Multimodal Semantics for Affordances and Actions

Lecture 3: Modeling Multimodal Common Ground

James Pustejovsky and Nikhil Krishnaswamy

ESSLLI 2022 Summer School

Galway, Ireland

August 8-19, 2022

Course Outline

- Monday: Components of Multimodal Communication
- Tuesday: Modeling Human-Object Interactions
- Wednesday: Modeling Multimodal Common Ground
- Thursday: Communicating with Multimodal Common Ground
- Friday: Reasoning with and about Affordances

Wednesday's Outline

- Recap - Shared Tasks
- VoxWorld and Embodied Interaction
- Embodiment within the Common Ground
- Accounting for Other Modalities: Gesture
- Aligning Language and Gesture

Recap . . . Embodied Communication

Mother and child interacting in a shared task

Situated Meaning in a Joint Activity

- Son: *Push on it (gesturing with hands)?*
- Mother: *Yes, press down.*
- Mother: *OK, that's enough. (co-attentional gaze)*
- Son: *Okay. (stops action)*
- Mother: *Now, let's sprinkle sugar on this.*

- Machine vision, language, gesture, action, common ground

Entity Type	**Examples**
Agents	mother, child
Shared goals	baking, icing
Beliefs, desires, intentions	Mother knows how to ice, bake, etc. Mother is teaching child
Objects	Mother, son, dough, counter, cutter, sugar pan, sugar, baking pan
Shared perception	the objects on the table
Shared Space	kitchen

Figure: Elements from the common ground.

Communicating in the Common Ground

1. Objects and events as we experience them are distinct from the way we refer to them with language.
2. The mechanisms in language allow us to package, quantify, measure, and order our experiences, creating rich conceptual reifications and semantic differentiations.
3. The surface realization of this ability is mostly manifest through our linguistic utterances, but is also witnessed through gestures.
4. By examining the nature of the common ground assumed in communication, we can study the conceptual expressiveness of these systems.

Common Ground - What is it?

- Defining Common Ground: Clark et al. (1991); Gilbert (1992); Traum (1994); Stalnaker (2002); Asher (1998); Tomasello and Carpenter (2007)
- The ability to understand another person in a shared context, through the use of co-situational and co-perceptual anchors, along with a means for identifying such anchors, using:
  - language
  - gesture
  - gaze
  - intonation.

Common Ground - Situated Experience

- Shared experiences (Co-situated, Co-perceptive)
  - witnessing a natural event
  - hearing a clap of thunder
  - feeling the earth tremor
- Agents in Shared Actions (Co-intention, Co-attention)
- Shared situated references
  - Objects and states are annotated by language and gesture
  - The communicative acts are now part of the shared experience

Different Models of Simulations

1. *Computational simulation modeling*. Variables in a model are set and the model is run, such that the consequences of all possible computable configurations become known.
2. *Situated embodied simulations*. Agent is embodied with a dynamic point-of-view or avatar in a virtual or simulated world.
3. *Embodied theories of mind*. The notion that agents carry a mental model of external reality in their heads.

VoxWorld and Embodied Interaction

- A contextualized 3D virtual realization of both the situational environment and the co-situated agents, as well as the most salient content denoted by communicative acts in a discourse.
- Built on the modeling language VoxML:
  - encodes objects with rich semantic typing and action affordances;
  - encodes actions as multimodal programs;
  - reveals the elements of the common ground in discourse between speakers;
- Offers a rich platform for studying the generation and interpretation of expressions, as conveyed through language and gesture;

The situated common ground consists of the following state information:

(1) 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.

(2)
$$\frac{\textbf{A:}a_1, a_2 \quad \textbf{B:}\Delta \quad \textbf{P:}b}{\mathcal{S}_{a_1} = \text{“You}_{a_2} \text{ see it}_b\text{”}}_{\mathcal{E}}$$

Public Announcement Logic

Plaza (1989), Baltag et al (1998), van Benthem et al (2006)

Modeling the knowledge of agents: d (Diana) and h (Human):

- $[a]p$: Agent a knows that p.
- Agent knowledge is encoded as sets of accessibility relations between situations: α.
- What is known is encoded as propositions in situations: ϕ.
- $\phi ::= \top \mid p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid [\alpha]\phi \mid [!\phi_1]\phi_2$
- $\alpha ::= a \mid ?\phi \mid \alpha_1; \alpha_2 \mid \alpha_1 \cup \alpha_2 \mid \alpha^*$
- Presupposition: $[(d \cup h)^*]\phi_p$

Multimodal Presuppositions in the Common Ground

Modeling the knowledge of agents: d (Diana) and h (Human):

- $[d]Point\_gesture$
- $[h]Diana\_at\_table$
- Presupposition: $[(d \cup h)^*]\phi_p$
- Assertion in the common ground: $[(d \cup h)^*]\phi_p \wedge \psi$
- "Move the blue block."
  $[!([(d \cup h)^*]Blue\_block \wedge [(d \cup h)^*]Grab\_gesture) \wedge Move\_block]$

Public Perception Logic 1/2

Modeling the perception of agents: d (Diana) and h (Human):

- Agent synthetic vision is encoded as sets of accessibility relations, α, between situations:
- What is seen in a situation is encoded as either a proposition, ϕ, an existential of an object, x, $\hat{x}$;
- $[a]_\sigma p$: Agent a perceives that p.
- $[a]_\sigma \hat{x}$: Agent a perceives that there is an x.
- $\neg[a]_\sigma \hat{x}$: Agent a does not perceive that there is an x.
- $\phi ::= \top \mid p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid [\alpha]_\sigma \phi \mid [!\phi_1]_\sigma \phi_2$
- $\alpha ::= a \mid ?\phi \mid \alpha_1; \alpha_2 \mid \alpha_1 \cup \alpha_2 \mid \alpha^*$

Public Perception Logic 2/2

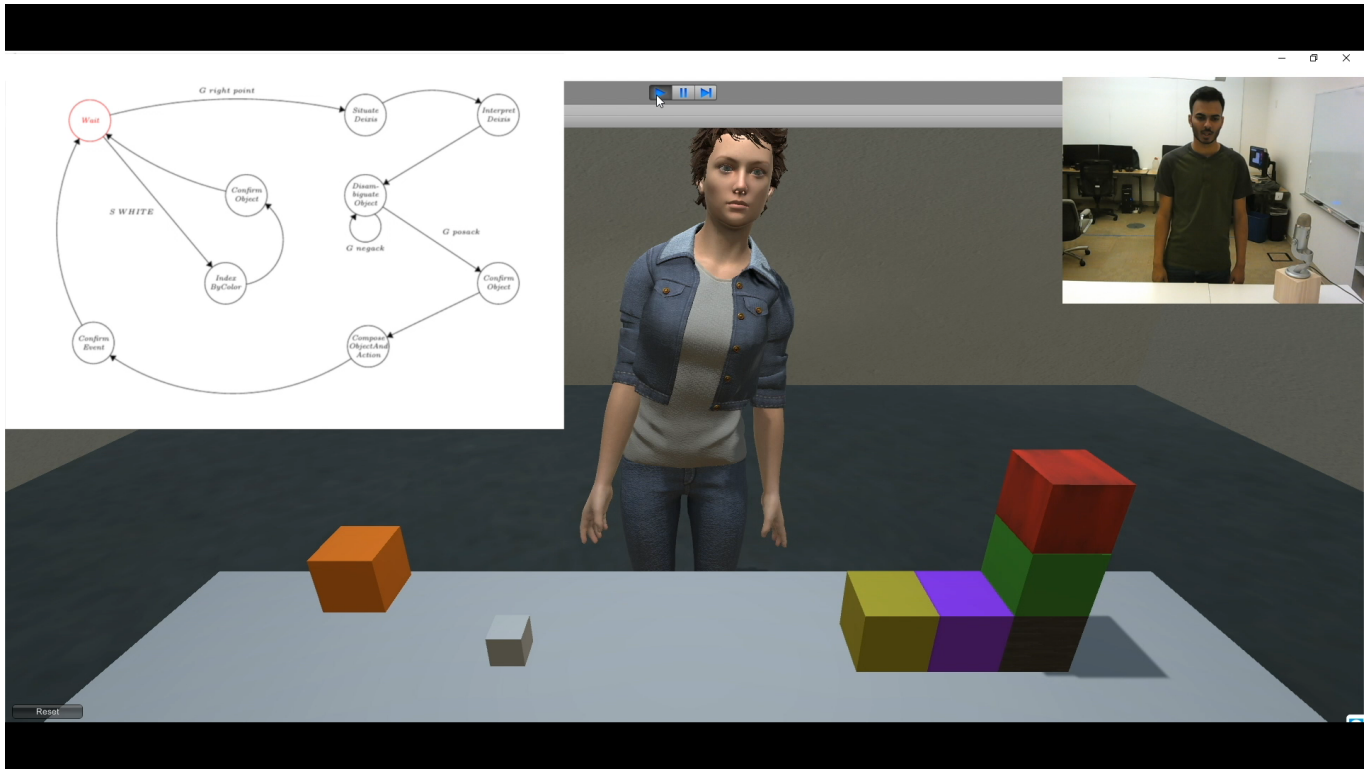
Common Ground involves co-perception:

- In order to co-attend, two agents direct gaze towards an object or event:
  $[a]_\sigma e_i$, $[b]_\sigma e_i$;
- Each agent sees the other attend;
  $[a]_\sigma([b]_\sigma e_i)$, $[b]_\sigma([a]_\sigma e_i)$.
- Each agent sees that the other agent sees her/him attend;
  $[b]_\sigma([a]_\sigma([b]_\sigma e_i))$, $[a]_\sigma([b]_\sigma([a]_\sigma e_i))$
- The co-perception for Diana and Human includes ϕ
  ("Everyone can see that ϕ.")
  $[(d \cup h)^*]_\sigma \phi$

Public Perception Logic 3/3

- Diana does not see the small purple block.
  $\neg[d]_\sigma\ Purple\_small$
- Everyone sees that the red block is on the black block.
  $[(d \cup h)^*]_\sigma on(Red, Black)$
- The small purple block is not visible to everyone.
  $\neg[(d \cup h)^*]_\sigma Purple\_small$

- Dialogue Manager PDA



▸ Link

The situated common ground consists of the following state information:

(3) 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.

(4)
$$\begin{array}{|c|}\hline \textbf{A}{:}a_1, a_2 \quad \textbf{B}{:}\Delta \quad \textbf{P}{:}b \\ \hline \\ \mathcal{S}_{a_1} = \text{“You}_{a_2} \text{ see it}_b\text{”} \\ \\ \hline \end{array}_{\mathcal{E}}$$

A state monad corresponds to those computations that read and modify a state in the discourse. **M** is a type constructor that constructs a function type taking a state as input and returns a pair of a value and a new or modified state as output.

- 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 $[e] \rightarrow [e] \rightarrow t$;

- State Monad: $M\alpha = State \rightarrow (\alpha \times State)$
- Given the current discourse, T, and a new expression, C, C updates D as follows:
- $[\![\overline{(\mathbf{T.C})}]\!]^{M,cg} = \lambda k.[\![\overline{\mathbf{T}}]\!](\lambda n.[\![\overline{\mathbf{C}}]\!](\lambda m.k(m\ n)))$
- $\mathbf{S}_0$: $[x, y, \dots]$ - Grab the blue block. $\Longrightarrow [\mathrm{b}_1, x, y, \dots]$
- $\mathbf{S}_1$: $[b_1, x, y, \dots]$ - Pick it_{b_1} up. $\Longrightarrow [\mathrm{b}_1, x, y, \dots]$

Continuations in Communicative Acts

Dialogue 1: Co-reference across multiple sentences

HUMAN$_1$: $\mathcal{S}$ = Pick up *a blue block*$_1$.
HUMAN$_1$: $\mathcal{S}$ = Move *it*$_1$ there.

- The information state is updated through a CPS transformation, creating the continuized type for each expression.
- Given the current discourse, D, and the new utterance, S, S integrates into D as follows:

(5) $[[\overline{(\mathbf{D}.\mathbf{S})}]]^{M,cg} = \lambda i \lambda k.[[\overline{\mathbf{D}}]]i(\lambda i'.[[\overline{\mathbf{S}}]]i'k)$

Unpacking the Continuation

(6) $[[\overline{(\mathbf{D.S})}]]^{M,cg} = \lambda i \lambda k.[[\overline{\mathbf{D}}]]i(\lambda i'.[[\overline{\mathbf{S}}]]i'k)$

- This states that the current discourse has two arguments, its left context i (where we are), and what is expected later in the discourse, k.
- The anaphoric pronoun (*it*) in the second sentence is interpreted relative to the introduction of the linguistic expression (*a blue block*) in the previous sentence.
- As a result, it has a logical antecedent that it can refer to.
- The first sentence is the context within which the second is interpreted, resulting in the pronoun *it* taking *a blue block* as its antecedent.

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 and gesture, G. Possible configurations in performing C:
  1. $C_a = \{(G), (S), (S, G)\}$
- These modal channels can be aligned or unaligned in the input.
- Monads allow for *informational distribution* among multimodal expressions being used in composition to form larger meanings.

Speech or Gesture in the Common Ground

(7) a.

A:a_1, a_2 **B**:Δ **P**:b $\mathcal{E} : E$
$\mathcal{G}_{a_1}$ = "grab it$_b$"

b.

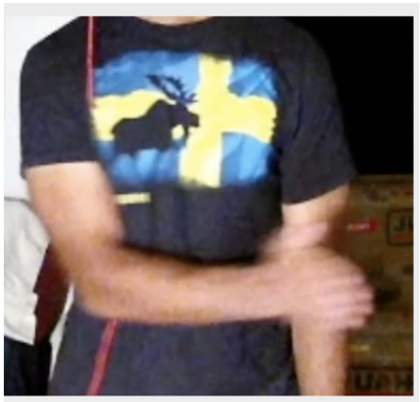
A:a_1, a_2 **B**:Δ **P**:b $\mathcal{E} : E$
$\mathcal{S}_{a_1}$ = "You$_{a_2}$ see it$_b$"

Modeling Action Composition in VoxWorld

- Object Model: State-by-state characterization of an object as it changes or moves through time.
- Action Model: State-by-state characterization of an actor's motion through time.
- Event Model: Composition of the object model with the action model.

Bidirectional Gesture Recognition and Generation

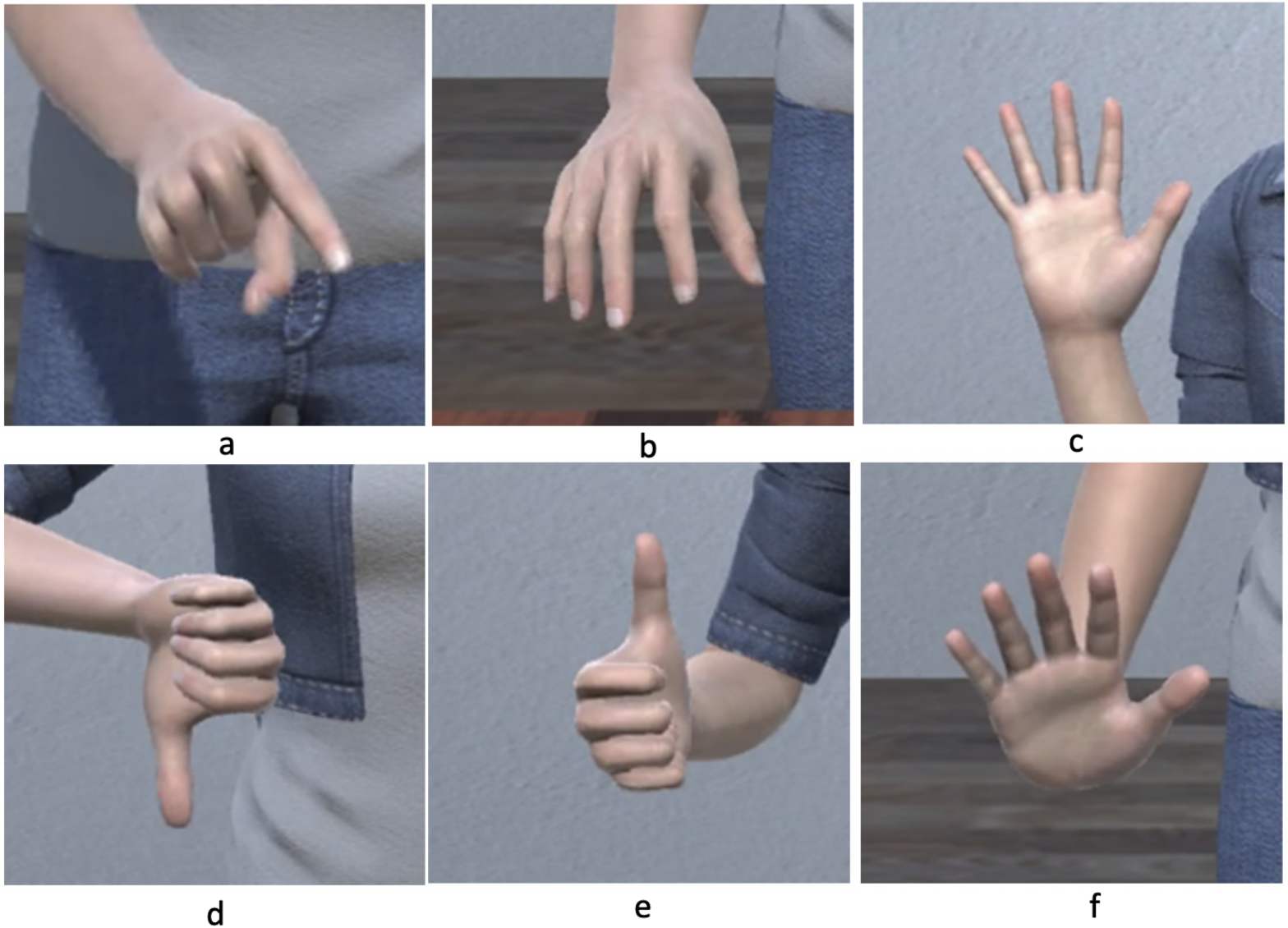
- On the left, a human is action gesturing to move an object to the left:
- On the right, the IVA is performing the identical gesture.



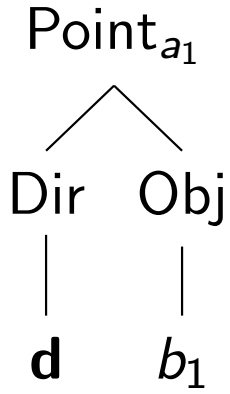
Gestures in Multimodal Interactions

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.

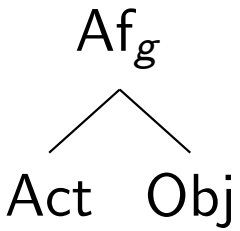
Gestures used in VoxWorld System



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



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



Gesture Grammar

(9) a. ACTION-OBJECT: e.g., *grab* [**Object**]

b. $GvP_1 \rightarrow G_{Af}\ D_{obj}$ (Action Focus)
$\rightarrow D_{obj}\ G_{Af}$ (Object Focus)

(10) a. ACTION-RESULT: e.g., *put* [**Object**] at [**Location**]

b. $GvP_2 \rightarrow G_{Af}\ D_{obj}\ D_{loc}$ (Action Focus)
$\rightarrow D_{obj}\ G_{Af}\ D_{loc}$ (Object Focus)
$\rightarrow D_{obj}\ D_{loc}\ G_{Af}$ (Transition Focus)

(11) a. ACTION-RESULT: e.g., *move* [**Object**] [**Direction**]

b. $GvP_3 \rightarrow G_{Af}\ D_{obj}\ D_{dir}$

Continuation-Style Semantics of Gesture

(12) a. $\mathbf{S}_G \rightarrow (\mathbf{NP})\ \mathbf{GvP}$
$[\![S]\!] = ([\![\mathbf{NP}]\!][\![\mathbf{GvP}]\!])$

b. $\mathbf{GvP}_1 \rightarrow \mathbf{G}_{af}\ \mathbf{D}_{Obj}$
$[\![\mathbf{GvP}_1]\!] = \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{G}_{af}]\!]j')j))$

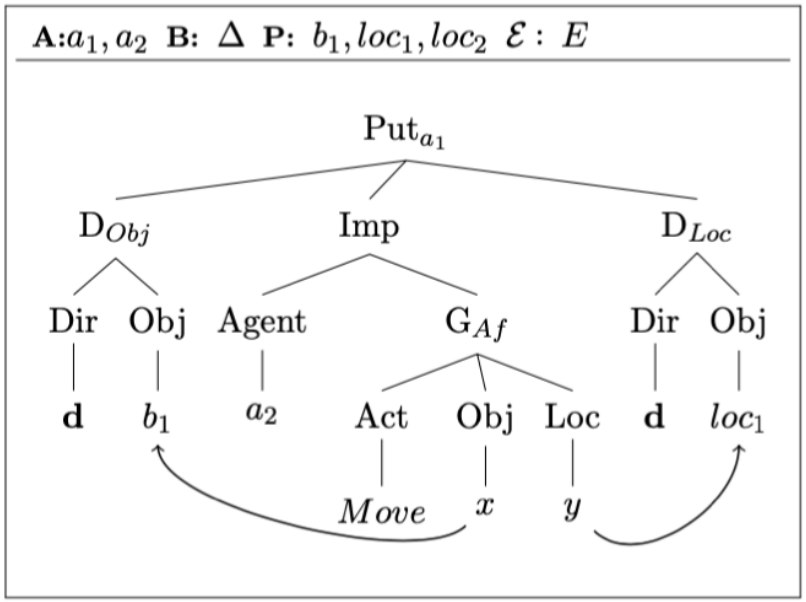
c. $\mathbf{GvP}_2 \rightarrow \mathbf{G}_{af}\ \mathbf{D}_{Obj}\mathbf{D}_{Loc}$
$[\![\mathbf{GvP}_2]\!] = \lambda k.([\![\mathbf{D}_{Loc}]\!]; \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{G}_{af}]\!]j')j)k))$

d. $\mathbf{GvP}_3 \rightarrow \mathbf{G}_{af}\ \mathbf{D}_{Obj}\mathbf{D}_{Dir}$
$[\![\mathbf{GvP}_3]\!] = \lambda k.([\![\mathbf{D}_{Dir}]\!]; \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{G}_{af}]\!]j')j)k))$

Gesture Sequence Denoting Command

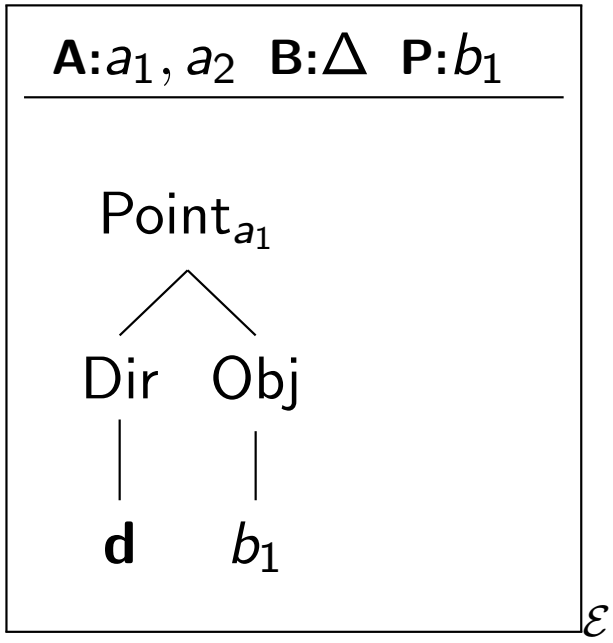
Single Modality (Gesture) Imperative

$$\text{HUMAN}_1\colon\ \mathcal{S} = \text{That}_{t1}$$
$$\mathcal{G} = [\textit{points to purple block}]_{t1}$$
$$\text{HUMAN}_2\colon\ \mathcal{G} = [\textit{makes grab gesture}]$$



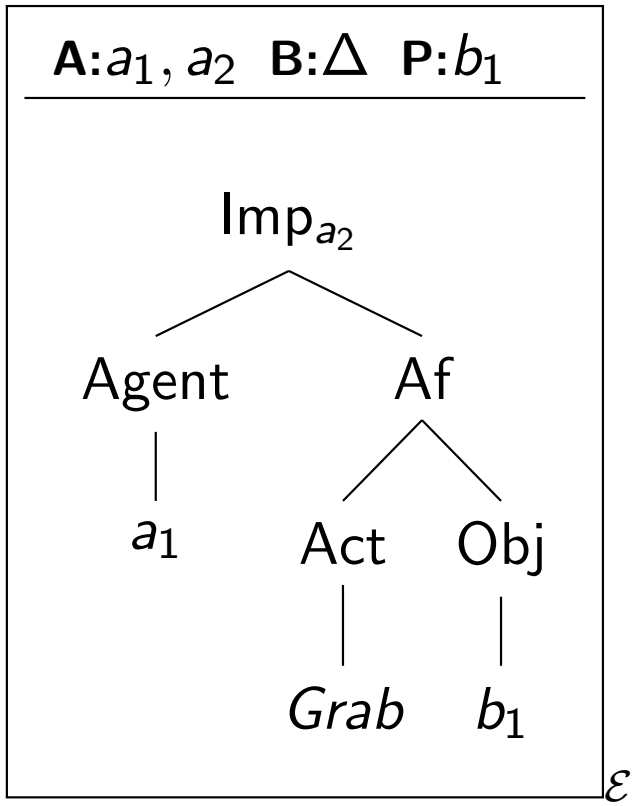
Gesture in the Common Ground

(13)



Gestures denoting Affordances

(14) a. $Grab_g \rightarrow Act \;\; Obj$
b. $Push_g \rightarrow Act \;\; Obj$
c. $Throw_g \rightarrow Act \;\; Obj$

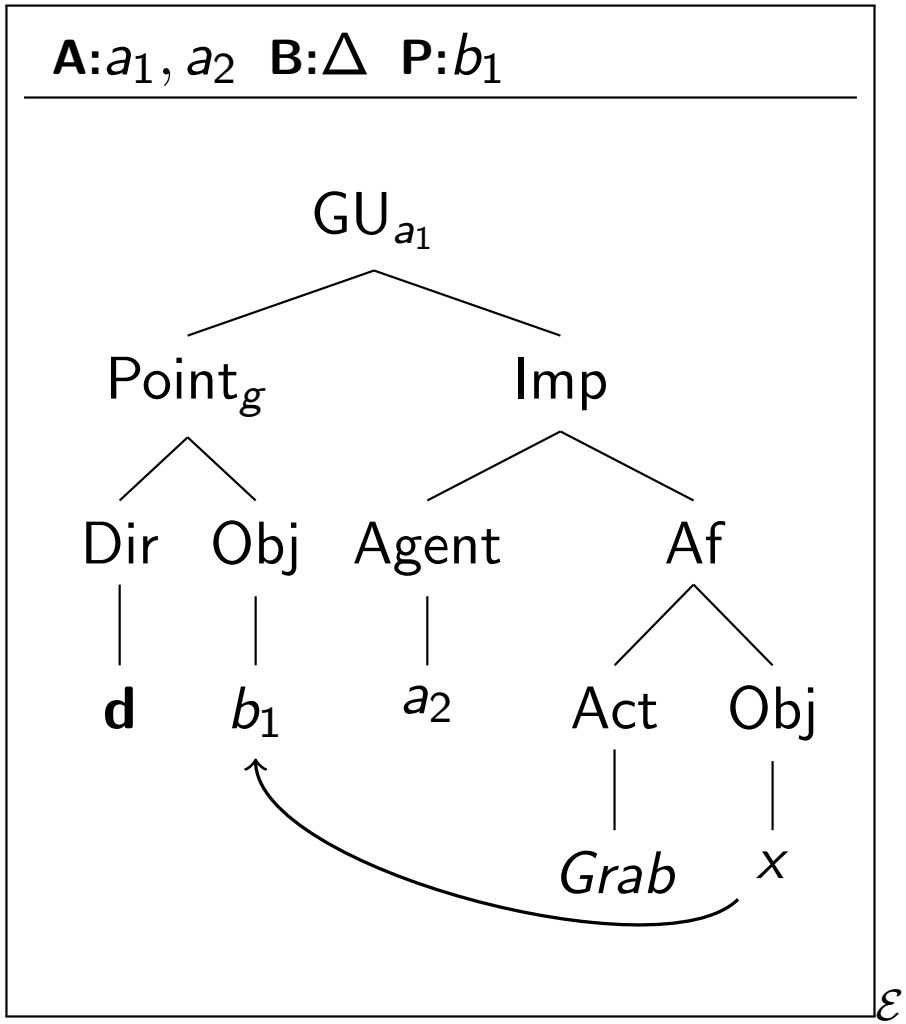


Object Affordances: Gibsonian and Telic

- Objects are antecedents to actions
  - block: Pick me up!, Move me!
  - cup: Pick me up!, Drink what's in me!
  - knife: Pick me up!, Cut that with me!
- Affordances are a subclass of continuations
  - $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(cup)$
    $grab \subseteq \textbf{sel}\ k_{Gib}$
    $drink \subseteq \textbf{sel}\ k_{Telic}$
  - $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(block)$
    $grab \subseteq \textbf{sel}\ k_{Gib}$
    $pick\_up \subseteq \textbf{sel}\ k_{Gib}$
    $move \subseteq \textbf{sel}\ k_{Gib}$

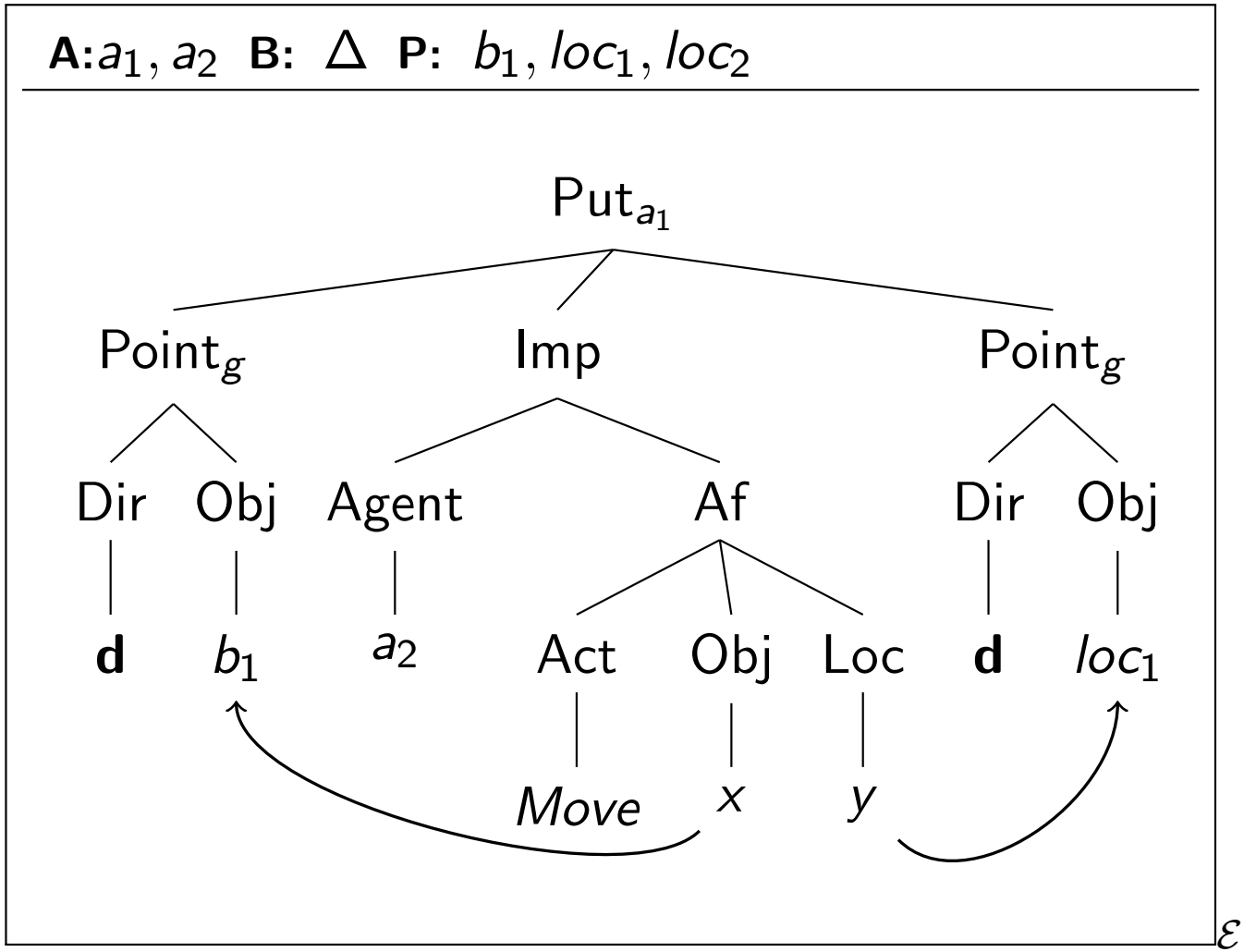
a_1: "That object b_1 grab b_1."

(15)



a_1: "That object b_1 move b_1 to there, the location loc_1."

(16)



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, and an explicit action A.
  - $C_a = \langle S, G, F, Z, A \rangle$
- These modal channels can be aligned or unaligned in the input.

Gestures used in VoxWorld

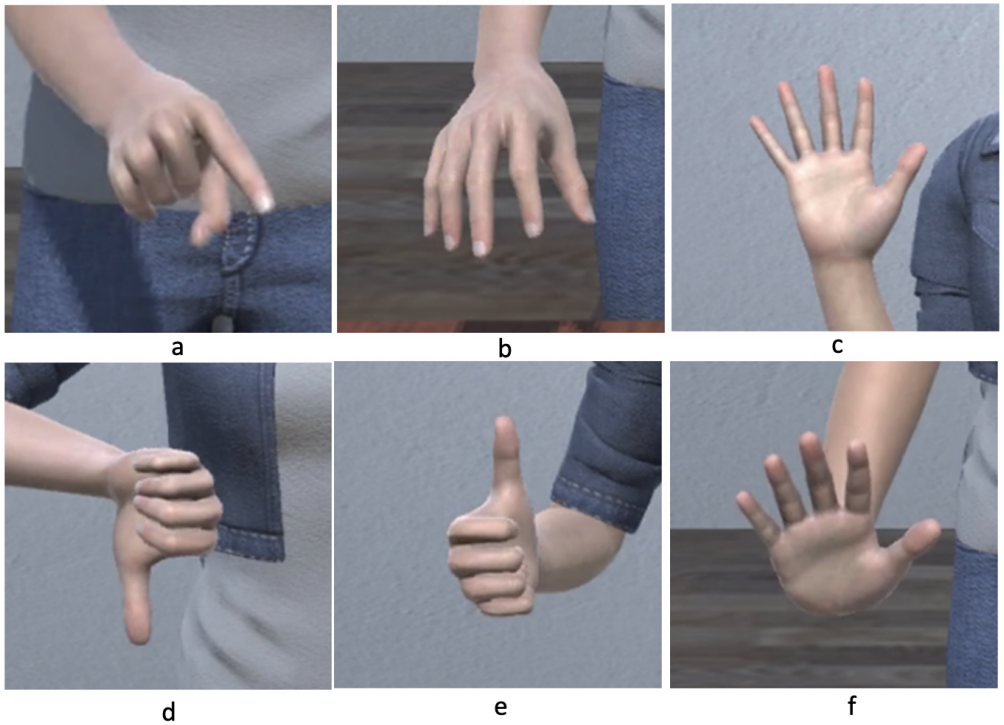
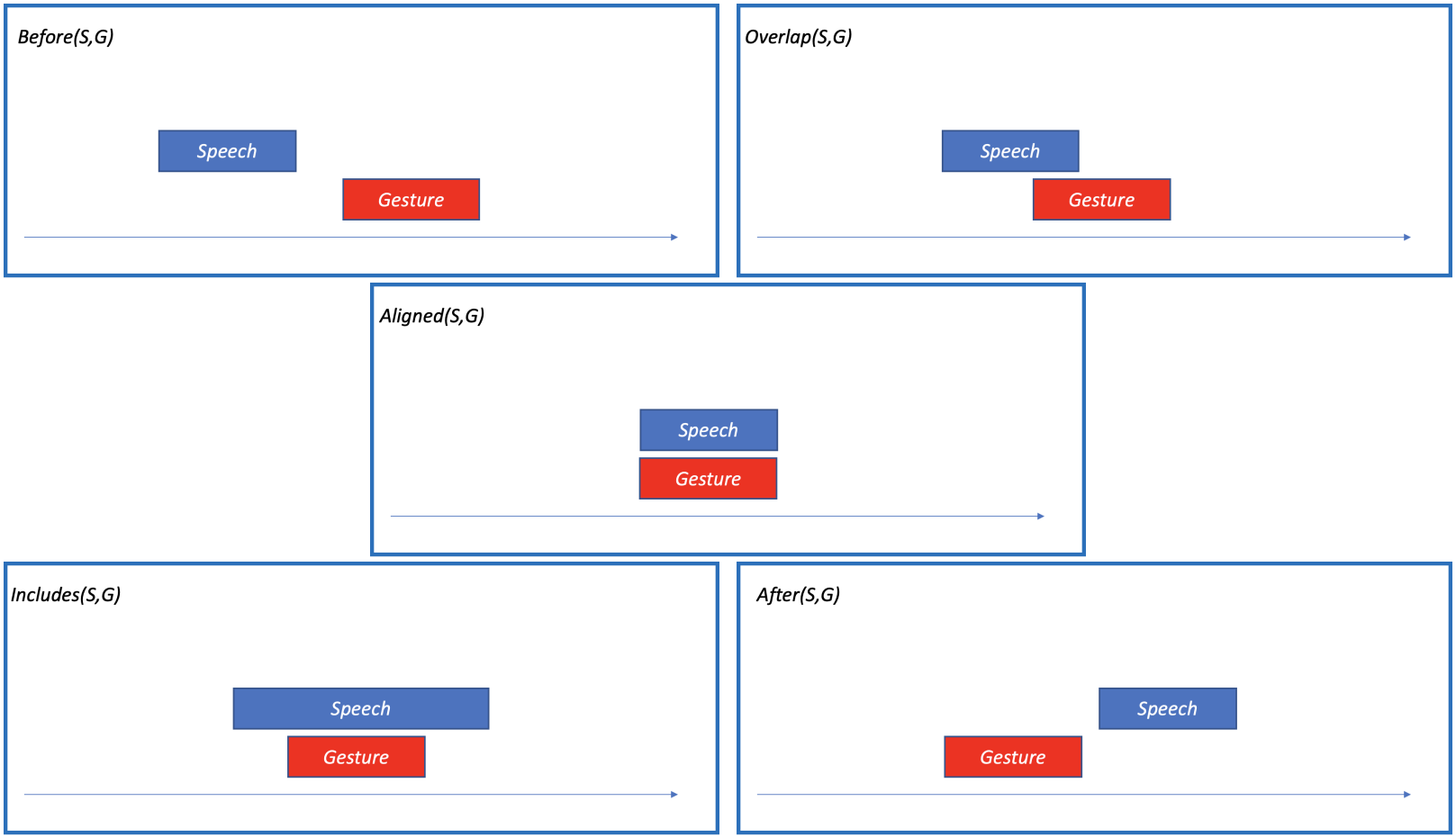


Figure: Some of the gestures generated by VoxWorld: pointing, grab, five, no, yes, push back.

Aligning Speech and Gesture in Dialogue



Aligning Speech and Gesture in Dialogue

A multimodal communicative act, C, consists of a sequence of gesture-language ensembles, (g_i, s_i), where an ensemble is temporally aligned in the common ground:

(17) **Co-gestural Speech Ensemble**: multimodal communication with Gesture, $\mathcal{G}$, and Speech, $\mathcal{S}$:

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

Each modal expresssion carries a continuation, k_g or k_s, and we denote the alignment of these two continuations as $k_s \otimes k_g$:

(18) $\lambda k_s.k_s([[\mathbf{s}]])$
$\lambda k_g.k_g([[\mathbf{g}]])$
$\lambda k_s \otimes k_g.k_s \otimes k_g([[\mathbf{(s,g)}]])$

Common-ground structure for **that** (ensemble) + **grab** (speech)

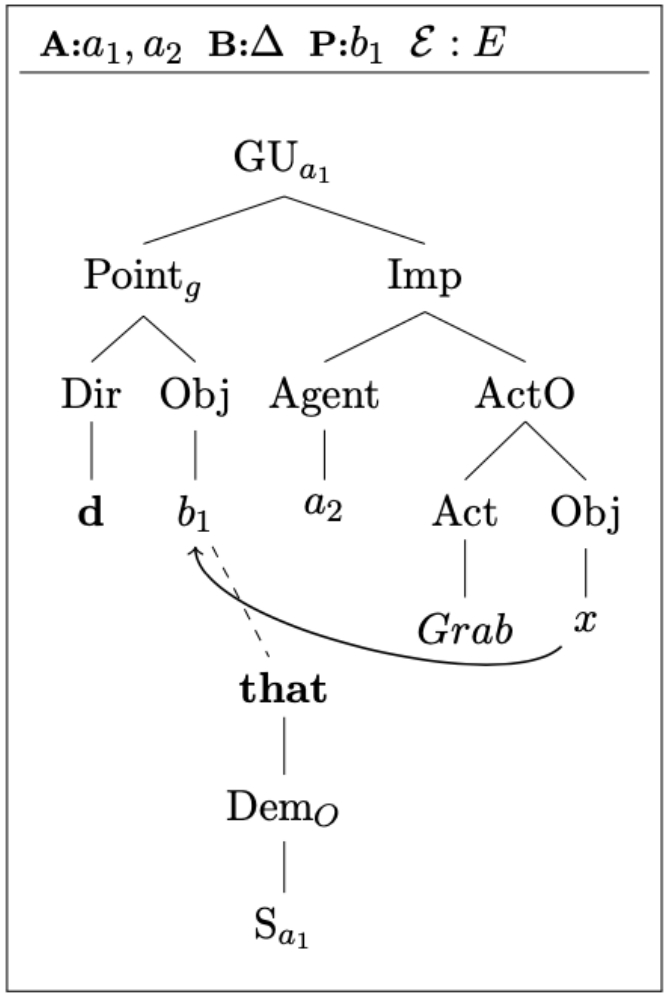


Figure: Common-ground structure for "that" (ensemble) + "grab"

Aligning Speech and Gesture in Dialogue

A multimodal communicative act, C, consists of a sequence of gesture-language ensembles, (g_i, s_i), where an ensemble is temporally aligned in the common ground:

(19) **Co-gestural Speech Ensemble**: multimodal communication with Gesture, $\mathcal{G}$, and Speech, $\mathcal{S}$:

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

Each modal expresssion carries a continuation, k_g or k_s, and we denote the alignment of these two continuations as $k_s \otimes k_g$:

(20) $\lambda k_s.k_s([\![\mathbf{s}]\!])$
$\lambda k_g.k_g([\![\mathbf{g}]\!])$
$\lambda k_s \otimes k_g.k_s \otimes k_g([\![\mathbf{(s,g)}]\!])$

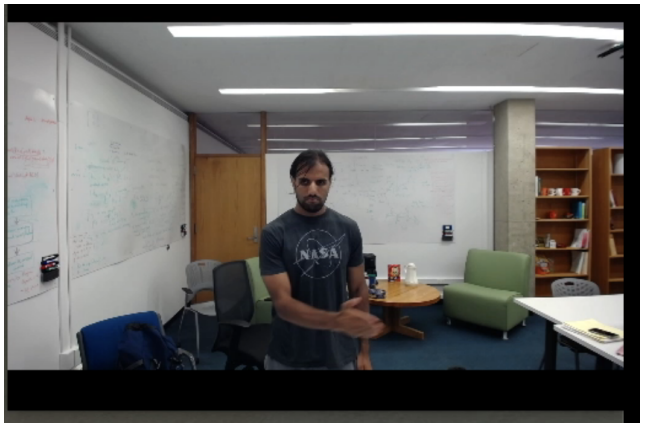
Single Modality (Gesture) Imperative

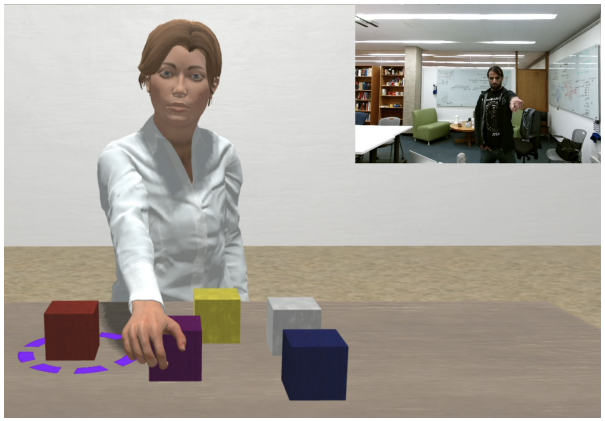
DIANA$_1$: $\mathcal{G} = [\textit{points to the purple block}]_{t1}$
DIANA$_2$: $\mathcal{G} = [\textit{makes move gesture}]_{t2}$
DIANA$_3$: $\mathcal{G} = [\textit{points to the blue block}]_{t3}$

Figure: Gesture generation for performing complex action.

Situated Meaning

Gesture sequence command





Single Modality (Gesture) Imperative

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

Figure: Gesture generation for performing complex action.

- $\mathbf{S}_0$ = on(red,table)
- **CA** = "Grab the red block."
- $\mathbf{S}_1$ = grasp(D,red)
- $\mathbf{S}_0$: $[x, y, \dots]$ - Grab the red block. $\Longrightarrow$ $[b_1, x, y, \dots]$

- $\mathbf{S}_0$ = on(red,table)
- **CA** = "Lift the red block."
- $\mathbf{S}_1$ = lift(D,red)
- **CA** = "Drop it."
- $\mathbf{S}_2$ = drop(D,red)
- $\mathbf{S}_0$: $[x, y, \ldots]$ - Lift the red block $[\emptyset]_{l_1}$. $\Longrightarrow$ $[b_1, l_1, x, y, \ldots]$
- $\mathbf{S}_1$: $[x, y, \ldots]$ - Drop it_{b_1}. $\Longrightarrow$ $[b_1, l_1, x, y, \ldots]$

- $\mathbf{S}_0$ = [on(red,table),on(blue,table)]
- **CA** = "Put the blue block in front of the red block."
- $\mathbf{S}_1$ = in_front(blue,red)
- **CA** = "Put the blue block behind the red block."
- $\mathbf{S}_2$ = behind(blue,red)
- $\mathbf{S}_0$: $[x, y, \dots]$ - Put the blue block in front of the red block$_{l_1}$. $\Longrightarrow$ $[b_1, b_2, l_1, x, y, \dots]$
- $\mathbf{S}_1$: $[b_1, b_2, l_1, x, y, \dots]$ - $[\emptyset]_{c_1}$ Put the blue block behind the red block$_{l_2}$. $\Longrightarrow$ $[b_1, b_2, l_1, l_2, x, y, \dots]$

- $\mathbf{S}_0$ = on(cup,table)
- **CA** = "Grab the cup."
- $\mathbf{S}_1$ = grasp(D,cup,m_1)
- **CA** = "Not like that."
- $\mathbf{S}_2$ = grasp(D,cup,m_2)
- **CA** = {"Yes.", "Slide the cup to the right"}
- $\mathbf{S}_3$ = l_1 := loc(cup); slide(D,cup,l_2)
- $\mathbf{S}_0$: $[x, y, \dots]$ - Grab the cup. $\Longrightarrow$ $[c_1, x, y, \dots]$
- $\mathbf{S}_1$: $[c_1, x, y, \dots]$ - $[\emptyset]_{c_1}$ Not like that$_{m_1}$.
  $\Longrightarrow$ $[c_1, m_1, m_2, x, y, \dots]$
- $\mathbf{S}_2$: $[c_1, x, y, \dots]$ - {Yes., Slide the cup to the right$_{d1}$}.
  $\Longrightarrow$ $[c_1, m_1, m_2, d_1, l_1, l_2, x, y, \dots]$

- $\mathbf{S}_0$ = on(knife,table)
- **CA** = "Grab the knife."
- $\mathbf{S}_1$ = grasp(D,knife,m_1)
- **CA** = "Not like that."
- $\mathbf{S}_1$ = grasp(D,knife,m_2)
- **CA** = {"Yes.", "Lift the knife"}
- $\mathbf{S}_1$ = lift(D,knife)
- $\mathbf{S}_0$: $[x, y, \dots]$ - Grab the knife. $\Longrightarrow$ $[k_1, x, y, \dots]$
- $\mathbf{S}_1$: $[k_1, x, y, \dots]$ - $[\emptyset]_{k_1}$ Not like that$_{m_1}$.
  $\Longrightarrow$ $[k_1, m_1, m_2, x, y, \dots]$
- $\mathbf{S}_2$: $[k_1, x, y, \dots]$ - {Yes., Lift the knife $[\emptyset]_{l_1}$}.
  $\Longrightarrow$ $[k_1, m_1, m_2, l_1, x, y, \dots]$

- $\mathbf{S}_0 =$ on(red,table)
- **CA** $= \mathit{Point}_{red}$
- $\mathbf{S}_1 =$ point(D,red)
- **CA** $= \mathit{Point}_{l_1}$
- $\mathbf{S}_2 =$ move(D,red,l_1)
- **CA** $= \mathit{Point}_{l_2}$
- $\mathbf{S}_3 =$ move(D,red,l_2)
- $\mathbf{S}_0$: $[x, y, \dots]$ - Point_{red}. $\Longrightarrow$ $[b_1, x, y, \dots]$
- $\mathbf{S}_1$: $[b_1, x, y, \dots]$ - Point_{l_1}. $\Longrightarrow$ $[b_1, l_1, x, y, \dots]$
- $\mathbf{S}_2$: $[b_1, l_1, x, y, \dots]$ - Point_{l_2}. $\Longrightarrow$ $[b_1, l_1, l_2, x, y, \dots]$

Affordance-denoting gestures

▸ Play!

Conclusion

- Situation context creates space for common ground
- Conversational acts (multimodal) populate common ground
- Common ground is dynamically updates
  - Implemented using continuation-based semantics
- Real-time negotiation of, e.g., perspective, alignment