Spatial Gesture Semantics

5. Lexicon-driven speech–gesture integration

Andy Lücking Alexander Henlein

Goethe University Frankfurt

July 28–August 01, 2025

Recap

Yesterday's lecture

- ML basics
- Multimodal AI
- Gesture detection and classification (drinking and eating gestures)

Today's lecture

- Given InfEval: how do speech and gesture integrate?
- Computing relation R
- We argue that usual dynamic semantic methods apply
- Main source: Frames

Recall: Conditioned interpretation

Conditioned interpretation:
If gesture γ is informationally evaluated to mean p, then the whole multimodal utterance α is interpreted as $\alpha[R(p, \beta)]$.

Minimized contexts

Bridging

- I can't ride my bike today. The back wheel's tire is flat.
- The footage shows a man running on stage and stabbing Adamowicz [...]. The assailant paces back and forth, arms aloft like a victorious boxer, still holding the 15cm (six-inch) knife.[1]

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

[2] H. H. Clark (1975). "Bridging". In: Proc. of the 1975 Workshop on Theoretical Issues in Natural Language Processing, 169–174

Bridging

- I can't ride my bike today. The back wheel's tire is flat.
- The footage shows a man running on stage and stabbing Adamowicz [...]. The assailant paces back and forth, arms aloft like a victorious boxer, still holding the 15cm (six-inch) knife.[1]

- The tire is understood as the tire of the bike.
- The knife is understood as the instrument of the stabbing event, and pacing back and forth the stabbing action.
- Such indirect anaphoric relations are known as bridging.[2]

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

[2] H. H. Clark (1975). "Bridging". In: Proc. of the 1975 Workshop on Theoretical Issues in Natural Language Processing, 169–174

Discourse referents in dynamic semantics

- "[an interpreter] must be able to recognize when a novel individual is mentioned in the input text and to store it along with its charactarization for future reference."[3]

[3] L. Karttunen (1969). "Discourse Referents". In: Proc. of the 1969 Conference on Computational Linguistics, 1–38

[4] H. Kamp and U. Reyle (1993). From Discourse to Logic. Kluwer Academic Publishers

Discourse referents in dynamic semantics

- "[an interpreter] must be able to recognize when a novel individual is mentioned in the input text and to store it along with its charactarization for future reference."[3]
- Bill owns a car. It is black.
- Bill doesn't own a car. #It is black.

[3] L. Karttunen (1969). "Discourse Referents". In: Proc. of the 1969 Conference on Computational Linguistics, 1–38

[4] H. Kamp and U. Reyle (1993). From Discourse to Logic. Kluwer Academic Publishers

Discourse referents in dynamic semantics

- "[an interpreter] must be able to recognize when a novel individual is mentioned in the input text and to store it along with its charactarization for future reference."[3]
- Bill owns a car. It is black.
- Bill doesn't own a car. #It is black.

- DRT: Discourse referents and conditions[4]
- $[x, y, z, z = y; \mathrm{Bill}(x), \mathrm{car}(y), \mathrm{own}(x, y), \mathrm{black}(z)]$
- $[x; \mathrm{Bill}(x), \neg[y; \mathrm{car}(y), \mathrm{own}(x, y)], \#\mathrm{black}(z)]$ (y not available for z)

[3] L. Karttunen (1969). "Discourse Referents". In: Proc. of the 1969 Conference on Computational Linguistics, 1–38

[4] H. Kamp and U. Reyle (1993). From Discourse to Logic. Kluwer Academic Publishers

Implicit discourse referents in dynamic semantics

- x *being healthy again*: $[x; \text{healthy}(x)]$
- Presupposition: x was ill and recovered
- ➔ Meaning postulate: $[v, x; \text{ailment}(v) \Rightarrow \text{recovered}(v, x)]$
- Now there is a new implicit discourse referent v![5]

Peter is healthy again.
The fever is gone.

[5] H. Kamp and A. Rossdeutscher (1994). "DRS-Construction and Lexically Driven Inference". In: Theoretical Linguistics 20, 165–235

Frames

Frames

- Frames can be conceived as stereotypical situation types which are connected to lexical items.
- A word form not only contributes its content, but it also evokes the frames it is connected to.
- Frame semantics is organized in a frame-base lexicon called FrameNet.

Framenet

- `https://framenet.icsi.berkeley.edu/frameIndex`
- Look up example entry *staircase.n* and the connecting_architecture frame

Frame Evocation[6]

Input:

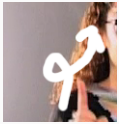
$$\lambda x. \begin{array}{|l|} \hline x \\ \hline \text{staircase}(x) \\ \hline \end{array}$$

Output:

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

[6] M. Irmer (2013). "Inferring Implicatures and Discourse Relations from Frame Information". In: Lingua 132. Special Issue: Implicature and Discourse Structure, 29–50

Example: affiliate *staircases*



- Inside the hall was an imposing staircase.
- InfEval: $R(\textit{spiral}, \text{staircase})$.

InfEval + Frame invocation

$\lambda x.$

x, z	e, y_1, y_2, y_3, y_4

e : connecting_architecture
$\text{Part}(e, x)$
$\text{staircase}(x)$
$\text{Creator}(e, y_1), y_1 = ?$
$\text{Descriptor}(e, y_2), y_2 = ?$
$\text{Direction}(e, y_3), y_3 = ?$
$\text{Material}(e, y_4), y_4 = ?$
$\text{spiral}(z)$
$R(\text{spiral}(z), \text{staircase}(x)), R = ?$

InfEval + Frame invocation

$$\lambda x. \begin{array}{|l|l|} \hline x, z & e, y_1, y_2, y_3, y_4 \\ \hline \multicolumn{2}{|l|}{\begin{array}{l} e : \text{connecting\_architecture} \\ \text{Part}(e, x) \\ \text{staircase}(x) \\ \text{Creator}(e, y_1), y_1 = ? \\ \text{Descriptor}(e, y_2), y_2 = ? \\ \text{Direction}(e, y_3), y_3 = ? \\ \text{Material}(e, y_4), y_4 = ? \\ \text{spiral}(z) \\ R(\text{spiral}(z), \text{staircase}(x)), R = ? \end{array}} \\ \hline \end{array}$$

- Since *spiral* is a shape predicate, the only plausible frame element to resolve R is $R =$ Descriptor.

Resolved multimodal meaning

$$\lambda x. \begin{array}{|l|l|} \hline x, z & e, y_1, y_2, y_3, y_4 \\ \hline \multicolumn{2}{|l|}{\begin{array}{l} e : \text{connecting\_architecture} \\ \text{Part}(e, x) \\ \text{staircase}(x) \\ \text{Creator}(e, y_1), y_1 = ? \\ \text{Descriptor}(e, y_2), y_2 = \text{spiral}(z), z = x \\ \text{Direction}(e, y_3), y_3 = ? \\ \text{Material}(e, y_4), y_4 = ? \end{array}} \\ \hline \end{array}$$

Benefits of being formally precise

Dowty (1979)

"[...] an important goal of formalization in linguistics is to enable subsequent researchers to see the defects of an analysis as clearly as its merits; only then can progress be made efficiently."[7]

[7] D. R. Dowty (1979). Word Meaning and Montague Grammar. Reidel, 322

Benefits of being formally precise

Dowty (1979)

"[...] an important goal of formalization in linguistics is to enable subsequent researchers to see the defects of an analysis as clearly as its merits; only then can progress be made efficiently."[7]

- There are some informal approaches to gesture and speech–gesture integration.
- Can frames say something more precise?

[7] D. R. Dowty (1979). Word Meaning and Montague Grammar. Reidel, 322

Slapping

- John [*slapping gesture*] punished his son.
- Non-at-issue conditional presupposition / local context:
  If John punished his son, then slapping would be involved.[8]

[8] P. Schlenker (2018). "Gesture projection and cosuppositions". In: Linguistics and Philosophy 41, 295–365, 318

Slapping

- John [*slapping gesture*] punished his son.
- Non-at-issue conditional presupposition / local context:
  If John punished his son, then slapping would be involved.[8]

- Can we reconstruct this with InfEval and conditioned interpretation?

[8] P. Schlenker (2018). "Gesture projection and cosuppositions". In: Linguistics and Philosophy 41, 295–365, 318

InfEval

- If we interpret the gesture as slapping, with *punished* being the lexical affiliate, then the multimodal information package '$R(\textit{slapped}, \text{punished})$' is obtained.
- The lexical unit *punish.v* evokes the Rewards_and_punishment frame.

Frame Evocation

$\lambda y.\lambda x.\lambda e.$

y, x, e
punish(e) agent(e, x) patient(e, y)

$\lambda y.\lambda x.\lambda e.$

y, x, e | z_1, z_2, z_3, z_4, z_5
e : rewards_and_punishment punish(e) Agent(e, x) Evaluee(e, y) Reason$(e, z_1), z_1 =?$ Degree$(e, z_2), z_2 =?$ Instrument$(e, z_3), z_3 =?$ Manner$(e, z_4), z_4 =?$ Means$(e, z_5), z_5 =?$

Punishing means

- Being an action-simulating gesture, slapping instantiates the non-core *Means* frame element: punish by slapping
- x punished y by slapping y.

$\lambda y.\lambda x.\lambda e.$

y, x, e, e', x', y'	z_1, z_2, z_3, z_4, z_5
e : rewards_and_punishment punish(e) Agent(e, x) Evaluee(e, y) Reason$(e, z_1), z_1 = ?$ Degree$(e, z_2), z_2 = ?$ Instrument$(e, z_3), z_3 = ?$ Manner$(e, z_4), z_4 = ?$ Means$(e, z_5), z_5 = \text{slap}(e'), e' = e$ agent$(e', x'), x' = x$ patient$(e', y'), y' = y$	

Punishing means

- But there is a different interpretation: x punished y by slapping $y'\neg y$.
- Think of John punishing his son by slapping the son's pet.

$\lambda y.\lambda x.\lambda e.$

y, x, e, e', x', y'	z_1, z_2, z_3, z_4, z_5
e : rewards_and_punishment punish(e) Agent(e, x) Evaluee(e, y) Reason$(e, z_1), z_1 = ?$ Degree$(e, z_2), z_2 = ?$ Instrument$(e, z_3), z_3 = ?$ Manner$(e, z_4), z_4 = ?$ Means$(e, z_5), z_5 = \text{slap}(e'), e' = e$ agent$(e', x'), x' = x$ patient$(e', y'), y' \neq y$	

Pro Dowty

- More precise analysis thanks to formal framework.
- What about speech–gesture mismatches?

Multimodal well-formedness

- Inside the hall was an imposing [*slapping gesture*] staircase.
- Local context: *every world w in which a staircase is in the hall is one in which slapping is involved*.
- Odd but possible.

Multimodal well-formedness

- Inside the hall was an imposing [*slapping gesture*] staircase.
- Local context: *every world w in which a staircase is in the hall is one in which slapping is involved*.
- Odd but possible.

- Slapping, denoting an action, is not a good candidate to fill any of the frame elements evoked by *staircase*
- Frame-based dynamic semantics algorithms would fail to integrate speech and gesture in this case and signal a mismatch.

Intermediate summary

- InfEval, conditioned interpretation, and, if required (i.e., $p \neq \beta$), frame-based integration of speech and gesture provides a systematic heuristic for analyzing iconic gesture in semantic research.
- it is computable,
- provides a notion of multimodal incongruence.

Mixed topics

Non-lexicalized percepts

The inscription looked like this: 

- What is the result of InfEval?
- No worries, there need not be one!
- A gesture that is such that it resists perceptual classification in terms of single words just contributes its iconic model (cf. spatial gesture semantics, Lect. 2).

“Energy spaces”

- force vectors instead of spatial place or path ones.[9]
- Example: the semantics of *climb* is captured in terms of two forces: one pulling downwards, one striving upwards.

[9] A. Goldschmidt and J. Zwarts (2016). “Hitting the nail on the head: Force vectors in verb semantics”. In: Semantics and Linguistic Theory, 433–450; L. Talmy (1988). “Force dynamics in language and cognition”. In: Cognitive Science 12, 49–100

“Energy spaces”

- force vectors instead of spatial place or path ones.[9]
- Example: the semantics of *climb* is captured in terms of two forces: one pulling downwards, one striving upwards.

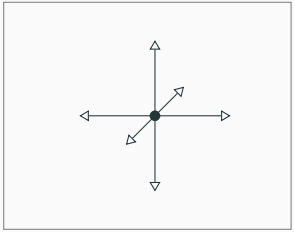
- Mathematical vector spaces are ontologically neutral.
- That is, the same formal devices can be used to model “energy spaces” consisting of force vectors.

[9] A. Goldschmidt and J. Zwarts (2016). “Hitting the nail on the head: Force vectors in verb semantics”. In: Semantics and Linguistic Theory, 433–450; L. Talmy (1988). “Force dynamics in language and cognition”. In: Cognitive Science 12, 49–100

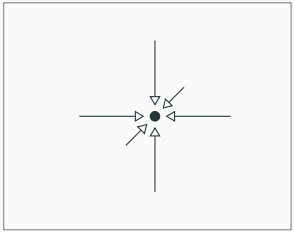
“Energy spaces”

- Speakers occupy the respective “center of gravity”

Repulsion space:



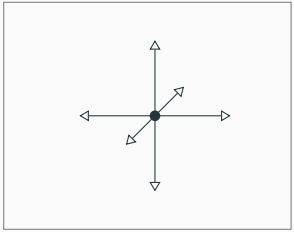
Attractor space:



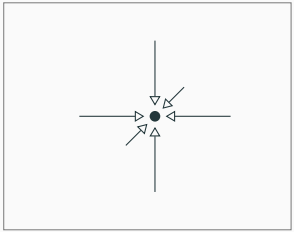
“Energy spaces”

- Speakers occupy the respective “center of gravity”
- *climbing*: an energy space spanned by the orthogonal projections of force vectors onto the downwards and upwards pulling ones in repulsion space.

Repulsion space:



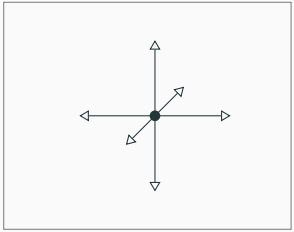
Attractor space:



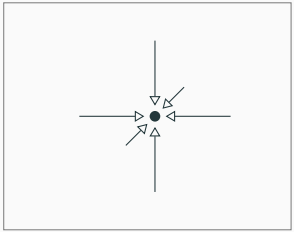
“Energy spaces”

- Speakers occupy the respective “center of gravity”
- *climbing*: an energy space spanned by the orthogonal projections of force vectors onto the downwards and upwards pulling ones in repulsion space.
- Force vectors are arguably involved in verbal construction like *on the one hand … on the other hand*: the two poles referred to are pulled apart by force vectors drawing in opposing directions.

Repulsion space:



Attractor space:



Repercussions for semantic theories

- The notion of meaning needed for InfEval is such that, when applied to an object, it returns a linguistic label.
- We spell this out in terms of perceptual classification, and the extemplification heuristic.
- Arguably, these components cannot be reconciled with a textbook possible worlds semantics.
- Are there alternatives?

The TTR, KoS, RTT "ecosystem"

A suitable candidate, to our minds, is a Type Theory with Records (TTR)[10]

- TTR incorporates words-as-classifiers[11]
- TTR includes frames as both, situations and situation types
- It also provides the ontology for the dialogue semantic theory KoS[12] (recall the importance of clarification interaction
- It underpins the most recent theory of pluralities and quantification, Referential Transparency Theory (RTT)[13]
- ...

[10] R. Cooper (2023). From Perception to Communication. A Theory of Types for Action and Meaning. Oxford UP

[11] S. Larsson (2015). "Formal Semantics for Perceptual Classification". In: Journal of Logic and Computation 25, 335–369

[12] J. Ginzburg (2012). The Interactive Stance: Meaning for Conversation. Oxford UP

[13] A. Lücking and J. Ginzburg (2022). "Referential transparency as the proper treatment of quantification". In: Semantics and Pragmatics 15, 1–58; A. Lücking and J. Ginzburg (2025). "Postmodern Quantification with Stuff". In: Proc. of *Sinn und Bedeutung*. Forthcoming

Appendix: Issueness

(Non-)at-issue

Mareike chooses the vegan pasta in the dining hall.

- at-issue: Mareike chooses the vegan pasta in the dining hall.
- non-at-issue (possible implicature): Mareike likes vegan pasta.
- non-at-issue (presupposition): There is vegan pasta in the dining hall.

(Non-)at-issue

Mareike does not choose the vegan pasta in the dining hall.

- ~~at-issue: Mareike chooses the vegan pasta in the dining hall.~~
- non-at-issue (possible implicature): Mareike likes vegan pasta.
- non-at-issue (presupposition): There is vegan pasta in the dining hall.

From Lect. 1: Denial

(1) [...] with a  roof over them

a. ?No, that's [?] not true. The roof (i) is not $\langle * \rangle$ / (ii) actually is $\langle * \rangle$

b. ?Wait a minute. The roof (i) is not $\langle * \rangle$ / actually is $\langle * \rangle$

[14] E.g. P. Schlenker (2018). "Gesture projection and cosuppositions". In: Linguistics and Philosophy 41, 295–365

From Lect. 1: Denial

(1) [...] with a  roof over them

a. ?No, that's [?] not true. The roof (i) is not $\langle * \rangle$ / (ii) actually is $\langle * \rangle$

b. ?Wait a minute. The roof (i) is not $\langle * \rangle$ / actually is $\langle * \rangle$

- We have already seen that gestures do not readily introduce linguistic predicates (only if this has been agreed upon in dialogue).
- And this neither at-issue ("No") and non-at-issue ("Wait a minute").
- Nonetheless, there has been claims that gestures are non-at-issue.[14]
- Can we shed more light on this?

[14] E.g. P. Schlenker (2018). "Gesture projection and cosuppositions". In: Linguistics and Philosophy 41, 295–365

Conditional meanings

- The conditioned interpretation heuristic literally puts the understanding of a multimodal utterance in the consequence of an indicative conditional ("If the gesture is InfEvaled to mean p, ...").
- Their consequences cannot be picked out by negation: The negation of a sentence of the form "If A then C" is either the conjunction "A and not C" or the conditional "If A then not C".[15].

[15] P. Egré and G. Politzer (2013). "On the negation of indicative conditionals". In: Proc. of the 19th Amsterdam Colloquium, 10–18

Conditional meainings

- A: If the staircase is spiral, it is an imposing one.
- # B: No, that's not true. The staircase is imposing.
- B: No, that's not true. The staircase is imposing even without being spiral.

- Hence, we would expect contexts of conditioned, but not explicitly agreed, gesture interpretation to involve nondeniable consequences.

Ex.: Staircases

- *If*  *is interpreted as "spiral", then* in the hall was an imposing spiral staircase.
- # No, that's not true. The staircase was actually straight
- No, that's not true. The staircase was actually straight, even if you interpret  as *spiral*.

(From "If A then C" and "A", "C" follows and can be negated.)

Further tests

- Ebert[16] distinguishes additional non-at-issue tests for co-speech gestures.
- Let us look at ellipsis.

[16] C. Ebert (2024). "Semantics of Gesture". In: Annual Review of Linguistics 10, 169–189

Ellipsis

- Co-speech gesture contribution is ignored in ellipsis constructions:
- In the hall was an imposing  staircase, and a window, too.

Ellipsis

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

- Of course, *spiral* does not need to take scope over *window*.

Upshot

InfEval of iconic gestures and conditioned interpretation can explain observations concerning the information status of linguistic descriptions of iconic gestures wrongly attributed to (non-)at-issueness elsewhere.