Visualizing Meaning: Modeling Communication through Multimodal Simulations

James Pustejovsky
Brandeis University

COLING 2018
Santa Fe, New Mexico
August 21, 2018

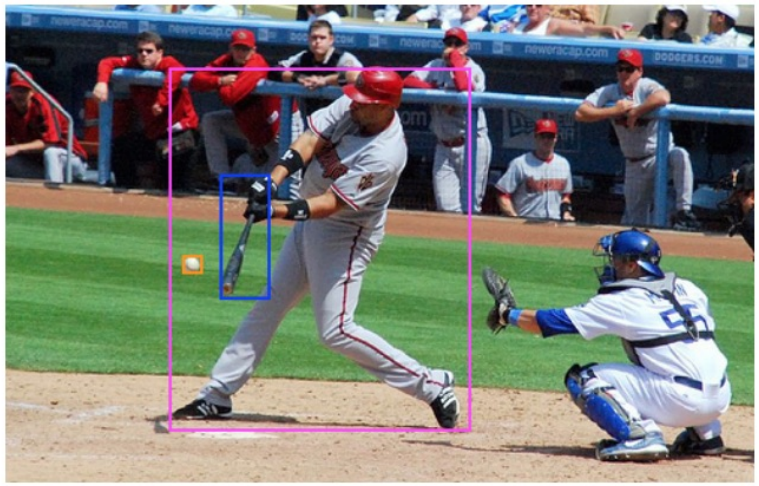
Major Themes of the Talk

1. Human-computer/robot 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.

Semantic Grounding 1/2

Visual Semantic Role Labeling

- Bounding region is identified and semantically labeled
- Region is linked to a linguistic expression in a caption
- Constraints on how visual semantic roles are grounded relative to each other



Semantic Grounding 2/2

Visual Semantic Role Labeling

- Jumping events with semantic role labels
- Im-Situ (Yatskar et al., 2016)

JUMPING			
ROLE	VALUE	ROLE	VALUE
AGENT	BOY	AGENT	BEAR
SOURCE	CLIFF	SOURCE	ICEBERG
OBSTACLE	-	OBSTACLE	WATER
DESTINATION	WATER	DESTINATION	ICEBERG
PLACE	LAKE	PLACE	OUTDOOR

Semantic grounding goes only so far . . .

- Understanding language is not enough;
- Situated grounding entails knowledge of situation and contextual entities.

HEY SIRI![1]

[1] Example thanks to Bruce Draper.

Our Approach

- A framework for studying interactions and communication between agents engaged in a shared goal or task (peer-to-peer communication).
- When two or more people are engaged in dialogue during a shared experience, they share a common ground, which facilitates situated communication.
- By studying the constitution and configuration of common ground in situated communication, we can better understand the emergence of decontextualized reference in communicative acts, where there is no common ground.

Mental Simulation and Mind Reading

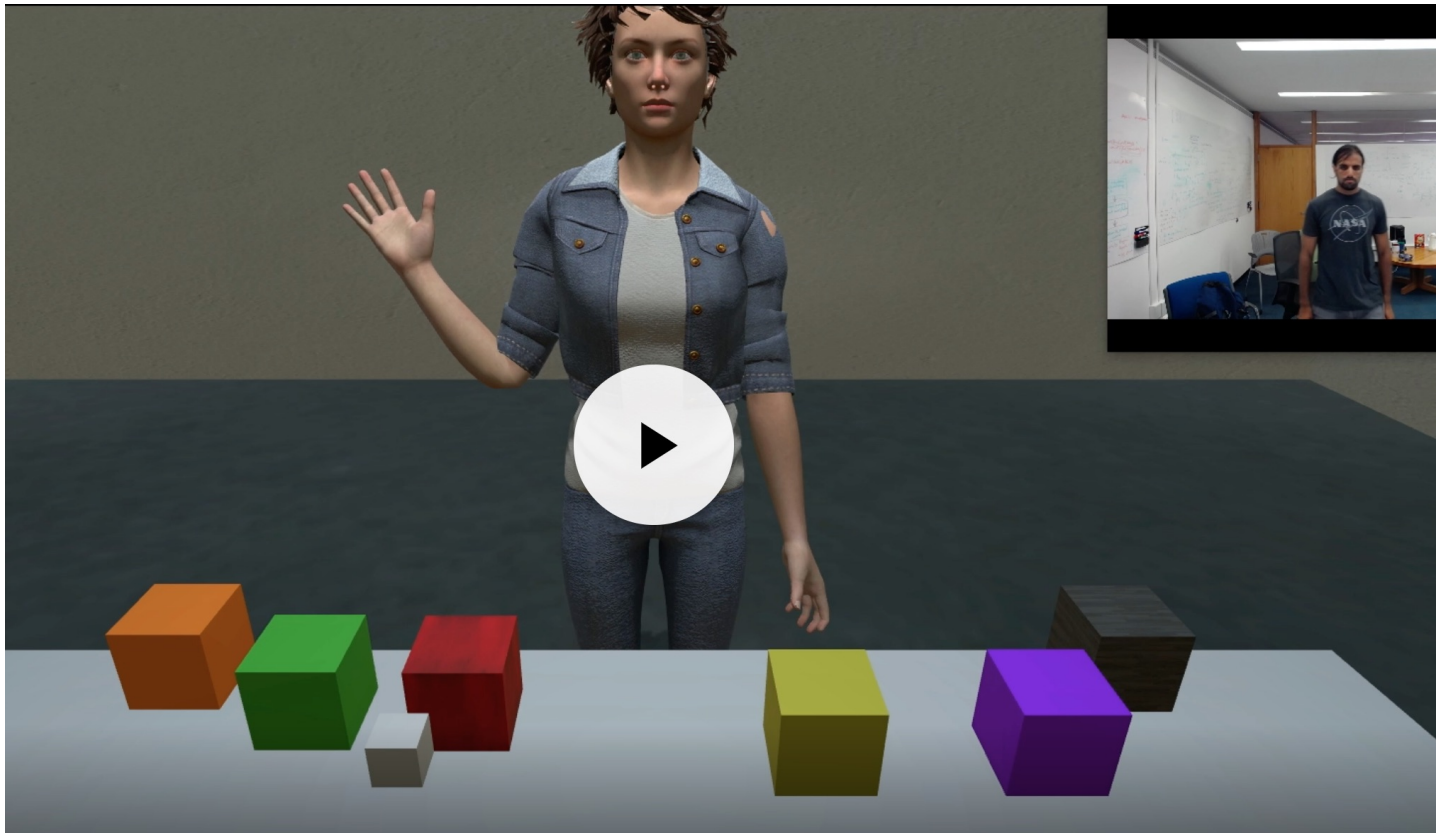
- Mental Simulations
  Graesser et al (1994), Barselou (1999), Zwaan and Radvansky (1998), Zwaan and Pecher (2012)
- Embodiment:
  Johnson (1987), Lakoff (1987), Varela et al. (1991), Clark (1997), Lakoff and Johnson (1999), Gibbs (2005)
- Mirror Neuron Hypothesis:
  Rizzolatti and Fadiga (1999), Rizzolatti and Arbib (1998), Arbib (2004)
- Simulation Semantics
  Goldman (1989), Feldman et al (2003), Goldman (2006), Feldman (2010), Bergen (2012), Evans (2013)

Multimodal Simulation

- 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;

Situated Grounding

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



Link

Areas Contributing to this Effort 1/2

- Multimodal parsing and generation: Johnston et al. (2005); Kopp et al. (2006); Vilhjàlmsson et al. (2007)
- Human Robot Interaction and Communication (HRI): Misra et al. (2015); She and Chai (2016); Scheutz et al. (2017); Henry et al. (2017); Nirenburg et al. (2018)
- Task-oriented dialogue and joint activities: Traum (2009); Gravano and Hirschberg (2011); Swartout et al. (2006); Marge et al. (2017)
- Semantic grounding of text to images and video: Chang et al. (2015); Lazaridou et al. (2015); Bruni et al. (2014), Yatskar et al. (2016)
- Gesture semantics and learning: Lascarides and Stone (2009); Clair et al. (2010); Anastasiou (2012); Matuszek et al (2014)

Areas Contributing to this Effort 2/2

- Visual reasoning with simulations: Forbus et al. (1991); Lathrop and Laird (2007); Seo et al. (2015); Lin and Parikh (2015); Goyal et al. (2018)
- Linking language to objects and actions: Liu and Chai (2015); Tellex et al. (2014); Artzi and Zettlemoyer (2013)
- Commonsense reasoning in virtual environments: Lugrin and Cavazza (2007); Wilks (2006); Flotyński and Walczak (2015)
- Learning by Communication with Robots: Cakmak and Thomaz (2012); She and Chai (2017)
- Logics of active perception: Musto and Konolige (1993); Bell and Huang (1998); Wooldridge and Lomuscio (1999)

Wordseye Coyne and Sproat (2001)

- Automatically converts text into representative 3D scenes.
- Relies on a large database of 3D models and poses to depict entities and actions
- Every 3D model can have associated shape displacements, spatial tags, and functional properties.

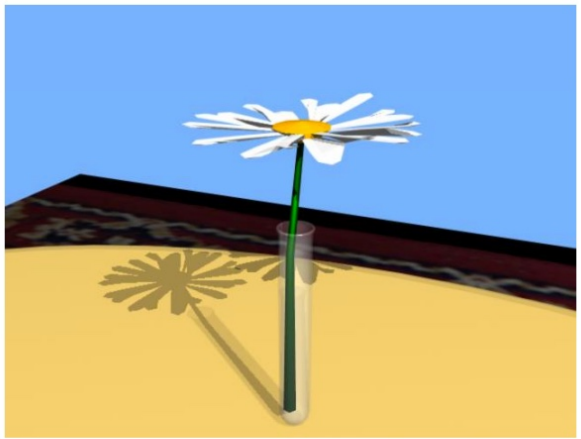


Figure 7: *The daisy is in the test tube.*

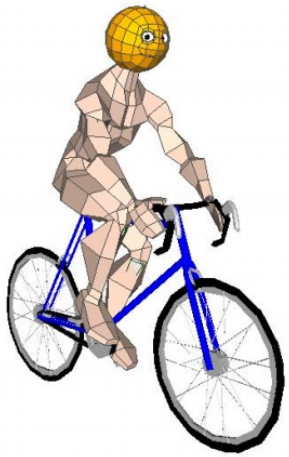
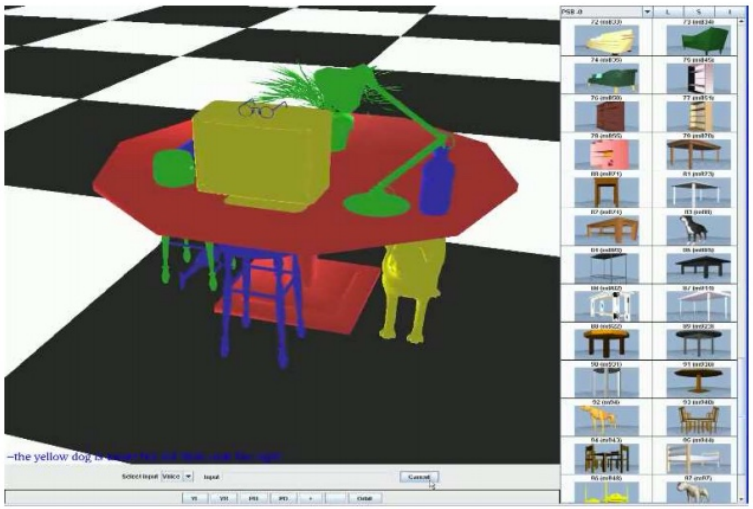


Figure 9: Usage pose for a 10-speed.

Automatic 3D scene generation Seversky and Yin (2006)

- The system contains a database of polygon mesh models representing various types of objects.
- composes scenes consisting of objects from the Princeton Shape Benchmark model database 2



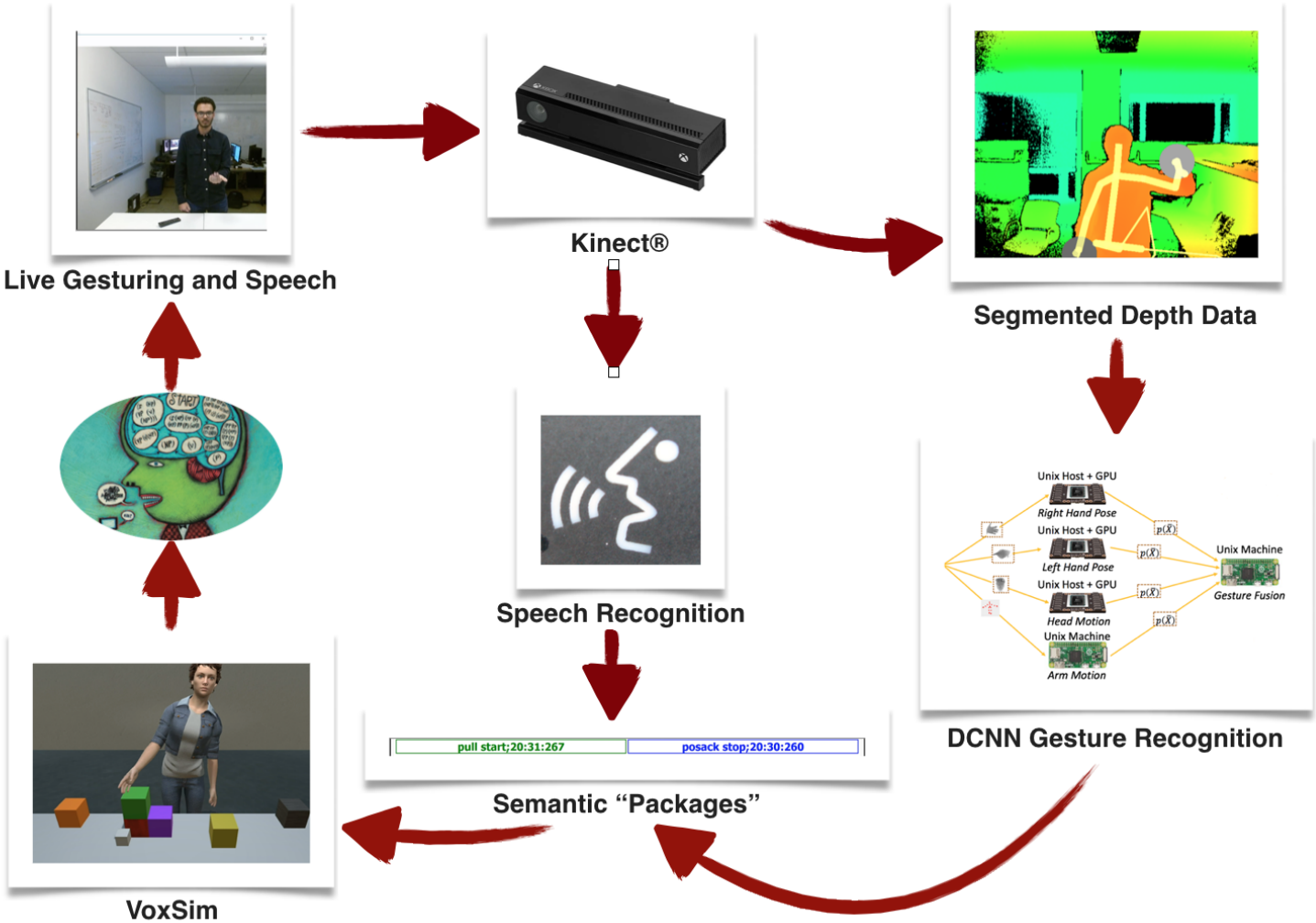
DARPA's Hallmarks of Communication

- Interaction has mechanisms to move the conversation forward (Asher and Gillies, 2003; Johnston, 2009)
- Makes appropriate use of multiple modalities (Arbib and Rizzolatti, 1996; Arbib, 2008)
- Each interlocutor can steer the course of the interaction (Hobbs and Evans, 1980)
- Both parties can clearly reference items in the interaction based on their respective frames of reference (Ligozat, 1993; Zimmermann and Freksa, 1996; Wooldridge and Lomuscio, 1999)
- Both parties can demonstrate knowledge of the changing situation (Ziemke and Sharkey, 2001)

DARPA's Hallmarks of Communication

- Makes appropriate use of multiple modalities
  Machine vision, language, gesture
- Interaction has mechanisms to move the conversation forward
  Dialogue Manager PDA
- Each interlocutor can steer the course of the interaction
  Human directs avatar towards goals; meanwhile avatar asks for clarification and teaches human what she understands
- Both parties can clearly reference items in the interaction based on their respective frames of reference
  Ensemble reference using deixis, language, and frame of reference
- Both parties can demonstrate knowledge of the changing situation
  Visualizing the epistemic state of the agents (EpiSim)

VoxWorld Architecture



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

Dynamic Interval Temporal Logic 1/2

Pustejovsky and Moszkowicz (2011)

- Event structure is integrated with first-order dynamic logic;
- Represents the attribute modified in the course of the event (the location of the moving entity, the extent of a created or destroyed entity, etc.);
- A complex event can be modeled as a sequence of frames;
- To adequately model events, the representation should track the change in the assignment of values to attributes in the course of the event.
- This includes making explicit any *predicative opposition* denoted by the verb:
  - *die* encodes going from $\neg dead(e_1, x)$ to $dead(e_2, x)$;
  - *arrive* encodes going from $\neg loc\_at(e_1, x, y)$ to $loc\_at(e_2, x, y)$.

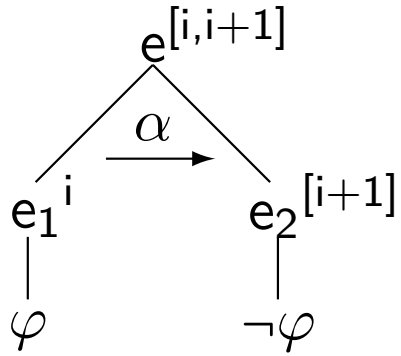
Dynamic Interval Temporal Logic 2/2

Pustejovsky and Moszkowicz (2011)

Two Primitive Event Types

State

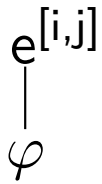
Simple Transition



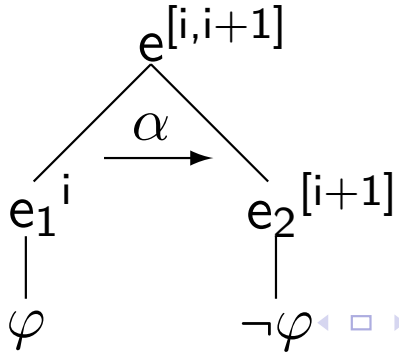
Derived Vendler Event Types

a. State

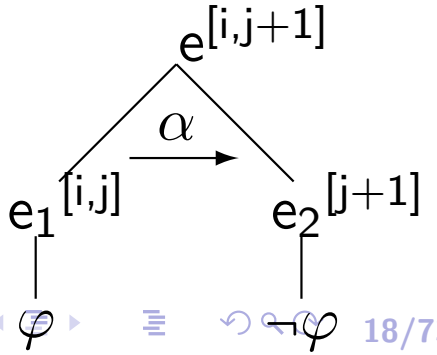
b. Process



c. Achievement



d. Accomplishment



Visual Object Concept Modeling Language (VoxML)

Pustejovsky and Krishnaswamy (2014, 2016)

- Modeling language for constructing 3D visualizations of concepts denoted by natural language expressions
- Used as the platform for creating *multimodal semantic simulations*
- Encodes dynamic semantics of nominals (objects) and events (programs) and adjectives (object properties)
- Platform independent framework for encoding and visualizing linguistic knowledge.

Visual Object Concept Modeling Language (VoxML)

Pustejovsky and Krishnaswamy (2014, 2016)

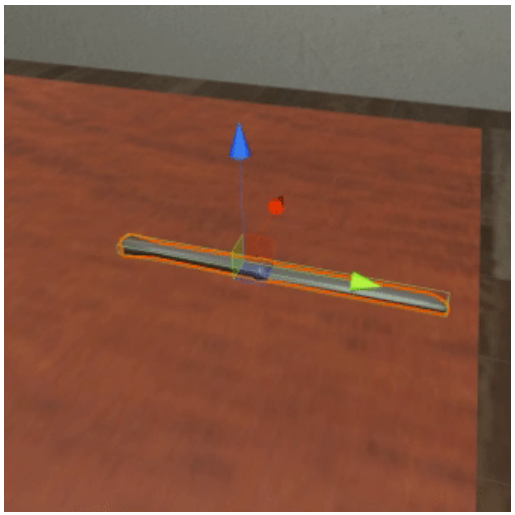
- Modeling and annotation language for "voxemes"
  - Visual instantiation of a lexeme
  - Lexemes may have many visual representation
- Scaffold for mapping from lexical information to simulated objects and operationalized behaviors
- Encodes afforded behaviors for each object
  - Gibsonian: afforded by object structure (Gibson,1977,1979)
    - grasp, move, lift, etc.
  - Telic: goal-directed, purpose-driven (Pustejovsky, 1995)
    - drink from, read, etc.

Visual Object Concept (Voxeme)

- Object Geometry Structure:
  Formal object characteristics in R3 space
- Habitat: Embodied and embedded object:
  Orientation
  Situated context
  Scaling
- Affordance Structure:
  What can one do to it
  What can one do with it
  What does it enable
- Voxicon: library of voxemes

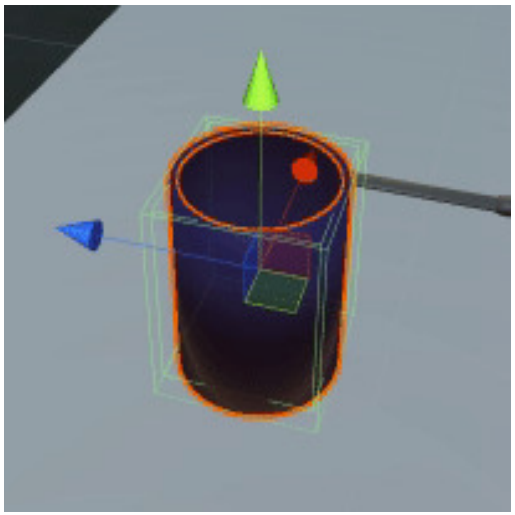
VoxML - knife

$$\begin{bmatrix} \textbf{knife} \\ \text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{knife} \\ \text{TYPE} = \textbf{physobj, artifact} \end{bmatrix} \\ \text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{rectangular\_prism[1]} \\ \text{COMPONENTS} = \textbf{handle[2], blade} \\ \text{CONCAVITY} = \textbf{flat} \\ \text{ROTATSYM} = nil \\ \text{REFLECTSYM} = \{XY\} \end{bmatrix} \\ \text{HABITAT} = \begin{bmatrix} \text{INTR} = {}_{[3]}\begin{bmatrix} \text{CONSTR} = \{X > Y, X \gg Z\} \\ \text{FRONT} = front(+X) \end{bmatrix} \\ \text{EXTR} = \ldots \end{bmatrix} \\ \text{AFFORD\_STR} = \begin{bmatrix} \text{A}_1 = H_{[3]} \rightarrow [grasp(x, [1])] \\ \text{A}_2 = H_{[3]} \rightarrow [grasp(x, [2]) \rightarrow grasp(x, [1])] \end{bmatrix} \\ \text{EMBODIMENT} = \begin{bmatrix} \text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true} \end{bmatrix} \end{bmatrix}$$



VoxML - cup

$$
\left[\begin{array}{l}
\textbf{cup} \\
\text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj, artifact}\end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{cylindroid[1]} \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATSYM} = \{Y\} \\ \text{REFLECTSYM} = \{XY, YZ\}\end{array}\right] \\
\text{HABITAT} = \left[\begin{array}{l}\text{INTR} = {}_{[2]}\left[\begin{array}{l}\text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y)\end{array}\right] \\ \text{EXTR} = {}_{[3]}\left[\text{UP} = align(Y, \mathcal{E}_{\perp Y})\right]\end{array}\right] \\
\text{AFFORD\_STR} = \left[\begin{array}{l}\text{A}_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x) \\ \text{A}_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x) \\ \text{A}_3 = H_{[2]} \rightarrow [grasp(x, [1])] \\ \text{A}_4 = H_{[3]} \rightarrow [roll(x, [1])]\end{array}\right] \\
\text{EMBODIMENT} = \left[\begin{array}{l}\text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true}\end{array}\right]
\end{array}\right]
$$

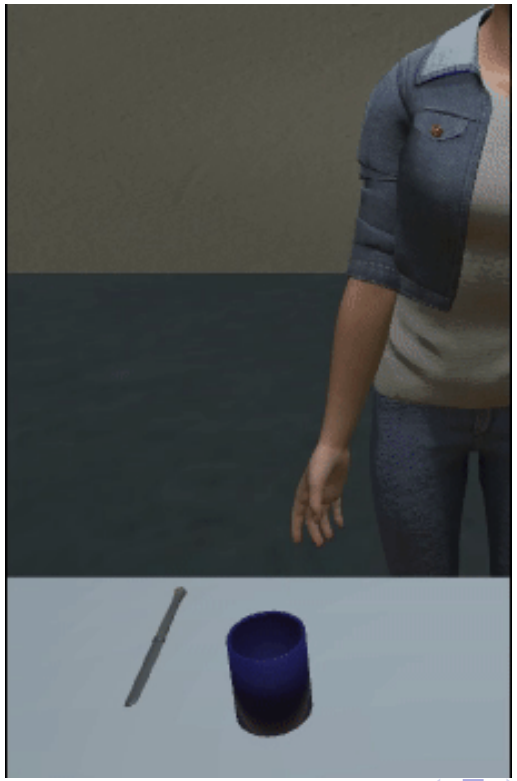


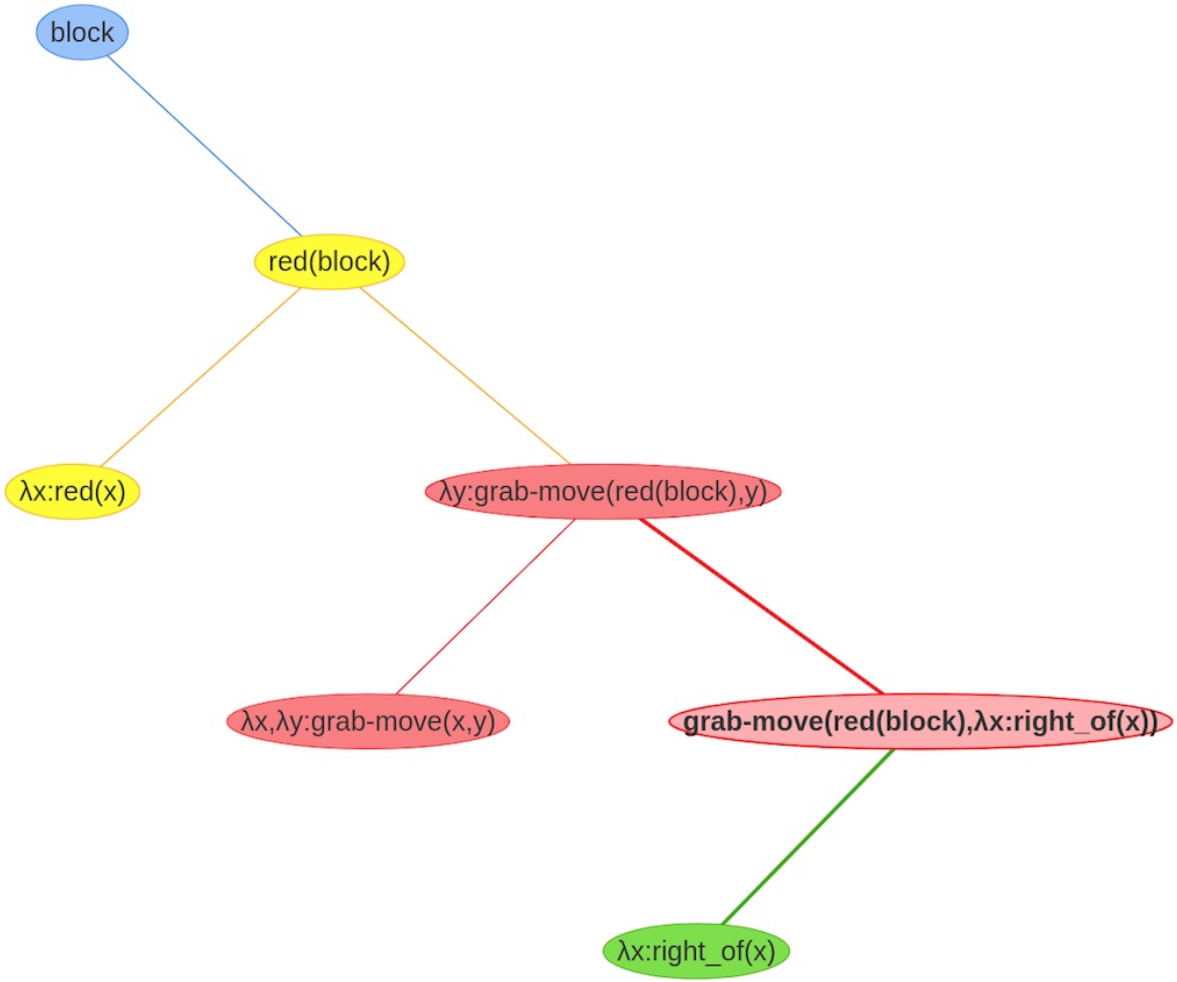
VoxML - grasp

$$\left[\begin{array}{l} \textbf{grasp} \\ \text{LEX} = \left[\begin{array}{l} \text{PRED} = \textbf{grasp} \\ \text{TYPE} = \textbf{transition\_event} \end{array}\right] \\ \text{TYPE} = \left[\begin{array}{l} \text{HEAD} = \textbf{transition} \\ \text{ARGS} = \left[\begin{array}{l} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \end{array}\right] \\ \text{BODY} = \left[\text{E}_1 = grasp(x, y)\right] \end{array}\right] \end{array}\right]$$

VoxML - grasp cup

- Continuation-style semantics for composition
- Used within conventional sentence structures and between sentences in discourse




block
red(block)
λx:red(x)
λy:grab-move(red(block),y)
λx,λy:grab-move(x,y)
grab-move(red(block),λx:right_of(x))
λx:right_of(x)

Modeling Action in VoxML

- 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.

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

Common Ground Structure

(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 conspecifics embody in the communication.

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

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.

Communicative Acts in Discourse 1/2

- Atomic (Warn, Invite, Greet)
  - Directly interpretable acts which reference the agents only
  - Hello, goodbye, watch out!
- Complex (Inform, Question, Command, Promise)
  - Operations over embedded expressions, which are then interpretable.
  - That is a banana.
  - Is this the one?
  - Move that block.
  - I promise I will come.

Communicative Acts in Discourse 2/2

- 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_ACT:

  1. $C\_ACT_a = (G)$
  2. $C\_ACT_a = (S)$
  3. $C\_ACT_a = (S, G)$

- For each of these configurations, we ask which communicative acts are expressible.

Communicating through Simulations

- Formal Models Provide a Reasoning Platform for the Computer
  - Minimal finite model enables inference; but ...
- They are not an effective medium for communicating with humans
  - Communication is facilitated through semiotic structures that are shared and understood by both partners.
- Multimodal semantic simulations are embodied representations of situations and events
  - Image schemas and visualizations of actions are core human competencies

Situated Communication - Under the hood 1/5

▸ Link

Situated Communication - Under the hood 1/5

- Speech recognition system
- Incremental parser
- Semantic interpretation of parsed output
- Referential grounding to something in context

Situated Communication - Under the hood 2/5

▸ Link

Situated Communication - Under the hood 2/5

- DCNN Gesture recognition from depth data (CSU)
- Incremental parser on DCNN output
- Contextual interpretation of received gesture signal
- Referential grounding to something in context

Events as Described by Gesture

Kendon (2004), Lascarides and Stone (2009)

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

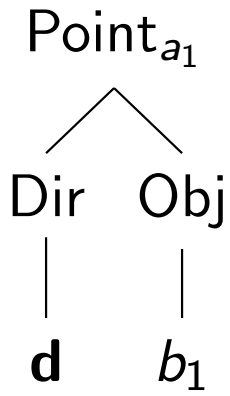
The stroke is the content-baring phase, $\mathbf{d}$, and in a pointing gesture, will convey the deictic orientational information.

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

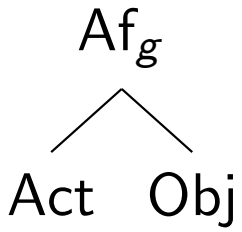
Gesture Grammar

Pustejovsky (2018)

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

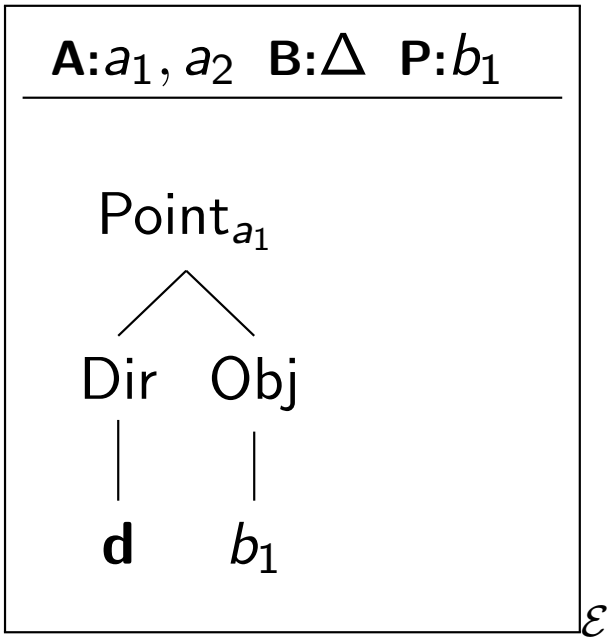


b. **Affordance**: $Af_g \rightarrow Act \;\; Obj$



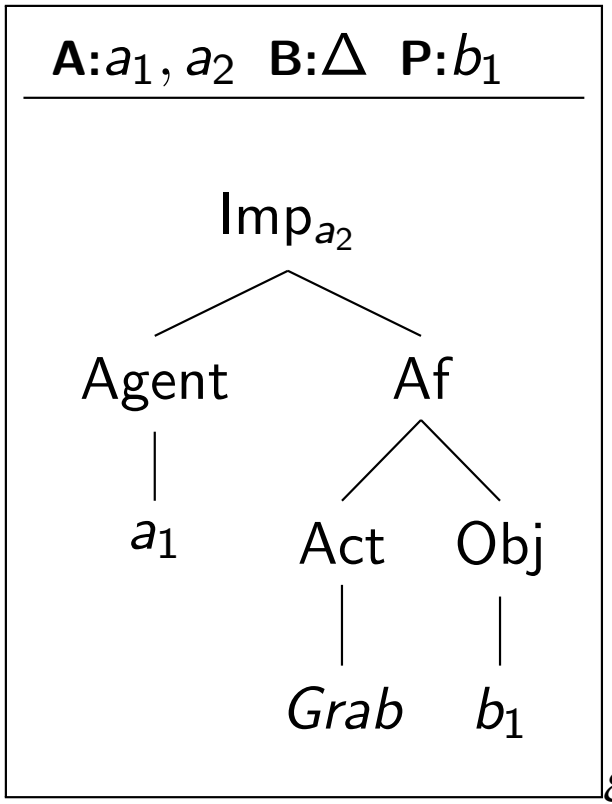
Gesture in the Common Ground

(4)



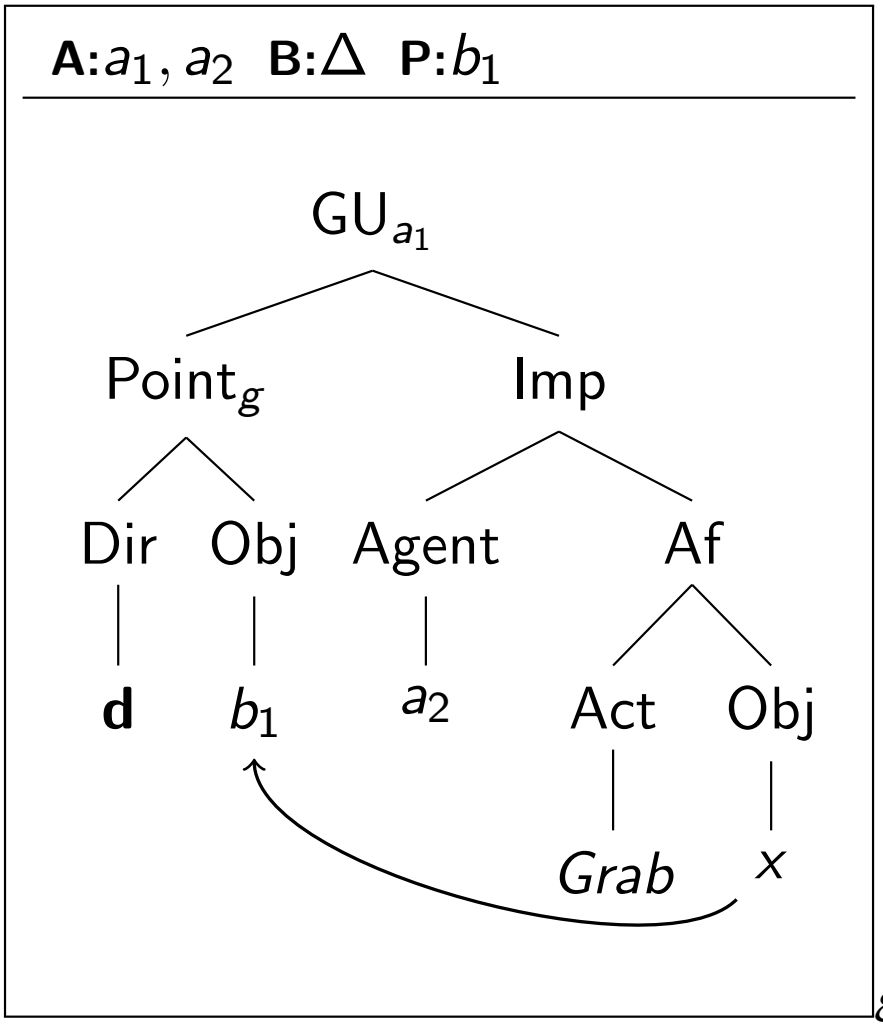
Gestures denoting Affordances

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



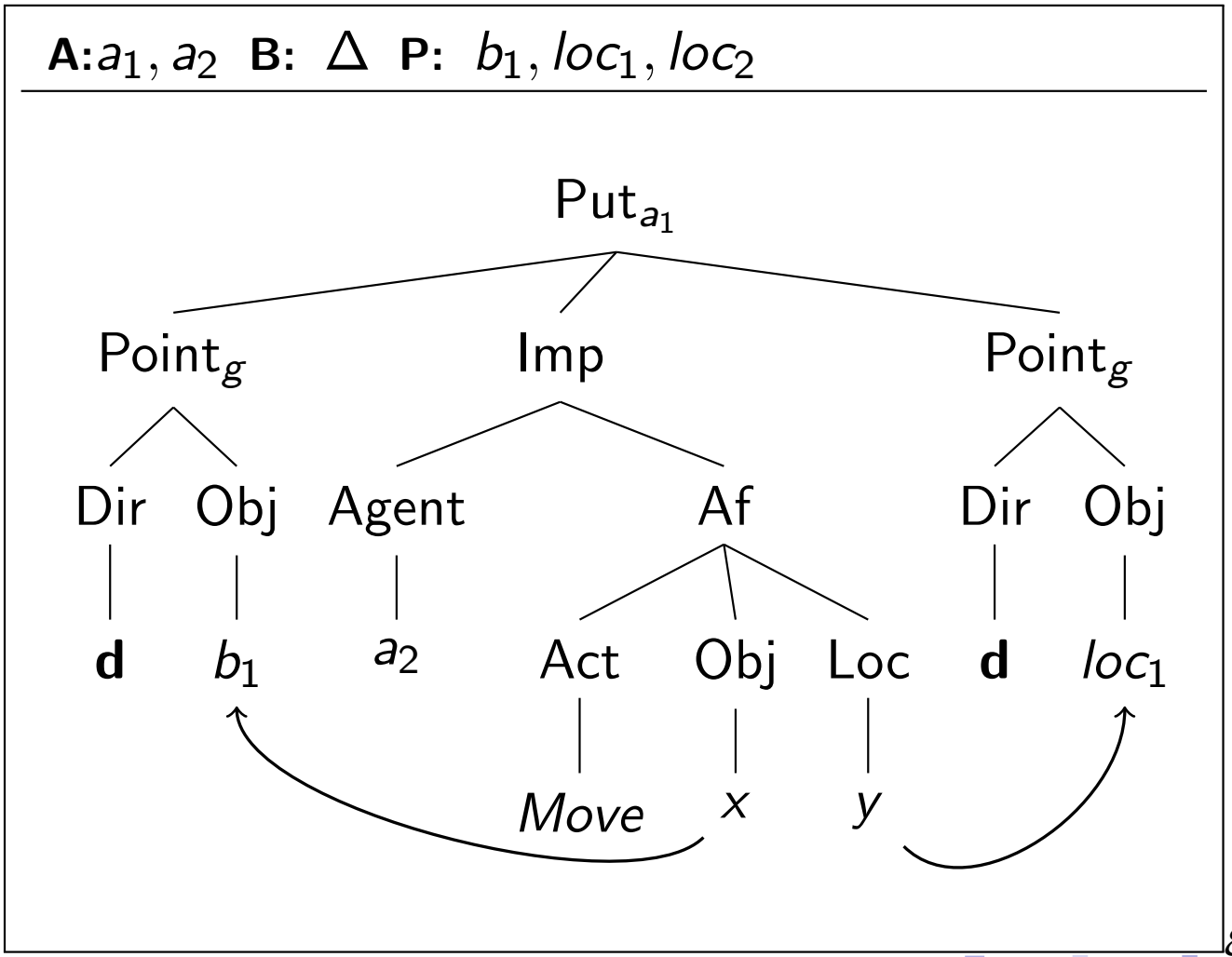
a_1: "That object b_1 grab b_1."

(6)



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

(7)



Situated Communication - Under the hood 3/5

Communication by Ensemble

▸ Link

Situated Communication - Under the hood 3/5

Communication by Ensemble

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:

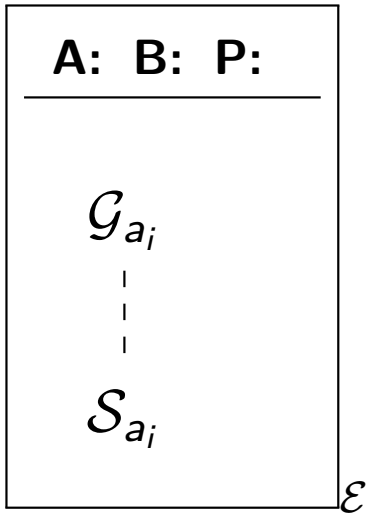
(8) $C = (g_1, s_1); \ldots; (g_i, s_i); \ldots; (g_n, s_n)$.

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

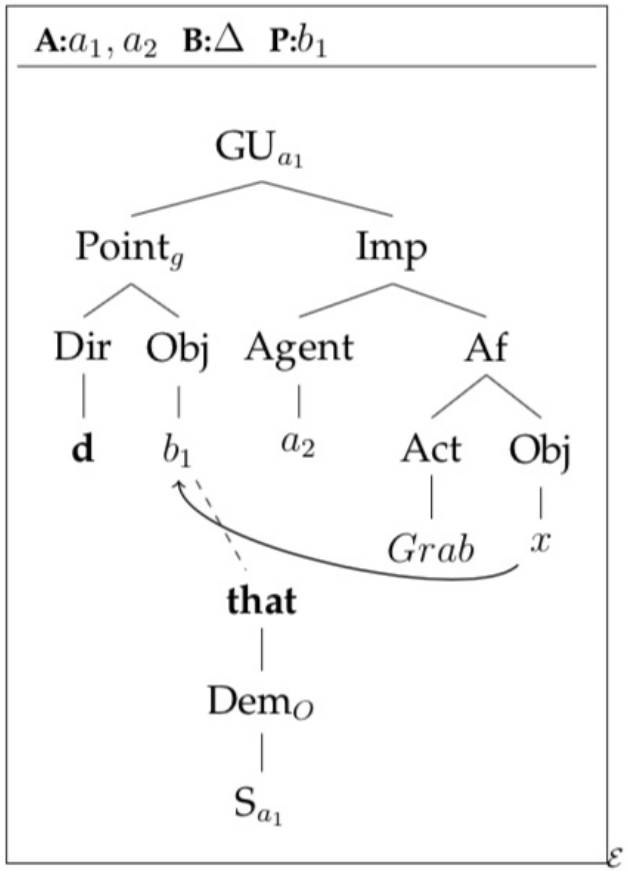
$$\begin{bmatrix} \mathcal{G} & g_1 & g_i & g_n \\ \mathcal{S} & s_1 & s_i & s_n \end{bmatrix}$$

Ensembles in the Common Ground

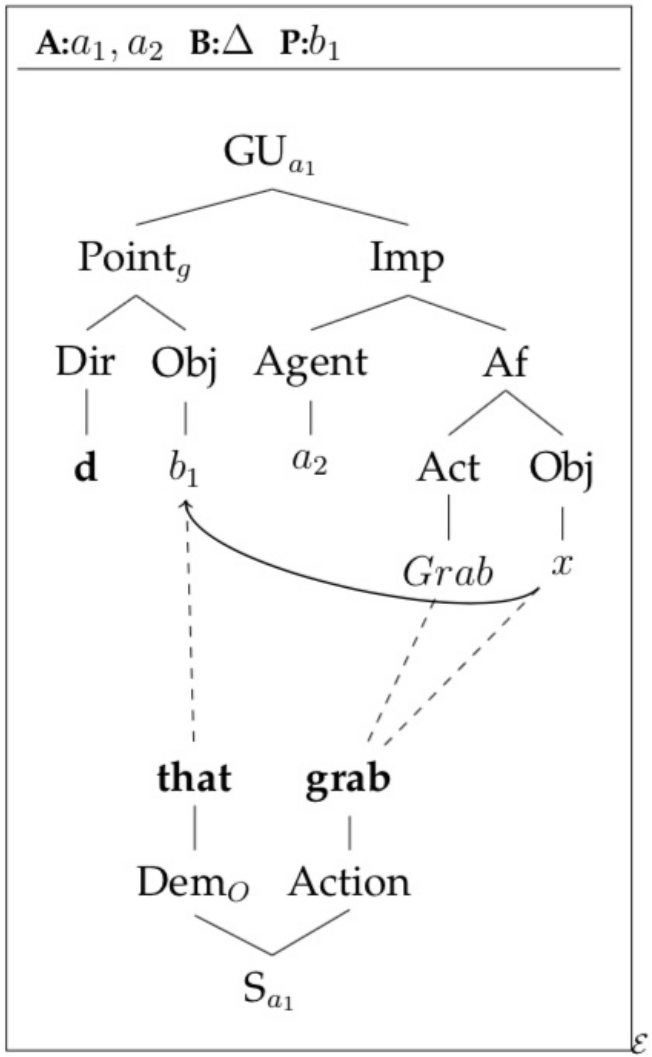
(10)



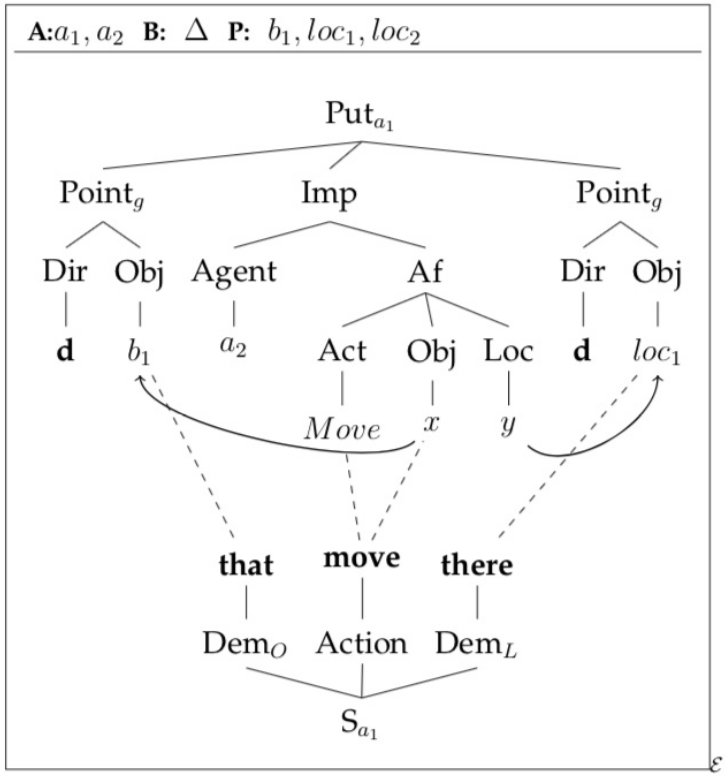
a_1: "**That** object b_1 grab b_1."



a_1: "**That** object b_1 **grab** b_1."



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



Situated Communication - Under the hood 4/5

Establishing Frame of Reference

▸ Link

Situated Communication - Under the hood 4/5

- Gesture is directly grounding to a orientation
- Human adopts the avatar's frame of reference for next command

Situated Communication - Under the hood 5/5

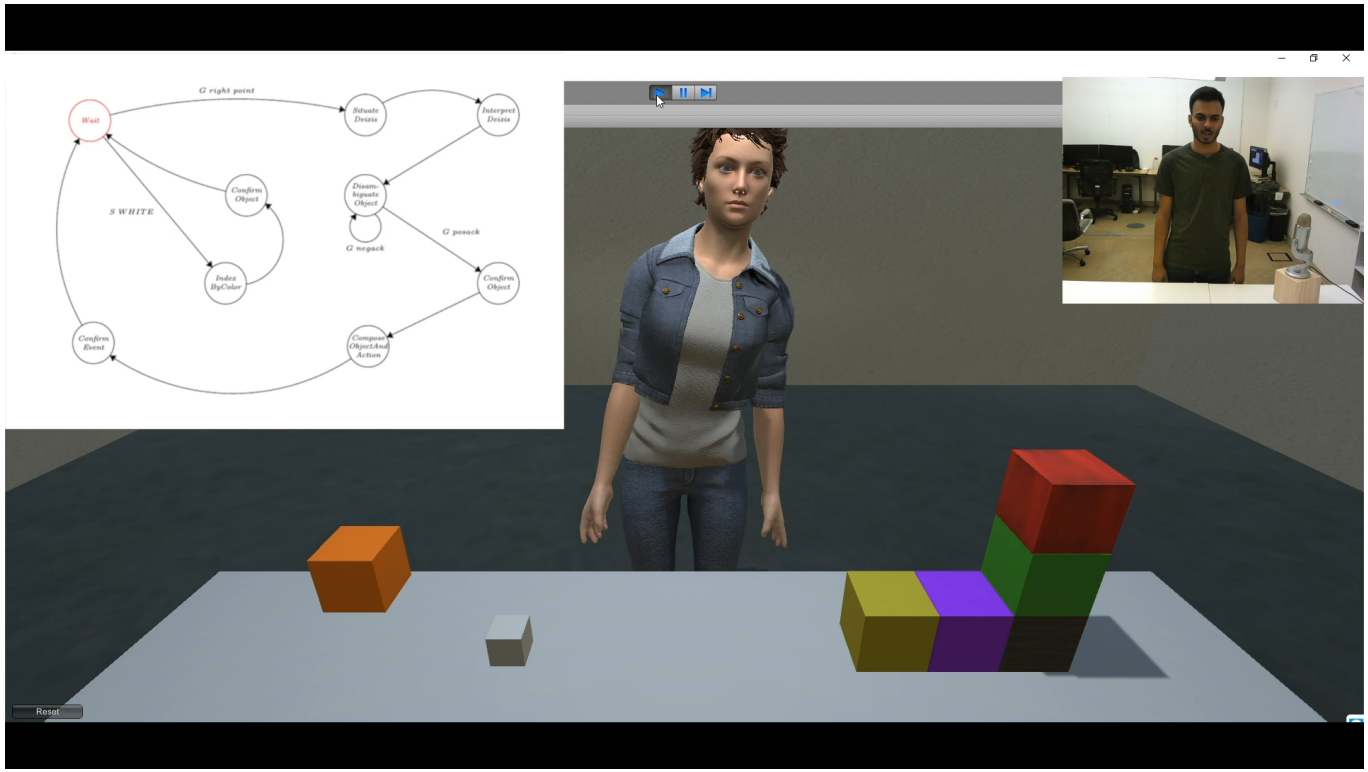
EpiSim: Epistemic State and Update

- Visualized what the agent knows and sees in the situated context;
- Public Announcement Logic
- Public Perception Logic

Dynamics of Communicative Interactions

Tracking moves in the Dialogue

- Dialogue Manager PDA



▸ Link

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.
- $\phi ::= \top \mid p \mid \neg\phi \mid \phi_1 \wedge \phi_2 \mid [\alpha]_\sigma \phi \mid [!\phi_1]_\sigma \phi_2$
- $\alpha ::= a \mid ?\phi \mid \alpha_1; \alpha_2 \mid \alpha_1 \cup \alpha_2 \mid \alpha^*$

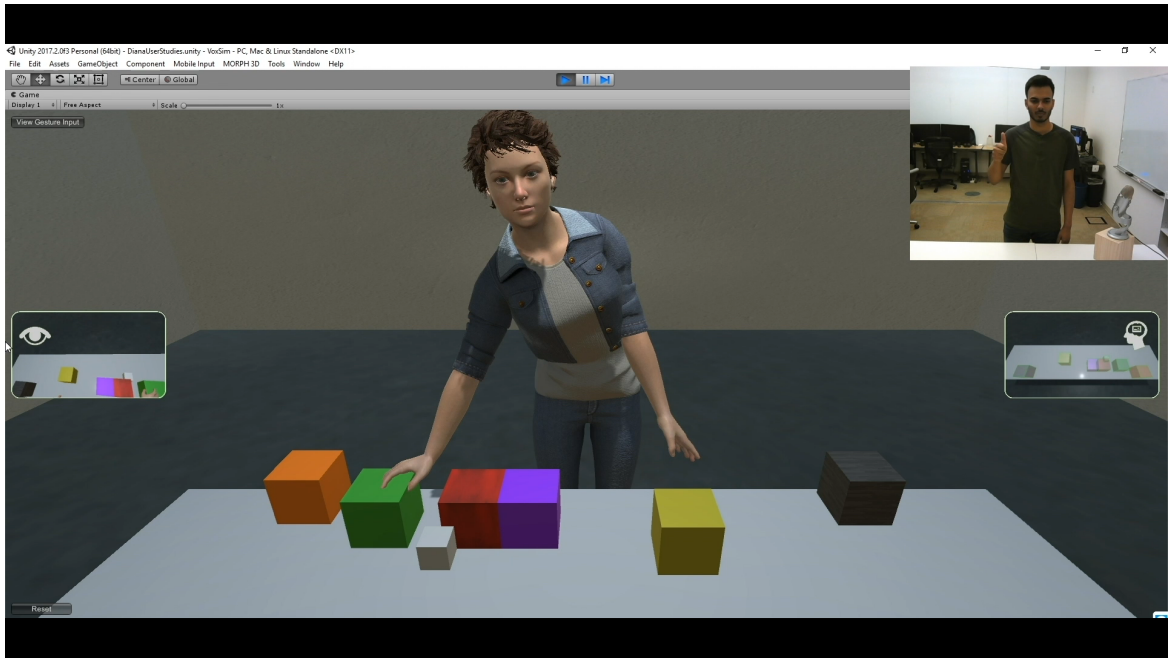
Public Perception Logic 2/2

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.")
  $[(d \cup h)^*]_\sigma \phi$

Situated Communication - Under the hood 5/5

EpiSim: Epistemic State and Update



▸ Link

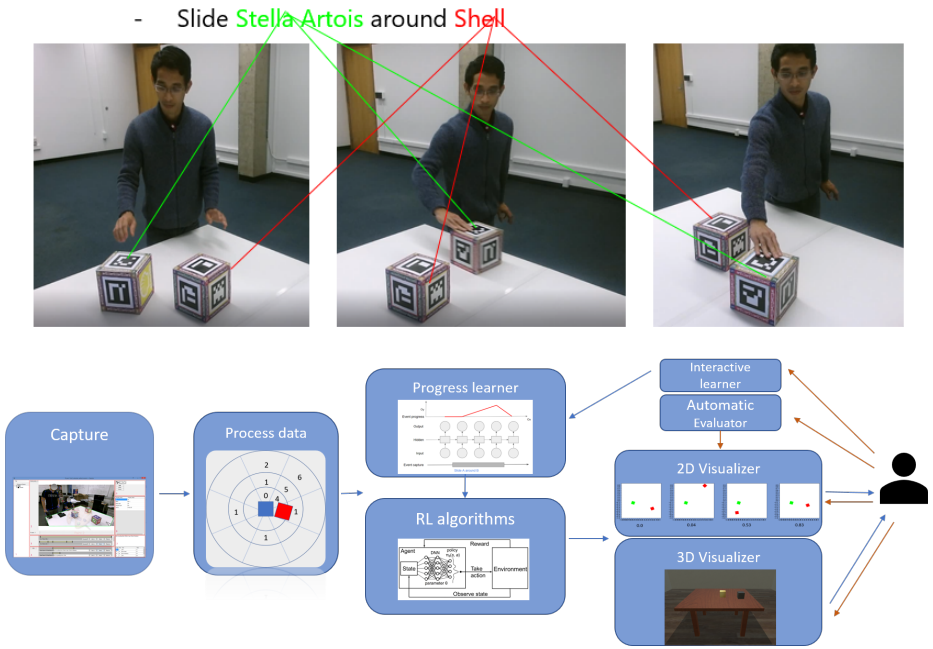
Learning by Communication

- Humans are able to recognize the consequences of their own actions as well as those performed by others.
- Recognizing new actions and learning novel events is critical for the communication of intentions and goals in conversation.
- We are experimenting with concept learning (object and event) through demonstration and observation in a simulation environment.

Learning from Demonstration

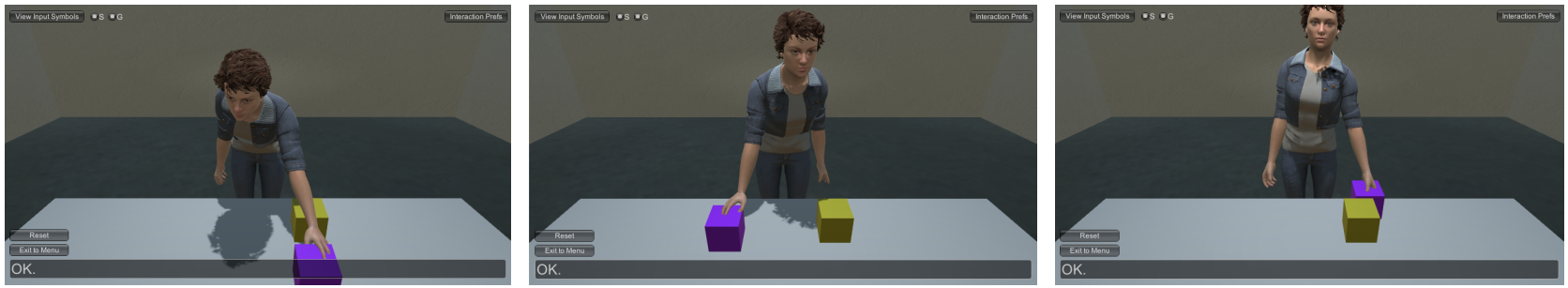
Case study: Learning "Slide Around" (Do, 2018)

- Capture and annotate human interaction with objects (2 performers, 20 clockwise and 20 counter-clockwise motions)
- Extract changing qualitative spatial relations between blocks.



Learning from Interaction

- Online corrections in the middle of a demonstration by clicking in a 2-D simulator to specify best locations after seeing agents demonstrate an action
- Clicking in the 2-D simulator maps to pointing in the 3-D interactive system



- Currently, we need more demonstrations to bootstrap the model; ongoing work on exclusively interactive learning

Evaluations: Structure Learning in VoxWorld

Krishnaswamy et al (forthcoming)

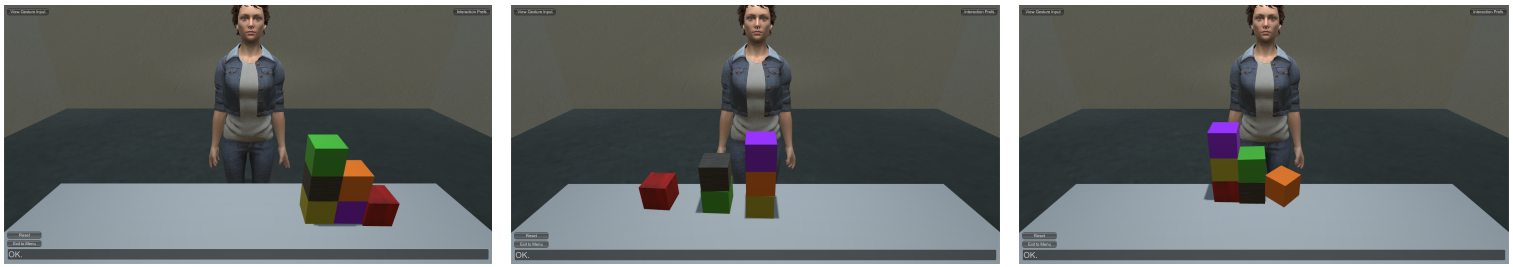


Figure: User-constructed staircases (3 shown of 17 samples)

- Learning:
  - CNN to predict most likely target configuration at current step
  - LSTM to choose remaining sequence of moves
  - Heuristic pruning on intersection of two sets to choose next legal move in current configuration

Evaluations: Structure Learning

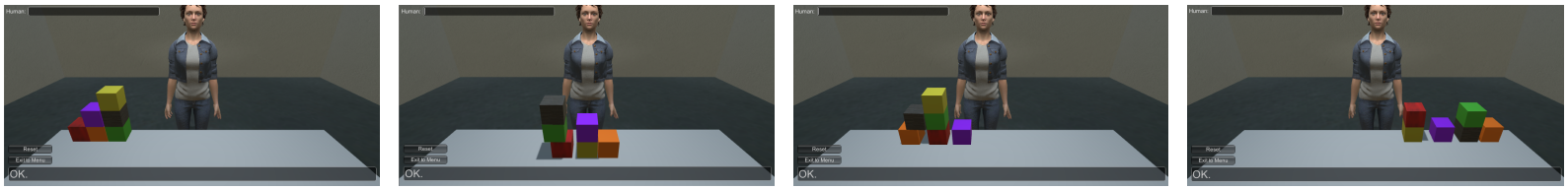


Figure: Example generated structures

Sample	Annotator Rating									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	μ	σ
1	6	9	8	9	10	9	8	9	8.5	1.1952
2	2	7	6	6	5	7	5	7	5.625	1.6850
3	4	8	7	8	8	9	5	8	7.125	1.7269
4	0	5	0	3	0	2	4	3	2.125	1.9594
5	2	4	3	5	8	5	3	5	4.375	1.8468
6	0	2	0	4	0	2	2	4	1.75	1.6690
7	6	9	8	9	9	9	4	9	7.875	1.8851
8	10	10	10	10	10	10	8	10	9.75	0.7071
9	5	8	7	8	1	9	7	8	6.625	2.5600
10	6	7	5	8	1	8	6	7	6	2.2678

Table: "On a scale of 0-10, how much does this resemble a staircase?"

Learned Staircase

Learning by Communicating

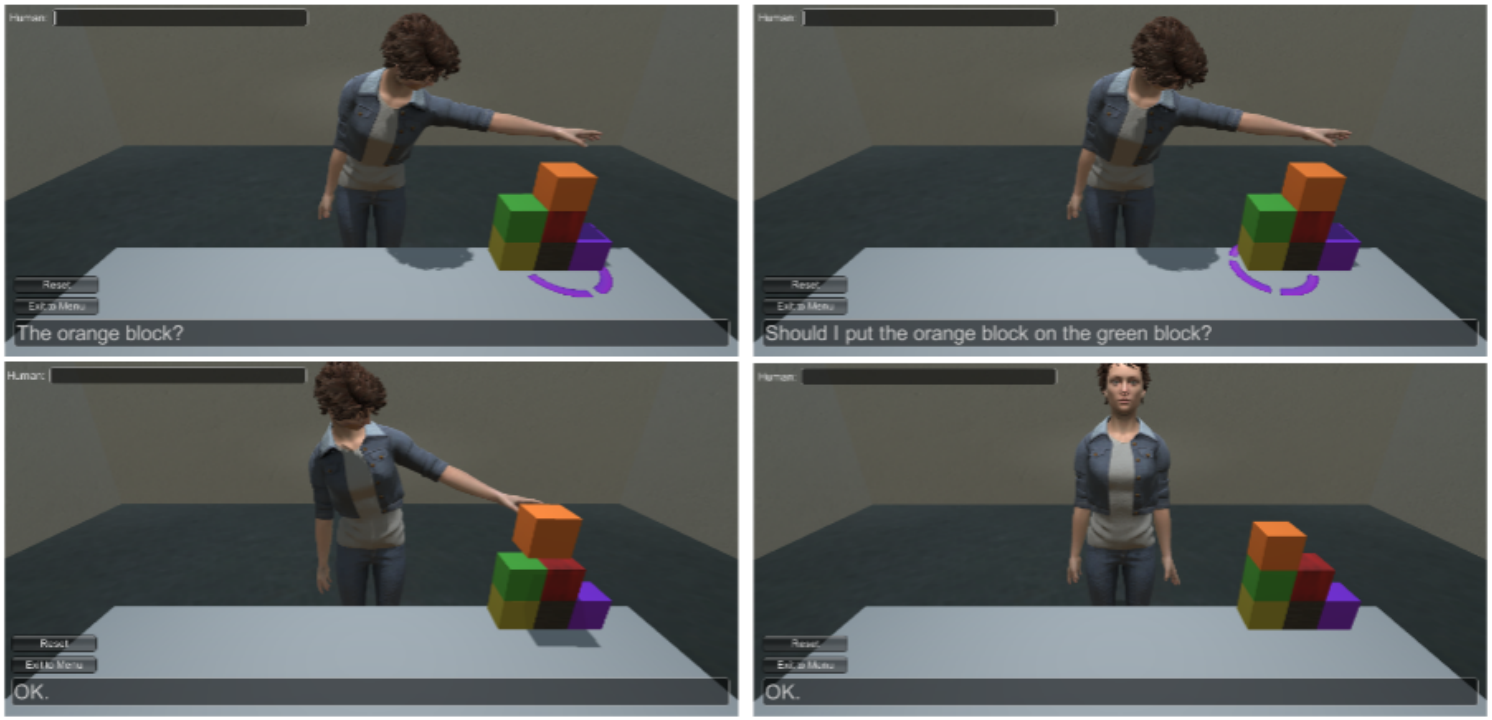
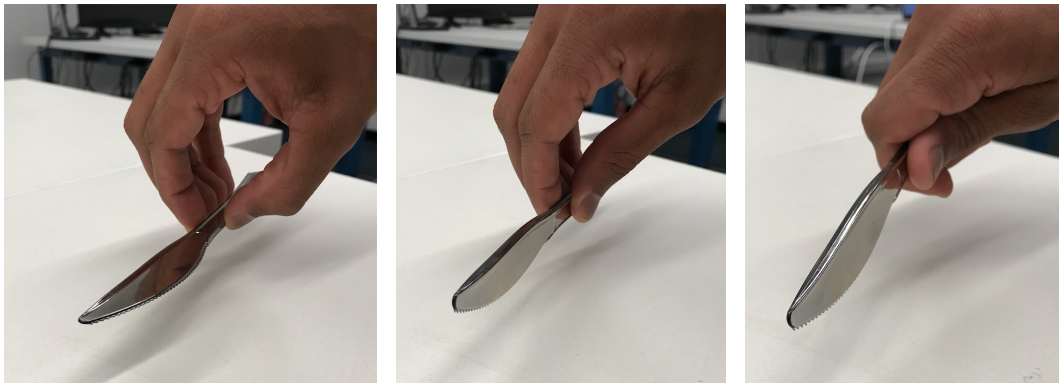
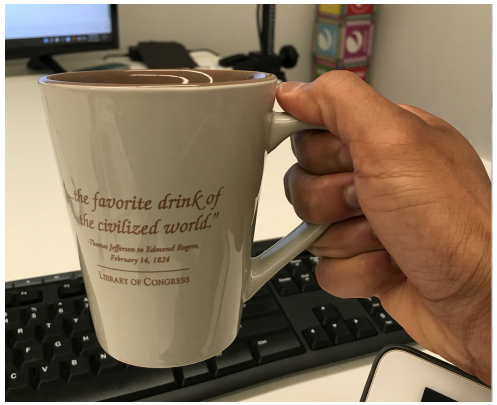


Fig. 4. Correcting a generated structure with multimodal communication.

Learning Affordances for Different Objects - Grasping 1/2



Learning Affordances for Different Objects - Grasping 2/2

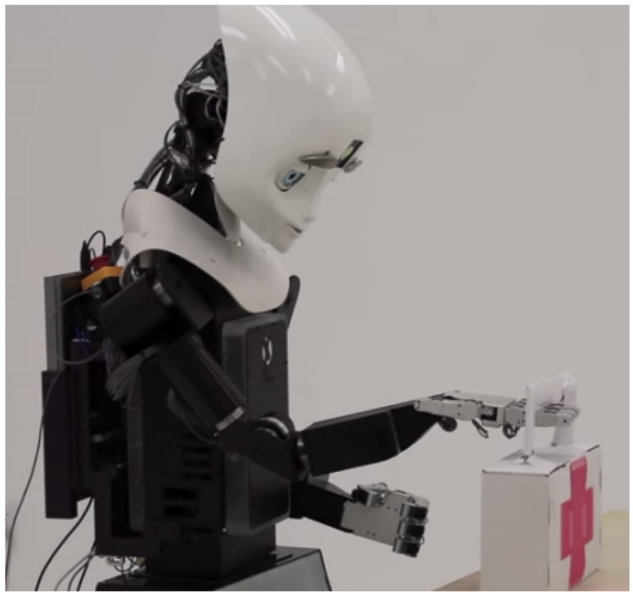


One-shot Action Learning 1/2

Scheutz (2017) "One-Shot Learning through Task-Based Natural Language Dialogues"

Human demonstrates picking up a

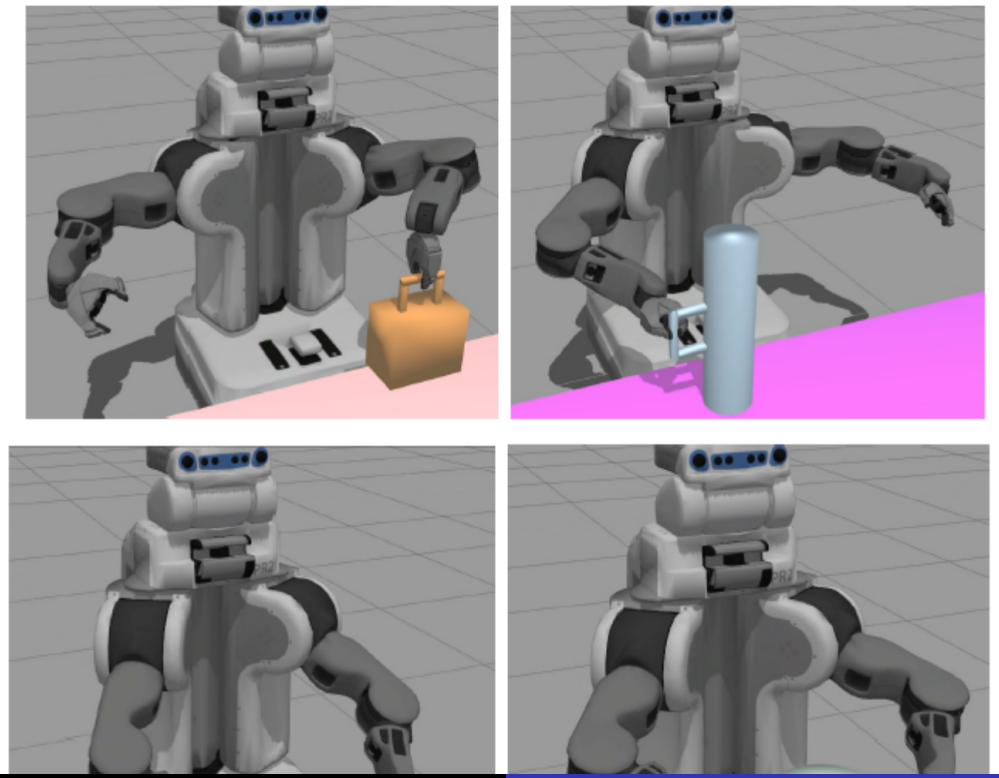
Robot learns to pick up a medkit



One-shot Action Learning 2/2

Scheutz (2017) "One-Shot Learning through Task-Based Natural Language Dialogues"

- Use multimodal simulations for Learning by Communication through dialogue
- Exploit the semantics and affordances of objects and their parts with VoxML



Conclusion

- Human-computer/robot communication requires deeper semantic models than we currently support: contextualizing and situating the interaction;
- Multimodal simulations provide both a model and a platform for studying the common ground for multimodal communicative interactions;
- Deeper semantic models also require more data:
  - leveraging existing language-image/video corpora for training models (ImSitu, ActivityNet, ImageNet, VisualGenome)
  - possible shared task surrounding a problem that can be mapped to VoxML, annotating object latent event structure as a way to capture object affordances
- We are collaborating with Matthias Scheutz to enrich HRI with situated grounding and contextualized semantics from VoxWorld simulations.

Thank You ☺

- Brandeis LLC lab members: Nikhil Krishnaswamy, Kyeongmin Rim, Tuan Do, Ken Lai, Kelley Lynch, Marc Verhagen
- CSU Vision lab members: Bruce Draper, Ross Beveridge, Pradyumna Narayana, Rahul Bangar
- University of Florida lab members: Jaime Ruiz, Isaac Wang
- Funded by a grant from DARPA within the CwC Program

Thank You ☺

- Brandeis LLC lab members: Nikhil Krishnaswamy, Kyeongmin Rim, Tuan Do, Ken Lai, Kelley Lynch, Marc Verhagen
- CSU Vision lab members: Bruce Draper, Ross Beveridge, Pradyumna Narayana, Rahul Bangar
- University of Florida lab members: Jaime Ruiz, Isaac Wang
- Funded by a grant from DARPA within the CwC Program

Publications

- Do, T., Krishnaswamy, N., & Pustejovsky, J. (2018). Teaching Virtual Agents to Perform Complex Spatial-Temporal Activities. In *Integrating Representation, Reasoning, Learning, and Execution for Goal Directed Autonomy*, AAAI Spring Symposium Series 2018.
- Krishnaswamy, N. & Pustejovsky, J. (2018). An Evaluation Framework for Multimodal Interaction. In *Proceedings of 11th LREC*, 2018.
- Krishnaswamy, N., Do, T., & Pustejovsky, J. (2018). Learning Actions from Events Using Agent Motions. In *Workshop on Annotation, Recognition, and Evaluation of Actions (AREA)*, 2018.
- Krishnaswamy, N. & Pustejovsky, J. (2018). Deictic Adaptation In a Virtual Environment. In *Proceedings of Spatial Cognition*, 2018.

Publications

- Pustejovsky, J. (2018). "From actions to events: Communicating through language and gesture", Michael A. Arbib (ed.) *How the Brain Got Language*, issues of *Interaction Studies*: 19:1/2.
- Pustejovsky, J. (2018). "Mapping from Surface to Abstract Event Structures in Language", *Journal of Linguistics*.