Proceedings of the Federated Conference on Computer Science and Information Systems pp. 383–392
DOI: 10.15439/2016F83
ACSIS, Vol. 8. ISSN 2300-5963

Modeling Co-Verbal Gesture Perception in Type Theory with Records

Andy Lücking
Goethe University Frankfurt
Robert-Mayer-Straße 10
D-60325 Frankfurt am Main, Germany
Email: luecking@em.uni-frankfurt.de

***Abstract*—In natural language *face to face* communication interlocutors exploit manifold non-verbal information resources, most notably hand and arm movements, i.e. gestures. In this paper, a type-theoretical approach using Type Theory with Records is introduced which accounts for iconic gestures within an information state update semantics. Iconic gestures are semantically exploited in two steps: firstly, their kinetic representations are mapped onto vector sequence representations from vector space semantics, modeling a perceptual gesture classification; secondly, these vectorial representations are linked to linguistic predicates, giving rise to a computational account to semantic-kinematic interfaces. Each of the steps involves reasoning processes, which are made explicit. The resulting framework shows how various resources have to be integrated in the update mechanism in order to deal with apparently simple multimodal utterances.**

I. Introduction

SEMANTIC theories, artificial intelligence and robotic systems developed for spoken face-to-face interaction eventually have to deal with non-verbal communication means like facial expressions, prosodic features, hand and arm gestures, or proxemic relations. This is because verbal and non-verbal means constitute an integrated communication system [1], [2]. Their tight coupling shows up strikingly in cases where non-verbal means are semantically significant, i.e. when they provide information beyond or even instead of the verbal one – respective data is given in Section II. We are concerned with *iconic* gestures in the sense of representational hand and arm movements in this paper, gestures which, roughly speaking, depict aspects of the scene talked about [3] (for an automatic gesture classification involving iconic ones see e.g. [4]). Such gestures are performed rather spontaneously and presumably do not obey formal constraints [2]. For this reason, they cannot be interpreted according to pre-defined lexical entries, contrary to emblematic gestures as, say, the *thumbs-up* symbol for indicating positive evaluation or agreement, or Karate postures [5] in physical instructions. Rather, the interpretation of such gestures is a challenge that is related to spatial perception [6]. Accordingly, a perceptually oriented iconic gesture classification is proposed which rests on an integration of various semantic resources, as envisaged by, e.g., [7]. The formal framework that provides a unified representational "home" for these resources is *Type Theory with Records* (TTR) [8]. In the following, TTR is applied to capture the semantic impact of co-verbal gestures by combining the following ingredients:

- a detailed kinematic gesture representation [9] – Section III-A;
- the gesture representation is mapped onto vector sequences from vector space semantics [10] – Section III-B;
- vector sequences are linked to the intensions of linguistic expressions along the lines of a formal semantics for perceptual classification [11], [12] – Section III-C;
- perceptual classification and linguistic semantics are finally related within a dynamic information state update semantics [11], [13] – Section III-C.

Finally, in Section IV the account is applied to the benchmark phenomena identified in Section II.

II. Some Data

The integration of speech and co-verbal gesture has been investigated, *inter alia*, by means of the (German) *Speech and Gesture Alignment Corpus* (SaGA) [9], from which the following examples are drawn. Examples are quoted according to their dialog number and start time (e.g. "V13, 3:36" means that the datum can be found in dialog V13 at minute 3:36 within the corresponding video file). Note that only the so-called *stroke* phase of gestures is considered here: it is assumed to be the "meaningful" part of a gestural movement, distinguished from a pre-stroke preparatory movement as well as from a post-stroke retraction movement, which may be required in order to bring hand and arms from an inactive rest position into an active position and back, respectively [14], [2]. The decorated screenshots are all taken from [12].

Fig. 1. *Staircases* (V10, 3:19)

978-83-60810-90-3/$25.00

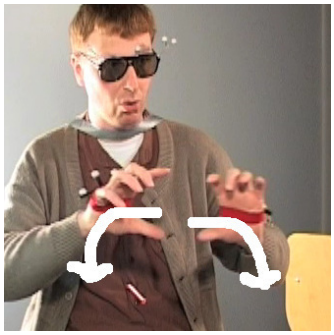


Fig. 2. *It has a concrete base* (V5, 0:39)

Fig. 3. *The house is like this.* (V11, 2:32)

In the first example – given in (1) –, which can be found in V10, 3:19, the speaker speaks about circular staircases. However, in her verbal description she just uses the hypernym *staircases*. The more specific circular information is provided by the affiliated gesture, which is shown in Fig. 1. 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 which 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 [15], [12].

(1) *Ich g[laube das sollen TREP]pen sein*
I think that should staircases be

'I think that should be staircases' + Fig. 1

Thus, gestures can be used to specify linguistic expressions to their hyponym meanings.

In a similar manner, the gesture shown in Fig. 2 indicates the shape of a concrete base, which is introduced into dialog in the following way:

(2) *die Skulptur die hat 'n [BeTONsockel]*
the sculpture it has a concrete base

'the sculpture has a concrete base' + Fig.2

From the gesture, but not from speech, we get that the concrete base of the sculpture has the shape of a flat cylinder – the gesture acts as a nominal modifier. Note, however, that the gesture is incomplete since it only depicts about half of a cylinder [16]. Thus, its interpretation interacts with a *good continuation* extension known from gestalt theory.

In the datum given in (3), the speaker speaks about a U-shaped building. However, no shape predicate is given verbally, instead the full shape-representing burden is delegated to the gesture. The gesture in turn is produced within the scope of a verbal demonstrative, which induces a shift in focus to the speaker's gesture. In contrast to examples (1) and (2), the utterance in (3) is not even interpretable without the gesture.

(3) *dann ist das Haus halt so []*
then is the house just this []

'then the house is like this' + Fig. 3

These examples illustrate that the informational enrichment by gestures includes

- invoking hyponymic meanings of affiliated expressions,
- indicating linguistically unexpressed properties,
- providing complete demonstrations.

In the following sections, a type-theoretical model is given that aims at accounting for these gestural enrichments. This account extends previous accounts (most notably [12]) in that it implements a connection to dynamic semantic theories and provides a means for dealing with *good continuations* in formal analyses of co-verbal gestures for the first time. The focus on spontaneous iconic gestures goes beyond functionally restricted *click* or *draw* gestures as predefined in multimodal grammars or dialog systems [17], [18], [19] – see [20] for a comparison of various multimodal approaches.

III. A Type-Theoretical Account to Gestures

Type Theory with Records (TTR) [8] has been developed as a formal framework for natural language semantics which integrates insights from situation semantics [21], Discourse Representation Theory [22] and Montagovian λ-calculus [23]. The basic notion of TTR is a *judgment* of the form $a : T$, meaning that object a is of type T. In order to account for more complex kinds of judgments, TTR develops the notions of *record* and *record type*. The former are matrices of labels and objects, the latter are matrices of labels and types. Record types can be used to regiment records: a record r is of record type RT, $r : RT$, just in case each label of the record type also occurs in the record (the record may contain more fields, though) and the label assignments from the record obey the type constraints imposed by the record type, as schematically exemplified in (4).

(4) $\begin{bmatrix} l_1 & = & o_1 \\ l_2 & = & o_2 \\ l_3 & = & o_3 \end{bmatrix} : \begin{bmatrix} l_1 & : & T_1 \\ l_3 & : & T_2(l_1) \end{bmatrix}$,
just in case $o_1 : T_1$ and $o_3 : T_2(o_1)$.

Record types may depend on other records or record types. For example, type T_2 in (4) depends on object o_1. Dependent types can be used in semantics, for instance, to capture existence

presuppositions imposed by proper names, as illustrated in (5) by example of the name "Max":

(5) $\lambda r : \left[\text{x} \quad : \mathit{Ind}\right] \; . \left[\text{c}_{\text{pn}} \quad : \quad \text{named}(r.\text{x}, \text{“Max”})\right]$

The type in (5) is a functional type in which the value of the range depends on the value of the domain (which in turn is constrained to be of type *Ind(ividual)*) – see [8] for this and various others formal TTR notions.

A. A String Theory of Gesture Events

TTR comes with a string theory of events based on work by [24]. Basically, a(n) (complex) event is segmented into event "snapshots" that are combined by the string concatenator '^'. For example, the event e of opening a door involves the sequence of an agent x gripping the door handle a, pressing the handle and pushing the door b:

(6) $([e : \text{grip}(x, a)]^\frown[e : \text{press}(x, a)]^\frown[e : \text{push}(x, b)])$

Now, gestures can be considered to be events [12]. To begin with, a simple gesture can be represented as a record straightforwardly, as in (7):[1]

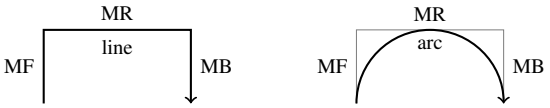
(7)
$$\begin{bmatrix}
\text{hand} & = & \text{right} \\
\text{hs} & = & \text{claw} \\
\text{carrier} & = & \begin{bmatrix}
\text{boh} & = & \text{none} \\
\text{plm} & = & \text{none} \\
\text{wrst} & = & \text{MR>MB>ML} \\
\text{move} & = & \text{line>line>line}
\end{bmatrix} \\
\text{sync} & = & \begin{bmatrix}
\text{sloc} & = & \text{CBR-F} \\
\text{eloc} & = & \text{CBR-N} \\
\text{stime} & = & \text{2:32} \\
\text{etime} & = & \text{2:33}
\end{bmatrix} \\
\text{rel} & = & \text{none}
\end{bmatrix}$$

The record in (7) represents the gesture shown in Fig. 3 ('claw' is used to label the American Sign Language (ASL) hand-shape *bent 5*). The values from (7) come from gesture annotation and are imported into TTR as objects of type *annotation predicate* (AP) – see [9] for an overview of a kinetic gesture annotation of this kind. In order to prevent value mismatches like hand-shapes occurring as directions, respective sub-typing of annotation predicates may be employed. For instance, using a type hierarchy from unification-based grammars, the type *AP* can be extended in terms of several sub-types, corresponding to the different kinds of values [12]. By this means, the gesture record entries can be regimented quite specifically, for instance, all *carrier* fields can be required to consist only of movement predicates. However, this rather technical detail is ignored in the following for the sake of brevity and the general type *AP* is used throughout.

The mnemonic labels introduce values for the handedness ('hand'; *left*, *right* or *both*), hand-shape ('hs' according to the *American Sign Language* alphabet), the movement path (where the movement is carried out by one or more 'carriers' [26]) and the relation to the other hand ('rel') as well as the temporal and locational properties ('sync'), where locations follow the gesture space model from Fig. 4.

Gestural movement can be brought about by one or more of three possible movement carriers: back-of-hand (boh), palm (plm) or wrist (wrst). A movement is captured in terms of a *direction* seen from the speaker (e.g. *move forward* (MF)) and a concatenation type which distinguishes straight ('line') from roundish ('arc') trajectories. For example, the same sequence of direction labels, MF>MR>MB, can give rise to an open rectangle or a semicircle, depending on the type of concatenation, as illustrated in (8):

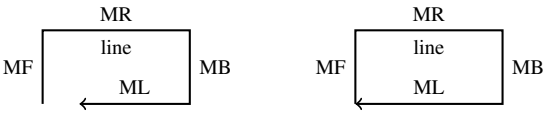
(8)



Note that the *sync*-feature's values allow to discriminate closed from incomplete shapes, which will be important for capturing gestalt properties (see Sec. IV below). For instance, the movement in (9) is underspecified with regard to the respective lengths of the movement parts. It can therefore represent both of the shapes illustrated in (10).

(9)
$$\begin{bmatrix}
\text{wrst} & = & \text{MF>MR>MB>ML} \\
\text{move} & = & \text{line>line>line>line}
\end{bmatrix}$$

(10)



The incomplete and the closed shape from (10) are distinguished in terms of their *sync* properties: closedness is defined as start (*sloc*) and end location (*eloc*) being the same, as expressed in the closeness condition **C-clos**.

C-clos:A gesture trajectory is closed iff start and end location are the same. The closure constraint has to distinguish one-handed from two-handed gestures:

- One-handed gesture: $\begin{bmatrix} \text{sloc} & : & AP \\ \text{eloc=sloc} & : & AP \end{bmatrix}$
- Two-handed gesture:

$$\begin{bmatrix}
\text{hands} & = & \text{both} \\
\text{lh} & = & \left[\text{sync} : \begin{bmatrix} \text{sloc} : AP \\ \text{eloc} : AP \end{bmatrix}\right] \\
\text{rh} & = & \left[\text{sync} : \begin{bmatrix} \text{sloc=lh.sync.sloc} : AP \\ \text{eloc=lh.sync.eloc} : AP \end{bmatrix}\right]
\end{bmatrix}$$

The basic representation format introduced above describes gesture events in terms of their kinetic sub-events and can be hooked to the "string theory of gesture events" straightforwardly. The outright string representation for the closed path gesture from (9), for example, is given in (11), where the gesture gets the event variable e:

[1]Matrix-based representations of gestures have been used in robotics at least since [25] and can be considered a standard representation format for gestures already.

$$(11)\quad e : \begin{bmatrix}\text{wrst} = \text{MF} \\ \text{sync} = \begin{bmatrix}\text{sloc} = \text{p1} \\ \text{eloc} = \text{p2}\end{bmatrix}\end{bmatrix} \widehat{\textit{line}} \begin{bmatrix}\text{wrst} = \text{MR} \\ \text{sync} = \begin{bmatrix}\text{sloc} = \text{p3} = \text{p2} \\ \text{eloc} = \text{p4}\end{bmatrix}\end{bmatrix} \widehat{\textit{line}}$$
$$\begin{bmatrix}\text{wrst} = \text{MB} \\ \text{sync} = \begin{bmatrix}\text{sloc} = \text{p5} = \text{p4} \\ \text{eloc} = \text{p6}\end{bmatrix}\end{bmatrix} \widehat{\textit{line}} \begin{bmatrix}\text{wrst} = \text{ML} \\ \text{sync} = \begin{bmatrix}\text{sloc} = \text{p7} = \text{p7} \\ \text{eloc} = \text{p8} = \text{p1}\end{bmatrix}\end{bmatrix}$$

By and large, the string notation using '$\frown$' and the gesture annotation using '>' are equivalent. That is, any record of the form shown in (7) can be translated into the string format illustrated in (11) without loss of information. In order to account for straight and bend movements, however, a small modification to string concatenation has to be made, however: the (temporal) string concatenation '$\frown$' is bifurcated into two spatial variants, '$\widehat{\textit{line}}$' and '$\widehat{\textit{arc}}$'. The string representation of gestures facilitates a rather detailed descriptive resolution. For instance, in order to decide whether movements fragments compose into one gesture event or belong to different gestures, convention **C-loc** can be employed:[2]

C-loc: If the start location of a movement part is identical to the end location of the previous movement part, both movement parts are concatenated within one gesture.

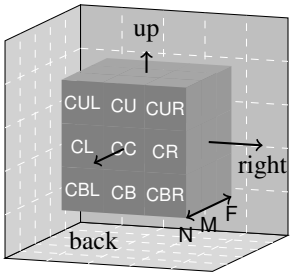
If **C-loc** is fulfilled, the more compact representation in (12) is preferred over the more detailed string representation (11), however:

$$(12)\quad \begin{bmatrix}\text{wrst} = \text{MF>MR>MB>ML} \\ \text{move} = \text{line>line>line>line} \\ \text{sync} = \begin{bmatrix}\text{sloc} = \text{p1} \\ \text{eloc} = \text{sloc} = \text{p1}\end{bmatrix}\end{bmatrix}$$

B. Trajectories within Vector Space

Given a kinetic representation format for gesture events, we need to spell out a semantic interpretation thereof in order to get access to the informativity of co-verbal gestures. To this end, a record type *Vec*(tor) is introduced, which provides an abstract model for configurations or trajectories. Via *Vec*, gesture representations are linked to a vector space semantics as developed by [28], [10]. The linking element in this mapping is the notion of *gesture space* [2], which refers to an inherently oriented space delimited by the reach of the speaker's arms. A respective gesture space model is illustrated in Fig. 4. The central anterior cube of this $3 \times 3 \times 3$-grid of cubes, which is labeled 'CC', is located right in front of the speaker's stomach. Each part of the space can be addressed by a name, which consists of a positioning plus a distance label. For instance, 'CUR-F' (*central upper right far*) is the topmost cube on the right, following the perspective employed in Fig. 4. These names can be used in order to provide values for the locational gesture representation fields conventionally labeled 'sloc' and 'eloc' – see example (7) above. The cube model also provides a regulating screw for the spatial granularity of the gesture space: the more cubes are employed, the higher the spatial resolution. The 27 different regions from the model in Fig. 4 are quite detailed already and are sufficient for present purposes.

[2]For a more sophisticated identity condition for gesture events see the criterion of [12], drawing on the event metaphysics of [27].



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

Fig. 4. *Gesture Space Model* seen from speaker's perspective

TABLE I
DIRECTIONAL CONSTRAINTS DERIVED FROM THE SAGITTAL PLANE OF THE GESTURE SPACE (EXTRACT). THE VECTOR TRANSLATION v OF THE BASIC CONFIGURATIONS IS ALSO GIVEN.

Configuration	=	Vector π_v	→	Constraints π_d
Handshape $\in$ {C, 5, B, O, Y}	=	$\{\mathbf{u}\}$	→	volume
{MF, MR, MB, ML}	=	$\mathbf{u}$	→	translational
$\varnothing$	=	–	→	–
MF>MR + line	=	$\mathbf{u} \perp \mathbf{v}$	→	orthogonal
MF>ML + line	=	$\mathbf{u} \perp \mathbf{v}$	→	orthogonal
MF>ML + arc	=	$\mathbf{u} \circ \mathbf{v}$	→	quadrant
MF>MR + arc	=	$\mathbf{u} \circ \mathbf{v}$	→	quadrant
...	=	...	→	...
MF + ...+ MB	=	$\mathbf{u}, \mathbf{u}^{-1}$	→	inverse
ML + ...+ MR	=	$\mathbf{u}, \mathbf{u}^{-1}$	→	inverse
sloc = eloc	=	$\mathbf{u}(0) = \mathbf{v}(1)$	→	closed
sloc $\neq$ eloc	=	$\mathbf{u}(0) \neq \mathbf{v}(1)$	→	open
lh.sloc = rh.sloc +	=	$\mathbf{u}(0) = \mathbf{v}(0)$		
lh.eloc = rh.eloc [two-handed]	=	$\mathbf{w}(1) = \mathbf{x}(1)$	→	closed
quadrant + quadrant + invers				semicircle
semicircle + semicircle				circle
orthogonal + orthogonal + invers				rectangular
rectangular + rectangular				rectangle
...				...
translational + crossing planes				diag(onal)

The gesture model spans along the three body planes, transverse (up-down), saggital (left-right) and frontal (front-back). For each plane, rules for inferring vectorial constraints from kinetic gesture representations can be formulated. Additionally, some hand shapes function as "line-thickness modifiers" of the gesture trajectory, giving rise to a three-dimensional body rather than to a two-dimensional sketch (cf. the work of [29]). Some rules that will be used below are collected in Table I by example of the sagittal plane. The rules are the back-bone of the gesture vectorization function π – see (14) below.

The primitive mathematical notion of a vector is used to model *paths*, where a path is a function $\mathbf{p} : [0, 1] \mapsto \mathbf{V}$, $\mathbf{V}$ being a three-dimensional vector space (for reasons of

simplicity we constrain ourselves to a purely spatial model; accounting also for temporal aspects would require $\mathbf{V}$ to be four-dimensional). Simple paths (i.e. lines) can be described by one vector, more complex paths arise out of vector sequences (like MF>MR>MB, respectively its vectorization). Each single vector $\mathbf{u}$ from a vector sequence has its own length in $[0, 1]$, ranging from the origin $\mathbf{u}(0)$ to the end $\mathbf{u}(1)$ ($\mathbf{u}(0.5)$ denotes the half of $\|\mathbf{u}\|$). In a vector sequence like $\mathbf{u} \perp \mathbf{v}$ it holds that $\mathbf{u}(1) = \mathbf{v}(0)$, that is, the described path is consecutive. Additionally, there are various kinds of vectors, discriminating, amongst others, vectorial representations of locations (needed for example for prepositional modifications like "3 cm above x") and of shapes (figuring as spatial denotations of predicates like "round") [30]. Finally, the core vectorial representations are extended with some shape-related features like being translational or circular. These features are due to the work of [31] (who uses partly differently named features, though), where they are used as lexical constraints on path shapes of (mainly) motion verbs. Accordingly, *Vec* consists of three basic fields: *Vtype* determines the kind of vector sequence in question (i.e., axis, path, ...), *Vpath* stores the path's shape and *Vshape* introduces shape-related constraints. For two-handed gestures, each hand gives rise to a record of type *Vec*. In that case, an additional field (conventionally labeled "comb") also of type *Vec* is introduced, which captures hand-crossing, combined information (a simple example is given in Sec. IV by means of the *concrete base*-gesture from Fig. 2).

$$(13)\quad \mathit{Vec} =_{\text{def}} \begin{bmatrix} \text{vt} & : & \text{Vtype} \\ \text{pt} & : & \text{Vpath} \\ \text{sh} & : & \text{set(Vshape)} \end{bmatrix}$$

There is a functional type π which maps records of annotations (i.e., records with entries of type *AP*) – labeled Rec^{AP} for short – onto vectors (*Vec*). According to the division of labor between vectorization and representational feature decomposition (cf. Table I and the above-given explanation), π consists of two sub-types, π_v and π_d:

$$(14)\quad \pi : [[\pi_v : \mathit{Rec}^{AP} \rightarrow \mathit{Vpath}] \rightarrow \pi_d : \mathit{Vpath} \rightarrow \text{set(Vshape)}] \rightarrow \mathit{Vec}$$

Keeping vectorization and feature decomposition apart will also play a role in accounting for the *good continuation* in Sec. IV below.

The vector function π exploits the constrains from Table I and works schematically as follows:

$$(15)\quad \text{if } r\text{:} \begin{bmatrix} \text{hand} & : & \text{AP} \\ \text{hs} & : & \text{AP} \\ \text{carrier} & : & \begin{bmatrix} \text{boh} & : & \text{AP} \\ \text{plm} & : & \text{AP} \\ \text{wrst} & : & \text{AP} \\ \text{move} & : & \text{AP} \end{bmatrix} \\ \text{sync} & : & \begin{bmatrix} \text{sloc} : \text{AP} \\ \text{eloc} : \text{AP} \end{bmatrix} \end{bmatrix},$$

$$\text{then } \pi_v(r) = \begin{cases} \text{pt1} = \begin{bmatrix} v(r.\text{wrst}, r.\text{move} \\ v(\text{sloc}, \text{eloc})) \end{bmatrix} : \text{Vpath} & \text{if } r.\text{wrst} \neq \varnothing \\ \text{pt2} = \begin{bmatrix} v(r.\text{plm}, r.\text{move} \\ v(\text{sloc}, \text{eloc})) \end{bmatrix} : \text{Vpath} & \text{if } r.\text{plm} \neq \varnothing \\ \text{pt3} = \begin{bmatrix} v(r.\text{boh}, r.\text{move} \\ v(\text{sloc}, \text{eloc})) \end{bmatrix} : \text{Vpath} & \text{if } r.\text{boh} \neq \varnothing \end{cases}$$

$$\text{then } \pi_d(\pi_v(r)) = \begin{cases} [\text{sh} = d(\text{pt1})] : \text{set(Vshape)} & \text{if pt1} \neq \varnothing \\ [\text{sh} = d(\text{pt2})] : \text{set(Vshape)} & \text{if pt2} \neq \varnothing \\ [\text{sh} = d(\text{pt3})] : \text{set(Vshape)} & \text{if pt3} \neq \varnothing \end{cases}$$

Since a single gesture may involve more than one movement path – an example is given in Sec. IV below – π_v just numbers the 'pt' values from the possible movement carriers. The procedure is incremental in that immediate pairs as well as skip pairs of directional APs are considered and compared to the table entries above: that is, neighboring APs are evaluated in terms of being quadrant or orthogonal, longer sequences of APs are inspected for fulfilling inversion. The features collected in the constraints of Table I then give rise to shape descriptions. For instance, by dint of the cooperation of Table I and π, the sample gesture from (7) is translated into the following perpendicular vector sequence:

$$(16)\quad \text{a.}\quad \pi_v\left(\begin{bmatrix} \text{wrst} = \text{MR>MB>ML} \\ \text{move} = \text{line>line>line} \\ \text{sync} = \begin{bmatrix} \text{sloc} = \text{p1} \\ \text{eloc} = \text{p2} \neq \text{p1} \end{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} \text{pt1} : \begin{bmatrix} \mathbf{u} \perp \mathbf{v} \perp \mathbf{w} \\ \mathbf{u}(0) \neq \mathbf{w}(1) \end{bmatrix} \end{bmatrix}$$

$$\text{b.}\quad \pi_d\left(\begin{bmatrix} \text{pt1} : \begin{bmatrix} \mathbf{u} \perp \mathbf{v} \perp \mathbf{w} \\ \mathbf{u}(0) \neq \mathbf{w}(1) \end{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} \text{sh} : \{\text{rectangular, open}\} \end{bmatrix}$$

Note that the *Vtype* field is underspecified in (16) – there is nothing within the kinetic gesture representation that determines the kind of the vector. Vector types can be instantiated in interaction with the linguistic lexicon when combined with speech.

C. Linking Gesture Perception to Reasoning

In representing meaning, a dynamic information state update semantics as developed in [11] is used. There, the context (e.g., presuppositions) is explicitly separated from the content of linguistic expressions in terms of *backgrounded* (bg) and *foregrounded* (f) information. For instance, the proper name example "Max" from (5) above is recaptured as follows:

$$(17)\quad [\![\text{Max}]\!] = \begin{bmatrix} \text{bg} & = & [\text{x} : \text{Ind}] \\ \text{f} & = & \lambda r : \text{bg}\left([\text{c}_{\text{pn}} : \text{named}(r.\text{x}, \text{"Max"})]\right) \end{bmatrix}$$

In order to implement the dynamic update of information, the functional type exemplified in (17) has to "accumulate" backgrounded information. This is expressed in terms of a *fixed point type* [8] – cf. also [11]. The fixed point type $\mathcal{F}$ corresponds to the unification of the domain and the range of a functional type. For instance, the fixed point type for the foregrounded meaning of "Max" from (17) is shown in (18):

(18) $\mathcal{F}([\![\text{Max}]\!].\text{f}) = \mathcal{F}(\lambda r : \left[\text{x} : \text{Ind}\right]\left(\left[\text{c}_{\text{pn}} : \text{named}(r.\text{x}, \text{“Max”})\right]\right))$
$= \begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{c}_{\text{pn}} & : & \text{named(x, “Max”)} \end{bmatrix}$

In general, context update proceeds as formulated in **C-upc**, where agents' contributions extend their current information state (i.e. their "take of the situation" – see [11, p. 8]):

C-upc:**(preliminary version)** If the current information state s_t is compatible to the background information of expression e, then $[\![\text{e}]\!]$ can be added to s_t to form information state s_{t+1}:
If $s_t \sqsubseteq [\![\text{e}]\!].bg$ then $s_{t+1} = s_t \wedge \mathcal{F}([\![\text{e}]\!].\text{f})$

Operation '$\wedge$' in (18) is the type-theoretic analog to unification – see [8] for details. Gestures constitute a *display situation* (DP) [32] and as such are part of the publicly available information state. The value of DP is a list: every newly produced gesture is put into initial position while all possibly already present gestures are stacked and remain available for eventual repair or clarification.

C-upg:If $G = [v : Vec]$ is the vector interpretation of a gesture g produced at state s_t, the display situation list of s_t gets updated with G:
If $\pi(g) = G$ at s_t, then s_{t+1}.dp.rest = s_t.dp and s_{t+1}.dp.first = G

How is the link between display situations and dynamic meaning implemented? Following the exemplification [33] account of [12], space-related predicates are equipped with a *conceptual vector meaning* (CVM) which extends their intensions by means of vectorial constraints (cf. also the classification approach of [11] and the lexical decomposition approach of [34]). In particular, CVM information uses representations that are of type *Vec*. A spatial predicate like *U-shaped*, for instance, has the following extended dynamic meaning:

(19) $[\![\text{U-shaped}]\!] =$

$$\begin{bmatrix} \text{bg} & = & \left[\text{x} : \text{Ind}\right] \\ \text{f} & = & \lambda r : \text{bg}\left(\left[\begin{array}{l} \text{c}_{\text{u}} : \text{U-shaped}(r.\text{x}) \\ \text{cvm} = \left[\text{vt} : \text{axis-path}(r.\text{x}, \text{pt}) \right. \\ \qquad \left. \text{pt} : \begin{bmatrix} \mathbf{u} \perp \mathbf{v} \perp \mathbf{w} \\ \mathbf{u}(0) \neq \mathbf{w}(1) \end{bmatrix} \right. \\ \qquad \left. \text{sh} : \left\{\text{rectangular, open}\right\} \right] : Vec \\ \text{c}_{\text{shape}} : \text{shape}(r.\text{x}, \text{cvm}) \end{array}\right)\right) \end{bmatrix}$$

That is, part of the descriptive meaning of 'U-shaped' is that its argument has a certain perceived shape, namely an axis that is of a particular rectangular configuration. Spatial representations like those in (19) are related to *imagistic description trees* employed in [35] in order to facilitate online recognition of iconic gestures within an artificial intelligence application. There, gesture tracking data is mapped onto (parts of) prototypical object shapes. A similar set of features from a three-axial system has also been employed in [36]. In contrast to such AI recognition approaches, the focus here is on linking multimodal utterances to linguistics and dialog theory.

The occurrence of an iconic gesture adds respective information to the DP of the current information state. Although it is clear that there is an informational interaction between CVM and DP to the effect that the merge operation ('$\wedge$') rules out combinations of predicates and gesture trajectories that are conflicting with regard to their respective *Vec* type information, this is not yet covered by the general context update **C-upc**. Due to the differences of labeling, the revised and final formulation explicitly has to take care of relating "dp" to "cvm":

C-upc:**(final version)** The display situation punctually affects the update mechanism in that DP information has to be merged with linguistic information:
If $s_t \wedge_{s_t.\text{dp.first} \wedge e.\text{cvm}} \sqsubseteq [\![\text{e}]\!].bg$ then $s_{t+1} = s_t \wedge \mathcal{F}([\![\text{e}]\!].\text{f})$

C-upc now correctly predicts that it is not well-formed to co-produce, say, a rectangular gesture and an adjective denoting roundness. The update mechanism is clocked by events: occurrences of speech (words) as well as of gestures trigger **C-upc**. Co-verbal gestures are integrated to the co-occuring linguistic expressions. This provides a co-occurrence based association between gesture and speech events that is not prone to functional ambiguities of the system in [19]. Such ambiguities arise if there are two possible attachment points in speech (say, due to signals involving deictic *here*) but only one attaching (pointing) gesture. Although **C-upc** goes beyond functional systems, it fails to do justice to the temporal and coherence relations that govern speech-gesture integration, in particular with regard to *hold* gestures [37], that are "frozen" gestural configurations that persist beyond the short life of affiliated verbal utterances. In order to provide a more sophisticated account to multimodal integration, event-based information state update mechanisms have to be extended by more detailed grammatical constraints incorporating temporal, intonational and semantic information [38], [12].

IV. Hyponymy, Good Continuation, and Demonstration

The TTR framework for iconic gestures sketched in the previous sections is in the following sections applied to the phenomena from Sec. II, which motivated the semantic integration of gestures in the first place.

The first example in (20), which is repeated from (1), introduces a spiral trajectory into the current information state.

(20) *Ich g[laube das sollen TREP]pen sein*
I think that should staircases be

'I think that should be staircases' + Fig. 1

The gesture event is represented as follows:

(21)
$$\begin{bmatrix} \text{hand} & = & \text{right} \\ \text{hs} & = & \text{G} \\ \text{carrier} & = & \begin{bmatrix} \text{wrst} = \text{MU} \\ \text{boh} = \text{MR>MF>ML>MB>MR} \\ \text{move} = \text{arc>arc>arc>arc} \end{bmatrix} \\ \text{sync} & = & \begin{bmatrix} \text{sloc} = \text{CR-N} \\ \text{eloc} = \text{CUR-N} \end{bmatrix} \end{bmatrix}$$

The spiral gesture has two motion aspects: a rotation carried out in the back of the hand and a translation performed by lifting the hand (i.e. wrist) – cf. the decomposition of 'spiralness' given in [31, p. 411] and of the spiral gesture in [12, p. 238]. The vector interpretation thus returns the following record of type *Vec*, where any segment '$\mathbf{u} \circ \mathbf{v}$' indicates that the vector sequence in question describes the fourth of a circle (cf. Table I):

(22)
$$\begin{bmatrix} \text{pt1} & : & \begin{bmatrix} \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \circ \mathbf{z} \circ \mathbf{y} \\ \mathbf{u}(0) \neq \mathbf{y}(1) \end{bmatrix} \\ \text{pt2} & : & \mathbf{a} \\ \text{sh} & : & \{\text{translational, circle, open}\} \end{bmatrix}$$

Note that the translational component introduced by pt2 distinguishes spirals from circles. The current context thus provides a gesticulated witness for the spiral type within the display situation:

(23)
$$s_t = \begin{bmatrix} \text{dp} = \begin{bmatrix} \text{pt1} & : & \begin{bmatrix} \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \circ \mathbf{z} \circ \mathbf{y} \\ \mathbf{u}(0) \neq \mathbf{y}(1) \end{bmatrix} \\ \text{pt2} & : & \mathbf{a} \\ \text{sh} & : & \{\text{translational, circle, open}\} \end{bmatrix} \end{bmatrix}$$

The noun "Treppen" (*staircases*) is unspecific with respect to kind and shape. It has a number of sub-kinds, however, that are distinguished according to their layout. This hyponymic relationship is captured within the inheritance type hierarchy in Fig. 5, where descendants provide the informational difference to their parent.

Given s_t from (23) and given the hyponymic relationships in Fig. 5, there are two possible updates to reach state s_{t+1}: one update uses the literally uttered hypernym *staircases*, the other the more specific *spiral staircases*, for which the context provides information due to the iconic gesture (updating with *straight staircases* is blocked, however, due to incompatible CVM constraints). Since more specific updates are generally preferred, the gesture allows us to infer the kind of staircases talked about. After applying **C-upc**, information state s_{t+1} looks as follows:

(24)
$$s_{t+1} = \begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{dp} & : & \begin{bmatrix} \text{pt1} : \begin{bmatrix} \mathbf{u} \circ \mathbf{v} \circ \mathbf{w} \circ \mathbf{z} \circ \mathbf{y} \\ \mathbf{u}(0) \neq \mathbf{y}(1) \end{bmatrix} \\ \text{pt2} : \mathbf{a} \\ \text{sh} : \{\text{translational, circle, open}\} \\ \text{vt1} : \text{axis-path(x, pt1)} \\ \text{vt2} : \text{axis-path(x, pt2)} \end{bmatrix} \\ \text{cvm=dp} & : & \textit{Vec} \\ c_{\text{u}} & : & \text{spiral-staircases(x)} \\ c_{\text{shape}} & : & \text{shape(x, cvm)} \end{bmatrix}$$

In example (2), re-given in (25), we observe a use of hand-shape that cuts out a volume rather than draws a trajectory on a plane (due to the "voluminous" hand shape 'C').

(25) *die Skulptur die hat 'n [Betonsockel]*
the sculpture it has a concrete base

'the sculpture has a concrete base' + Fig.2

The gesture that is part of (25) is produced with both hands, so kinetic features are distributed over the left and the right hand. Accordingly, the *Vec* type of the gesture has double entries, partitioned according to the carriers of the respective hand. Gesture representation and interpretation is shown in (26) and (27), respectively.

(26)
$$\begin{bmatrix} \text{hands = both} \\ \text{rh} = \begin{bmatrix} \text{hand} & = & \text{right} \\ \text{hs} & = & \text{C} \\ \text{carrier} & = & \begin{bmatrix} \text{wrst = MR>MF} \\ \text{move = arc} \end{bmatrix} \\ \text{sync} & = & \begin{bmatrix} \text{sloc=lh.sync.sloc = CC-N} \\ \text{eloc = CR-M} \end{bmatrix} \end{bmatrix} \\ \text{lh} = \begin{bmatrix} \text{hand} & = & \text{left} \\ \text{hs} & = & \text{C} \\ \text{carrier} & = & \begin{bmatrix} \text{wrst = ML>MF} \\ \text{move = arc} \end{bmatrix} \\ \text{sync} & = & \begin{bmatrix} \text{sloc = CC-N} \\ \text{eloc = CL-M} \end{bmatrix} \end{bmatrix} \\ \text{rel = axisymmetric} \end{bmatrix}$$

(27)
$$\begin{bmatrix} \text{pt1lh} & = & \begin{bmatrix} \{\mathbf{u} \circ \mathbf{v}\} \\ \mathbf{u}(0) \neq \mathbf{v}(1) \end{bmatrix} \\ \text{pt1rh} & = & \begin{bmatrix} \{\mathbf{w} \circ \mathbf{x}\} \\ \mathbf{w}(0) \neq \mathbf{x}(1) \end{bmatrix} \\ \text{comb} & = & \begin{bmatrix} \text{pt} = \begin{bmatrix} \mathbf{u}(0) = \mathbf{w}(0) \\ \mathbf{v}(1) \neq \mathbf{x}(1) \\ \mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \\ \mathbf{a}(0) \neq \mathbf{c}(1) \end{bmatrix} \\ \text{sh} = \{\text{semicircle, volume, open}\} \end{bmatrix} \end{bmatrix}$$

The vector representation in the 'comb'-field introduces combined path information for both hands. In the example above, the two quartercircles from the axisymmetrical movement of the hands combine to a semicircle. Modeling two-handed gestures thus adds another level of complexity which can only be mentioned here (with the exception of the closure contraint in **C-clos** – see also Table I).

The lexical entry for *concrete-base* is underspecified with respect to shape and does not have any shape-related hyponyms:

(28)
$$\begin{bmatrix} \text{bg} & = & \begin{bmatrix} \text{x} : \text{Ind} \end{bmatrix} \\ \text{f} & = & \lambda r : \text{bg}\left(\begin{bmatrix} c_{\text{cb}} : \text{concrete-base}(r.\text{x}) \end{bmatrix}\right) \end{bmatrix}$$

The descriptive meaning in (28) imposes no top-down constraints on CVM or DP. Additionally, the gesture path violates the closure constraint **C-clos**. For these reasons, it is likely

$$[\![\text{staircases}]\!] = \begin{bmatrix} \text{bg} & = & \left[\text{x} : \text{Ind}\right] \\ \text{f} & = & \lambda r : \text{bg}\left(\left[\text{c} : \text{staircases}(r.\text{x})\right]\right) \end{bmatrix}$$

$$[\![\text{spiral-staircases}]\!] = \left[\text{f} = \lambda r : \text{bg}\left(\left[\begin{array}{l} \text{cvm} = \left[\begin{array}{l} \text{pt1} : \left[\begin{array}{l}\mathbf{u}\circ\mathbf{v}\circ\mathbf{w}\circ\mathbf{z}\circ\mathbf{y} \\ \mathbf{u}(0)\neq\mathbf{y}(1)\end{array}\right] \\ \text{pt2} : \mathbf{a} \\ \text{sh} : \left\{\text{translational, circle, open}\right\} \\ \text{vt1} : \text{shape-path}(r.\text{x, pt1}) \\ \text{vt2} : \text{axis-path}(r.\text{x, pt2}) \end{array}\right] : \mathit{Vec} \\ \text{c}_{\text{shape}} : \text{shape}(r.\text{x, cvm}) \end{array}\right)\right]$$

$$[\![\text{straight-staircases}]\!] = \left[\text{f} = \lambda r : \text{bg}\left(\left[\begin{array}{l} \text{cvm} = \left[\begin{array}{l} \text{pt} : \mathbf{u} \\ \text{sh} : \left\{\text{translational, diag}\right\} \\ \text{vt} : \text{axis-path}(r.\text{x, pt}) \end{array}\right] : \mathit{Vec} \\ \text{c}_{\text{shape}} : \text{shape}(r.\text{x, cvm}) \end{array}\right)\right]$$

Fig. 5. Type hierarchy showing the hypernym "staircases" and two of its hyponyms.

that the gesture path is elliptical. There are several (in fact, infinitely many) ways to close the path so that the vector sequence has at least two vectors whose coordinates coincide. Out if these options only one is a good continuation, namely that one which brings about the shortest closure while maintaining the concatenation type (*arc*, in this case). In other words, 'good continuation', *GoCont*, is a function mapping record types *Vec* to record types *Vec*, i.e. $\mathit{GoCont} : \mathit{Vec} \rightarrow \mathit{Vec}$. Since *GoCont* changes the vector sequence representation (but not the movement annotation, of course) of a gesture, it gives rise to a re-labeling by means of π_d (cf. (14) above). The above-given features of *GoCont* can be formulated as constraints over an input display situation 'dp_{in}' and an output display situation 'dp_{out}', both being of type *Vec*.

(29) $\mathit{GoCont} =_{\text{def}}$

$$\lambda r : \begin{bmatrix} \text{ap1} & = & \text{open} : \mathit{AP} \\ \text{cc} & = & \{\circ\,|\perp\} : \text{Vpath} \\ \text{dp}_{\text{in}} & : & \begin{bmatrix} \text{sh} : \text{set(AP)} \\ \text{pt} : \text{Vpath} \\ \text{vt} : \text{Vtype} \end{bmatrix} \\ \text{c}_{\text{memb}} & : & \text{member(ap1, dp}_{\text{in}}\text{.sh)} \\ \text{c}_{\text{conc}} & : & \text{member(cc, dp}_{\text{in}}\text{.pt)} \\ \text{cvm} & : & \varnothing \end{bmatrix}$$

$$\left(T = \begin{bmatrix} \text{spt} & : & \text{Vpath} \\ \text{c}_{\text{cond}} & : & \text{member}(r.\text{cc, spt}) \\ \text{dp}_{\text{out}} & : & \left[\begin{array}{l}\text{pt} = \left[\begin{array}{l} r.\text{dp}_{\text{in}}\text{.pt}\ r.\text{cc}\ \text{spt} \\ r.\text{dp}_{\text{in}}\text{.pt}(0) = \text{spt}(1)\end{array}\right] \\ \text{vt} = r.\text{dp}_{\text{in}}\text{.vt} : \text{Vtype}\end{array}\right] \end{bmatrix}\right) . \ \pi_d(T),$$

where "$\{\circ\,|\perp\}$" means that the concatenation type is either arc-like ($\widehat{arc}$) or straight ($\widehat{line}$).

Applying (the two-handed extension of) *GoCont* to the path from (28) gives rise to a voluminous circle, that is, a cylinder:

(30)

$$\mathit{GoCont}\left(\left[\begin{array}{l} \text{dp}_{\text{in}} = \left[\begin{array}{lcl} \text{pt1lh} & = & \left[\begin{array}{l}\left\{\mathbf{u}\circ\mathbf{v}\right\} \\ \mathbf{u}(0)\neq\mathbf{v}(1)\end{array}\right] \\ \text{pt1rh} & = & \left[\begin{array}{l}\left\{\mathbf{w}\circ\mathbf{x}\right\} \\ \mathbf{w}(0)\neq\mathbf{x}(1)\end{array}\right] \\ \text{comb} & = & \left[\begin{array}{l}\text{pt} = \left[\begin{array}{l}\mathbf{u}(0)=\mathbf{w}(0) \\ \mathbf{v}(1)\neq\mathbf{x}(1) \\ \mathbf{a}\circ\mathbf{b}\circ\mathbf{c} \\ \mathbf{a}(0)\neq\mathbf{c}(1)\end{array}\right] \\ \text{sh} = \left\{\text{semicircle, volume, open}\right\}\end{array}\right] \\ \text{cvm} & = & \varnothing \end{array}\right] \end{array}\right]\right)$$

$$\rightarrow \left[\begin{array}{l} \text{dp}_{\text{out}} = \left[\begin{array}{lcl} \text{pt1lh} & = & \left[\begin{array}{l}\left\{\mathbf{u}\circ\mathbf{v}\circ\mathbf{y}\right\} \\ \mathbf{u}(0)\neq\mathbf{y}(1)\end{array}\right] \\ \text{pt1rh} & = & \left[\begin{array}{l}\left\{\mathbf{w}\circ\mathbf{x}\circ\mathbf{z}\right\} \\ \mathbf{w}(0)\neq\mathbf{z}(1)\end{array}\right] \\ \text{comb} & = & \left[\begin{array}{l}\text{pt} = \left[\begin{array}{l}\mathbf{u}(0)=\mathbf{w}(0) \\ \mathbf{y}(1)=\mathbf{z}(1) \\ \mathbf{a}\circ\mathbf{b}\circ\mathbf{c}\circ\mathbf{d}\circ\mathbf{e} \\ \mathbf{a}(0)=\mathbf{e}(1)\end{array}\right] \\ \text{sh} = \left\{\text{circle, volume, closed}\right\}\end{array}\right] \end{array}\right] \end{array}\right]$$

The good continuation is accomplished by the combined path of both hands.

Since a cylinder is a regular shape that has a lexicalized verbalization, its CVM makes the connection between the trajectory from (30) and the intension of *cylinder* explicit:

(31) $[\![\text{cylinder}]\!] =$

$$\left[\begin{array}{lcl} \text{bg} & = & \left[\text{x} : \text{Ind}\right] \\ \text{f} & = & \lambda r : \text{bg}\left(\left[\begin{array}{l} \text{c}_{\text{cy}} : \text{cylinder}(r.\text{x}) \\ \text{cvm} = \left[\begin{array}{lcl} \text{vt} & = & \text{shape-path}(r.\text{x, cvm}) \\ \text{pt} & = & \left[\begin{array}{l}\left\{\mathbf{a}\circ\mathbf{b}\circ\mathbf{c}\circ\mathbf{d}\circ\mathbf{e}\right\} \\ \mathbf{a}(0)=\mathbf{e}(1)\end{array}\right] \\ \text{sh} & = & \left\{\text{circle, volume, closed}\right\}\end{array}\right] : \mathit{Vec} \\ \text{c}_{\text{shape}} : \text{shape}(r.\text{x, cvm}) \end{array}\right)\right] \end{array}\right]$$

Given the good continuation and the DP triggering of the lexical resource from (31), the most informative information state update s_{t+1} is the following (using only combined paths):

(32)
$$s_{t+1} = \begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{dp} & = & \textit{GoCont}\left(\begin{bmatrix} \text{pt} = \left\{\mathbf{a} \circ \mathbf{b} \circ \mathbf{c}\right\} \\ \text{sh} = \left\{\text{semicircle, volume, open}\right\} \end{bmatrix}\right) \\ & & \rightarrow \begin{bmatrix} \text{vt} & = & \text{shape-path(x, cvm)} \\ \text{pt} & = & \begin{bmatrix} \left\{\mathbf{a} \circ \mathbf{b} \circ \mathbf{c} \circ \mathbf{d} \circ \mathbf{e}\right\} \\ \mathbf{a}(0) = \mathbf{e}(1) \end{bmatrix} \\ \text{sh} & = & \left\{\text{circle, volume, closed}\right\} \end{bmatrix} \\ \text{cvm=dp} & : & \textit{Vec} \\ \text{c}_{\text{cb}} & : & \text{concrete-base(x)} \\ \text{c}_{\text{cy}} & : & \text{cylinder(x)} \\ \text{c}_{\text{shape}} & : & \text{shape(x, cvm)} \end{bmatrix}$$

The final example also concerns the depiction of shape that is not realized in speech. The difference, however, being that in the datum given in (33), replicating example (3), the shape of the house talked about is explicitly delegated to the co-verbal gesture due to the demonstrative "so":

(33) *dann ist das Haus halt so []*
then is the house just this []

'then the house is like this' + Fig. 3

For this reason, the DP triggers the (initially empty) shape field of the target noun *house* directly, rather then detouring *via* a collateral expression as in (32), although the resulting information state does not reflect this subtle difference any more.

The lexical entry for house is a standard one-place predicate whose shape-field is unfilled:

(34) $[\![\text{house}]\!] =$
$$\begin{bmatrix} \text{bg} & = & \left[\text{x} : \text{Ind}\right] \\ \text{f} & = & \lambda r : \text{bg}\left(\begin{bmatrix} \text{c}_{\text{hs}} : \text{house } (r.\text{x}) \\ \text{cvm} : \textit{Vec} \\ \text{c}_{\text{shape}} : \text{shape}(r.\text{x, cvm}) \end{bmatrix}\right) \end{bmatrix}$$

The information state after processing the noun has the following public information:

(35)
$$s_{t+1} = \begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{c}_{\text{hs}} & : & \text{house (x)} \\ \text{cvm} & : & \textit{Vec} \\ \text{c}_{\text{shape}} & : & \text{shape(x, cvm)} \end{bmatrix}$$

The gesture, which has been described and related to the predicate *U-shaped* in (16) and (19) above triggers a further update **C-upc**, leading to identifying the gesture's trajectory and the exemplified predicate with the shape description of the house. The resulting state s_{t+2} is shown in (36):

(36)
$$s_{t+2} = \begin{bmatrix} \text{x} & : & \text{Ind} \\ \text{c}_{\text{hs}} & : & \text{house (x)} \\ \text{cvm=dp} & : & \textit{Vec} \\ \text{c}_{\text{shape}} & : & \text{shape(x, cvm)} \\ \text{dp} & = & \begin{bmatrix} \text{pt} : \begin{bmatrix} \mathbf{u} \perp \mathbf{v} \perp \mathbf{w} \\ \mathbf{u}(0) \neq \mathbf{w}(1) \end{bmatrix} \\ \text{sh} : \left\{\text{rectangular, open}\right\} \end{bmatrix} : \textit{Vec} \\ \text{c}_{\text{u}} & : & \text{U-shaped(x)} \end{bmatrix}$$

Note, however, that the *type* of the vector (the field labeled 'vt') with regard to the shape of the house is not specified. Since the lexical entry for *house* leaves shape information underspecified and the trajectory is neutral about its type, no type information is available from these resource. The most likely type 'axis-path' has to be inferred once again, but this inference is beyond the descriptive coverage of this paper.

V. Conclusion

A formal model has been sketched for relating the perception of iconic gestures to language within an artificial intelligence-oriented information state update framework. The interface between low-level perceptual features and semantic predicates is spelled out in terms of TTR, a large-scale formal theory for language, perception and interaction [39], [40]. The model accounts for semantic key phenomena of multimodal discourse by example of speech-gesture integration like meaning specification, speech-gesture mismatches and semantic enrichments. Following the general framework of [12], a characteristics is that iconic gestures are not interpreted extensionally directly, but rather related to percepts and intensional features of natural language predicates. As the dialog proceeds, gestures are stacked onto the "gesture storage", which allows to approximate the temporal interplay of speech and gesture. Contrary to spoken language, gestures can be "frozen" and kept persistent over a period of talking. Relating both communication means via dynamic information state update mechanisms paves the way for integrating a more detailed time-based notion like that of "communication channels" [37]. Further extensions include the integration of grammatical and dialog-interactive representations of information states as well as a broadened empirical range of non-verbal behavior. In particular two-handed gestures and the combined paths they give rise to need special treatment. Further descriptive extensions might involve methodological extensions as well: the gesture perception part is a typical application area for machine learning approaches.

References

[1] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge, MA: Cambridge University Press, 2004.

[2] D. McNeill, *Hand and Mind – What Gestures Reveal about Thought*. Chicago: Chicago University Press, 1992.

[3] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[4] M. Refice, M. Savino, M. Caccia, and M. Adduci, "Automatic classification of gestures: a context-dependent approach," in *Proceedings of the 2011 Federated Conference on Computer Science and Information System*, 2011, pp. 743–750.

[5] T. Hachaj, M. R. Ogiela, and M. Piekarczyk, "Dependence of Kinect sensors number and position on gestures recognition with gesture description language semantic classifier," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. IEEE, 2013, pp. 571–575.

[6] M. W. Alibali, "Gesture in spatial cognition: Expressing, communicating, and thinking about spatial information," *Spatial Cognition and Computation*, vol. 5, pp. 307–331, 2005.

[7] R. Cooper and A. Ranta, "Natural languages as collections of resources," in *Language in Flux: Relating Dialogue Coordination to Language Variation, Change and Evolution*, ser. Communication, Mind and Language, R. Cooper and R. Kempson, Eds. London: College Publications, 2008.

[8] R. Cooper, "Type theory and semantics in flux," in *Philosophy of Linguistics*, ser. Handbook of Philosophy of Science, R. Kempson, T. Fernando, and N. Asher, Eds. Oxford and Amsterdam: Elsevier, 2012, vol. 14, pp. 271–323.

[9] A. Lücking, K. Bergmann, F. Hahn, S. Kopp, and H. Rieser, "The Bielefeld speech and gesture alignment corpus (SaGA)," in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, ser. LREC 2010. Malta: 7th International Conference for Language Resources and Evaluation, 2010. doi: 10.13140/2.1.4216.1922 pp. 92–98.

[10] J. Zwarts and Y. Winter, "Vector space semantics: A model-theoretic analysis of locative prepositions," *Journal of Logic, Language, and Information*, vol. 9, no. 2, pp. 169–211, 2000.

[11] S. Larsson, "Formal semantics for perceptual classification," *Journal of Logic and Computation*, 2013. doi: 10.1093/logcom/ext059

[12] A. Lücking, *Ikonische Gesten. Grundzüge einer linguistischen Theorie*. Berlin and Boston: De Gruyter, 2013, zugl. Diss. Univ. Bielefeld (2011).

[13] S. Dobnik, R. Cooper, and S. Larsson, "Modelling language, action and perception in Type Theory with Records," in *Proceedings of the 7th International Workshop on Constraint Solving and Language Processing*, ser. CSLP'12, 2012, pp. 51–62.

[14] A. Kendon, "Gesticulation and speech: Two aspects of the process of utterance," in *The Relationship of Verbal and Nonverbal Communication*, ser. Contributions to the Sociology of Language, M. R. Key, Ed. The Hague: Mouton, 1980, vol. 25, pp. 207–227.

[15] D. Loehr, "Apects of rhythm in gesture in speech," *Gesture*, vol. 7, no. 2, pp. 179–214, 2007.

[16] H. Rieser, "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, 2008, pp. 167–174.

[17] M. Johnston, P. R. Cohen, D. McGee, S. L. Oviatt, J. A. Pittman, and I. Smith, "Unification-based multimodal integration," in *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics*, European Chapter Meeting of the ACL. Madrid, Spain: Association for Computational Linguistics, 1997, pp. 281–288.

[18] M. Johnston, "Deixis and conjunction in multimodal systems," in *Proceedings of the 18th Conference on Computational Linguistics – Volume I*, International Conference On Computational Linguistics. Saarbrücken, Germany: Association for Computational Linguistics, 2000, pp. 362–368.

[19] B. Bringert, R. Cooper, P. Ljunglöf, and A. Ranta, "Multimodal dialogue system grammars," in *Proceedings of the 9th Workshop on the Semantics and Pragmatics of Dialogue*, ser. Dialor'05, 2005.

[20] A. Lücking, H. Rieser, and M. Staudacher, "Multi-modal integration for gesture and speech," in *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, ser. Brandial'06, D. Schlangen and R. Fernández, Eds. Potsdam: Universitätsverlag Potsdam, 2006, pp. 106–113.

[21] J. Barwise and J. Perry, *Situations and Attitudes*, ser. The David Hume Series of Philosophy and Cognitive Science Reissues. Stanford: CSLI Publications, 1983.

[22] H. Kamp and U. Reyle, *From Discourse to Logic*. Dordrecht: Kluwer Academic Publishers, 1993.

[23] R. Montague, *Formal Philosophy: Selected Papers*. New Haven: Yale University Press, 1974.

[24] T. Fernando, "Observing events and situations in time," *Linguistics and Philosophy*, vol. 30, pp. 527–550, 2007. doi: 10.1007/s10988-008-9026-1

[25] S. Kopp, P. Tepper, and J. Cassell, "Towards integrated microplanning of language and iconic gesture for multimodal output," in *Proceedings of the International Conference on Multimodal Interfaces (ICMI'04)*. ACM Press, 2004, pp. 97–104.

[26] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception & Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[27] L. B. Lombard, *Events: A Metaphysical Study*. London: Routledge & Kegan Paul, 1986.

[28] J. Zwarts, "Vectors as relative positions: A compositional semantics of modified PPs," *Journal of Semantics*, vol. 14, no. 1, pp. 57–86, 1997.

[29] H. Rieser, "On factoring out a gesture typology from the Bielefeld speech-gesture-alignment corpus," in *Proceedings of GW 2009: Gesture in Embodied Communication and Human-Computer Interaction*, S. Kopp and I. Wachsmuth, Eds. Berlin and Heidelberg: Springer, 2010, pp. 47–60.

[30] J. Zwarts, "Vectors across spatial domains: From place to size, orientation, shape, and parts," in *Representing Direction in Language and Space*, ser. Explorations in Language and Space. Oxford, NY: Oxford University Press, 2003, vol. 1, ch. 3, pp. 39–68.

[31] M. Weisgerber, "Decomposing path shapes: About an interplay of manner of motion and 'the path'," in *Proceedings of the Annual meeting of the Gesellschaft für Semantik*, ser. Sinn und Bedeutung 10, C. Ebert and C. Endriss, Eds. Berlin: Zentrum für allgemeine Sprachwissenschaft, 2006, pp. 405–419.

[32] A. Lücking, "The display situation," Towards a formal description of gesture and the speech-gesture interface: Panel the 6th conference of the International Society for Gesture Studies (ISGS) at the University of California, San Diego.

[33] N. Goodman, *Languages of Art. An Approach to a Theory of Symbols*, 2nd ed. Idianapolis: Hackett Publishing Company, Inc., 1976.

[34] M. Weisgerber, "Where lexical semantics meets spatial description: A framework for "Klettern" and "Steigen"," in *Proceedings of Sinn und Bedeutung 2005*, ser. SuB9, E. Maier, C. Bary, and J. Huitink, Eds., 2005, pp. 507–521.

[35] T. Sowa, *Understanding Coverbal Iconic Gestures in Shape Descriptions*. Berlin: Akademische Verlagsgesellschaft, 2006, zugl. Diss. Univ. Bielefeld.

[36] K. Barczewska and A. Drozd, "Comparison of methods for hand gesture recognition based on dynamic time warping algorithm," in *Proceedings of the 2013 Federated Conference on Computer Science and Information Systems*, M. P. M. Ganzha, L. Maciaszek, Ed. IEEE, 2013, pp. 207–210.

[37] H. Rieser, "When hands talk to mouth. gesture and speech as autonomous communicating processes," in *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue*, ser. SEMDIAL 2015 goDIAL, C. Howes and S. Larsson, Eds., Gothenburg, Sweden, 2015, pp. 122–130.

[38] K. Alahverdzhieva and A. Lascarides, "Analysing language and co-verbal gesture in constraint-based grammars," in *Proceedings of the 17th International Conference on Head-Driven Phase Structure Grammar (HPSG)*, S. Müller, Ed., Paris, 2010, pp. 5–25.

[39] J. Ginzburg, *The Interactive Stance: Meaning for Conversation*. Oxford, UK: Oxford University Press, 2012.

[40] R. Cooper and J. Ginzburg, "TTR for natural language semantics," in *The Handbook of Contemporary Semantic Theory*, 2nd ed., S. Lappin and C. Fox, Eds. Oxford, UK: Wiley-Blackwell, 2015, ch. 12, pp. 375–407.