A Uniform Meaning Representation for NLP Systems:

Lecture 1: Formal Foundations of UMR; Extensions beyond AMR

Martha Palmer and James Pustejovsky

Joint work with Jens Van Gysel, Meagan Vigus, Jin Zhao, Nianwen Xue, Jayeol Chun, Kenneth Lai, Sara Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James Martin, Stephan Oepen, Rosa Vallejos

ESSLLI 2023 Summer School

Ljubljana, Slovenia

August 7-11, 2023

Course Outline

- Monday: Formal Foundations of UMR and Extensions beyond AMR
- Tuesday: UMR Mechanisms for Quantification and Discourse Anaphora
- Wednesday: Annotation in UMR for Multiple Languages and Parsing UMRs
- Thursday: Extensions of UMR for Multimodal Communication and Situated Grounding
- Friday: UMR for Knowledge Grounding and Logical Inference

The Components of Multimodal Communication

- Achieving and maintaining common ground
  - shared conceptual space
- Context-aware interpretation of communicative acts
  - language, gesture, gaze
- Recognizing Object-specific knowledge and behavior
- Objects and actions are situated in the interaction
- Agents are embodied in the interaction:
  - all actions (communicative or physical) are interpreted through embodiment.

Situated Semantic Grounding and Embodiment

- Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication.
- Situated semantic grounding assumes shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.
- VoxWorld : a multimodal simulation framework for modeling Embodied Human-Computer Interactions and communication between agents engaged in a shared goal or task.

Situated Meaning: shared task of icing cupcakes

Situated Meaning in a Joint Activity

- Son: *Put it there (gesturing with co-attention)?*
- Mother: *Yes, go down for about two inches.*
- Mother: *OK, stop there. (co-attentional gaze)*
- Son: *Okay. (stops action)*
- Mother: *Now, start this one (pointing to another cupcake).*

Situated Meaning

Elements from the Common Ground

Agents	mother, son
Shared goals	baking, icing
Beliefs, desires, intentions	Mother knows how to ice, bake, etc. Mother is teaching son
Objects	Mother, son, cupcakes, plate, knives, pastry bag, icing, gloves
Shared perception	the objects on the table
Shared Space	kitchen

The Challenge of Situated Grounding

1. Human-Human/Computer interactions require at least the following capabilities:
   - Robust recognition and generation within multiple modalities
     - language, gesture, vision, action;
   - understanding of contextual grounding and co-situatedness;
   - appreciation of the consequences of behavior and actions.

2. Multimodal simulations provide an approach to modeling human-computer communication by situating and contextualizing the interaction, thereby visually demonstrating what the computer/robot sees and believes.

The Meaning of Embodiment in Communication

- Agent has situated meaning for the objects and actions in the environment;
- Recognition of the human's embodiment; agent has awareness of people's linguistic and gestural expressions, facial expressions, and actions.
- Self-embodiment of the agent: the agent has "spatial presence" within the domain of the interaction

Shared Conceptual Space

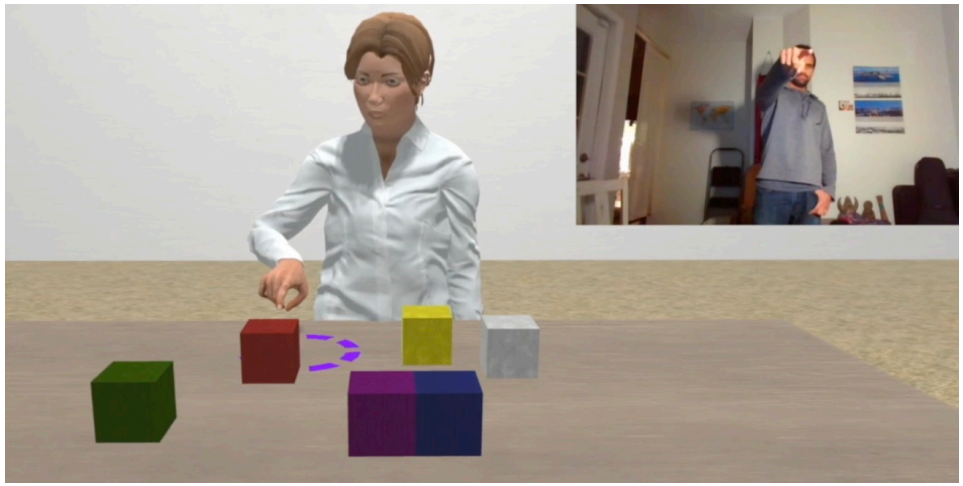


Figure: *Left:* Human-human collaborative interaction; *Right:* Human-avatar interaction.

Embodiment and Situated Meaning

- Elements of Situated Meaning
  - Identifying the *actions and consequences* associated with objects in the environment.
  - Encoding a multimodal expression contextualized to the *dynamics of the discourse*
  - *Situated grounding*: Capturing how multimodal expressions are anchored, contextualized, and situated in context
- Modalities Deployed
  - gesture recognition and generation
  - language recognition and generation
  - affect, facial recognition, and gaze
  - action generation

Recognition of Human's Embodiment

Awareness of the partner's:

- linguistic and gestural expressions
- gestural expressions
- facial expressions
- gaze and eye tracking
- actions

The agent continuously constructs and maintains a representation of the embodiment of its human partner.

Intelligent Virtual Agents

Embodied Environments

A non-verbal interaction between a human and IVA using gesture, gaze, and action.

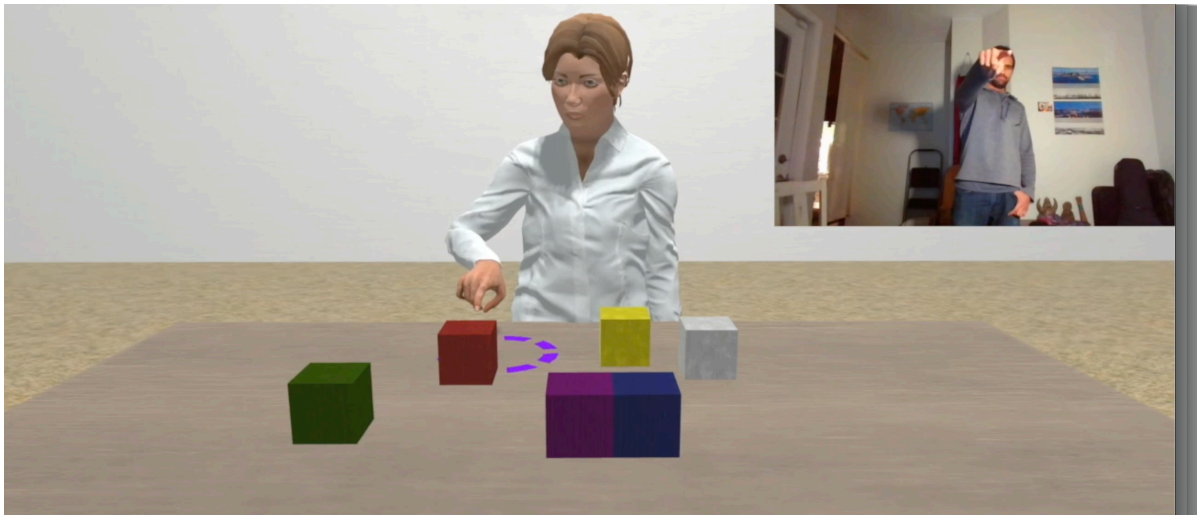


Figure: IVA Diana engaging in an embodied HCI with a human user.

▸ Link

Modeling Human Object Interactions (HOIs)

- The objects in a dialogue carry much more semantic information than conventionally assumed.
- This includes knowledge for how the objects can be manipulated by an agent in space and time, their *Gibsonian affordances*, and how they can be used, their *Telic affordances*.
- Such information also includes knowledge of how an object is situated in the environment relative to an agent for specific purposes and actions, that is, its *habitat*.
- Affordance encoding and recognition can improve object and action classification in HCI tasks.

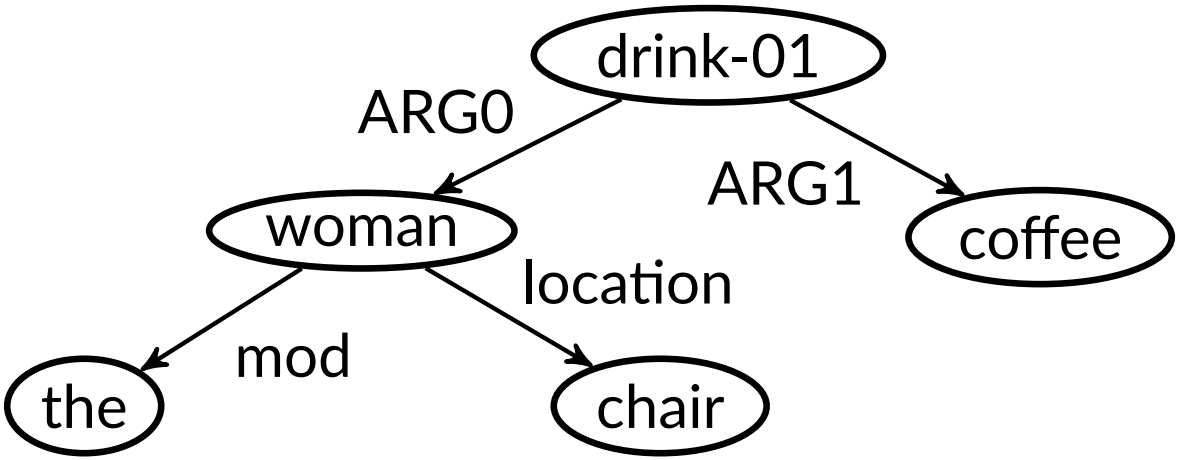
AMRs Don't Describe Human-Object Interactions

Neither do conventional semantic representations

"A woman drinks coffee."

(1) a. $drink(w, c)$
b. $\exists x \exists y[\text{woman}(x) \wedge coffee(y) \wedge \text{drink}(x, y)]$
c. $\text{event}(drink) \wedge \text{agent}(woman) \wedge \text{patient}(coffee)$

AMR is Context Unaware



- A upright seated woman is holding in her hand, a cup filled with coffee while she drinks it.
- The cup is upright so the container portion (inside) is able to hold coffee.
- She is holding the cup by an attached handle.
- The cup is tilted towards her and touches her partially open mouth, in order to allow drinking.

Dynamic Discourse Interpretation

- Common Ground Structure
  - Co-belief
  - Co-perception
  - Co-situatedness
- Multimodal communication act:
  - language
  - gesture
  - action
- Dynamic tracking and updating of dialogue with:
  - Discourse Structure
  - Gesture Grammar
  - Action Interpretation

Multimodal Communicative Acts

- A communicative act, performed by an agent, a, is a tuple of expressions from the modalities available to a, involved in conveying information to another agent.
- We restrict this to the modalities of speech, S, gesture, G, facial expression F, gaze Z, an explicit action A.
  1. $C_a = \langle S, G, F, Z, A \rangle$
- These modal channels can be aligned or unaligned in the input.

Visual Object Concept Modeling Language (VoxML)

Pustejovsky and Krishnaswamy (2016)

- Encodes afforded behaviors for each object
  - Gibsonian: afforded by object structure (Gibson,1977,1979)
    - grasp, move, lift, etc.
  - Telic: goal-directed, purpose-driven (Pustejovsky, 1995, 2013)
    - drink from, read, etc.
- Voxeme
  - Object Geometry: Formal object characteristics in R3 space
  - Habitat: Conditioning environment affecting object **affordances** (behaviors attached due to object structure or purpose);
  - Affordance Structure:
    - What can one do to it
    - What can one do with it
    - What does it enable

VoxWorld Architecture

Pustejovsky and Krishnaswamy (2016), Krishnaswamy (2017), Pustejovsky et al (2017), Narayana et al (2018)

- Dynamic interpretation of actions and communicative acts:
  - Dynamic Interval Temporal Logic (DITL)
  - Dialogue Manager
- VoxML: Visual Object Concept Modeling Language
- EpiSim: Visualizes agent's epistemic state and perceptual state in context;
  - Public Announcement Logic
  - Public Perception Logic
- VoxSim: 3D visualizer of actions, communicative acts, and context.
  - Built on Unity Game Engine

Co-belief and Co-perception in the Common Ground

- *Public announcement logic* (PAL)
  - $[\alpha]\varphi$ denotes that an agent "α *knows* φ".
  - Public Announcement: $[!\varphi_1]\varphi_2$
  - Any proposition, φ, in the common knowledge held by two agents, α and β, is computed as: $[(\alpha \cup \beta)^*]\varphi$.
- *Public perception logic* (PPL)
  - $[\alpha]_\sigma\varphi$ denotes that agent "α perceives that φ".
  - $[\alpha]_\sigma\hat{x}$ denotes that agent "α perceives that there is an *x*."
  - Public Display: $[!\varphi_1]_\sigma\varphi_2$
  - The co-perception by two agents, α and β includes φ : $[(\alpha \cup \beta)^*]_\sigma\varphi$

Multimodal Semantics for Common Ground

Common Ground Structure (CGS)

The situated common ground consists of the following state information:

(2) a. **A**: The agents engaged in communication;
b. **B**: The shared belief space;
c. **P**: The objects and relations that are jointly perceived in the environment;
d. $\mathcal{E}$: The embedding space that both agents occupy in the communication.

(3)
$$\frac{\textbf{A}{:}a_1, a_2 \quad \textbf{B}{:}\Delta \quad \textbf{P}{:}b}{\mathcal{S}_{a_1} = \text{“You}_{a_2} \text{ see it}_b\text{”}}_{\mathcal{E}}$$

Multimodal Semantics for Common Ground

Modeling the Current Context

- State Monad: $M\alpha = State \rightarrow (\alpha \times State)$
- Context is a stack of items and the type of left contexts is a list of entities, $[e]$.
- Right contexts will be interpreted as continuations: a discourse that requires a left context to yield a truth value., of type $[e] \rightarrow t$.
- Hence, context transitions are of type $[e] \rightarrow [e] \rightarrow t$;
- Given the current discourse, T, and a new expression,C, C updates T as follows:
- $[\![\overline{(\mathbf{T}.\mathbf{C})}]\!]^{M,cg} = \lambda k.[\![\overline{\mathbf{T}}]\!](\lambda n.[\![\overline{\mathbf{C}}]\!](\lambda m.k(m\ n)))$

Adding Gesture to Common Ground

Multimodal Contextualized Reference

- Representing how gestures denote
- Encoding co-perception of situated objects under reference
- Situated alignment of expressions from distinct modalities

Gesture Types in Multimodal Interactions

1. Deixis (pointing) gestures, generated to request information regarding an object, a location, or a direction when performing a specific action;
2. Iconic action gestures, generated to request clarification on how (what manner of action) to perform a specific task;
3. Affordance-denoting gestures, generated to describe how the agent can interact with an object, even when it does not know what it is or what it might be used for;
4. Direct situated actions, where the agent responds to a command or request by acting in the environment directly.

Gestures used in VoxWorld

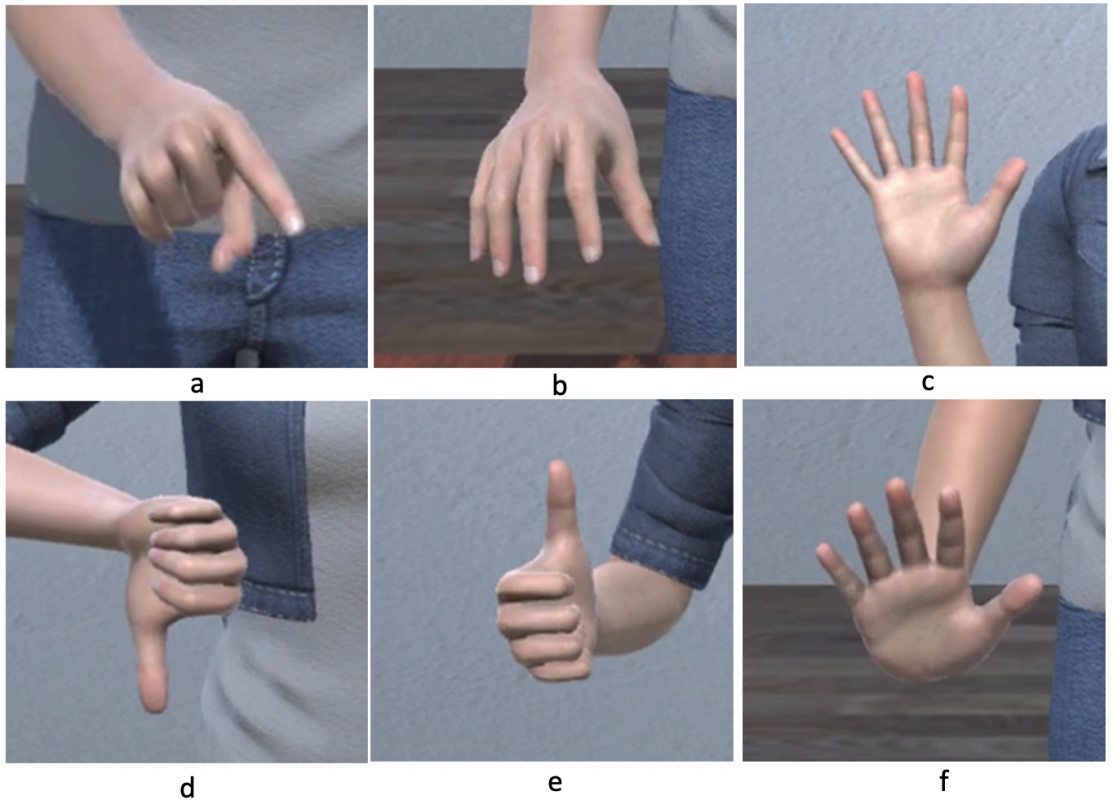


Figure: Some of the gestures generated by VoxWorld: pointing, grab, five, no, yes, push back.

Bidirectional Gesture Recognition and Generation

- On the left, a human is action gesturing to move an object to the left:
- On the right, the IVA is performing the identical gesture.

Gestures in Multimodal Interactions

1. Deixis (pointing) gestures, generated to request information regarding an object, a location, or a direction when performing a specific action;
2. Iconic action gestures, generated to request clarification on how (what manner of action) to perform a specific task;
3. Affordance-denoting gestures, generated to describe how the IVA can interact with an object, even when it does not know what it is or what it might be used for;
4. Direct situated actions, where the IVA responds to a command or request by acting in the environment directly.

Situated Meaning

Gesture sequence command

Single Modality (Gesture) Imperative

diana$_1$: $\mathcal{G}$ = [*points to the purple block*]$_{t1}$
diana$_2$: $\mathcal{G}$ = [*makes move gesture*]$_{t2}$
diana$_3$: $\mathcal{G}$ = [*points to the blue block*]$_{t3}$

Situated Meaning

Gesture sequence command

Figure: Gesture generation for performing complex action.

Common Ground Structure (CGS)

(4) State Monad: $M\alpha = \textit{State} \rightarrow (\alpha \times \textit{State})$

(5) a. **A**: The agents engaged in communication;
b. **B**: The shared belief space;
c. **P**: The objects and relations that are jointly perceived in the environment;
d. $\mathcal{E}$: The embedding space that both agents occupy in the communication.

(6)

A:a_1, a_2 **B:**Δ **P:**b
$\mathcal{S}_{a_1}$ = "You$_{a_2}$ see it$_b$"

$\mathcal{E}$

Multimodal Configurations

- A communicative act, performed by an agent, a, is a tuple of expressions from the modalities available to a, involved in conveying information to another agent.
- We restrict this to the modalities of a linguistic utterance, S (either an intonational contour or speech), and a gesture, G. There are three possible configurations in performing C:
  1. $C_a = (G)$
  2. $C_a = (S)$
  3. $C_a = (S, G)$
- These modal channels can be aligned or unaligned in the input.

Actions as Described by Gesture

Kendon (2004), Lascarides and Stone (2009)

- $G = (prep); (prestrokehold); stroke; retract$

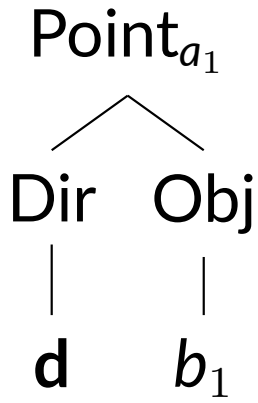
The stroke is the content-baring phase, **d**, and in a pointing gesture, will convey the deictic orientational information.

- $[\![point]\!] = [\![End(cone(\mathbf{d}))]\!]$
- Gestures can denote a range of primitive action types, including: grasp, hold, pick up, move, throw, pull, push, separate, and put together.

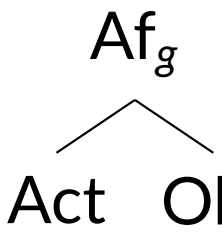
Gesture Grammar

Pustejovsky (2018)

(7) a. **Deixis**: $Point_g \rightarrow Dir \;\; Obj$



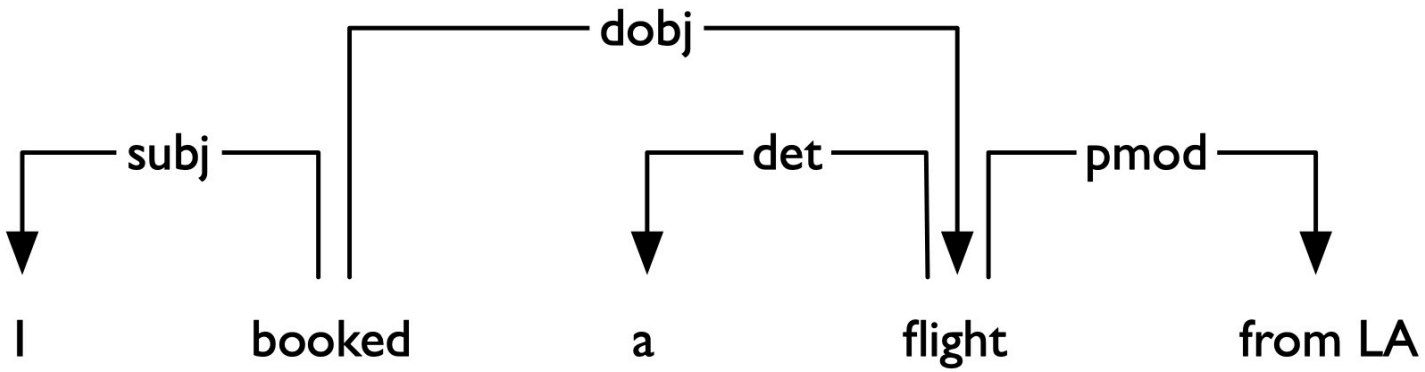
b. **Affordance**: $Af_g \rightarrow Act \;\; Obj$



Gesture Grammar

- action-object: e.g., *grab* [**Object**]
- $GP_1 \rightarrow G_{Af}\ D_{obj}$ (Action Focus)
  $\rightarrow\ D_{obj}\ G_{Af}$ (Object Focus)
- action-result: e.g., *put* [**Object**] at [**Location**]
- $GP_2 \rightarrow G_{Af}\ D_{obj}\ D_{loc}$ (Action Focus)
  $\rightarrow\ D_{obj}\ G_{Af}\ D_{loc}$ (Object Focus)
  $\rightarrow\ D_{obj}\ D_{loc}\ G_{Af}$ (Transition Focus)
- action-result: e.g., *move* [**Object**] [**Direction**]
- $GP_3 \rightarrow G_{Af}\ D_{obj}\ D_{dir}$

Syntactic Dependency Tree

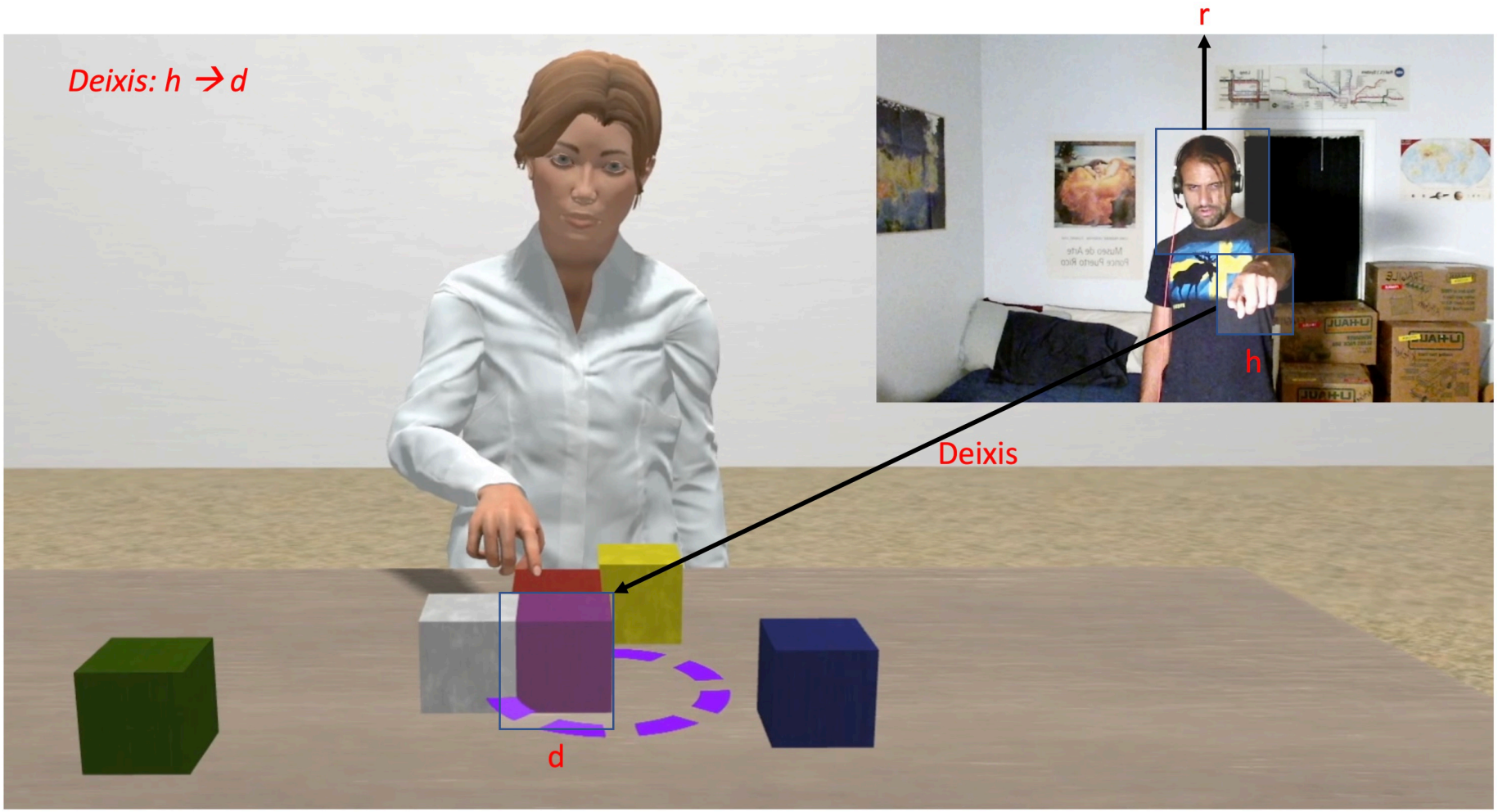


- In an arc $h \rightarrow d$, the word h is called the head, and the word d is called the dependent.
- The arcs form a rooted tree.
- Each arc has a label, l, and an arc can be described as (h, d, l)

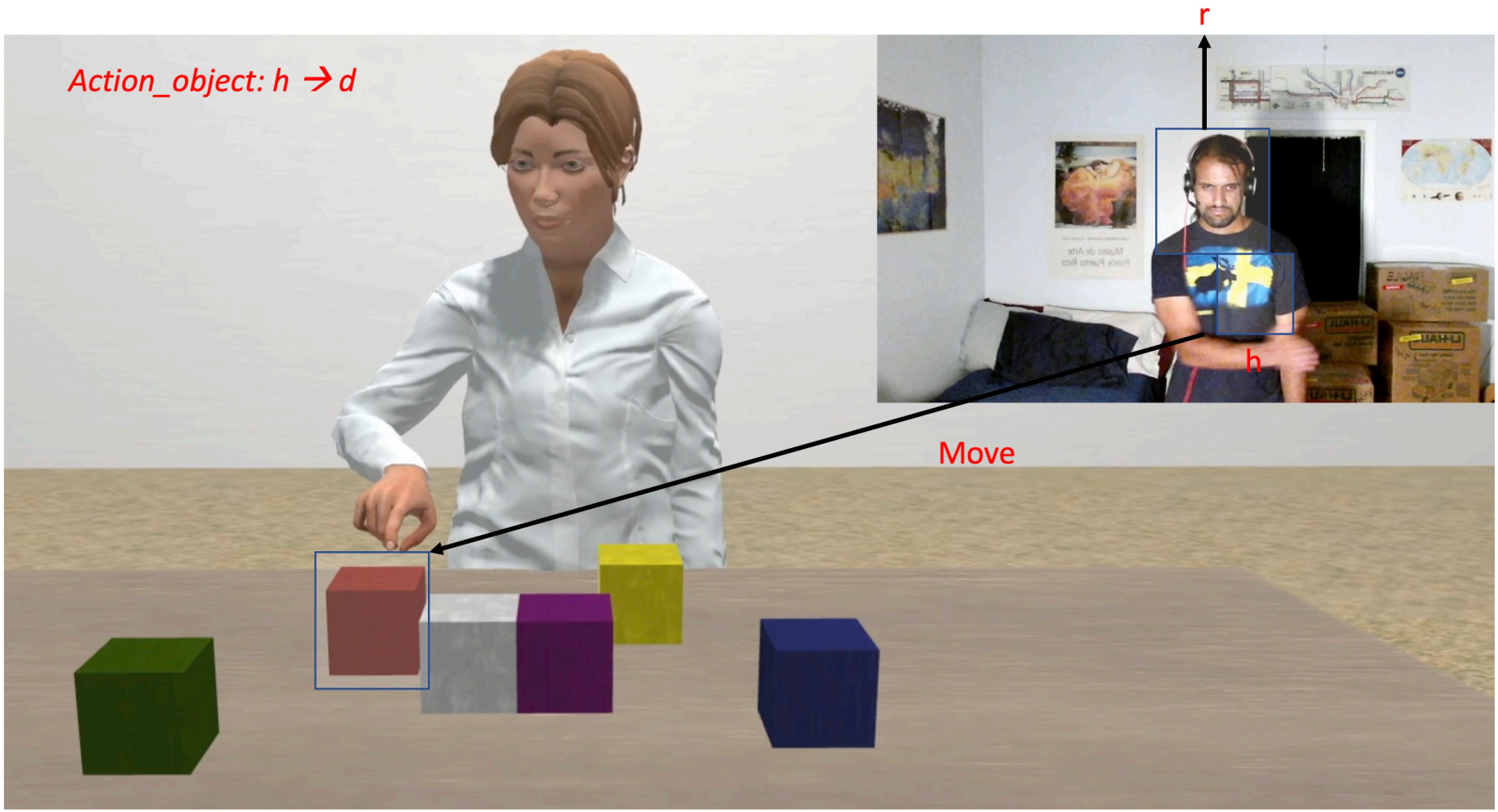
Gesture Dependency Structure

- The agent is the root.
- The gesture stroke is the head, h and its referent is the dependent, d.
- Arcs form a rooted tree.
- Each arc has a label, l, indicating its type.
- Example: (h, d, l)
  (hand,block,deixis)

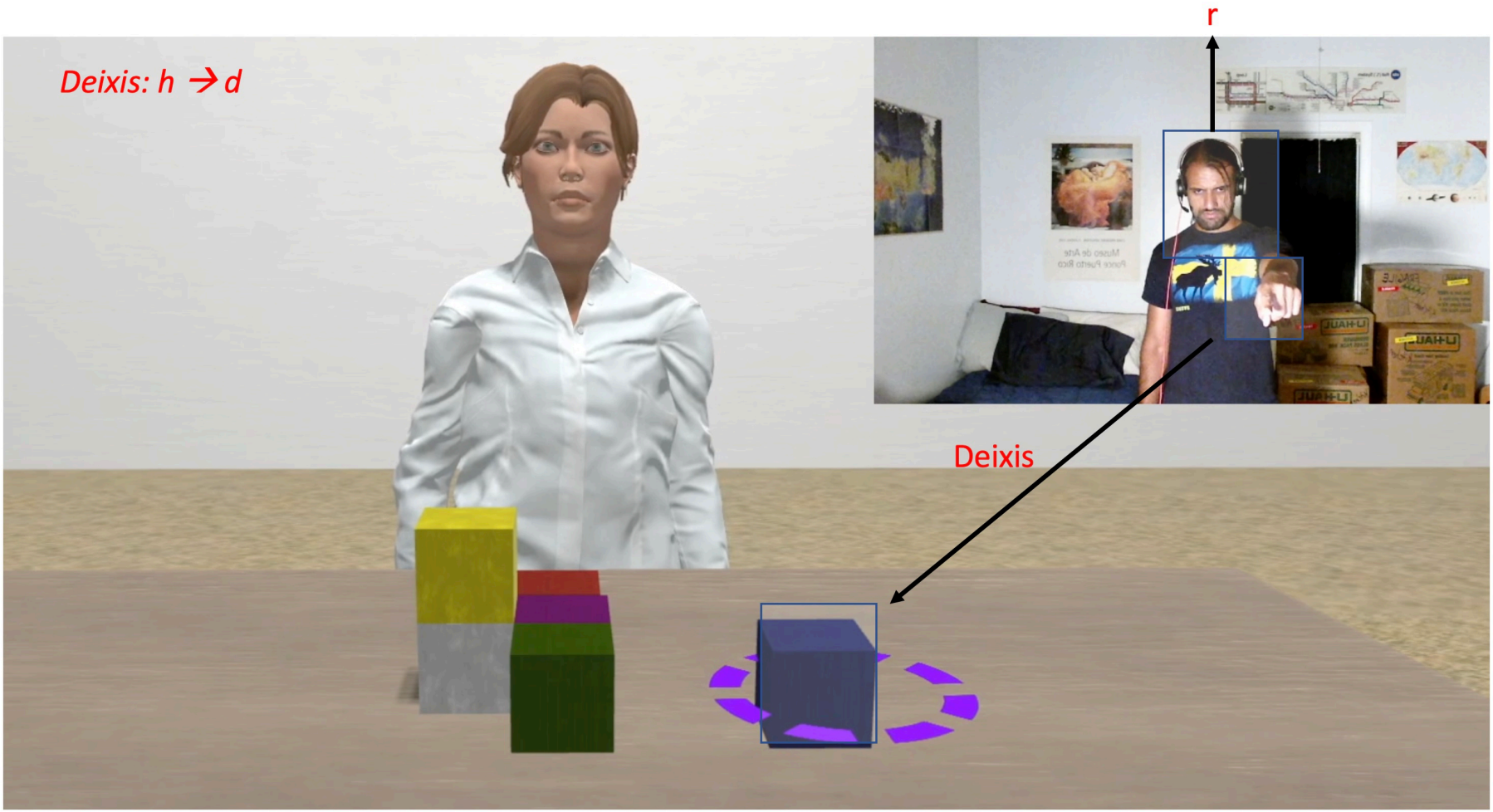
Gesture Dependency Structure



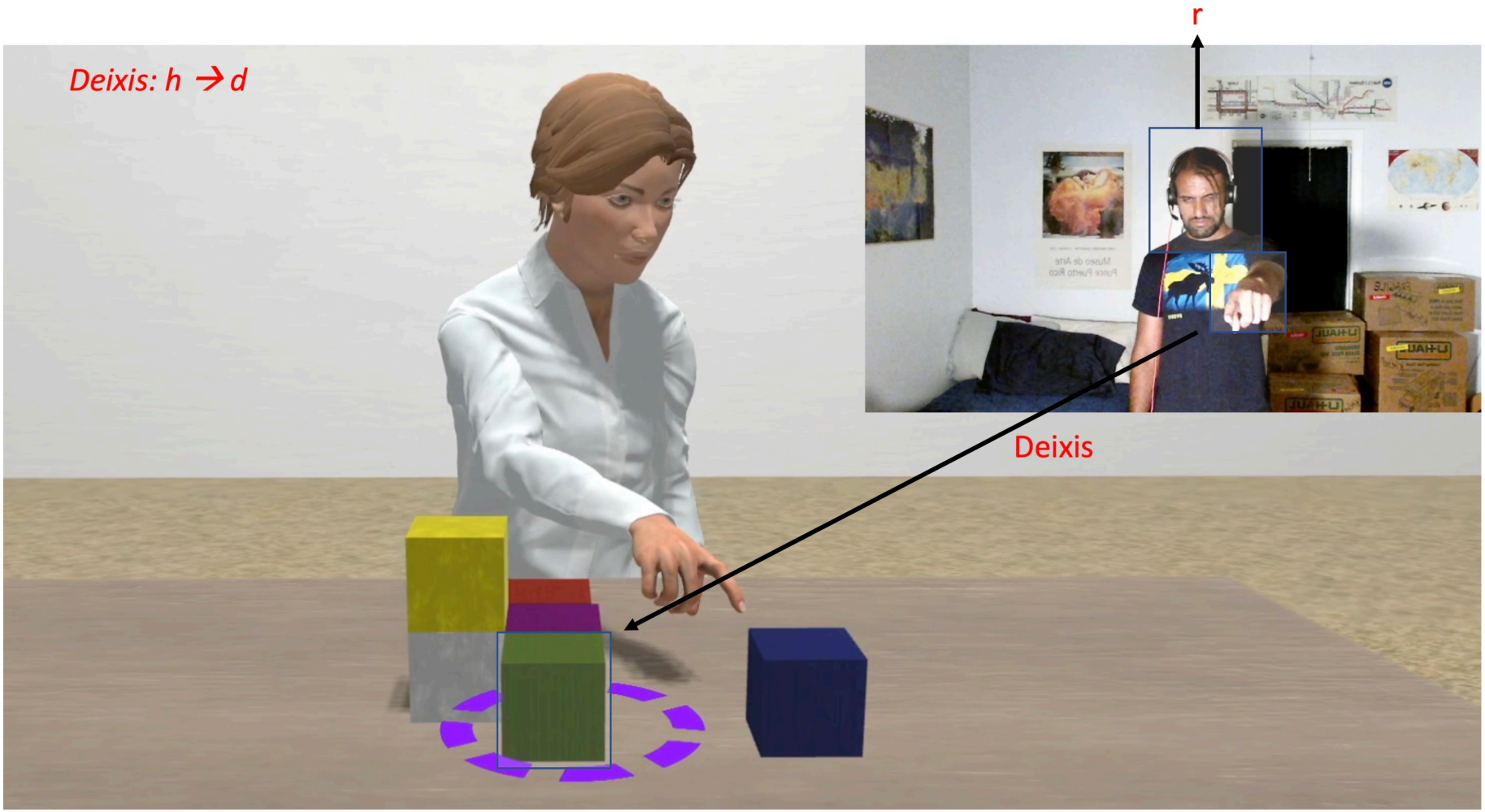
Gesture Dependency Structure



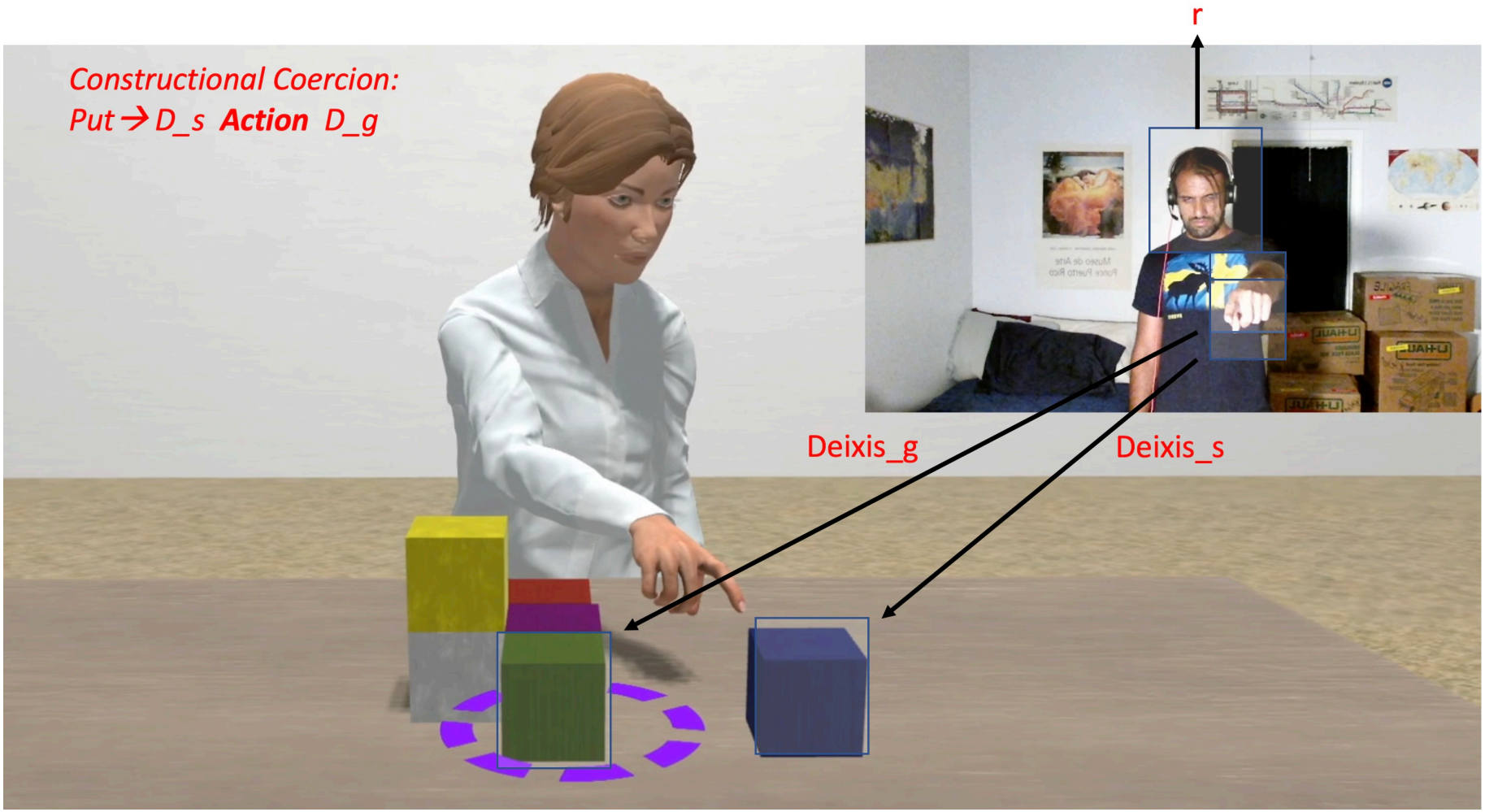
Gesture Dependency Structure



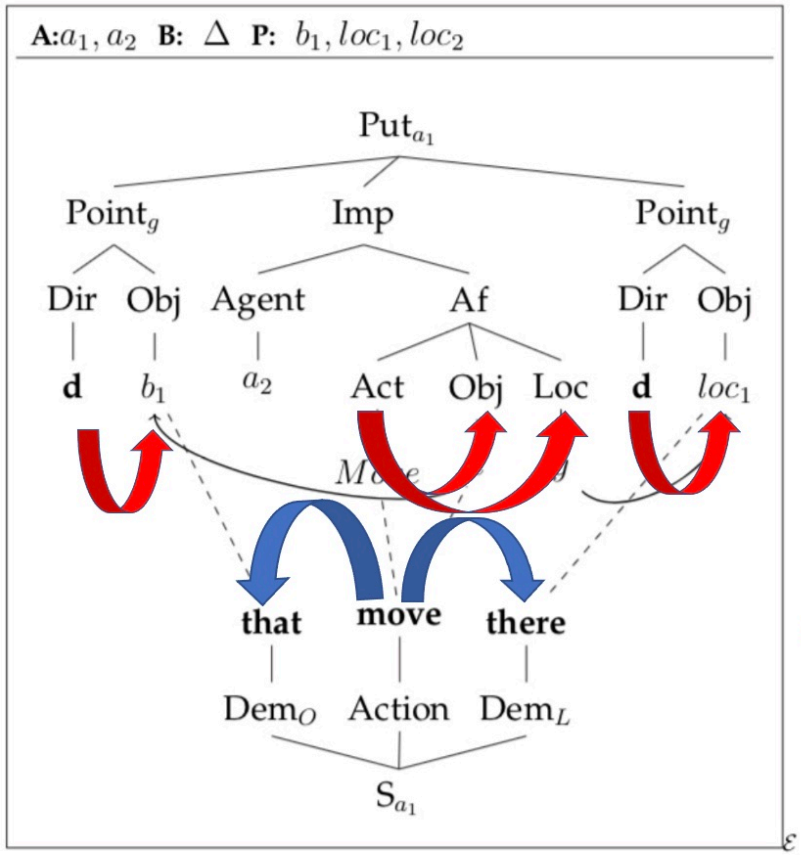
Gesture Dependency Structure



Gesture Dependency Structure



Linking Dependency Structures



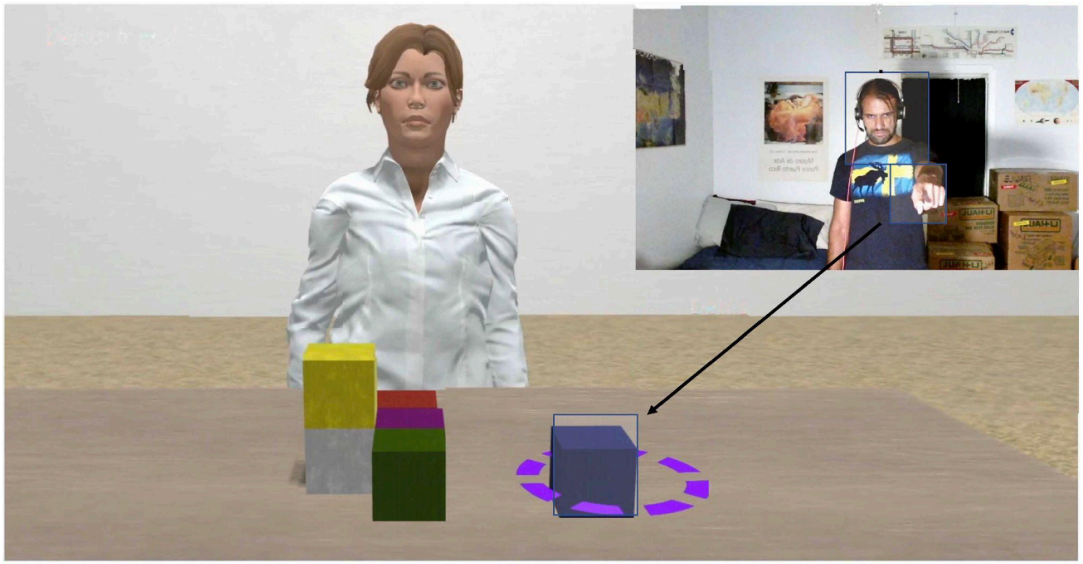
Gesture Dependency Structure

Syntactic Dependency Structure

Motivating Gesture AMR

- Exploits the natural association between Dependency Structures and DAGs (e.g., AMR graphs);
- Facilitates the alignment to Speech, when represented as AMRs or UMRs;
- Much easier to annotate for creating large datasets of multimodal dialogues
  - for theoretical investigations;
  - for training ML systems.

Gesture AMR (GAMR)

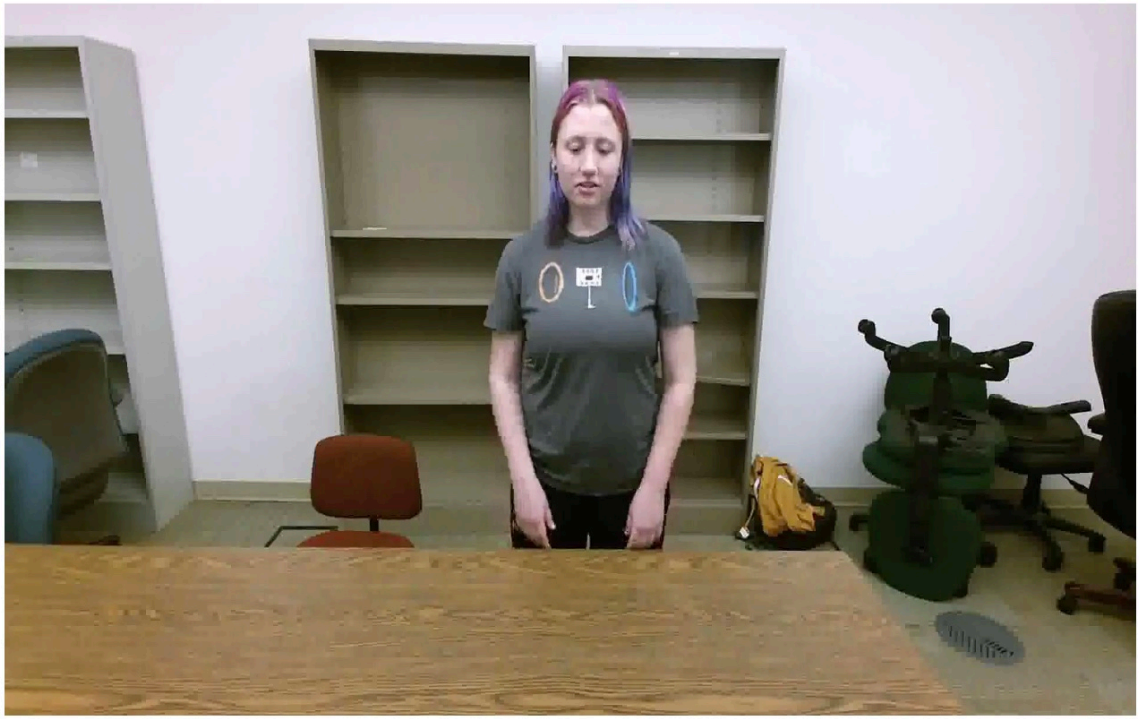


- Needs to represent:
  - situated meaning
  - common ground, objects, participants
  - modes of communication
- While abstracting away from physical descriptions

Gesture AMR (GAMR)

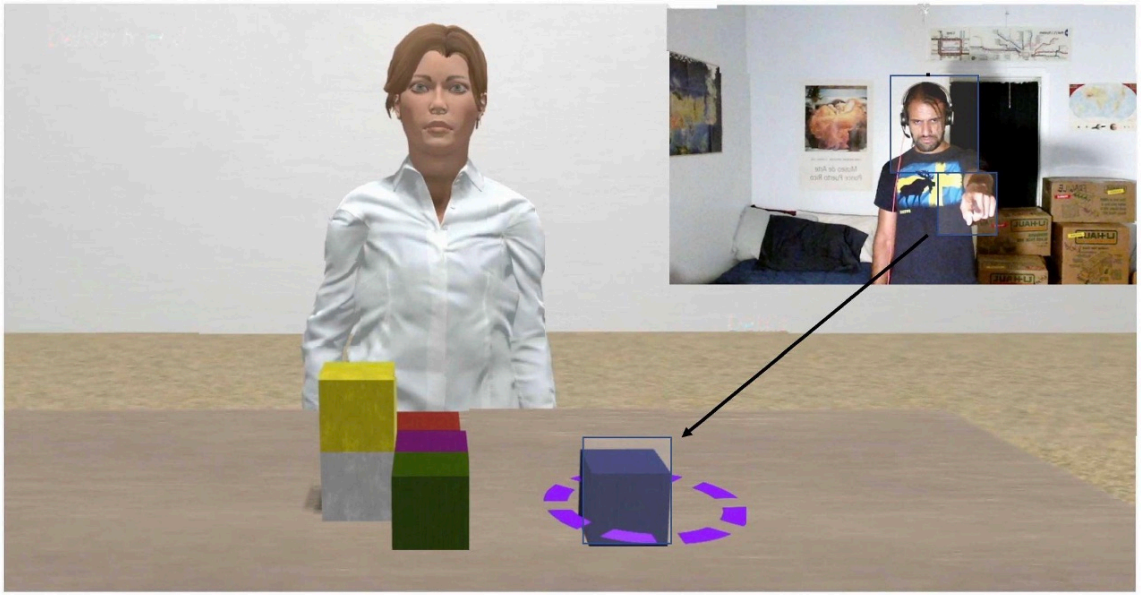
- 8 hours of video
- 40 participants
- Shared task of arranging block structures
- *Signaler* gives instructions
- *Actor* interprets and builds structures
- Wang et al, 2017

Gesture AMR (GAMR)



- Gesture: how speakers move hands when they communicate information
- Existing schemes typically focus on hand shape and movement
- We focus on gesture tied to speech: iconic, deictic, metaphoric, & emblematic

Gesture AMR (GAMR)

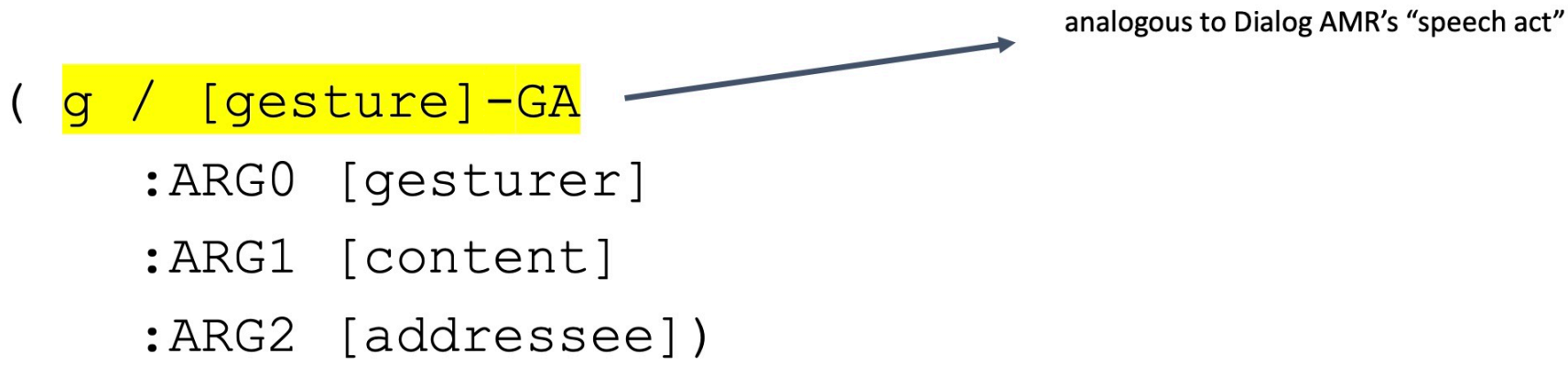


- Needs to represent:
  - situated meaning
  - common ground, objects, participants
  - modes of communication
- While abstracting away from physical descriptions

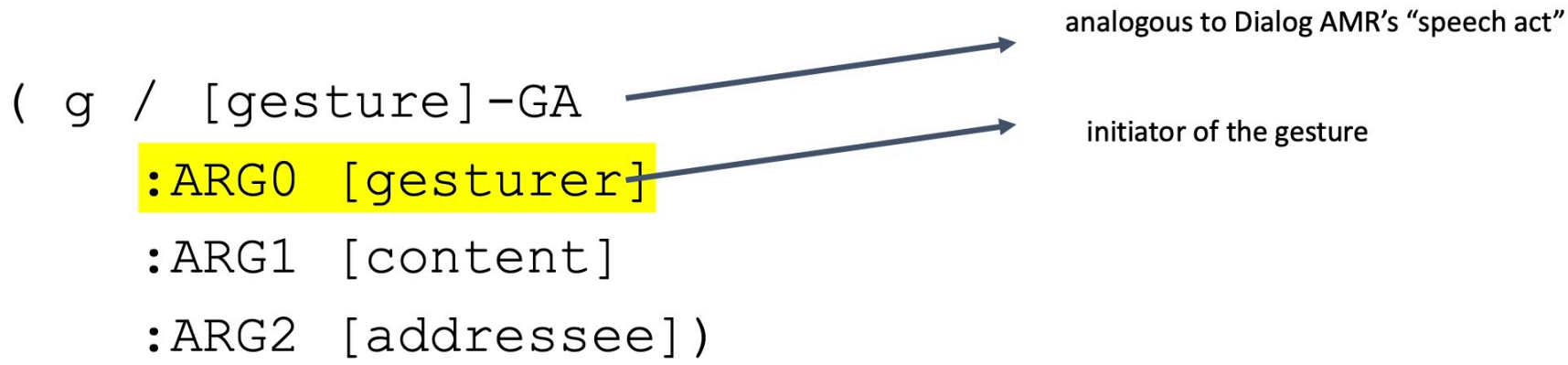
Gesture AMR (GAMR)

```
( g / [gesture]-GA
     :ARG0 [gesturer]
     :ARG1 [content]
     :ARG2 [addressee])
```

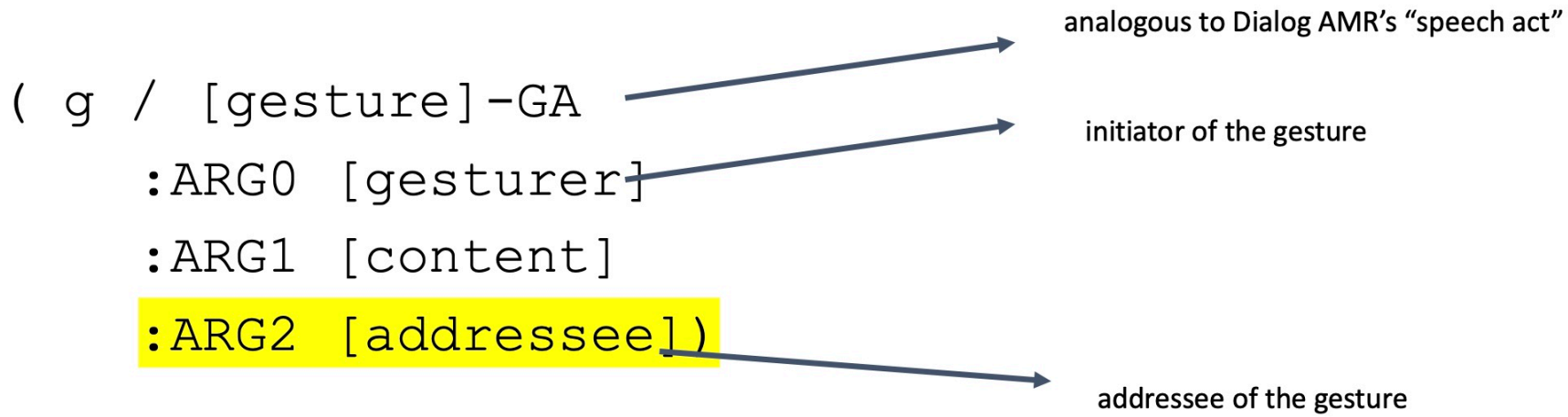
Gesture AMR (GAMR)



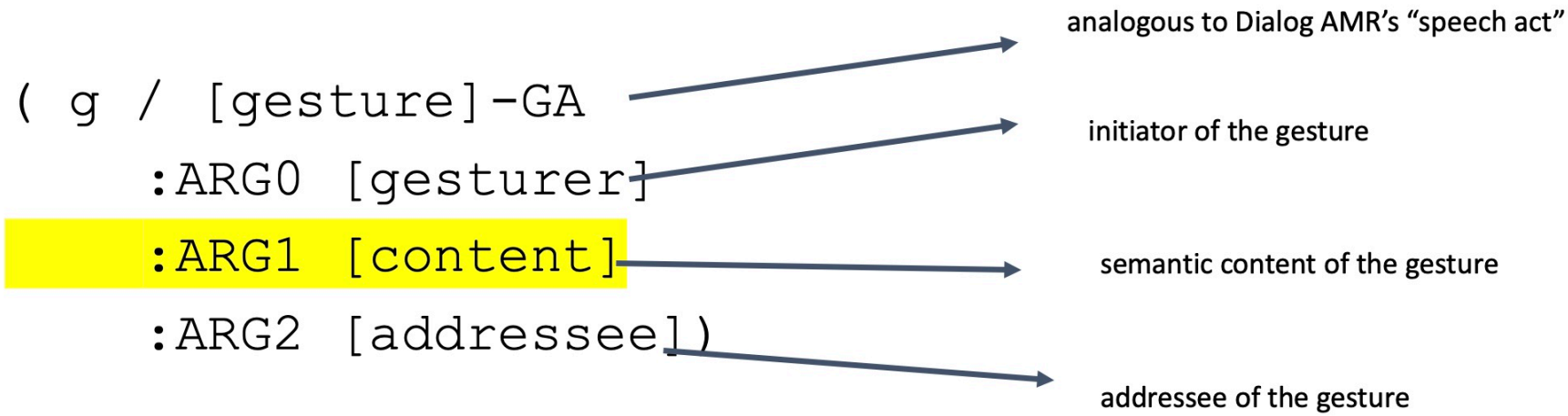
Gesture AMR (GAMR)



Gesture AMR (GAMR)



Gesture AMR (GAMR)

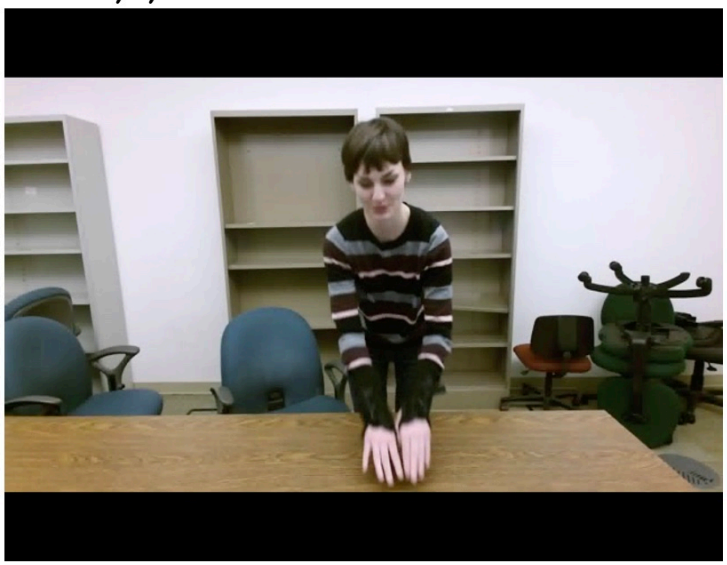


Gesture AMR - Deixis

```
(d / deixis-GA
      :ARG0 (g / gesturer)
      :ARG1 (b / block)
      :ARG2 (a / addressee
```

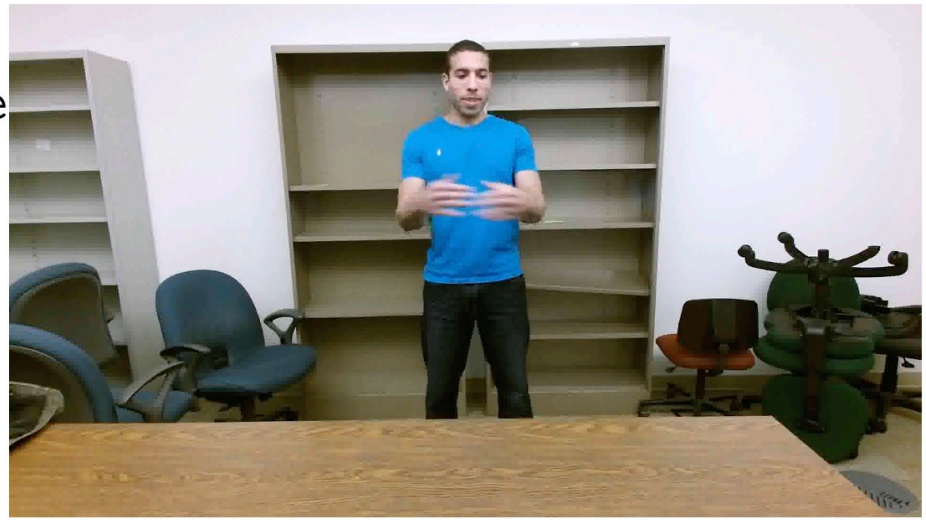
Gesture AMR - Iconic

```
(i / icon-GA
      :ARG0 (g / gesturer)
      :ARG1 (s / slide-01)
            :direction (f / forward))
      :ARG2 (a / addressee))
```



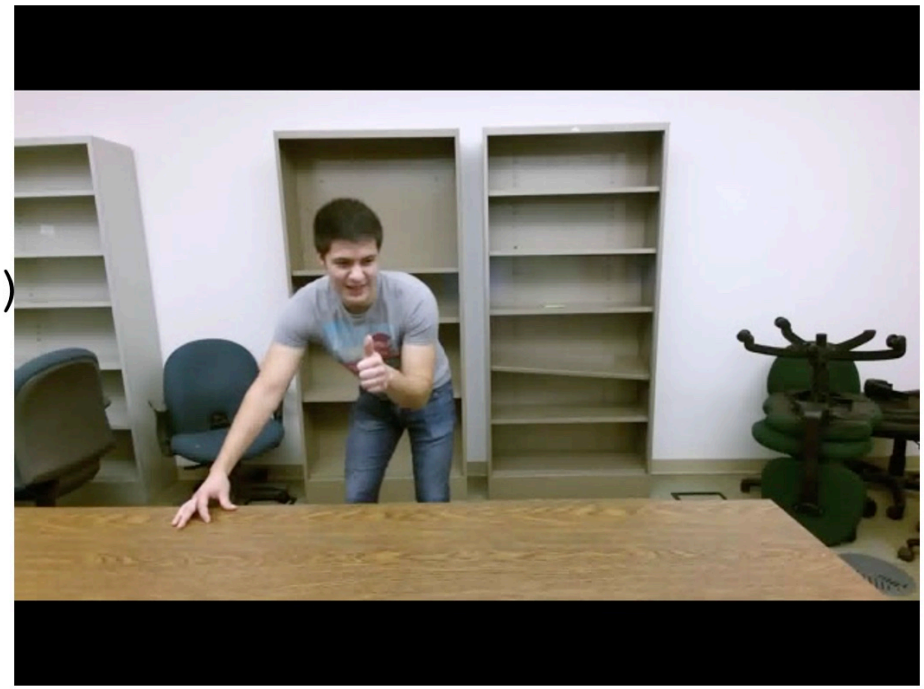
Gesture AMR - Metaphor

```
(m / metaphor-GA
      :ARG0 (g / gesturer)
      :ARG1 (s / sound wave
      :ARG2 (a / addressee)
```



Gesture AMR - Emblem

```
(e / emblem-GA
     :ARG0 (g / gesturer)
     :ARG1 (y / yes)
     :ARG2 (a / addressee))
```



Gesture Sequences - Units

```
(g / gesture-unit
     :op1 (i / icon-GA
          :ARG0 (g2 / gesturer)
          :ARG1 (b / block)
          :ARG2 (a / addressee))
     :op2  (d / deixis-GA
          :ARG0 g2
          :ARG1 (l/ location)
          :ARG2 a))
```

Gesture Sequences - Units

```
(g / gesture-unit
     :op1 (i / icon-GA
          :ARG0 (g2 / gesturer)
          :ARG1 (b / block)
          :ARG2 (a / addressee))
     :op2  (d / deixis-GA
          :ARG0 g2
          :ARG1 (l/ location)
          :ARG2 a))
```

Gesture Sequences - Units

```
(g / gesture-unit
     :op1 (i / icon-GA
          :ARG0 (g2 / gesturer)
          :ARG1 (b / block)
          :ARG2 (a / addressee))
     :op2  (d / deixis-GA
          :ARG0 g2
          :ARG1 (l/ location)
          :ARG2 a))
```

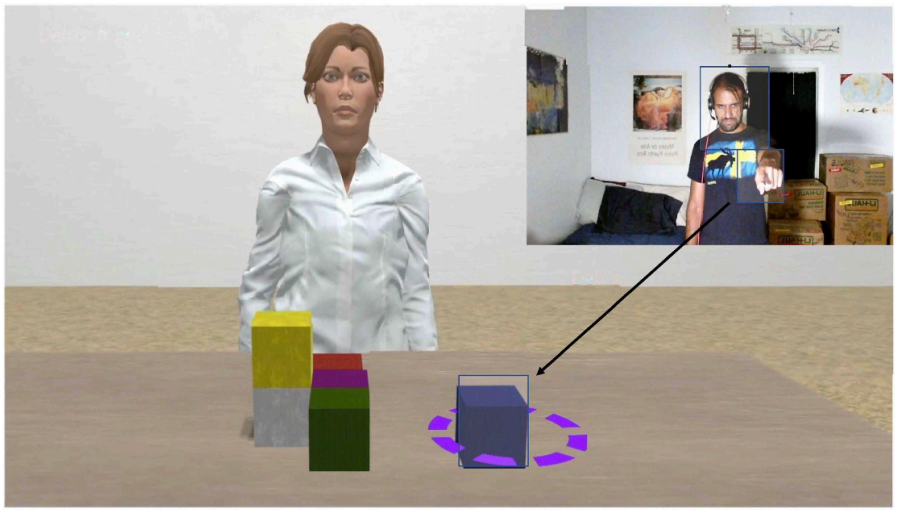
Aligning Gesture Units

(2) **Co-gestural Speech Ensemble**:

$$\begin{bmatrix} \mathcal{G} & g_1 & \dots & g_i & \dots & g_n \\ \mathcal{S} & s_1 & \dots & s_i & \dots & s_n \end{bmatrix}$$

CO-GESTURAL SPEECH

HUMAN: s_1 = Put
$g_1 = \varnothing$
HUMAN: s_2 = [that block]
g_2 = [*points to the blue block*]
HUMAN: s_3 = there.
g_3 = [*points to the purple block*]



Annotating situated actions in dialogue

- Why are actions important for dialogue?
- Actions contribute to context
  - Anaphora
    - [*lifts pencil*] “I used this for the sketch.”
    - “My brother said ‘thumbs up’!”
  - Bridging
    - “I went to the store today.” [*takes fruit out of a grocery bag*]

Annotating situated actions in dialogue

- Why are actions important for dialogue?
- Actions change the state of the world
  - Actions can add, modify, or delete items in the common ground
  - Tracking objects and actions in the environment is necessary for situated grounding