Spatial Gesture Semantics

2. Spatial Gesture Semantics

Andy Lücking Alexander Henlein

Goethe University Frankfurt

July 28–August 01, 2025

Recap

Yesterday's lecture

- Different dimensions of classifying gestures
- Focus on iconic gestures
- Two levels of meaning: symbolic vs. visual
- Basic vector space semantics

- Affiliate
- Gesture phases, stroke
- Kendon's Continuum

Followup: Gestural Categories[1]

emblems, illustrators, affect displays, regulators, adaptors

- Emblems: conventionalized
- Illustrators: accompany speech, bound up with the narrative (e.g., iconic)
- Affect displays: convey emotion, often occur involuntarily (e.g., facial expressions)
- Regulators: discourse management (e.g., backchannel signals)
- Adaptors: self-regulation, often reflect nervousness or stress (e.g., tapping on the table)

[1] P. Ekman and W. V. Friesen (1969). "The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding". In: Semiotica 1, 49–98, frame

Representing Gestures

Problem

- The alphabet provides a ready-made transcription system for written text, and phonetic transcription systems for spoken language.
- But how to represent iconic gestures?

Annotation schema

$$
\begin{bmatrix}
\textit{right/left hand} & \\
\text{handshape } 0 & \\
\text{palm} & \begin{bmatrix} \text{orient } 0 \\ \text{path } 0 \\ \text{dir } 0 \end{bmatrix} \\
\text{boh} & \begin{bmatrix} \text{orient } 0 \\ \text{path } 0 \\ \text{dir } 0 \end{bmatrix} \\
\text{wrist} & \begin{bmatrix} \text{path } 0 \\ \text{dir } 0 \\ \text{extent } 0 \end{bmatrix} \\
\text{sync} & \begin{bmatrix} \text{config } 0 \\ \text{rel-mov } 0 \\ \text{s-loc } 0 \\ \text{e-loc } 0 \end{bmatrix}
\end{bmatrix}
$$

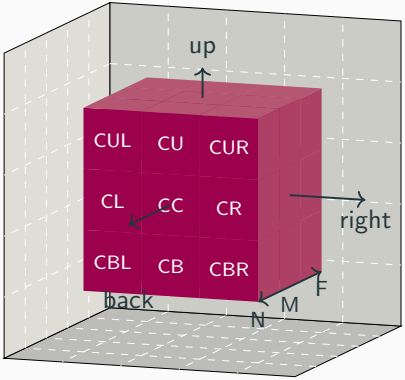
- Kinematic gesture representation along a hand's palm orientation, back of hand orientation, wrist position and movement, and relation to other hand (sync)
- The values indicated by "0" have to be filled with obvious descriptive labels

Handshapes

Deictic	D	index finger pointing
Gun	G	index finger and thumb out
Fist	F	all fingers curled in
Relaxed	R	all fingers loose with no intention
Open	O	all fingers spread outward
Cup	C	all fingers curled midway, claw-like, high intention
Knife	K	fingers straight and flat
Angled	A	like, knife, bent
Pursed	P	fingers pointed together
Hole	H	fingers curved in, forming cylinder shape
Okay	Q	iconic OK shape
Two	T	index and middle finger pointed out
Loose	L	relaxed shape with fingers curled in
Steepled	S	two open hands, fingertips touching
Wall	W	two open hand forming a wall or barrier
Jailed	J	two hands forming a barrier with fingers not pressed against each other

- Handshape notation according to M3d
- `https://m3d.upf.edu/`

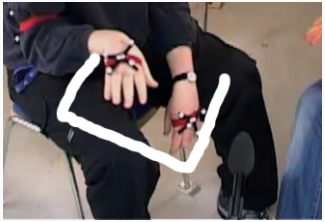
Gesture space



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

- For sloc and eloc
- Extent of movement: small – medium – large

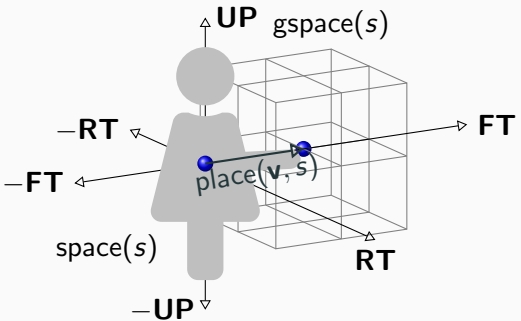
Ex.: U-shape



$$
\begin{bmatrix}
\textit{right hand} & \\
\text{handshape} & \text{O} \\
\text{palm} & \begin{bmatrix} \text{orient} & \text{PDN} \\ \text{path} & 0 \\ \text{dir} & 0 \end{bmatrix} \\
\text{boh} & \begin{bmatrix} \text{orient} & \text{BUP} \\ \text{path} & 0 \\ \text{dir} & 0 \end{bmatrix} \\
\text{wrist} & \begin{bmatrix} \text{path} & \text{line}>\text{line}>\text{line} \\ \text{dir} & \text{MR}>\text{MB}>\text{ML} \\ \text{extent} & \text{large} \end{bmatrix} \\
\text{sync} & \begin{bmatrix} \text{config} & \text{RHA} \\ \text{rel-mov} & \text{none} \\ \text{s-loc} & \text{CBR-F} \\ \text{e-loc} & \text{CBR-N} \end{bmatrix}
\end{bmatrix}
$$

Interpreting Gesture Representations

Gesture space is an oriented vector space



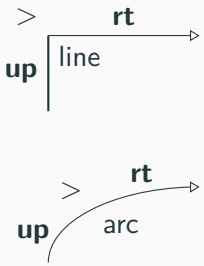
- Recall that every entity is assigned a vector space
- Every speaker s has a gesture space 'gspace(s)'

Spatial gesture semantics

- Strategy: Gesture representation is translated into vector sequences
- Gestural vector sequences add spatial meaning to interpretation

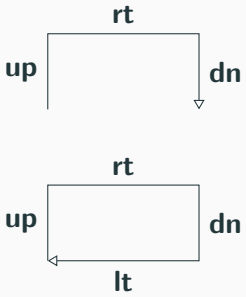
Lines and arcs

- The concatenation of movement annotation labels distinguishes between line and arc,
- They distinguish roundish from angular paths.
- A minimal example is shown to the right, where the iconic models emerging from vector sequence description **up** $>_{\text{line}}$ **rt** respectively **up** $>_{\text{arc}}$ **rt** are given.
- We notate $>_{\text{line}}$ as $\perp$ and $>_{\text{arc}}$ as $\circ$.



Open and closed paths

- If a gesture ends in the location in gesture space where it started, the path is closed; otherwise open
- Closed path is represented in gesture annotation as: s-loc = e-loc



Gesture vectorization

(1) Gesture vectorization function

a. $\text{vec}(\text{handshape } \ulcorner\alpha\urcorner) = \left[\text{hs}\ulcorner\alpha\urcorner\right]$

b. $\text{vec}(\mathbf{u} >_{\text{line}} \mathbf{v}) = \left[\text{traj}\,\mathbf{u} \perp \mathbf{v}\right]$

c. $\text{vec}(\mathbf{u} >_{\text{arc}} \mathbf{v}) = \left[\text{traj}\,\mathbf{u} \circ \mathbf{v}\right]$

d. $\text{vec}(\text{s-loc},\text{e-loc}) =$
$$\begin{cases} \text{sync traj}[a] = \text{traj}[z] & \text{if s-loc} = \text{e-loc} \\ \text{sync traj}[a] \neq \text{traj}[z] & \text{else} \end{cases}$$

- Handshape is copied (a).
- Vectorization applies progressively over movement annotations (b,c).
- Condition (d) checks whether a given movement trajectory brings about a closed or an open path.

Gesture vectorization

(1) Gesture vectorization function

a. $\text{vec}(\text{handshape } \ulcorner\alpha\urcorner) = \left[\text{hs}\ulcorner\alpha\urcorner\right]$

b. $\text{vec}(\mathbf{u} >_{\text{line}} \mathbf{v}) = \left[\text{traj}\,\mathbf{u} \perp \mathbf{v}\right]$

c. $\text{vec}(\mathbf{u} >_{\text{arc}} \mathbf{v}) = \left[\text{traj}\,\mathbf{u} \circ \mathbf{v}\right]$

d. $\text{vec}(\text{s-loc},\text{e-loc}) =$
$$\begin{cases} \text{sync traj}[a] = \text{traj}[z] & \text{if s-loc} = \text{e-loc} \\ \text{sync traj}[a] \neq \text{traj}[z] & \text{else} \end{cases}$$

- Handshape is copied (a).
- Vectorization applies progressively over movement annotations (b,c).
- Condition (d) checks whether a given movement trajectory brings about a closed or an open path.
- We call a vectorized gesture an **iconic model**
- An iconic model is an AVM with three reserved features: handshape ('hs'), trajectory ('traj'), and syncronization ('sync').

Iconic models: Examples

Let's look at some examples of constructing iconic models from gestures.

Ex.: Roof

poles with a roof

over them

$$\text{vec}\left(\begin{bmatrix} \textit{right hand} & \\ \text{handshape} & \text{D} \\ \text{wrist} & \begin{bmatrix} \text{path} & \text{line} \\ \text{dir} & \text{MR} \\ \text{extent} & \text{small} \end{bmatrix} \\ \text{sync} & \begin{bmatrix} \text{config} & \text{RHA} \\ \text{rel-mov} & \text{none} \\ \text{s-loc} & \text{CC-M} \\ \text{e-loc} & \text{CR-M} \end{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} \text{hs} & \text{D} \\ \text{traj} & \textbf{rt} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

Ex.: Wheel

they go on that wheel

$$\text{vec}\left(\begin{bmatrix} \textit{right hand} & \\ \text{handshape} & \text{D} \\ \text{wrist} & \begin{bmatrix} \text{path} & \text{arc>arc>arc>arc} \\ \text{dir} & \text{MU>MF>MD>MB} \\ \text{extent} & \text{medium} \end{bmatrix} \\ \text{sync} & \begin{bmatrix} \text{config} & \text{BHA} \\ \text{rel-mov} & \text{none} \\ \text{s-loc} & \text{CC-M} \\ \text{e-loc} & \text{CC-M} \end{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} \text{hs} & \text{D} \\ \text{traj} & \textbf{up} \circ \textbf{fw} \circ \textbf{dn} \circ \textbf{bw} \\ \text{sync} & \text{traj[a]} = \text{traj[z]} \end{bmatrix}$$

Ex.: Like this

is like this

,

$$\text{vec}\left(\begin{bmatrix} \textit{right hand} & \\ \text{handshape} & \text{O} \\ \text{wrist} & \begin{bmatrix} \text{path} & \text{line>line>line} \\ \text{dir} & \text{MR>MB>ML} \\ \text{extent} & \text{large} \end{bmatrix} \\ \text{sync} & \begin{bmatrix} \text{config} & \text{RHA} \\ \text{rel-mov} & \text{none} \\ \text{s-loc} & \text{CB-F} \\ \text{e-loc} & \text{CB-N} \end{bmatrix} \end{bmatrix}\right) = \begin{bmatrix} \text{hs} & \text{C} \\ \text{traj} & \textbf{rt} \perp \textbf{bw} \perp \textbf{lt} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

Interim summary

- Iconic models are vector sequences with handshapes and are derived from gesture annotations by means of vectorization function ‘vec’.
- Iconic models are the semantic contributions of gestures and impose spatial constraints on the evaluation of multimodal utterance, to which we turn shortly.
- In most cases, however, the iconic models do not apply verbatim, that is, in exactly the orientation and size as they are represented by the gesture in gesture space.

Rotation, Scaling, and Perspective

The problem and its solution

- Gestures do not depict the real-world sizes of the objects and events talked about.

➔ think of the *the house is like this* example

- The orientation of iconic models in gesture space does not need to map directly onto the described situation

The problem and its solution

- Gestures do not depict the real-world sizes of the objects and events talked about.

➔ think of the *the house is like this* example

- The orientation of iconic models in gesture space does not need to map directly onto the described situation

Iconic models can be object to two mathematical operations

- scaling
- rotation

Scaling and rotation

Scaling

- Scaling is just multiplication of the three-dimensional gesture vector $\mathbf{v}$ with a scalar k.
- $\mathbf{v} = \langle x, y, z \rangle, k \in \mathbb{N}$, then $\mathbf{v}k = \langle xk, yk, zk \rangle$
- We notate scalar multiplication of an iconic model as $\text{vec}(\gamma).\text{traj} \cdot k$

Scaling and rotation

Scaling

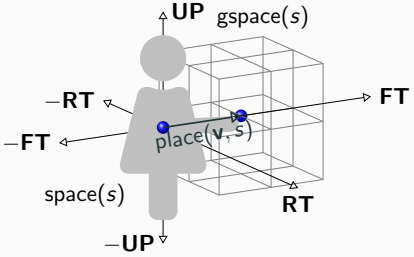
- Scaling is just multiplication of the three-dimensional gesture vector $\mathbf{v}$ with a scalar k.
- $\mathbf{v} = \langle x, y, z \rangle, k \in \mathbb{N}$, then $\mathbf{v}k = \langle xk, yk, zk \rangle$
- We notate scalar multiplication of an iconic model as $\text{vec}(\gamma).\text{traj} \cdot k$

Rotation

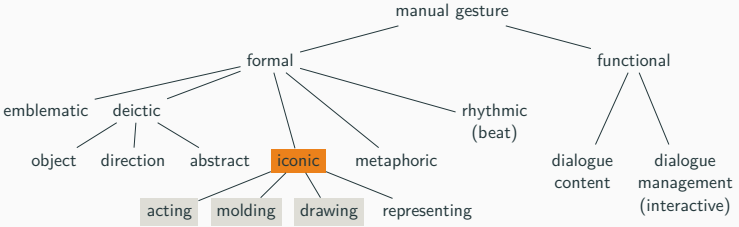
- A vector is rotated by multiplying it with a rotation matrix
- There is a rotation matrix for each level ($-$**FT**/**FT**, $-$**RT**/**RT**, and $-$**UP**/**UP**)
  - $R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$
  - $R_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin & 0 & \cos\theta \end{bmatrix}$
  - $R_z(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$
- We notate the rotation of an iconic model as follows: $\text{vec}(\gamma).\text{traj} \cdot R_d(\theta)$, where d is one of the dimensions x, y, z.

Perspective

- Speakers and the origins of their gesture spaces are connected by a place vector that is aligned with the **FT** level.
- Accordingly, the orientation of the place vector in relation to the anatomical planes already determines speaker perspective and defines the indexical reference frame for relative locations.
- If the perspective is fixed by the speaker's viewpoint, then rotation is blocked and the intersection of the gesture vector or vector sequence and the spatial domain is orientationally faithful to the iconic model.
- A perspectival iconic model is defined as follows: $\text{vec}(\gamma).\text{traj} \cdot R_d(0)$ (i.e., a model with zero rotation).



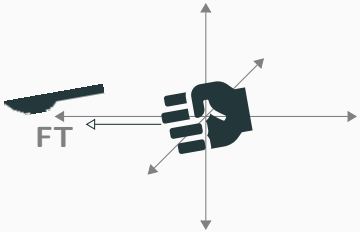
Frame Title

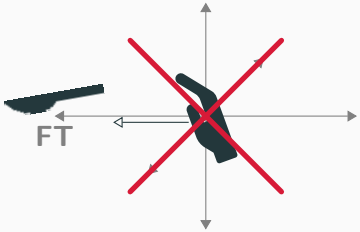


- Vector sequences seem to be sufficient for drawing and molding
- What about acting gestures?

Acting and handshapes

- *throwing a dagger*  .
- Not only movement, but also manner (handshape)





Miming

- The point of miming is that the mime uses his physical actions to denote physical actions of the same kind.
- That is, miming is a form of direct quotation.

- Handshape quotation from sign language semantics[2]
- $[\![\text{HSQ}]\!] = \lambda g.\lambda e[\mathit{demonstration}(g, e)]$
- g is the actual gesture and e is the handshape of the quoted action.

[2] K. Davidson (2015). "Quotation, demonstration, and iconicity". In: Linguistics and Philosophy 38, 477–520

Completing the example

$$\begin{bmatrix} \text{hs} & \text{P} \\ \text{traj} & \textbf{fw} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

- $[\![\text{HSQ}]\!](P) = \lambda e[\mathit{demonstration}(P, e)]$,
- $\rightsquigarrow$ the set of events that "are like" 'P'.
- Handshape quotation is expressed for iconic models as follows: $[\![\text{HSQ}]\!](\text{vec}(\gamma).\text{hs})$.

Interim summary

- We are now in the position to interpret (some kinds of) manual gestures.

- $\text{vec}(\gamma) \mapsto$ iconic model
- $\text{vec}(\gamma).\text{traj} \cdot k$ [scaling]
- $\text{vec}(\gamma).\text{traj} \cdot R_d(\theta)$ [rotation]
- $[\![\text{HSQ}]\!](\text{vec}(\gamma).\text{hs})$ [handshape quotation]

Interim summary

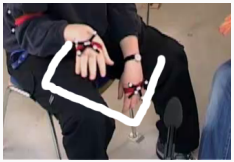
- We are now in the position to interpret (some kinds of) manual gestures.
- Final step: compositional speech–gesture integration.

- $\text{vec}(\gamma) \mapsto$ iconic model
- $\text{vec}(\gamma).\text{traj} \cdot k$ [scaling]
- $\text{vec}(\gamma).\text{traj} \cdot R_d(\theta)$ [rotation]
- $[\![\text{HSQ}]\!](\text{vec}(\gamma).\text{hs})$ [handshape quotation]

Informal examples

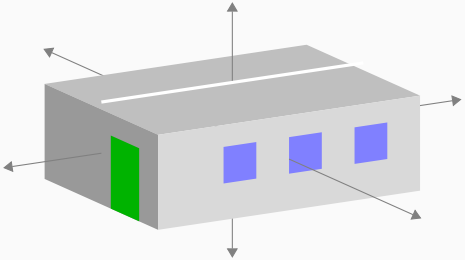
- Let us look at some of our yesterday's examples.
- We will see the gesture, its corresponding iconic model, and a "positive" and a "negative situation"

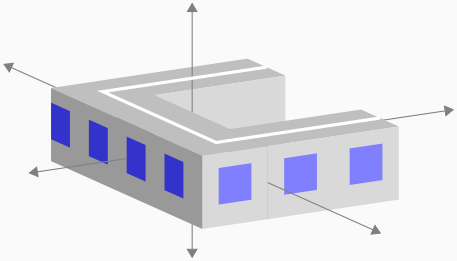
Ex: The house is like this



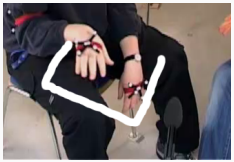
$$\begin{bmatrix} \text{hs} & \text{C} \\ \text{traj} & \textbf{rt} \perp \textbf{bw} \perp \textbf{lt} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

$\mapsto$ axis-path of house



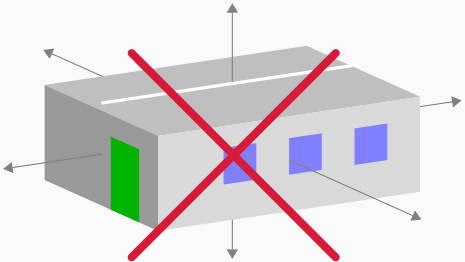


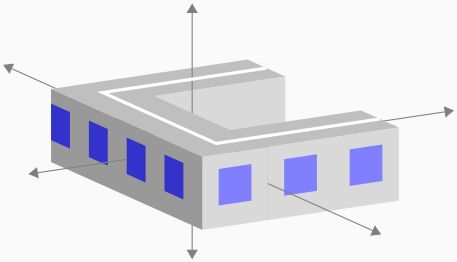
Ex: The house is like this



$$\begin{bmatrix} \text{hs} & \text{C} \\ \text{traj} & \textbf{rt} \perp \textbf{bw} \perp \textbf{lt} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

$\mapsto$ axis-path of house

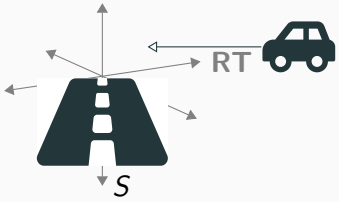


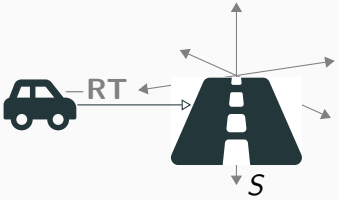


Ex.: Car is pulling out

$$\begin{bmatrix} \text{hs} & \text{K} \\ \text{traj} & \textbf{lt} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

$\rightsquigarrow$ place-path of car, speaker viewpoint (= no rotation)

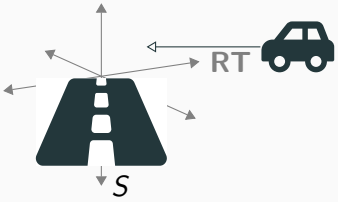


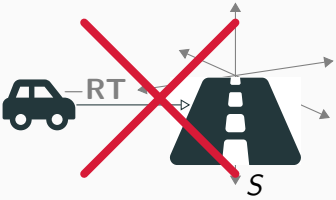


Ex.: Car is pulling out

$$\begin{bmatrix} \text{hs} & \text{K} \\ \text{traj} & \textbf{lt} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

$\rightsquigarrow$ place-path of car, speaker viewpoint ($=$ no rotation)

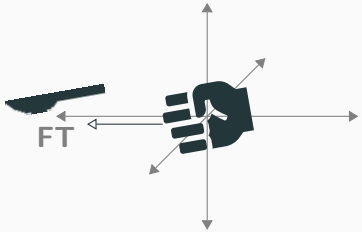


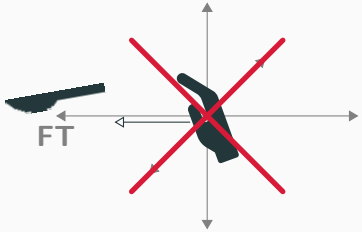


Ex.: Throw a dagger

$$\begin{bmatrix} \text{hs} & \text{P} \\ \text{traj} & \textbf{fw} \\ \text{sync} & \text{traj[a]} \neq \text{traj[z]} \end{bmatrix}$$

$\rightsquigarrow$ place-path of throwing, HSQ





Visual level of meaning

- Result from yesterday's lecture: keep apart linguistic and gestural contributions to meaning
- We do so directly by splitting meaning into a linguistic (as usual) and a visual level (vectors)
- There is much cognitive motivation for this separation from Dual Coding and lexical semantics

Visual level of meaning

- Result from yesterday's lecture: keep apart linguistic and gestural contributions to meaning
- We do so directly by splitting meaning into a linguistic (as usual) and a visual level (vectors)
- There is much cognitive motivation for this separation from Dual Coding and lexical semantics

- Example: lexical entry for dagger:
- [ling] $\lambda x.\mathit{dagger}(x)$
  [vis] $\{\lambda \mathbf{u} \in \text{space}(x)[\text{axis-path}(\mathbf{u}, x)]\}$

Affiliation

- A gesture attaches to a "docking point" in speech, the affiliate[3]
- Hints: Temporal alignment, stressed intonation, semantic constraints
- Ex.: *[$_{prep}$and] [$_{stroke}$'throw the dagger]* (' indicates secondary stress)

- Affiliate is a is a lexical item in about 70% of cases[4]

➔ Grammaticalization

[3] E. A. Schegloff (1984). "On some Gestures' Relation to Talk". In: Structures of Social Action. Studies in Conversational Analysis. Ed. by J. M. Atkinson and J. Heritage, 266–296

[4] A. Mehler and A. Lücking (2012). "Pathways of Alignment between Gesture and Speech: Assessing Information Transmission in Multimodal Ensembles". In: Proc. of the International Workshop on Formal and Computational Approaches to Multimodal Communication under the auspices of ESSLLI 2012, Opole, Poland, 6-10 August

Exceptions

- pro-speech gestures

  and  *the dagger*

- post-speech gestures

  and throw the dagger — 

- holds

  and  *throw the dagger —*  ...

- We will ignore these here, but note that holds might require more sophisticated multimodal composition techniques[5]

[5] H. Rieser (2024). "Multi-modal Anaphora and Broadcasting of Information by Gestural Post-holds". In: Dialogue & Discourse 15, 36–84

Backbone: Multimodal grammars

There are two HPSG frameworks for compositional speech–gesture integration:

- Multiple Recursion Semantics (MRS)[6]
- Unification and semantic role-structures[7]

- But standard in semantics: Functional application and lambda calculus
- Problem: we need kinematic–phonetics interface

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

[7] A. Lücking (2013). Ikonische Gesten. Grundzüge einer linguistischen Theorie. Zugl. Diss. Univ. Bielefeld (2011). De Gruyter

AVMs and Unification

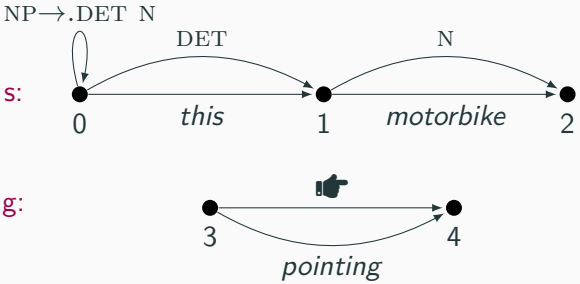
Linguistic entities are modeled as typed feature structures, represented by Attribute–Value Matrices (AVMs):

$$\begin{bmatrix} \text{PHON} & \langle \text{walks} \rangle \\ \text{HEAD} & \begin{bmatrix} \textit{verb} \\ \text{AGR}\ \boxed{1}\begin{bmatrix} \text{PER} & \text{3rd} \\ \text{NUM} & \text{sg} \end{bmatrix} \end{bmatrix} \\ \text{VAL} & \begin{bmatrix} \text{SPR}\left\langle \begin{bmatrix} \text{NP} \\ \text{AGR}\ \boxed{1} \end{bmatrix} \right\rangle \end{bmatrix} \end{bmatrix}$$

Boxed numbers (e.g., $\boxed{1}$) indicate structure sharing (unification):

- $\begin{bmatrix} \text{PER} & \text{3rd} \\ \text{NUM} & \text{sg} \end{bmatrix} \sqcup \begin{bmatrix} \text{GEND} & \text{fem} \end{bmatrix} = \begin{bmatrix} \text{PER} & \text{3rd} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \end{bmatrix}$
- $\begin{bmatrix} \text{PER} & \text{3rd} \\ \text{NUM} & \text{sg} \end{bmatrix} \sqcup \begin{bmatrix} \text{PER} & \text{1st} \end{bmatrix} = \bot$

Multichart parser[8]

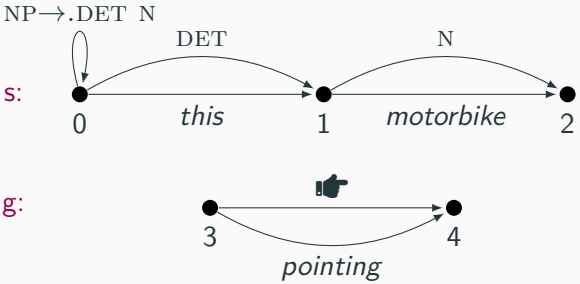


Possible multicharts:

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

[8] M. Johnston (1998). "Unification-based Multimodal Parsing". In: Proc. of the 36th Annual Meeting on Association for Computational Linguistics – Volume I, 624–630

Multichart parser[8]



Possible multicharts:

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

But which one? → phon + sem

[8] M. Johnston (1998). "Unification-based Multimodal Parsing". In: Proc. of the 36th Annual Meeting on Association for Computational Linguistics – Volume I, 624–630

Adding Information Structure

$$\begin{bmatrix} \text{PHON} & \begin{bmatrix}\text{ACCENT accent}\end{bmatrix} \\ \textit{context} & \\ \text{INFO-STRUC} & \begin{bmatrix} \textit{info-struc} & \\ \text{FOCUS} & \text{set(content)} \\ \text{GROUND} & \begin{bmatrix} \textit{ground} \\ \text{LINK set(content)} \\ \text{TAIL set(sontent)} \end{bmatrix} \end{bmatrix} \end{bmatrix}^{a}$$

[a] E. Engdahl and E. Vallduví (1996). "Information Packaging in HPSG". In: Edinburgh Working Papers in Cognitive Science. Ed. by E. Engdahl and E. Vallduví, 1–31.

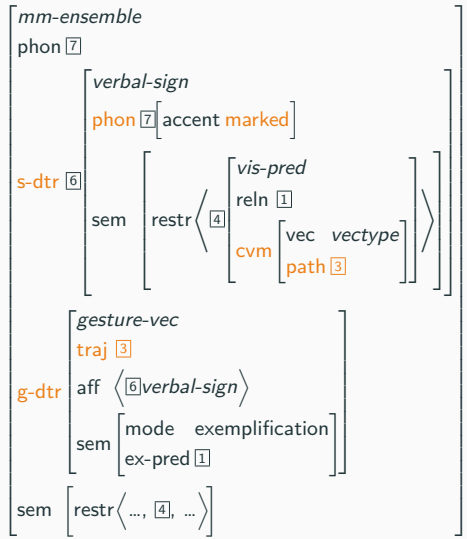
information packaging:
"A-stressed" constituents are coindexed with FOCUS elements, and "B-stressed" are coindexed with LINK elements.

$$\boxed{1}\begin{bmatrix} \textit{word} & \\ \text{PHON} & \begin{bmatrix}\text{ACCENT A}\end{bmatrix} \\ \text{INFO-STRUC} & \begin{bmatrix}\text{FOCUS } \boxed{1}\end{bmatrix} \end{bmatrix}$$

MM Ensemble[9]

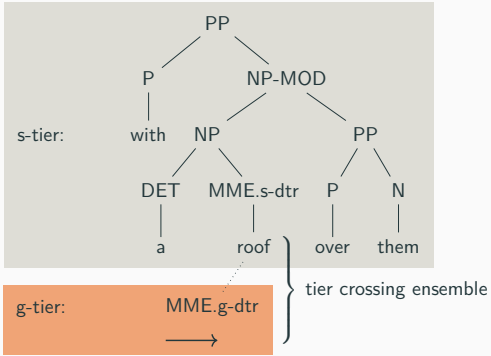
putting it all together:

- tier-crossing construction: mm-ensemble
- s-dtr is phonetically marked
- s-dtr carries a VIS (originally called conceptual vector meaning (CVM))
- g-dtr fills the path value of the VIS



[9] Slightly adapted from A. Lücking (2013). Ikonische Gesten. Grundzüge einer linguistischen Theorie. Zugl. Diss. Univ. Bielefeld (2011). De Gruyter

Derivation in MM Grammar[10]



- usual compositional derivation of speech
- "pointwise" multimodal integration into VIS

→ extended truth-conditions for spatially extended models

[10] K. Alahverdzhieva, A. Lascarides, and D. Flickinger (2017). "Aligning speech and co-speech gesture in a constraint-based grammar". In: Journal of Language Modelling 5, 421–464; A. Lücking (2013). Ikonische Gesten. Grundzüge einer linguistischen Theorie. Zugl. Diss. Univ. Bielefeld (2011). De Gruyter

Composing speech and gesture

- A multimodal utterance $\alpha[\beta/\gamma]$ consisting of a sentence α, a gesture γ and its affiliate β is true, iff α is true and there is an embedding of the iconic model of γ – possibly transformed by scaling or rotation, and possibly additionally constrained by perspective or quotation – into the spatial configuration 'space($[\![\beta]\!]^e$)' projected from $[\![\beta]\!]^e$.
- A sentence α is true in a situation *s* iff *s* is part of the proposition (= set of situations) expressed by α.

Standard composition for speech...

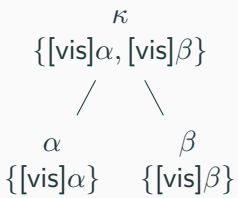
- If [ling]$[\![\beta]\!]$ is a function whose domain contains [ling]$[\![\alpha]\!]$, then [ling]$[\![\kappa]\!]$ = [ling]$[\![\beta]\!]$([ling]$[\![\alpha]\!]$).
- Ex.:
  $\beta = \lambda x \in D_e.\lambda e \in D_s[\mathit{throw\_a\_dagger}(x)]$
  $\alpha = \mathit{andy}$
  $\lambda x \in D_e.\lambda e \in D_s[\mathit{throw\_a\_dagger}(x)](\mathit{andy}) =$
  $\lambda e \in D_s[\mathit{throw\_a\_dagger}(\mathit{andy})]$

...and for speech and gesture

- $[\![\gamma]\!] = \text{vec}(\gamma)$
- If [vis]$[\![\beta]\!]$ is a function whose domain contains [vis]$[\![\gamma]\!]$, then [vis]$[\![\text{MM}]\!]$ = [vis]$[\![\beta]\!]$([vis]$[\![\gamma]\!]$).
- Ex.1:
  $\beta = \lambda \mathbf{v} \in \text{space}(x)[\text{axis-path}(\mathbf{v}, x)]$
  $\gamma = \mathbf{u}$
  $\lambda \mathbf{v} \in \text{space}(x)[\text{axis-path}(\mathbf{v}, x)](\mathbf{u}) = [\text{axis-path}(\mathbf{u}, x)]$
- Ex.2:
  $\beta = \lambda \mathbf{v} \in \text{space}(x)[\text{place-path}(\mathbf{v}, x)]$
  $\gamma = \mathbf{fw} \perp \mathbf{rt} \perp \mathbf{bw} \cdot k \cdot R_z(\theta)$
  $\lambda \mathbf{v} \in \text{space}(x)[\text{place-path}(\mathbf{v}, x)](\mathbf{fw} \perp \mathbf{rt} \perp$
  $\mathbf{bw} \cdot k \cdot R_z(\theta)) = [\text{place-path}(\mathbf{fw} \perp \mathbf{rt} \perp \mathbf{bw} \cdot k \cdot R_z(\theta), x)]$

What happens with β-converted [vis] conditions?

- Outside of MM Ensemble there is no functional dependency between the [vis] conditions of daughters.
- In this case, the vector representations of the daughters are merged into the set of visual meanings of the mother node.
- $[\text{vis}][\![\kappa]\!] = [\text{vis}][\![\alpha]\!] \cup [\text{vis}][\![\beta]\!]$



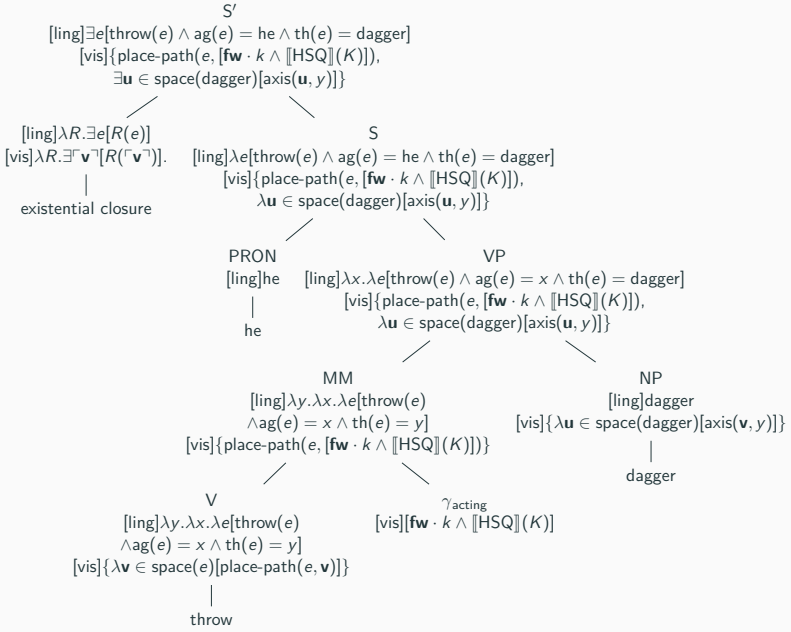
Accounting for modes of representation

[vis]$[\![\gamma]\!]$ is in the domain of [vis]$[\![\alpha]\!]$ if [vis]$[\![\gamma]\!] \subset D_v$ and

- the vector space of α is of type space$(e), e \in D_s$ and γ is a drawing or acting gesture
- the vector space of α is of type space$(x), x \in D_e$ and γ is a drawing or molding gesture

Capturing mismatches

- First account of multimodal well-formedness resp. mismatch
- A speech–gesture mismatch occurs if there is no **v** which embeds the iconic model into the spatial configuration projected from the verbal affiliate.
  - Embedding is empty (there is no such event which "looks like" the gesture in our model)
  - There is a conflict between the affiliate's lexically specified [vis] and the iconic model of a gesture (e.g., a rectangular gesture and the circular axis-path of *disk*)



- Usual compositional derivation of [ling]
- Gesture adds [vis]

→ two-dimensional truth conditions

Summary

- Independently motivated vector space semantics
- Lexical extensions for speech–gesture integration ([vis], or CVM)[11]
- Vectorization of (some kinds of) gestures
  - Scaling [drawing, modeling]
  - Rotation [drawing, modeling]
  - Handshape quotation [acting]

- Captures the "semantic innocence" of gestures
- Well-behaved truth-functional and compositional semantics

[11] Argued for early on by, e.g., H. Rieser (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), 167–174

Remaining Challenges

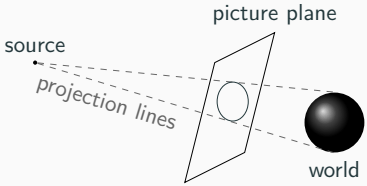
- Spatial gesture semantics is making important advances in terms of iconicity and multimodal compositionality.
- But that does not mean that everything has solved.

- Two-handed gestures (single gesture vs. two separate gestures)
- Static representing gestures
- Gesture holds
- More fine-grained kinematic and temporal interpretation (expressiveness, intensifiers, ...)
- Other (non-spatial?) kinds of gestures
- ...

Appendix

Projections[12]

- 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.
- Formally: $[\![P]\!]_{S,c} \subseteq \{\langle w, v\rangle \mid proj_S(w, v) = P\}$



[12] G. Greenberg (2021). "Semantics of Pictorial Space". In: Review of Philosophy and Psychology 12, 847–887

Projections are inadequate for gesture semantics

Semiotic shortfall

- The gesture plays the role of the picture plane which displays the projection source.
- The content of a gesture is the set of world–viewpoint pairs where a gesturer performs that gesture, seen from the viewpoint in question.

➔ The content of a gesture is in turn a gesture!

➔ Projection semantics fails to distinguish between the gesture as a physical action and its content.

[13] J. Bressem (2013). "A linguistic perspective on the notation of form features in gestures". In: Body – Language – Communication. Ed. by C. Müller et al. Vol. 1, 1079–1098

Projections are inadequate for gesture semantics

Semiotic shortfall

- The gesture plays the role of the picture plane which displays the projection source.
- The content of a gesture is the set of world–viewpoint pairs where a gesturer performs that gesture, seen from the viewpoint in question.

→ The content of a gesture is in turn a gesture!

→ Projection semantics fails to distinguish between the gesture as a physical action and its content.

Perceptual shortfall

- Ex.: *Rolling*
- part of the rotation movement of the index finger that runs backwards[13]
- This configuration, however, is a purely perceptual one; it can never be projected onto a physical movement.

[13] J. Bressem (2013). "A linguistic perspective on the notation of form features in gestures". In: Body – Language – Communication. Ed. by C. Müller et al. Vol. 1, 1079–1098