261–283

Krifka M (1998) The origins of telicity. In: Rothstein S (ed) Events and Grammar. No. 70 in Studies in Linguistics and Philosophy, Springer, Dordrecht, https://doi.org/org/10.1007/978-94-011-3969-4_9

Krifka M (2012) Some remarks on event structure, conceptual spaces and the semantics of verbs. Theoretical Linguistics 38(3-4):223–236. https://doi.org/10.1515/tl-2012-0014

Laparle S (2021) Tracking discourse topics in co-speech gesture. In: Duffy VG (ed) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior. Springer International Publishing, Cham, no. 12777 in Lecture Notes in Computer Science, pp 233–249

Lappin S (2015) Curry typing, polymorphism, and fine-grained intensionality. In: The Handbook of Contemporary Semantic Theory. John Wiley & Sons, Ltd, chap 13, p 408–428, https://doi.org/10.1002/9781118882139.ch13

Larsson S (2015) Formal semantics for perceptual classification. Journal of Logic and Computation 25(2):335–369. https://doi.org/10.1093/logcom/ext059

Larsson S (2020) Extensions are indeterminate if intensions are classifiers. In: Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue – Full Papers. SEMDIAL, Virtually at Brandeis, Waltham, New Jersey, URL http://semdial.org/anthology/Z20-Larsson_semdial_0019.pdf

Lascarides A, Stone M (2009) A formal semantic analysis of gesture. Journal of Semantics 26(4):393–449. https://doi.org/10.1093/jos/ffp004

Lewis D (1969) Convention: A Philosophical Study. Harvard University Press, Cambridge, MA

Lewis D (1970) General semantics. Synthese 22(1/2):18–67

Loehr D (2007) Apects of rhythm in gesture in speech. Gesture 7(2):179–214

Lücking A (2013a) Ikonische Gesten. Grundzüge einer linguistischen Theorie. De Gruyter, Berlin and Boston

Lücking A (2013b) Multimodal propositions? From semiotic to semantic considerations in the case of gestural deictics. In: Fernández R, Isard A (eds) Poster Abstracts of the Proceedings of the 17th Workshop on the Semantics and Pragmatics of Dialogue, Amsterdam, SemDial 2013, pp 221–223

Lücking A (2016) Modeling co-verbal gesture perception in type theory with records. In: Ganzha M, Maciaszek L, Paprzycki M (eds) Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, Annals of Computer Science and Information Systems, vol 8. IEEE, pp 383–392, https://doi.org/10.15439/2016F83

Lücking A (2018) Witness-loaded and witness-free demonstratives. In: Coniglio M, Murphy A, Schlachter E, et al (eds) Atypical Demonstratives. Syntax, Semantics and Pragmatics. No. 568 in Linguistische Arbeiten, De Gruyter, Berlin and Boston, p 255–284

Lücking A (2021) Gesture. In: Müller S, Abeillé A, Borsley RD, et al (eds) Head Driven Phrase Structure Grammar: The handbook. No. 9 in Empirically Oriented Theoretical Morphology and Syntax, Language Science Press, Berlin, chap 27, p 1201–1250, https://doi.org/10.5281/zenodo.5543318, URL https://langsci-press.org/catalog/book/259

Lücking A (2022) Aspects of multimodal communication. Habilitation, Université Paris Cité

Lücking A, Ginzburg J (2022) Referential transparency as the proper treatment of quantification. Semantics and Pragmatics 15(4):1–58. https://doi.org/10.3765/sp.15.4

Lücking A, Ginzburg J (2023) Leading voices: Dialogue semantics, cognitive science, and the polyphonic structure of multimodal interaction. Language and Cognition 15(1):148–172. https://doi.org/10.1017/langcog.2022.30

Lücking A, Bergmann K, Hahn F, et al (2010) The Bielefeld speech and gesture alignment corpus (SaGA). In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality, 7th International Conference for Language Resources and Evaluation, Malta, LREC 2010, pp 92–98, https://doi.org/10.13140/2.1.4216.1922

Lücking A, Bergman K, Hahn F, et al (2013) Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment Corpus (SaGA) and its applications. Journal on Multimodal User Interfaces 7(1-2):5–18. https://doi.org/10.1007/s12193-012-0106-8

Lücking A, Pfeiffer T, Rieser H (2015) Pointing and reference reconsidered. Journal of Pragmatics 77:56–79. https://doi.org/10.1016/j.pragma.2014.12.013

Marr D (2010) Vision. The MIT Press, https://doi.org/10.7551/mitpress/9780262514620.001.0001

Marr D, Vaina L (1982) Representation and recognition of the movements of shapes. Proceedings of the Royal Society of London Series B Biological Sciences 214(1197):501–524. https://doi.org/10.1098/rspb.1982.0024

McNeill D (1985) So you think gestures are nonverbal? Psychological Review 92(3):350–371

McNeill D (1992) Hand and Mind – What Gestures Reveal about Thought. Chicago University Press, Chicago

McNeill D (2005) Gesture and Thought. University of Chicago Press, Chicago

McNeill D, Duncan SD (2000) Growth points in thinking-for-speaking. In: McNeill D (ed) Language and Gesture. Cambridge University Press, Cambridge, MA, chap 7, p 141–161

McNeill D, Cassell J, Levy ET (1993) Abstract deixis. Semiotica 95(1-2):5–19. https://doi.org/10.1515/semi.1993.95.1-2.5

Mehler A, Lücking A (2012) Pathways of alignment between gesture and speech: Assessing information transmission in multimodal ensembles. In: Giorgolo G, Alahverdzhieva K (eds) Proceedings of the International Workshop on Formal and Computational Approaches to Multimodal Communication under the auspices of ESSLLI 2012, Opole, Poland, 6-10 August

Miletić V, Stuart L, Tantau T, et al (2024) The beamer package. A LaTeX class for producing presentations and slides. URL https://ctan.org/pkg/beamer

Mittelberg I, Waugh LB (2014) Gestures and metonymy. In: Müller C, Cienki A, Fricke E, et al (eds) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbooks of Linguistics and Communication Science, vol 2. De Gruyter Mouton, Berlin and Boston, chap 132, p 1747–1766, https://doi.org/10.1515/9783110302028.1747

Morzycki M (2009) Degree modification of gradable nouns: size adjectives and adnominal degree morphemes. Natural Language Semantics 17:175–203. https://doi.org/10.1007/s11050-009-9045-7

Müller C (2014) Gestural modes of representation as techniques of depiction. In: Müller C, Cienki A, Fricke E, et al (eds) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbooks of Linguistics and Communication Science, vol 2. De Gruyter Mouton, Berlin and Boston, chap 128, p 1687–1702

Mulligan K, Simons P, Smith B (1984) Truth-makers. Philosophy and Phenomenological Research 44:287–321

Muskens RA (1996) Combining Montague semantics and discourse representation. Linguistics and Philosophy 19:143–186

Müller C (2017) How recurrent gestures mean: Conventionalized contexts-of-use and embodied motivation. Gesture 16(2):277–304. https://doi.org/10.1075/gest.16.2.05mul

O'Keefe J (1996) The spatial prepositions in English, vector grammar, and the cognitive map theory. In: Bloom P, Peterson MA, Nadel L, et al (eds) Language and Space. MIT Press, Cambridge, MA, chap 7, p 277–316

O'Keefe J (2003) Vectors grammar, places, and the functional role of the spatial prepositions in English. In: Representing Direction in Language and Space. No. 1 in Explorations in Language and Space, Oxford University Press, Oxford, NY, chap 4, p 69–85

Özyürek A (2014) Hearing and seeing meaning in speech and gesture: Insights from brain and behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences 369(1651). https://doi.org/10.1098/rstb.2013.0296

Paivio A (1986) Mental Representations: A Dual Coding Approach. Oxford University Press, New York

Parsons T (1990) Events in the Semantics of English. No. 19 in Current Studies in Linguistics Series, MIT Press, Cambridge

Partee BH (1977) Possible worlds semantics and linguistic theory. The Monist 60(3):303–326

Petruck MRL, Dodge EK (2016) MetaNet: Repository, identification system, and applications. In: Birch A, Zuidema W (eds) Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts. Association for Computational Linguistics, Berlin, Germany, URL https://aclanthology.org/P16-5008

Poesio M, Yu J, Paun S, et al (2023) Computational models of anaphora. Annual Review of Linguistics 9(1):561–587. https://doi.org/10.1146/annurev-linguistics-031120-111653

Rescher N (1999) How many possible worlds are there? Philosophy and Phenomenological Research 59(2):403–420

Rieser H (2004) Pointing in dialogue. In: Proceedings of the Eighth Workshop on the Semantics and Pragmatics of Dialogue, Barcelona, Catalog '04, pp 93–100

Rieser H (2008) Aligned iconic gesture in different strata of mm route-description. In: LonDial 2008: The 12th Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL), King's College London, pp 167–174

Rieser H (2011) How to disagree on a church window's shape using gesture. In: Hölker K, Marello C (eds) Dimensionen der Analyse von Texten und Diskursen. Festschrift für János Sándor Petőfi. LIT Verlag, Berlin, p 231–246

Rieser H (2015) When hands talk to mouth. Gesture and speech as autonomous communicating processes. In: Howes C, Larsson S (eds) Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue, Gothenburg, Sweden, SEMDIAL 2015: goDIAL, pp 122–130

Rieser H, Poesio M (2009) Interactive gestures in dialogue: a PTT model. In: Proceedings of the 10th Annual Meeting of the Special Interest Group in Discourse and

Dialogue, SIGDIAL 2009, pp 87–96

de Ruiter JP (2000) The production of gesture and speech. In: McNeill D (ed) Language and Gesture. Cambridge University Press, Cambridge, UK, chap 14, p 284–311

Schegloff EA (1984) On some gestures' relation to talk. In: Atkinson JM, Heritage J (eds) Structures of Social Action. Studies in Conversational Analysis. Studies in Emotion and Social Interaction, Cambridge University Press, Cambridge, MA, chap 12, p 266–296

Schlenker P (2018) Gesture projection and cosuppositions. Linguistics and Philosophy 41(3):295–365. https://doi.org/10.1007/s10988-017-9225-8

Schlenker P (2019) Gestural semantics. Natural Language & Linguistic Theory 37:735–784

Schlenker P, Chemla E (2018) Gestural agreement. Natural Language & Linguistic Theory 36:587–625. https://doi.org/10.1007/s11049-017-9378-8

Shimojima A (2015) Semantic Properties of Diagrams and Their Cognitive Potentials. Studies in the Theory and Applications of Diagrams, CSLI Publications, Stanford, CA

Sowa T (2006) Understanding Coverbal Iconic Gestures in Shape Descriptions. Akademische Verlagsgesellschaft, Berlin

Steels L, Belpaeme T (2005) Coordinating perceptually grounded categories through language: A case study for colour. Behavioral and Brain Sciences 28(4):469–529

Streeck J (2008) Depicting by gesture. Gesture 8(3):285–301. https://doi.org/10.1075/gest.8.3.02str

Talmy L (1988) Force dynamics in language and cognition. Cognitive Science 12(1):49–100. https://doi.org/10.1016/0364-0213(88)90008-0

Ullman S (1996) High-Level Vision. A Bradford Book, MIT Press, Cambridge, MA

Warglien M, Gärdenfors P, Westera M (2012) Event structure, conceptual spaces and the semantics of verbs. Theoretical Linguistics 38(3-4):159–193. https://doi.org/10.1515/tl-2012-0010

Weisgerber M (2006) Decomposing path shapes: About an interplay of manner of motion and 'the path'. In: Ebert C, Endriss C (eds) Proceedings of Sinn und Bedeutung, SuB 10, pp 405–419, https://doi.org/10.18148/sub/2006.v10i2.741

Zeevat H (1989) A compositional approach to discourse representation theory. Linguistics and Philosophy 12(1):95–131

Zimmermann TE (2011) Model-theoretic semantics. In: Maienborn C, Heusinger Kv, Portner P (eds) Semantics: An International Handbook of Natural Language Meaning, Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK), vol 1. De Gruyter Mouton, Berlin and Boston, chap 33, p 762–802, https://doi.org/10.1515/9783110226614.946

Zimmermann TE (2019) Intensions, types and models. In: Altshuler D, Rett J (eds) The Semantics of Plurals, Focus, Degrees, and Times: Essays in Honor of Roger Schwarzschild. Springer International Publishing, Cham, p 41–56, https://doi.org/10.1007/978-3-030-04438-1_3

Zimmermann TE (2022) Extensions in compositional semantics. In: Özgün A, Zinova Y (eds) Language, Logic, and Computation. Springer International Publishing, Cham, pp 148–172

Zwarts J (1997) Vectors as relative positions: A compositional semantics of modified PPs. Journal of Semantics 14(1):57–86

Zwarts J (2003) Vectors across spatial domains: From place to size, orientation, shape, and parts. In: van der Zee E, Slack J (eds) Representing Direction in Language and Space. No. 1 in Explorations in Language and Space, Oxford University Press, Oxford, NY, chap 3, p 39–68

Zwarts J (2017) Spatial semantics: Modeling the meaning of prepositions. Language and Linguistics Compass 11(5):e12241–n/a. https://doi.org/10.1111/lnc3.12241, e12241 LNCO-0444.R2

Zwarts J, Gärdenfors P (2016) Locative and directional prepositions in conceptual spaces: The role of polar convexity. Journal of Logic, Language and Information 25(1):109–138

Zwarts J, Winter Y (2000) Vector space semantics: A model-theoretic analysis of locative prepositions. Journal of Logic, Language, and Information 9(2):169–211