Multimodal Semantic Grounding for Communicating with Computers

James Pustejovsky, Nikhil Krishnaswamy
Keigh Rim, Mark Hutchens, Ken Lai,
Katherine Krajovic, Daeja Showers, Eli Goldner
Brandeis University

CwC PI Meeting
DARPA@Virtual
July 21, 2020

Diana's World Team


James Pustejovsky


Nikhil Krishnaswamy


Ross Beveridge


Francisco R. Ortega


Lisa Daunhauer


Jaime Ruiz


Daniel Delgado

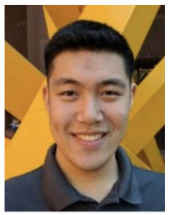

Matt Dragan


Katie Krajovik


Dhruva Patil


Joe Strout


Heting Wang


Isaac Wang


David White

Where We Started

Context in Communication - Paul Cohen

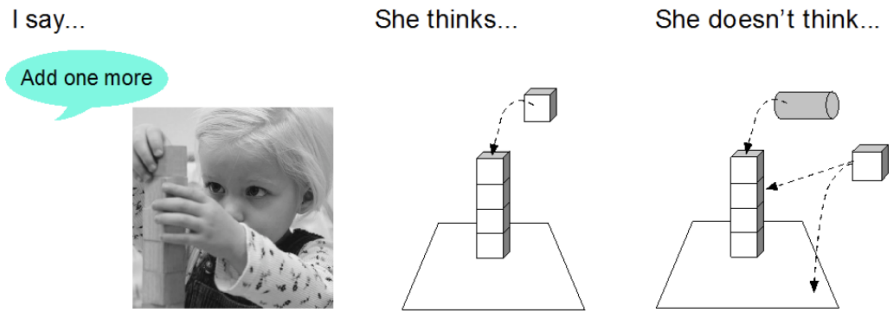
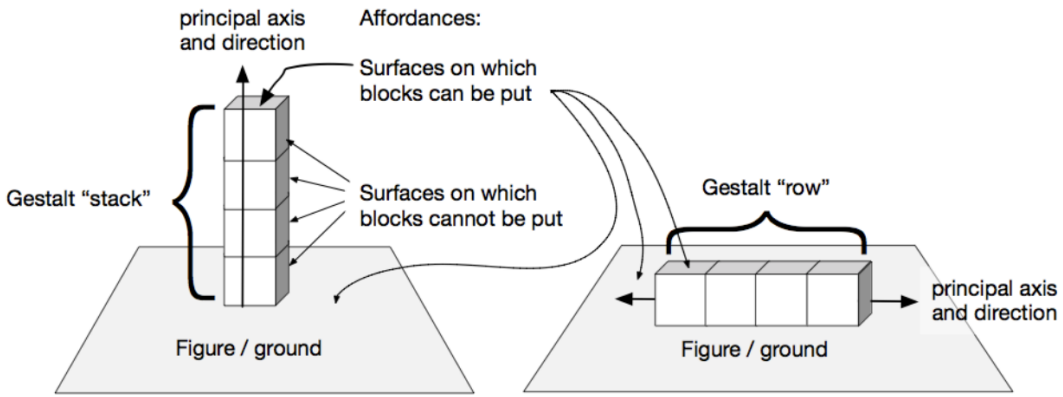
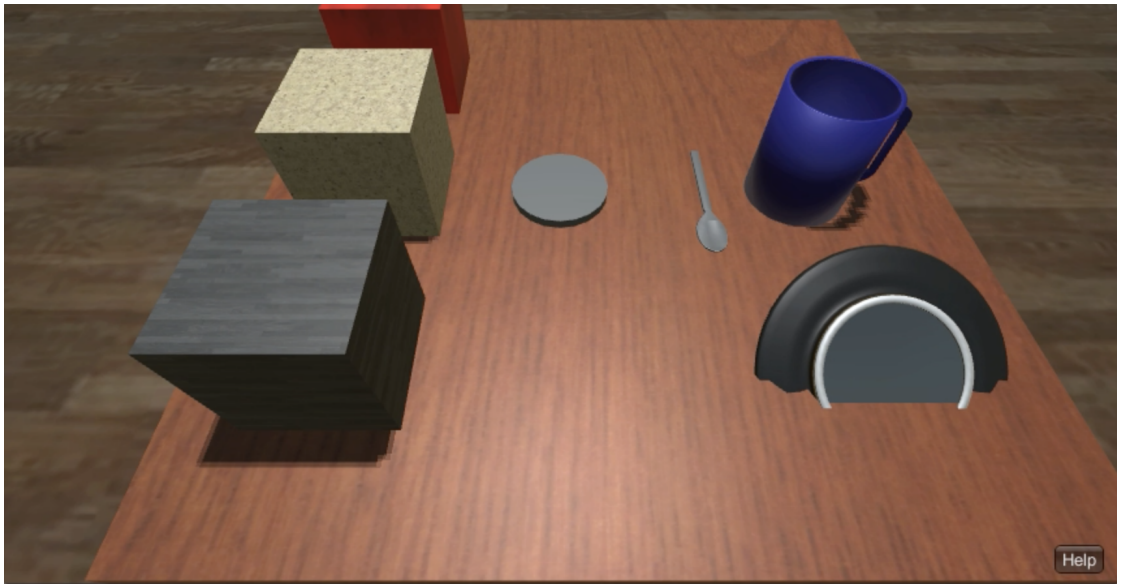


Figure 2. "Add one more" is ambiguous out of context, but given context it is remarkably precise.



VoxWorld (January 2016)

Freak Out - Taking Simulation Physics Seriously

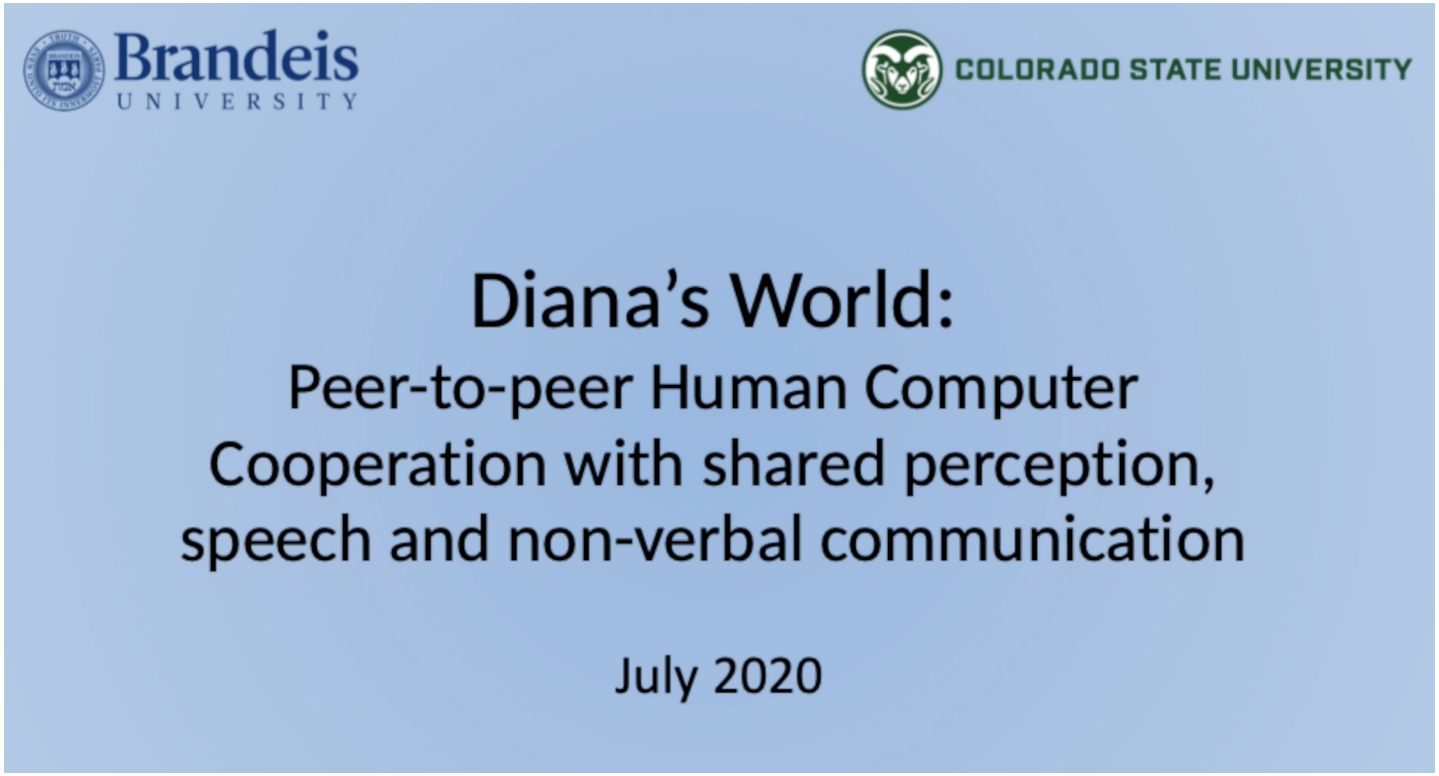


▸ Link

Outline

- Diana's World Demo
- Model of *situated meaning*
  - Situated Grounding
  - Dynamic Discourse Interpretation
  - Affordance and Goal Recognition
- Transfer Learning for affordance and concept acquisition
- Improvements to Speech Recognition
- Adopting VoxWorld to Mobile Robotics
- Jarvis and data visualization and navigation

Diana's World



▸ Link

Accomplishments

- Providing a multimodal semantics for communication -
  - Situated Meaning
  - Common ground
- Continuation Semantics for Multimodal Communication
- Handling asynchronous input from multiple modalities
- VoxWorld enabled as a platform architecture for multiple applications
  - Adopting VoxWorld to Agent Bob
  - Adopting VoxWorld to Mobile Robot

Situated Meaning

Mother and son interacting in a 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

Our Approach - Situated Meaning

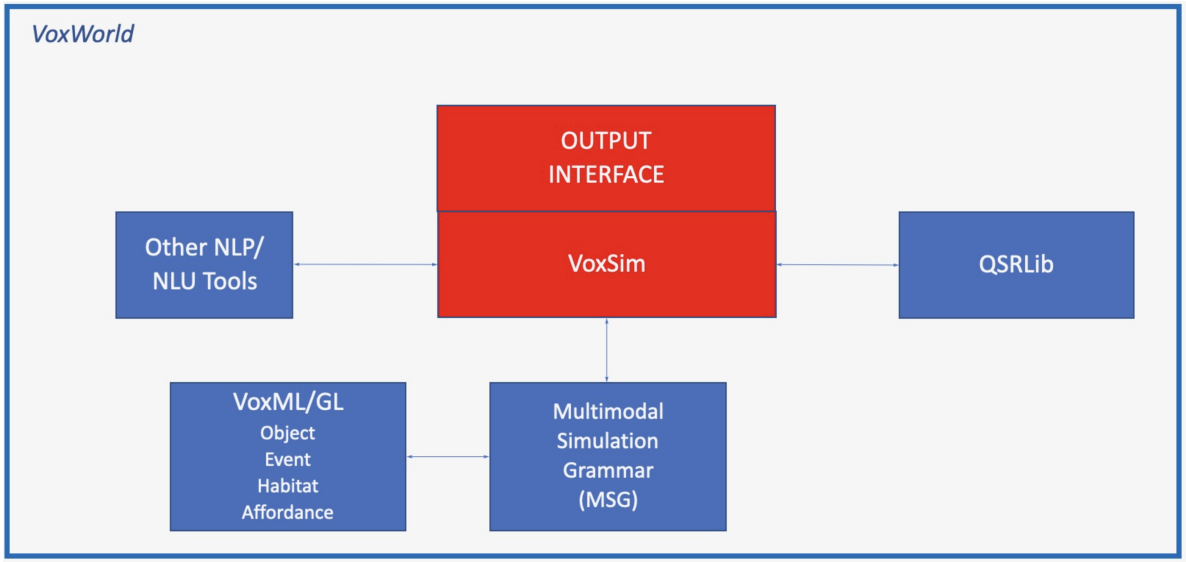
- Identifying the *actions and consequences* associated with objects in the environment.
- Encoding a multimodal expression contextualized to the *dynamics of the discourse*
- *Situated grounding*: Capturing how multimodal expressions are anchored, contextualized, and situated in context

- Shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.
- Multiple modalities, such as language and gesture, need to be aligned or coordinated in oder to be interpretable.
  - An utterance, gesture, or action can be interrupted;
  - Channels can convey signals asynchronously

VoxWorld: A Platform for Multimodal Simulations

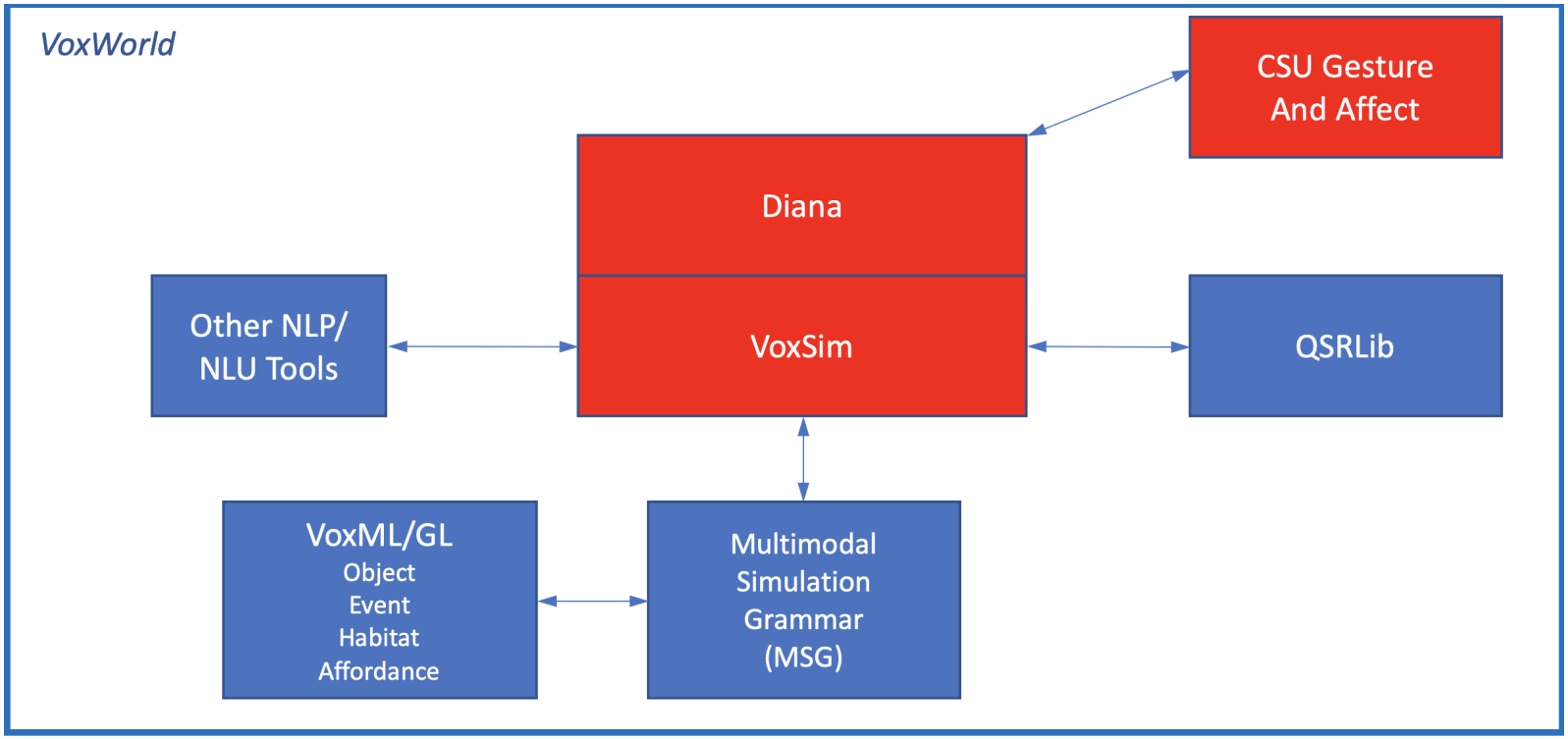
1. *Computational simulation modeling*. Variables in a model are set and the model is run, such that the consequences of all possible computable configurations become known.
2. *Situated embodied simulations*. Agent is embodied with a dynamic point-of-view or avatar in a virtual or simulated world.
3.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

VoxWorld: A Platform for Multimodal Simulations



VoxWorld: A Platform for Multimodal Simulations

Interfacing Diana to CSU Gesture and Affect Systems



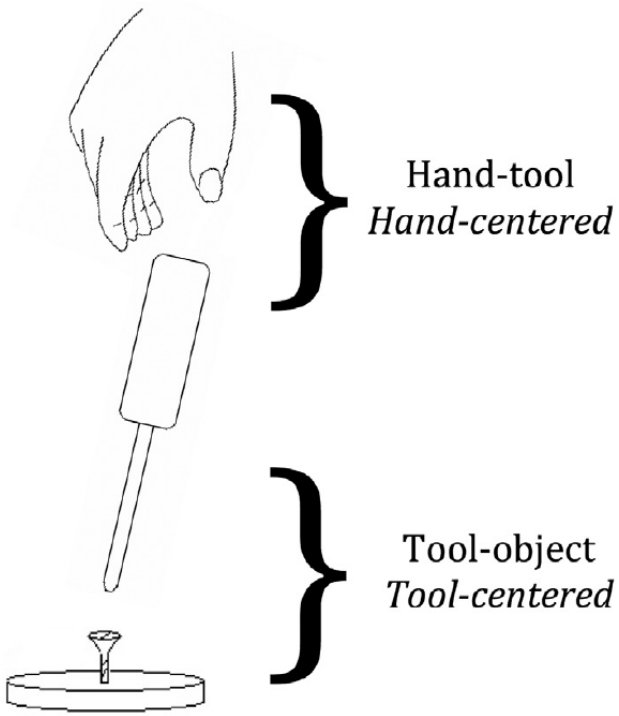
Affordance and Goal Recognition

1. Perceived purpose is an integral component of how we interpret situations and reason about utterances in communicative contexts.
   - Events are purposeful and directed;
   - Places are functional;
   - Objects are usable and manipulable.
2. Affordances are latent action structures of how an agent interacts with objects in the environment, in different modalities:
   - language, gesture, vision, action;
3. Qualia Structure provides a link to such latent actions structures associated with objects in utterances and the context.

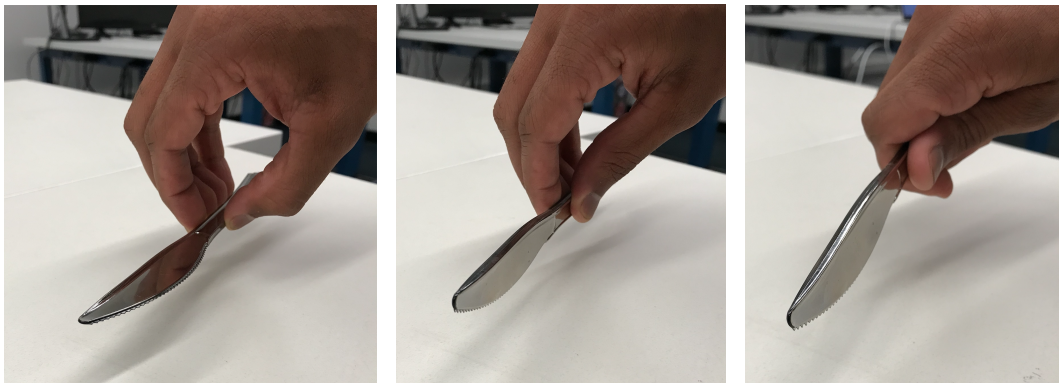
Visual Object Concept Modeling Language (VoxML)

- 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: Orientation, Situated context, Scaling
  - Affordance Structure:
    - What can one do to it
    - What can one do with it
    - What does it enable

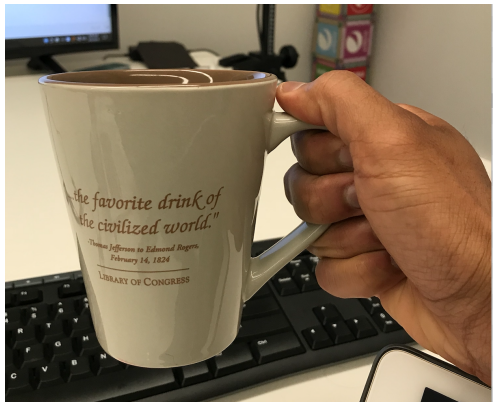
Habitats Encode Reference Frames in Affordances



Learning Affordances for Different Objects - Grasping 1/2

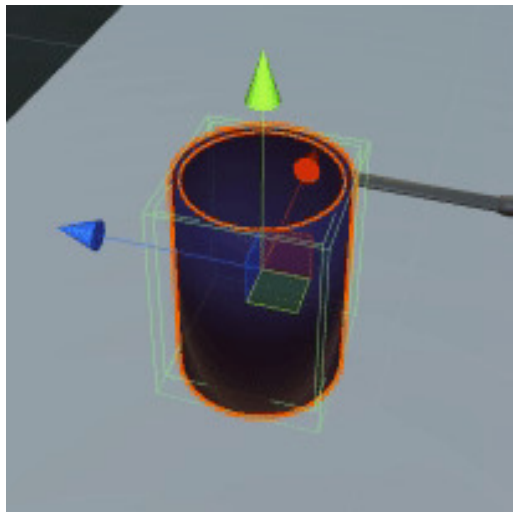


Learning Affordances for Different Objects - Grasping 2/2



VoxML - cup

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

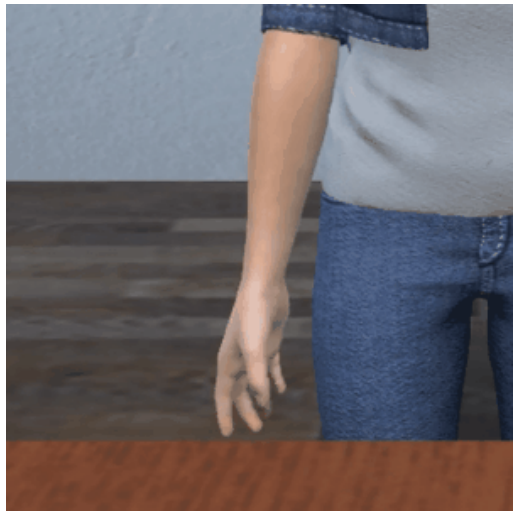


$$
\begin{bmatrix}
\textbf{put} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix}
\text{HEAD} = \textbf{transition} \\
\text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{bmatrix} \\
\text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while(hold(x,y), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)] \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\textbf{on} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{on} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix}
\text{CLASS} = \textbf{config} \\
\text{VALUE} = \textbf{EC} \\
\text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:3D} \\ \text{A}_2 = \textbf{y:3D} \end{bmatrix} \\
\text{CONSTR} = \textbf{y}\rightarrow\text{HABITAT}\rightarrow\text{INTR}[align]
\end{bmatrix}
\end{bmatrix}
$$

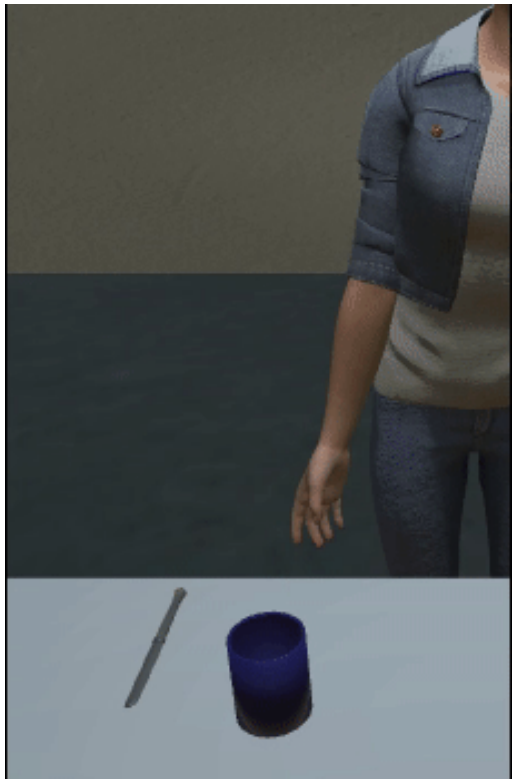
VoxML - grasp

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



VoxML - grasp cup

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



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

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 : $[(\alpha \cup \beta)^*]_\sigma\varphi$

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

- 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 D as follows:
- $[\![\overline{(\mathbf{T}.\mathbf{C})}]\!]^{M,cg} = \lambda k.[\![\overline{\mathbf{T}}]\!](\lambda n.[\![\overline{\mathbf{C}}]\!](\lambda m.k(m\ n)))$

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.

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

Actions as Described by Gesture

Kendon (2004), Lascarides and Stone (2009)

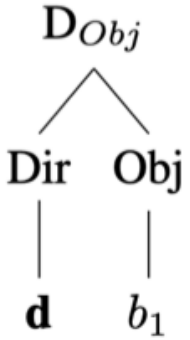
- $G \rightarrow (Prep)\ (Pre\_stroke\ Hold)$ **Stroke** $Retract$

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

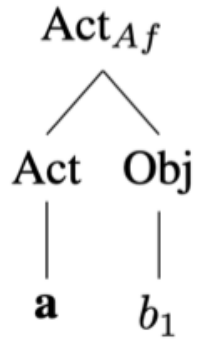
- $[[point]] = [[End(cone(\mathbf{d}))]]$

Interpreted Gesture Tree:

a. **Deixis**: $D_{obj} \rightarrow Dir\ \ Obj$



b. **Action**: $G_{Af} \rightarrow Act\ \ Obj$



Gesture Grammar

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

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

(4) 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)

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

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

Continuation-Style Semantics of Gesture

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

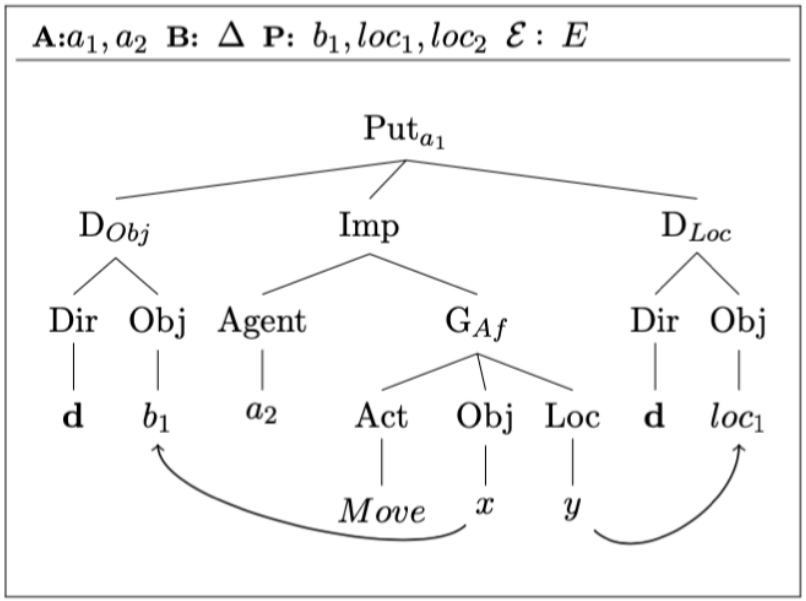
Gesture Sequence Denoting Command

SINGLE MODALITY (GESTURE) IMPERATIVE

HUMAN$_1$: $\mathcal{G} = [\textit{points to the purple block}]_{t1}$

HUMAN$_2$: $\mathcal{G} = [\textit{makes move gesture}]_{t2}$

HUMAN$_3$: $\mathcal{G} = [\textit{points to the red block}]_{t3}$



$$[\![\mathbf{D}_{Obj}.\mathbf{Move}.\mathbf{D}_{Loc}]\!] = \lambda k.([\![\mathbf{D}_{Loc}]\!]; \lambda j.([\![\mathbf{D}_{Obj}]\!]; \lambda j'.(([\![\mathbf{Move}]\!]j')j)k))$$

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:

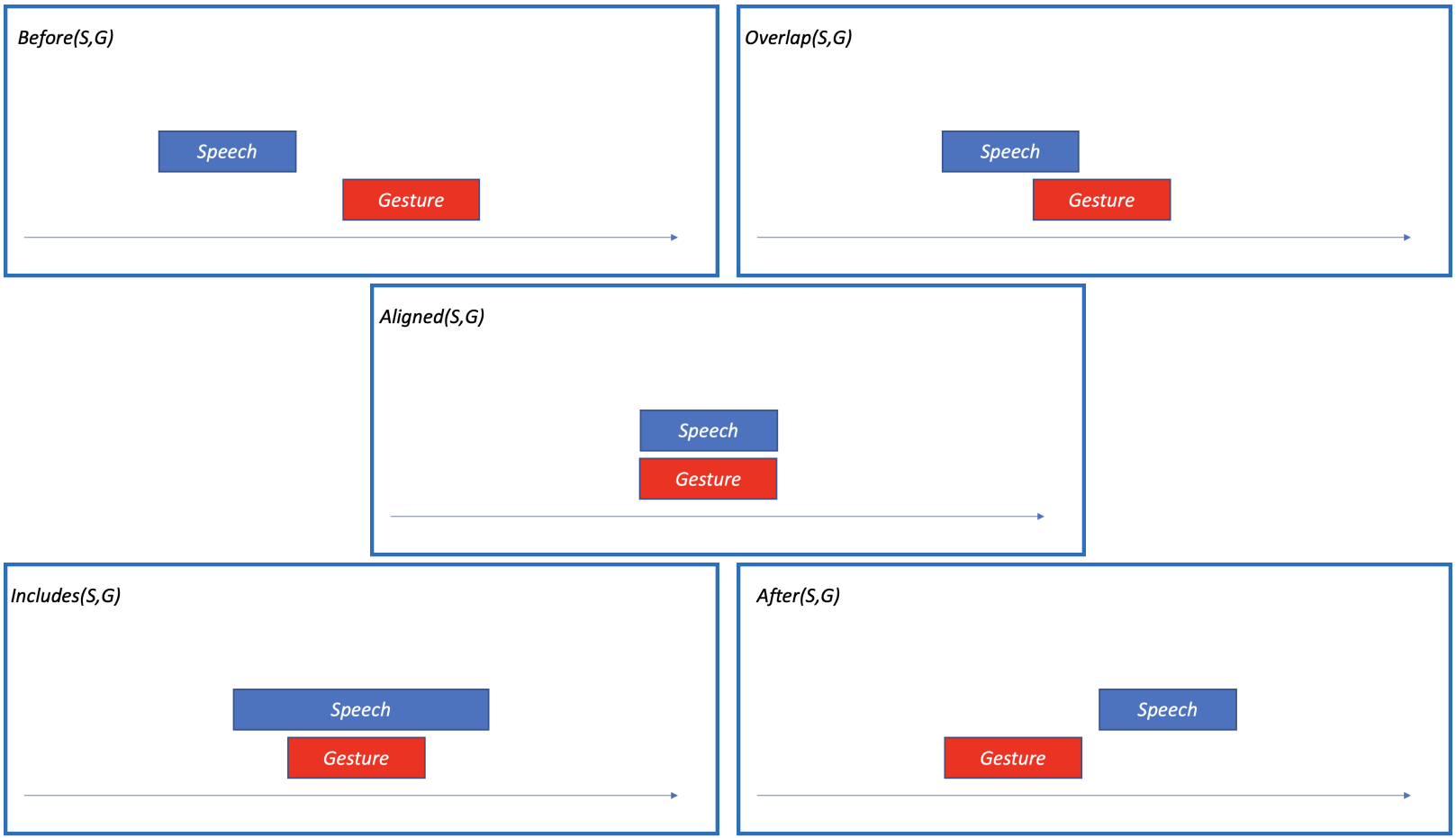
(7) **Co-gestural Speech Ensemble**: multimodal communication with Gesture, $\mathcal{G}$, and Speech, $\mathcal{S}$:

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

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

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

Aligning Speech and Gesture in Dialogue

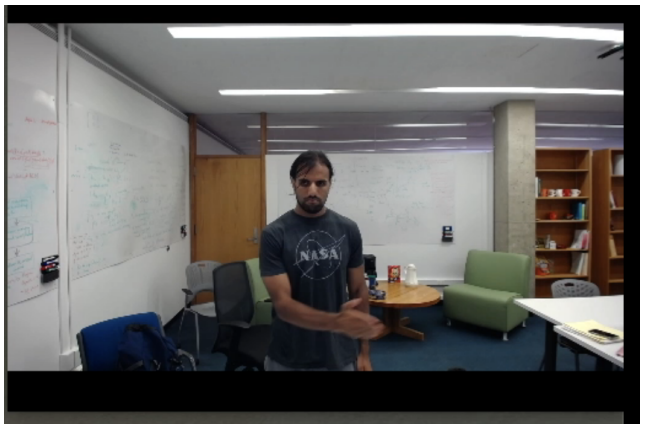


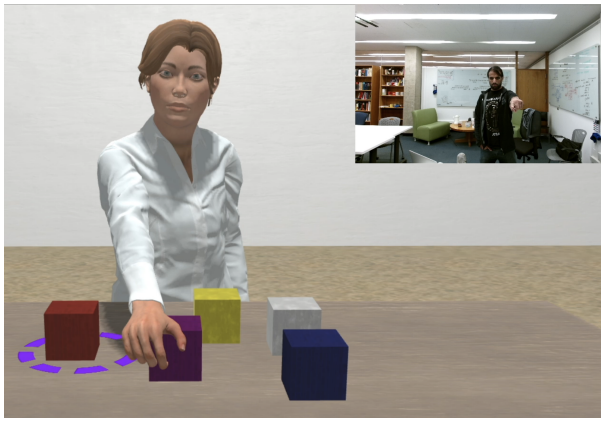
Object Affordances: Gibsonian and Telic

- 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
  - $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(cup)$
    $grab \subseteq \textbf{sel}\ k_{Gib}$
    $drink \subseteq \textbf{sel}\ k_{Telic}$
  - $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(block)$
    $grab \subseteq \textbf{sel}\ k_{Gib}$
    $pick\_up \subseteq \textbf{sel}\ k_{Gib}$
    $move \subseteq \textbf{sel}\ k_{Gib}$

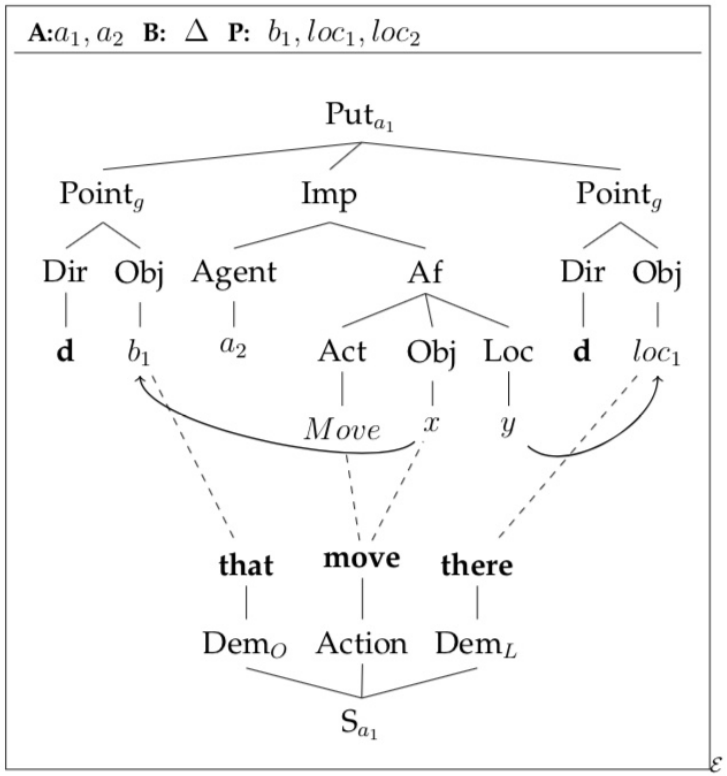
Situated Meaning

Gesture and co-gestural speech imperative





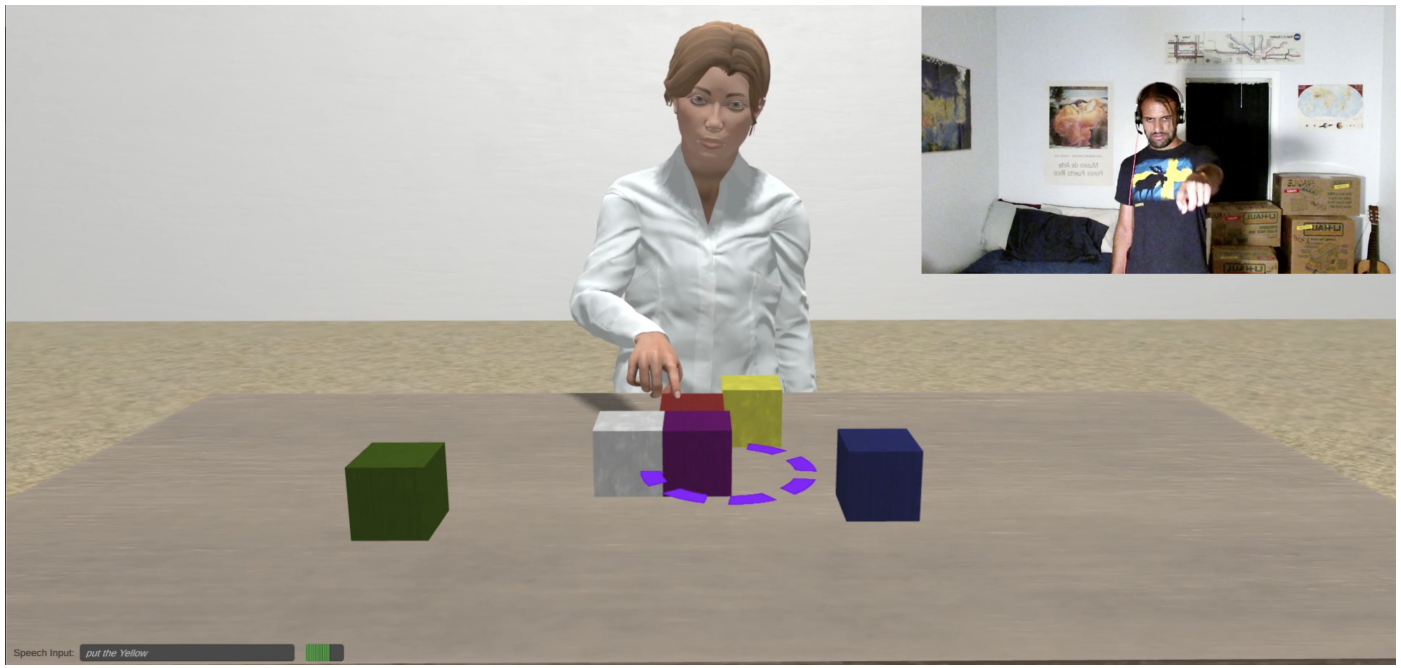
a_1: "**That** object b_1 **move** b_1 to **there**, location loc_1."



$\lambda k'_s \otimes k'_g.\overline{(\langle \textbf{that}, Point_1\rangle\langle \textbf{move}, Move\rangle)}(\lambda r_s \otimes r_g.\overline{\langle \textbf{that}, Point_2\rangle}$
$(\lambda k_s \otimes k_g.k'_s \otimes k'_g(k_s \otimes k_g r_s \otimes r_g)))$

Interruption during Dialogue

Correcting and Undoing Parameter Binding in Actions



▸ Link

Interruption during Dialogue - Under the Hood

Correcting and Undoing Parameter Binding in Actions

$\lambda k.\overline{C_1}(\lambda n.\overline{C_2}(\lambda m.k(m\ n)))$

- $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(block)$
- $grab \subseteq \textbf{sel}\ k_{Gib}$
- $put \subseteq \textbf{sel}\ k_{Gib}$
- $\lambda k.k(put) \Longrightarrow M, cg_1 \vDash on(yellow, purple)$

- *"Wait, on the white one."*
- **undo** $k = \lambda k.k(put)$
- **Rewind** the state monad and **Reassign**:
- $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(block)$
- $put \subseteq \textbf{sel}\ k_{Gib}$
- $\lambda k.k(put) \Longrightarrow$
- $M, cg_1 \vDash on(yellow, white)$

Improvements to Speech Recognition

Switched from IBM Watson to Google ASR

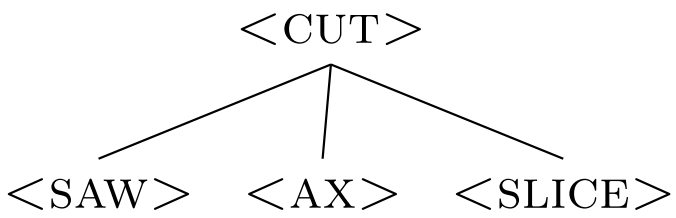
- We condition the recognition on existing context
  - If an object is already in focus, we bias recognition toward PPs
  - if a location is already in focus, we bias recognition toward NPs
  - if blackboard is empty, we give equal weight to NPs and VPs at the expense of PPs
- We generate the content of the possible syntactic category at runtime from the domain vocabulary
  - we tested both lists and hashtables in the generation step; lists were faster by about 6%
  - all of the new prediction methods were about 2x faster than the old recognition method
- We added non-Diana vocabulary for our mobile robotics work: there was no appreciable difference in recognition quality or time (~ 0.001 seconds), meaning we can use the same syntactic category method with vocabulary from multiple domains across multiple projects.

▸ Link

Transfer Learning of Object Affordances

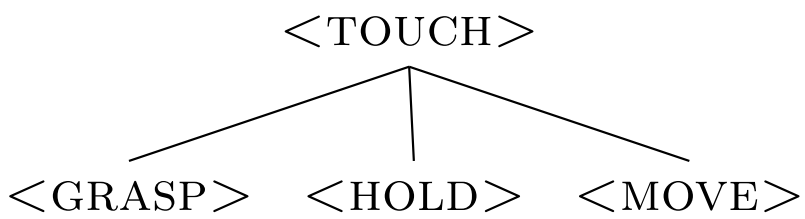
- Gibsonian/Telic affordances are associated with abstract properties:
  - spheres roll, sphere-like entities probably do too;
  - small cups are graspable, small cylindroid-shaped objects probably are too.
- Similar objects have similar habitats/affordances:
- This informs the way you can talk about items in context:
  - Q: "What am I pointing at?"
  - A: "I don't know, but it looks like {a ball/a container/etc.}

(9)

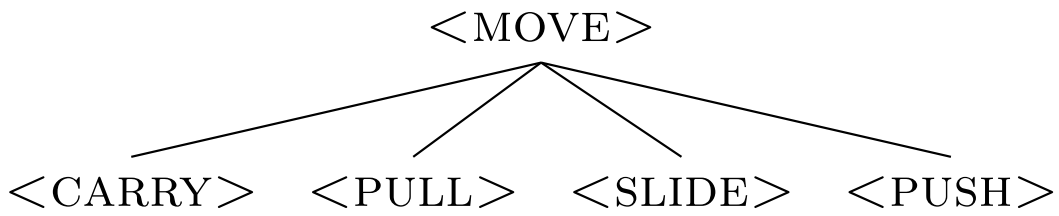


(10) a. This object is **cuttable** (the {wood/ tree/ bread}).
b. The wood is **sawable** (=‘cuttable with a saw’).
c. The tree is **axable** (=‘cuttable with an ax’).
d. The bread is **sliceable** (=‘cuttable with a knife’).

(11)



(12)



Transfer Learning of Object Affordances

Exploits the linkages between affordances and objects in VoxML

- Train over a sample of 17 different objects: blocks, KitchenWorld objects (apple, grape, banana, book, etc.)
- Trained 200 dimensional affordance and habitat embeddings using a Skip-Gram model, for 50,000 epochs with a window size of 3:
  - These embeddings serve as the inputs to the object prediction architectures
- Using the affordance embeddings in vector space, predict which object they belong to: using a 7-layer MLP; a 4-layer CNN with 1D convolutions

Transfer Learning of Object Affordances

- The architectures:

MLP	CNN
Input	Input
Dense (32 $\times$ $tanh$**)**	**Conv1D (64** $\times$ $ReLU$**)**
20% Dropout	$ReLU$
Dense (196 $\times$ $ReLU$**)**	20% Dropout
20% Dropout	**Conv1D (250** $\times$ $ReLU$**)**
Dense (92 $\times$ $tanh$**)**	Global Max Pooling 1D
20% Dropout	20% Dropout
Dense (196 $\times$ $tanh$**)**	**Dense (196)**
Dense (92 $\times$ $ReLU$**)**	20% Dropout
Dense (32 $\times$ $tanh$**)**	$ReLU$
Output $(softmax)$	**Output** $(softmax)$
70,913 params	100,923 params

- Ground truth clusters generated by k-means clustering over human-annotated object similarity. Sample aggregate results:

Model	% predictions in correct cluster	% predictions always in correct cluster
MLP (Habitats)	78.82	27.06
MLP (Affordances)	**84.71**	38.82
CNN (Habitats)	78.82	27.06
CNN (Affordances)	81.18	**40.00**

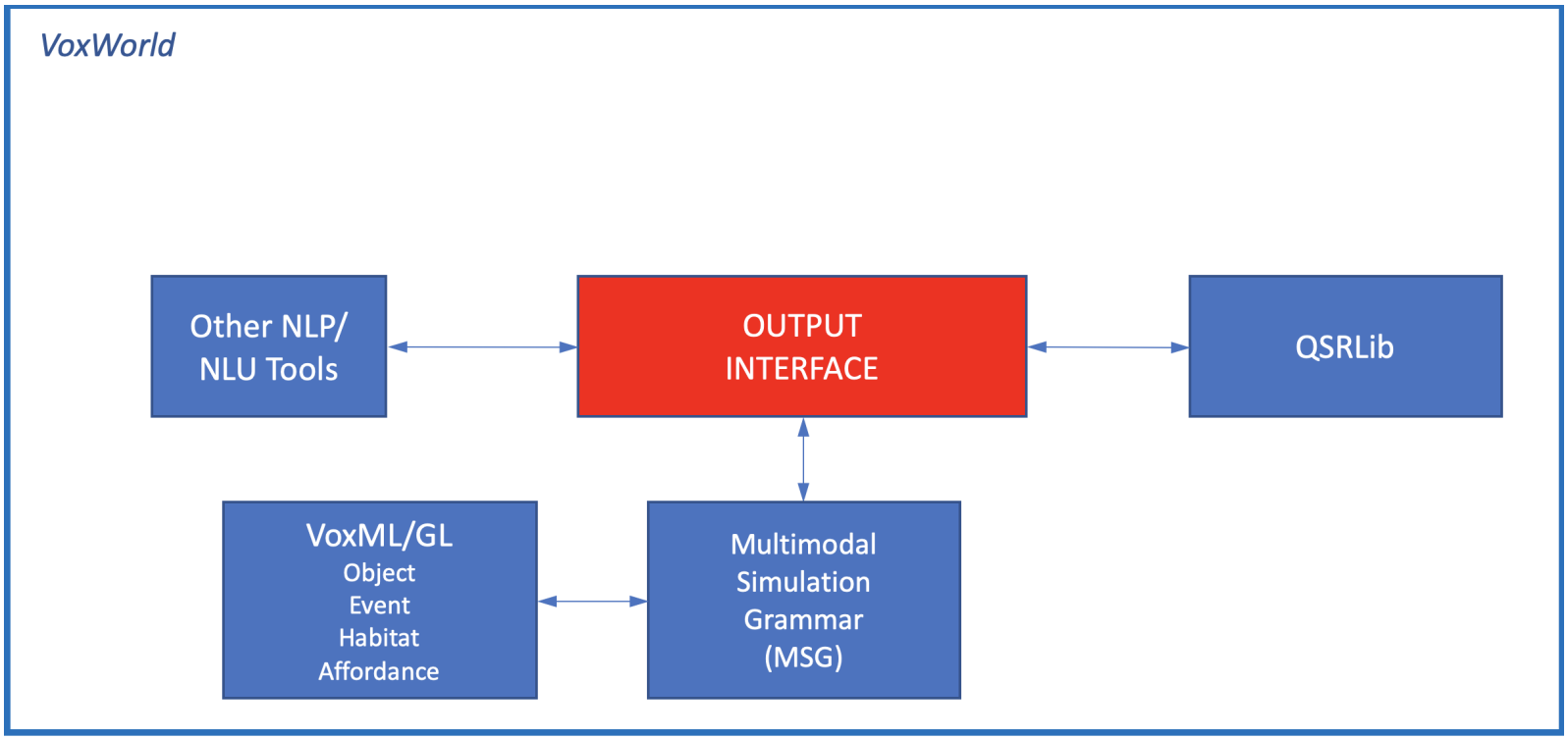
- Object specific results (input: vectorized affordances for plate)

MLP (Habitats)	MLP (Affordances)	CNN (Habitats)	CNN (Affordances)
book, cup, bowl, bottle	cup, bottle, apple	book	cup, bottle

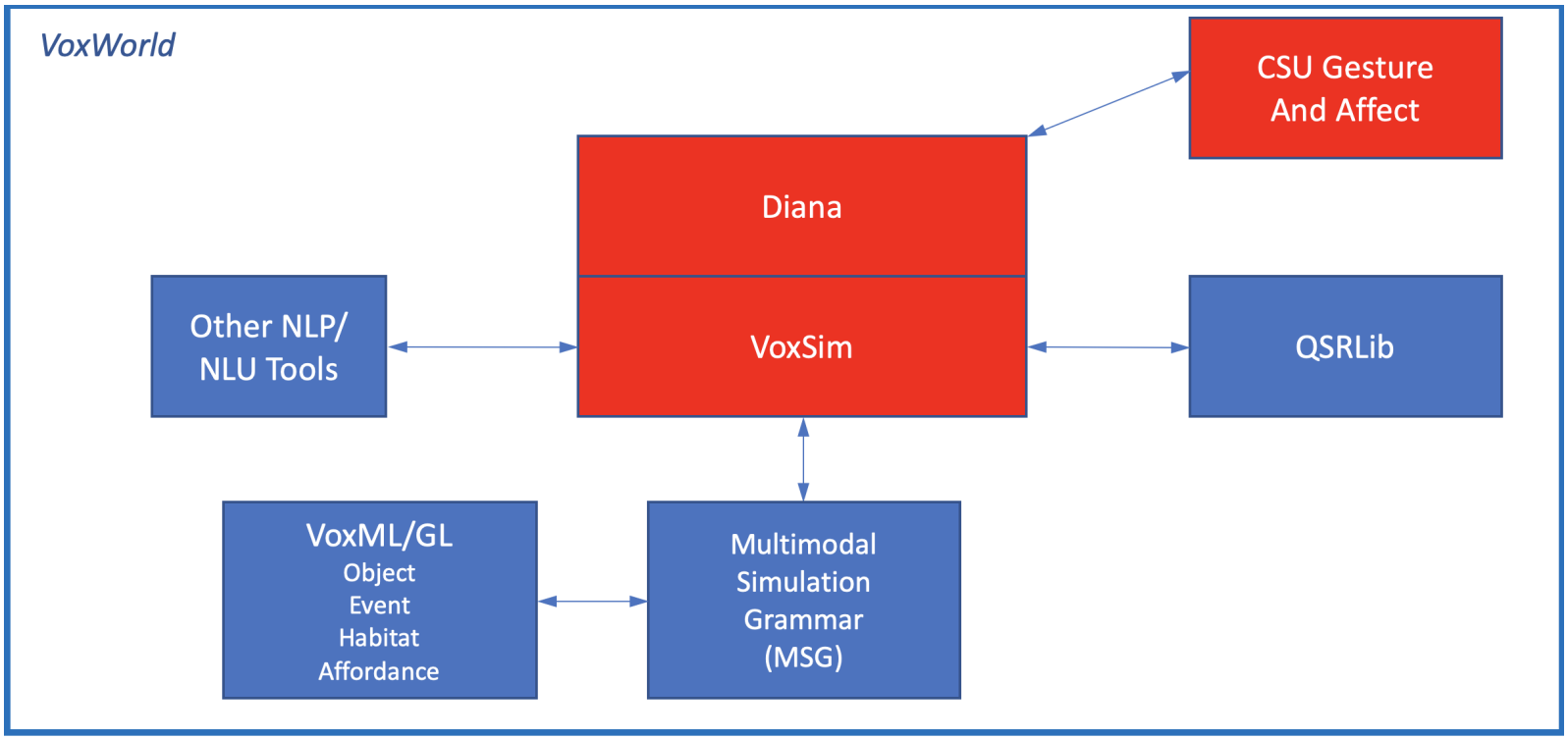
Transfer Learning of Object Affordances

▸ Play!

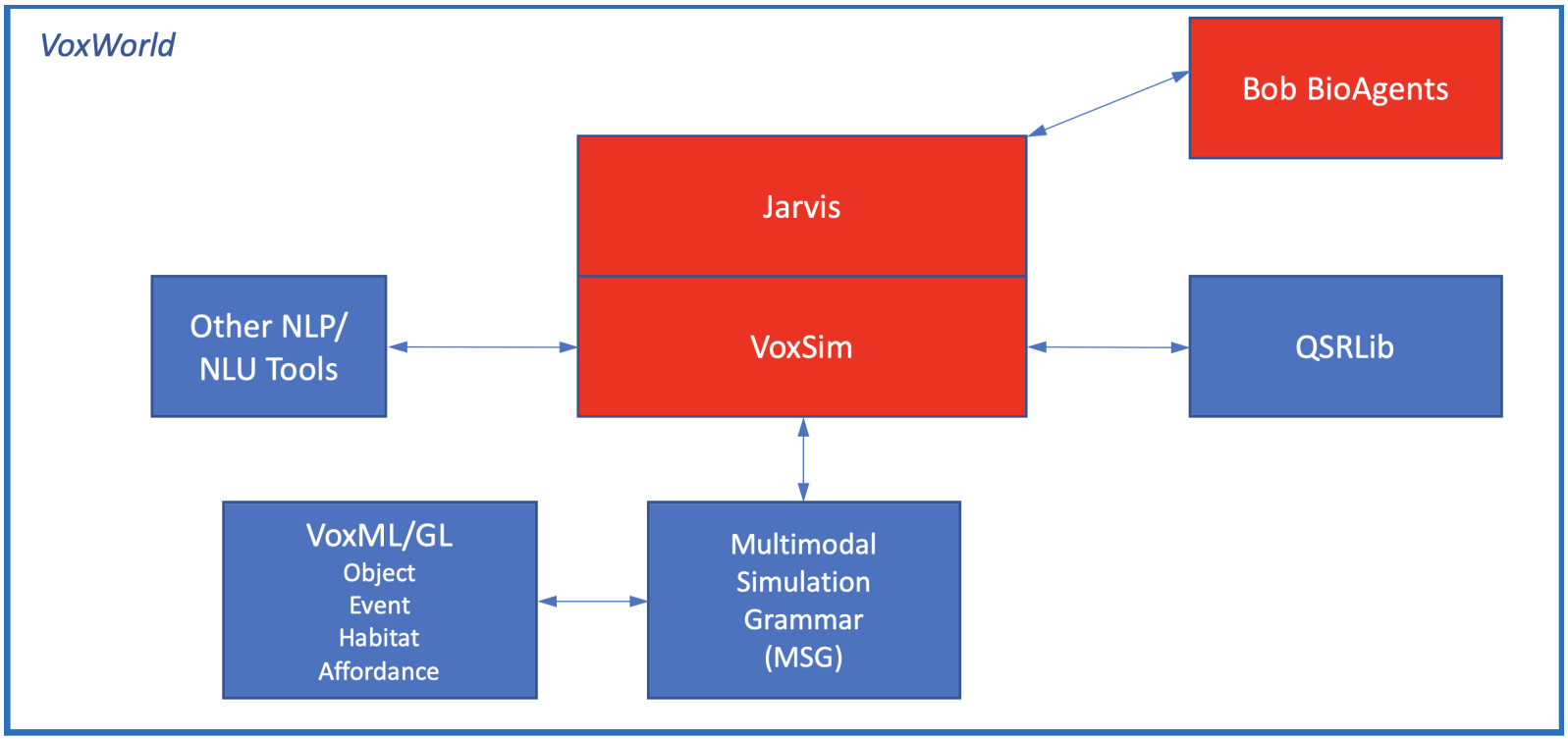
Refactoring VoxWorld as a Platform with Components



Refactoring VoxWorld as a Platform with Components

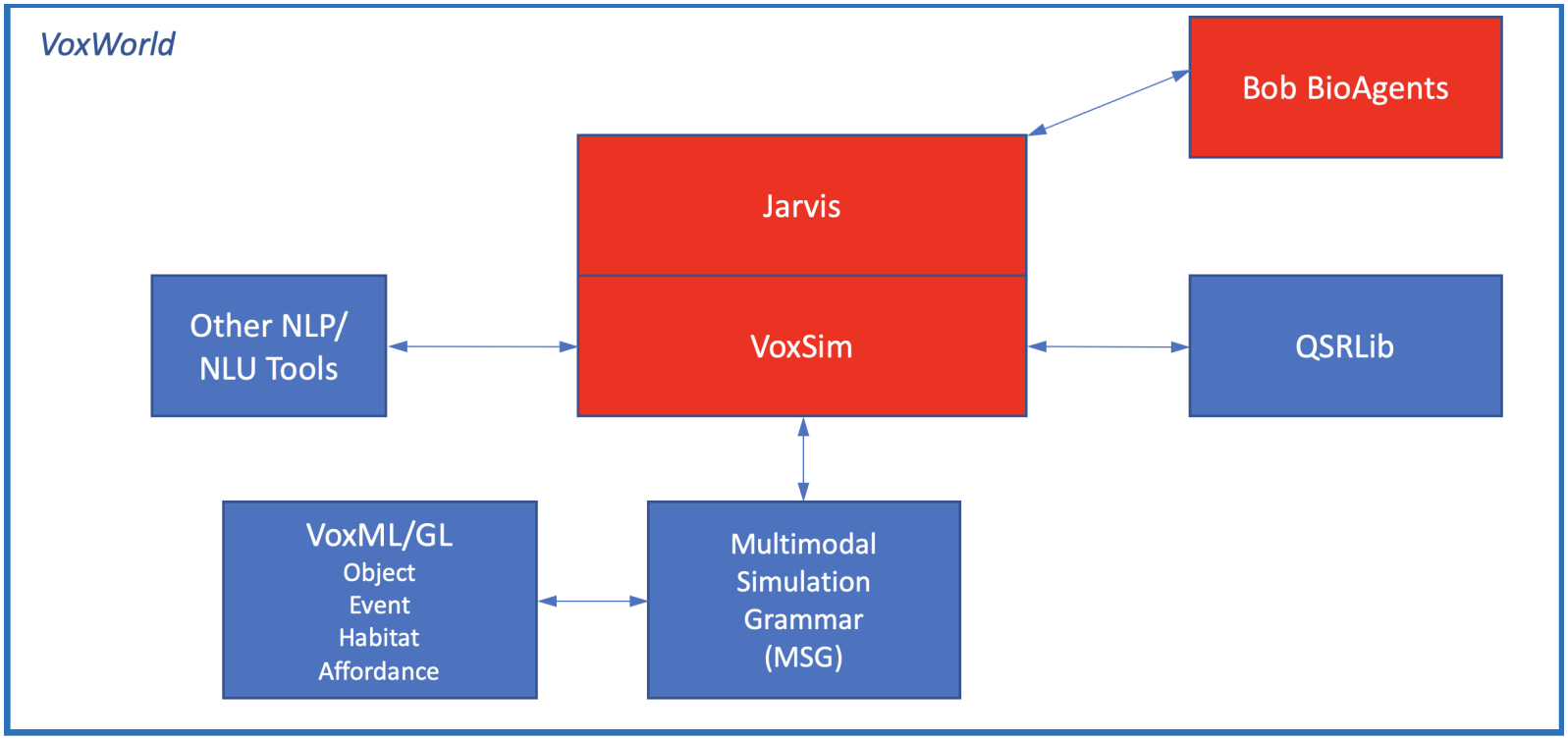


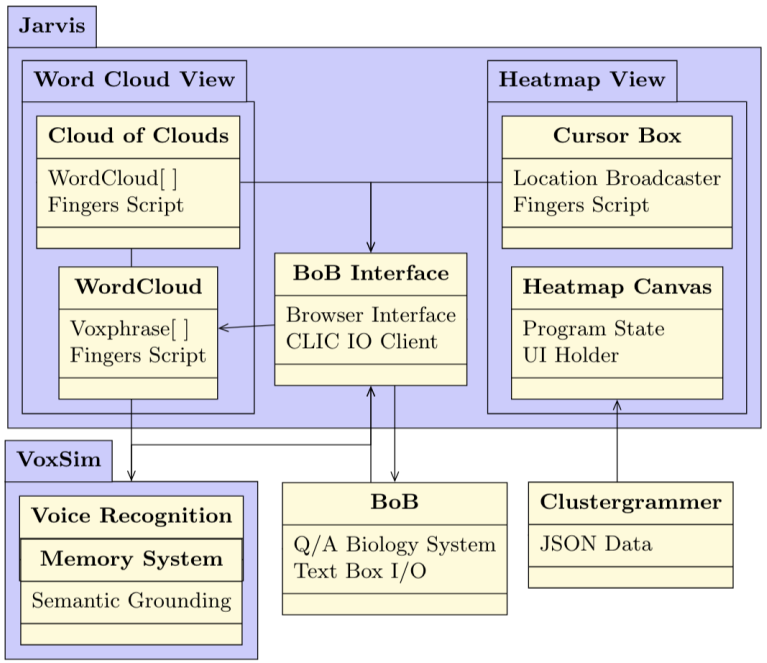
Refactoring VoxWorld as a Platform with Components



Interfacing VoxWorld to Bob through Haptics - *Jarvis*

Collaboration with SIFT and Tufts University

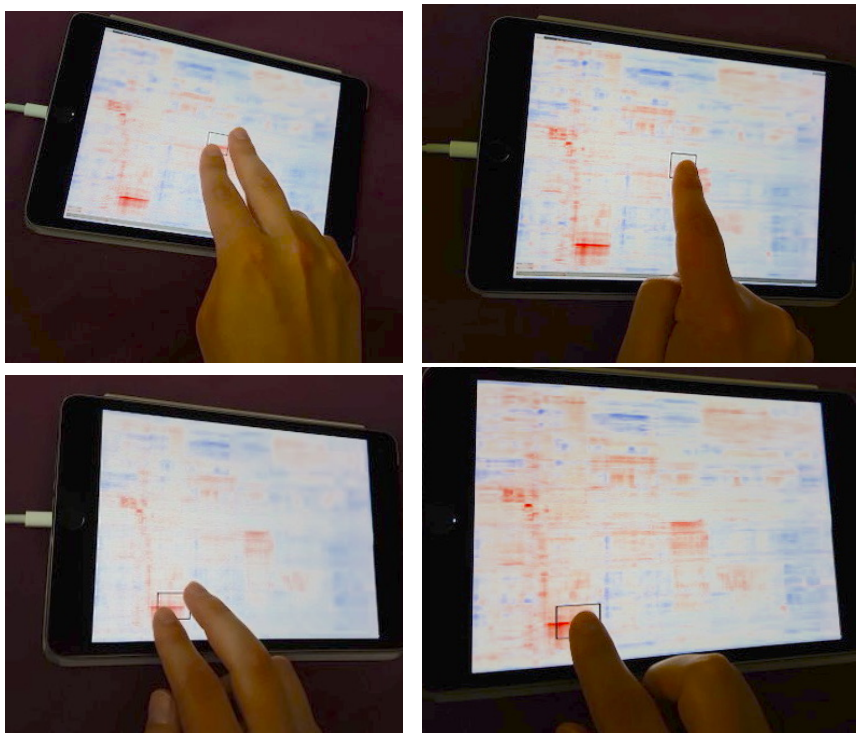




Gesture	Heatmap Interpretation	Word Cloud Interpretation
Tap	—	Semantically ground word
Swipe	Swap View	Swap View
Pan	Move selection box	—
Scale	Resize selection box	Zoom camera
Rotate	—	Rotate Word Cloud
Long Press	—	Semantically ground cloud

Table 2. Gestures available in Jarvis

- Jarvis with Projective Gesture (2019) ▸ Link
- Jarvis with Haptic Gesture (2020)



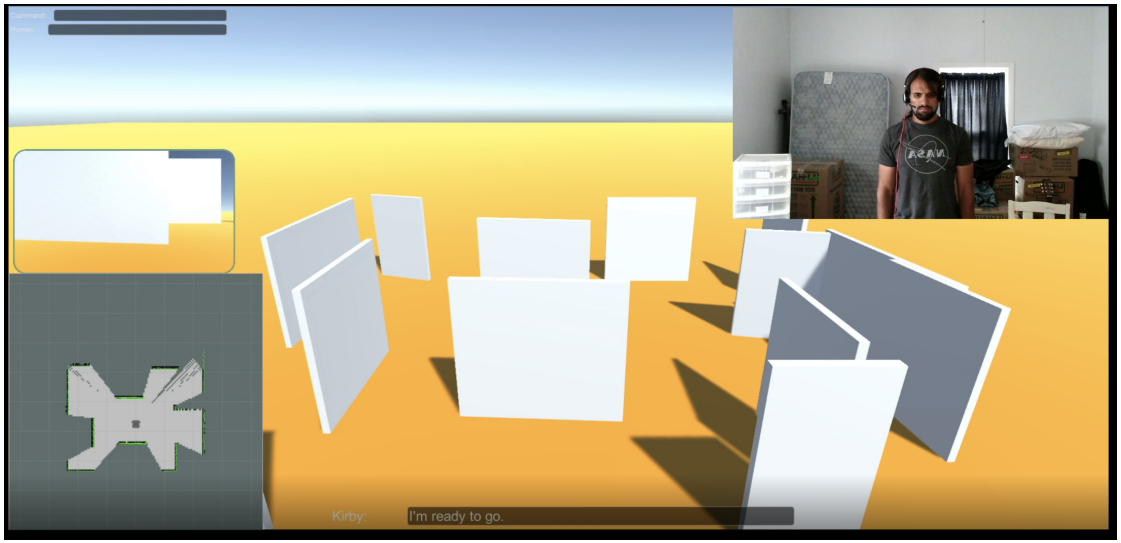
- Gesture and language communication with a Turtlebot-3:

- Fiducials represent registered proxies for object sorts in the environment:

Figure 2: Two fiducials.

Refactoring VoxWorld for Robot Navigation and Control

Kirby's World



▸ Link

Conclusion - Robust Human-Computer communication

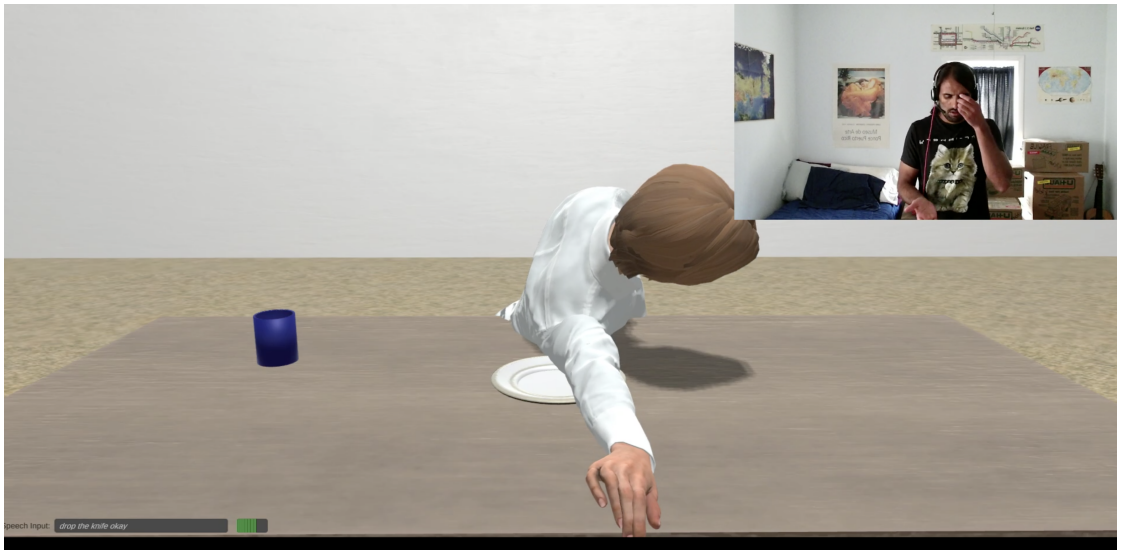
- In CwC we developed the notion of simulation semantics into situated meaning and situated grounding
  - A multimodal contextual encoding method based dynamic semantics and action and affordance recognition
  - Semantic simulations of an agent's utterance (Mind reading) and demonstrated understanding by enactment of that interpretation
- Multimodality provides a way to examine different models
  - Each modality provides a way to probe the models of other modalities
  - Provides task-appropriate data for diverse tasks (interaction, concept learning, dialogue, etc.)
- Deployed within framework of Neurosymbolic AI

Moving Forward and Future Work

- Continue developing Simulation Semantics and VoxWorld
  - collaboration between Brandeis and CSU
  - now with Nikhil Krishnaswamy at CSU
- Continue Dialogue Modeling with ARL HRI Corpus (SCOUT)
- Begin adapting Kirby's World to Scheutz's DIARC for PR2
- New NSF Funding to deploy VoxWorld and Simulation Semantics in developing Artificial General Intelligence
  - with Josh Hartshorne at Boston College
- Explore Virtual Reality and Augmented Reality implementations of VoxWorld and simulation semantics

VoxWorld (July 2020)

Diana being Diana – Still work to be done



▸ Link

CwC Related Publications

- Pustejovsky, J. and Krishnaswamy, N. (2020). Situated Meaning in Multimodal Dialogue: Human-Robot and Human-Computer Interaction. (Under Review)
- Krishnaswamy, N. and Pustejovsky, J. (2020). Neurosymbolic AI for Situated Language Understanding. In Annual Conference on Advances in Cognitive Systems (ACS). Cognitive Systems Found.
- Krishnaswamy, N. and Pustejovsky, J. (2020). A Formal Analysis of Multimodal Referring Expressions Under Common Ground. International Conference on Language Resources and Evaluation (LREC).
- Pustejovsky, J. and Krishnaswamy, N. (2020). Multimodal Communication with Computers and Robots, Avios Conversational Interaction Conference 2020, San Jose, CA.
- Krishnaswamy, N., Narayana, P., Bangar, R., Rim, K., Patil, D., McNeely-White, D., Ruiz, J., Draper, B., Beveridge, R., and Pustejovsky, J. (2020). Diana's World: A Situated Multimodal Interactive Agent. In AAAI Conference on Artificial Intelligence (AAAI): Demos Program.
- Hutchens, M., Krishnaswamy, N., Cochran, B., and Pustejovsky, J. (2020). Jarvis: A Multimodal Visualization Tool for Bioinformatic Data. In International Conference on Human-Computer Interaction (HCII): Late-Breaking Papers. Springer.

CwC Related Publications

- Krajovic, K., Krishnaswamy, N., Dimick, N. J., Salas, R. P., and Pustejovsky, J. (2020). Situated Multimodal Control of a Mobile Robot: Navigation through a Virtual Environment. In Special Session on Situated Dialogue with Virtual Agents and Robots (RoboDIAL): Late-Breaking Papers. ACL.
- Krishnaswamy, N. and Pustejovsky, J. (2019). Situated Grounding Facilitates Multimodal Concept Learning for AI. In Visually Grounded Interaction and Language Workshop (ViGIL). Neural Information Processing Systems.
- Krishnaswamy, N. and Pustejovsky, J. (2019). Multimodal Continuation-style Architectures for Human-Robot Interaction. In Workshop on Cognitive Vision: Integrated Vision and AI for Embodied Perception and Interaction. Cognitive Systems.
- Krishnaswamy, N. and Pustejovsky, J. (2019). Generating a Novel Dataset of Multimodal Referring Expressions. In International Workshop on Computational Semantics (IWCS). ACL.
- Lai, Kenneth, and James Pustejovsky. "A Dynamic Semantics for Causal Counterfactuals." In Proceedings of the 13th International Conference on Computational Semantics, pp. 1-8. 2019.

CwC Related Publications

- Krishnaswamy, N., Friedman, S., and Pustejovsky, J. (2019). Combining Deep Learning and Qualitative Spatial Reasoning to Learn Complex Structures from Sparse Examples with Noise. In AAAI Conf. on AI.
- Pustejovsky, J. and Krishnaswamy, N. (2019). Situational Grounding within Multimodal Simulations. In AAAI Workshop on Games and Simulations in AI (GameSim). AAAI.
- Krishnaswamy, N. and Pustejovsky, J. (2018). Deictic Adaptation in a Virtual Environment. In Spatial Cognition XI: International Conference on Spatial Cognition. Springer.
- Pustejovsky, J. and Krishnaswamy, N. (2018). The Role of Event Simulation in Spatial Cognition. In Workshop on Models and Representations in Spatial Cognition (MRSC). Springer.
- Narayana, P., Krishnaswamy, N., Wang, I., Bangar, R., Patil, D., Mulay, G., Rim, K., Beveridge, R., Ruiz, J., Pustejovsky, J., and Draper, B. (2018). Cooperating with Avatars Through Gesture, Language and Action. In Intelligent Systems Conference, IEEE.
- Do, T., Krishnaswamy, N., Rim, K., and Pustejovsky, J. (2018). Multimodal Interactive Learning of Primitive Actions. In AAAI Fall Symposium: Artificial Intelligence for Human-Robot Interaction. AAAI.
- Pustejovsky, J. and Krishnaswamy, N. (2018). Every Object Tells a Story. In Workshop on Events and Stories in the News (EventStory). ACL.

Thank You ☺

- Brandeis LLC lab members: Nikhil Krishnaswamy, Kyeongmin Rim, Mark Hutchens, Ken Lai, Katherine Krajovic, Daeja Showers, Eli Goldner, Kelley Lynch
- CSU Vision lab members: Ross Beveridge, Bruce Draper, Rahul Bangar, David White, Pradyumna Narayana, Dhruva Patil
- University of Florida lab members: Jaime Ruiz, Isaac Wang
- Funded by a grant from DARPA within the CwC Program