Situated Meaning in Multimodal Dialogue: Human-Robot and Human-Computer Interactions

James Pustejovsky* — **Nikhil Krishnaswamy****

* *Department of Computer Science, Brandeis University*
** *Department of Computer Science, Colorado State University*

ABSTRACT. *The demand for more sophisticated natural human-computer and human-robot interactions is rapidly increasing, as users become more accustomed to conversation-like interactions with their devices. This requires not only the robust recognition and generation of expressions through multiple modalities (language, gesture, vision, action), but also the encoding of situated meaning: (a) the situated grounding of expressions in context; (b) an interpretation of the expression contextualized to the dynamics of the discourse; and (c) an appreciation of the actions and consequences associated with objects in the environment. In this paper, we introduce VoxWorld, a multimodal simulation platform for modeling human-computer interactions. It is built on the language VoxML, and offers a rich platform for studying the generation and interpretation of expressions, as conveyed through multiple modalities, including: language, gesture, and the visualization of objects moving and agents acting in their environment.*

RÉSUMÉ. *La demande d'interactions naturelles homme-ordinateur et homme-robot plus sophistiquées augmente rapidement, car les utilisateurs s'habituent davantage aux interactions de type conversation avec leurs appareils. Cela nécessite non seulement la reconnaissance et la génération robustes d'expressions à travers de multiples modalités (langage, geste, vision, action), mais aussi l'encodage du sens situé : (a) l'ancrage situé des expressions dans le contexte; (b) une interprétation de l'expression contextualisée à la dynamique du discours; et (c) une appréciation des actions et des conséquences associées aux objets dans l'environnement. Nous présentons VoxWorld, une plateforme de simulation multimodale pour la modélisation des interactions homme-machine. Il est construit sur le langage VoxML et offre une plate-forme riche pour étudier la génération et l'interprétation d'expressions, telles qu'elles sont véhiculées à travers de multiples modalités, notamment : le langage, le geste et la visualisation des objets en mouvement et des agents agissant dans leur environnement.*

KEYWORDS: *Multimodal dialogue, affordances, qualia structure, continuations, gesture, simulations, common ground, situated meaning, semantic grounding, referring expressions.*

MOTS-CLÉS : *Dialogue multimodal, affordances, structure qualia, continuations, geste, simulations, terrain d'entente, sens situé, ancrage sémantique, expressions de référence.*

1. Introduction

When humans communicate with each other through language, there is a shared understanding of both an utterance meaning (content) and the speaker's meaning in the specific context (intent). The ability to link these two is the act of situationally grounding meaning to the local context, typically referred to as "establishing the common ground" between interlocutors (Stalnaker, 2002; Asher, 1998). Language use may reflect only a subset of all properties of the current situation, where a full description may be impossible or at least unwieldy. Some kinds of information may in fact be more efficiently communicated using other modalities, such as gesture (e.g., deixis for pointing), demonstration or action, images, or some other visual modality. A central component to the contextualized interpretation of meaning in a discourse is the situational determination of the meanings of expressions given the common ground. It is this notion of *situated meaning* that is missing in most current human-computer and human-robot interaction models, and the focus of the present paper.

In this paper, we argue that the problem of situational awareness and the creation of *situated meaning* in discourse involves at least three components: (a) the situated *grounding* of expressions in context; (b) an interpretation of the expression contextualized to the *dynamics* of the discourse; and (c) an appreciation of the *actions and consequences* associated with objects in the environment. In Section 2, we expand on these aspects of meaning in some detail, and then in Section 3, we adopt the modeling language, VoxML, designed to encode non-linguistic, multimodal aspects of meaning associated with concepts. In section 4, we present a computational framework, VoxWorld, within which these components are operationalized to facilitate multimodal communication between humans and robots or computers. Section 5 outlines a framework within which to interpret multimodal expressions, while Section 6 presents experimental evidence from single and mixed modality dialogues, illustrating the different ways in which meaning is situated in goal-directed dialogues.

2. Interactions in the Common Ground

There has been a growing interest in the Human-Robot Interaction community on how to contextually resolve ambiguities that may arise from communication in situated dialogues, from earlier discussions on how HRI dialogues should be designed (Fischer, 2011; Scheutz *et al.*, 2011), how perception and grounding can be integrated into language understanding (Landragin, 2006), to recent work on task-oriented dialogues (Williams *et al.*, 2019). This is the problem of identifying and modifying the *common ground* between speakers (Clark and Brennan, 1991; Stalnaker, 2002; Asher, 1998). It has long been recognized that an utterance's meaning is subject to contextualized interpretation; this is also the case with gestures in task-oriented dialogues. E.g., depending on the situation, an oriented hand gesture could refer either to an action request ("move it") or a dismissive response ("forget it") (Williams *et al.*, 2019). Even a request for action can be underspecified, denoting either a continuous movement or a movement to a specific location. Similarly, depending on the situation, the definite description in the command "Open the box." may uniquely refer or not, depending on how many boxes are in the context. These and similar miscommunications or the need for clarification in dialogue have been

called *situated grounding problems* (Marge and Rudnicky, 2013), and can be viewed as problematic in a model that appeals to and encodes both a visual modality and situational information into the dialogue state. What the occurrence of these issues makes apparent is the complexity underlying the interpretation of referential expressions in actual situated dialogues. The richness provided by situationally grounding computer or robot behaviors brings to the surface interpretive questions similar to those of a human in the same scenario.

Some recent efforts have been made to provide contextual grounding to linguistic expressions. For example, work on "multimodal semantic grounding" within the natural language processing and image processing communities has resulted in a number of large corpora linking words or captions with images (cf. Chai *et al.* (2016)). In this paper, we argue that language understanding and linking to abstract instances of concepts in other modalities is insufficient; *situated grounding* entails knowledge of situation and contextual entities beyond that provided by a multimodal linking approach (cf. Kennington *et al.* (2013)).

Actual situated meaning is much more involved than aligning captions and bounding boxes in an image: e.g., Hunter *et al.* (2018) discuss the contribution of non-linguistic events in situated discourse, and also whether they can be the arguments to discourse relations. Similarly, it is acknowledged that gesture is part of either the direct content of the utterance (Stojnić *et al.*, 2019) or cosuppositional content (Schlenker, 2020). Hence, we must assume that natural interactions with computers and robots have to account for interpreting and generating language and gesture.

Figure 1. *Mother and son interacting in a shared task of icing cupcakes.*

Consider the joint activity shown in Fig. 1 above between a mother and her son, where they are engaged in icing cupcakes in a kitchen setting. The dialogue in Fig. 2 illustrates some possible multimodal expressions used in such a context of joint activity between two agents.

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).*

Figure 2. *Dialogue.*

Viewed as a multi-agent collaborative task interaction, there are some obvious elements constituting the common ground between the two agents in Fig. 1. These include reference to: the participants (agents); shared beliefs and assumptions; shared goals and intentions; the accompanying objects in the situation; the shared perception

of these objects; and the surrounding space within which the situation unfolds. Some of these elements are given below in Fig. 3.

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

Figure 3. *Elements from the common ground for Figure 1.*

From this example, it is apparent that we can identify three core aspects of meaning that contribute to the common ground in a multimodal dialogue:

1) *co-situatedness* and *co-perception* of the agents, such that they can interpret the same situation from their respective frames of reference. This might be a human and an avatar perceiving the same virtual scene from different perspectives; or a human sharing the perspective of a robot as it navigates through a disaster zone;

2) *co-attention* of a shared situated reference, which allows more expressiveness in referring to the environment (i.e., through language, gesture, visual presentation, etc.). The human and avatar might refer to objects in multiple modalities with a common model of differences in perspective-relative references (e.g., "your left, my right"); or the human sharing the robot's perspective might be able to direct its motion using reference in natural language ("go through the second door on the left") or gesture ("go this way," with pointing);

3) *co-intent* of a common goal, such that misaligned relationships between agents reflect a breakdown in the common ground. A human and avatar interacting around a table might seek to collaborate to build a structural pattern known to one or both of them; or the human and robot sharing perspective both have a goal to free someone trapped behind a door in a fire. The robot informs the human about the situation and the human helps the robot problem-solve in real time until the goal is achieved.

What this suggests is that any robust communication between humans and computers or robots will require at least three capabilities: (a) a robust recognition and generation within multiple modalities; (b) an understanding of contextual grounding and co-situatedness in the conversation; and (c) an appreciation of the consequences of behavior and actions taking place throughout the dialogue. To this end, in our work, we have developed a platform making use of semantically interpreted *multimodal simulations*, which provides an approach to modeling human-computer communication by both situating and contextualizing the interaction, thereby visually demonstrating what the co-agent computer or robot is hearing, seeing, thinking, and doing. This platform is based on VoxML, a modeling language for encoding traditionally non-linguistic, multimodal, aspects of meaning associated with the objects that we encounter, manipulate, and explore in our environment. We turn to this discussion in the next section.

3. VoxML: Encoding Knowledge of Action and Behavior

Here we argue that a significant part of any model for situated communication is an encoding of the semantic type, functions, purposes, and uses introduced by the objects under discussion. I.e., a semantic model of perceived *object teleology*, as introduced by Qualia Structure, for example (Pustejovsky, 1995), as well as *object affordances* (Gibson, 1977) is needed to help ground expression meaning to speaker intent.

Objects under discussion in discourse (cf. Ginzburg (1996)) can be partially contextualized through their semantic type and their qualia structure: e.g., a food item has a TELIC value of eat, a pencil, a TELIC of $write$, a box, a CHAIR of sit_in, and so forth. However, while an artifact may be designed for a specific purpose, this can only be achieved under specific circumstances. To account for this context-dependence, Pustejovsky (2013) enriches the lexical semantics of words denoting artifacts (the TELIC role specifically) by introducing the notion of an object's *habitat*, which encodes these circumstances. For example, an object, x, within the appropriate context $\mathcal{C}$, performing the action π will result in the intended or desired resulting state, $\mathcal{R}$, i.e., $\mathcal{C} \rightarrow [\pi]\mathcal{R}$. That is, if the habitat $\mathcal{C}$ (a set of contextual factors) is satisfied, then every time the activity of π is performed, the resulting state $\mathcal{R}$ will occur. The precondition context $\mathcal{C}$ is necessary to specify, since this enables the local modality to be satisfied.

The habitat for an object is situated within an *embedding space* and then contextualized within it. For example, in order to use a glass to drink from, the concavity has to be oriented upward, the interior must be accessible, and so on. Similarly, a chair must also be oriented up, the seat must be free and accessible, it must be large enough to support the user, etc. An example of what the resulting knowledge structure for the habitat of a chair is shown below, where these constraints are superscripted with " *".

These distinctions in habitats facilitate both Gibsonian and telic affordances and transfer learning of Gibsonian affordances relies on information taken from telic affordances (its use for sitting), and vice versa (see Section 6.4): below, the F and C values specify size and part structure, respectively.

$$(1)\ \lambda x \left[\begin{array}{l} \textbf{chair(x)} \\ \text{F} = [phys(x), on(x, y_1)^*, in(x, y_2)^*, clear(x_1)^*, orient(x, up)^*, \\ \qquad support(x_1, y_3)^*] \\ \text{C} = [seat(x_1), back(x_2), legs(x_3)] \\ \text{T} = \lambda z \lambda e[\mathcal{C} \rightarrow [sit(e, z, x)]\mathcal{R}_{sit}(x)] \\ \text{A} = [made(e', w, x)] \end{array}\right]$$

The notion of habitat and the attached behaviors that are associated with an object are further developed in Pustejovsky and Krishnaswamy (2016), where an explicit connection to Gibson's ecological psychology is made, along with a direct encoding of the *affordance structure* for the object (Gibson, 1977). The affordance structure available to an agent, when presented with an object, is the set of actions that can be performed with it. We refer to these as GIBSONIAN affordances, and they include "grasp", "move", "hold", "turn", etc. This is to distinguish them from more goal-directed, intentionally situated activities, what we call TELIC affordances.

VoxML (Visual Object Concept Modeling Language) is a modeling language for constructing 3D visualizations of concepts denoted by natural language expressions,

and is being used as the platform for creating multimodal semantic simulations in the context of human-computer and human-robot communication (Pustejovsky and Krishnaswamy, 2016; Krishnaswamy and Pustejovsky, 2016). It adopts the basic semantic typing for objects and properties from Generation Lexicon and the dynamic interpretation of event structure developed in Pustejovsky and Moszkowicz (2011), along with a continuation-based dynamic interpretation for both sentence and discourse composition (De Groote, 2001; Barker and Shan, 2014; Asher and Pogodalla, 2010).

VoxML forms the scaffolding we use to encode knowledge about objects, events, attributes, and functions by linking lexemes to their visual instantiations, termed the "visual object concept" or *voxeme*. Voxemes representing humans or IVAs are lexically typed as *agents*, but agents, due to their embodiments, ultimately inherit from physical objects and so fall under objects in the taxonomy. In parallel to a lexicon, a collection of voxemes is termed a *voxicon*. There is no requirement on a voxicon to have a one-to-one correspondence between its voxemes and the lexemes in the associated lexicon, which often results in a many-to-many correspondence. That is, the lexeme *plate* may be visualized as a [[SQUARE PLATE]], a [[ROUND PLATE]], or other voxemes, and those voxemes in turn may be linked to other lexemes such as *dish* or *saucer*. Each voxeme is linked to either an object geometry, a program in a dynamic semantics, an attribute set, or a transformation algorithm, which are all structures easily exploitable in a rendered simulation platform.

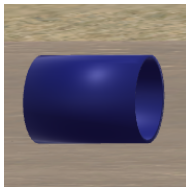


Figure 4. *Cup in habitat allowing rolling.*

An OBJECT voxeme's semantic structure provides *habitats*, which are situational contexts or environments conditioning the object's *affordances*, which may be either "Gibsonian" affordances (Gibson, 1977) or "Telic" affordances (Pustejovsky, 1995; Pustejovsky, 2013). A habitat specifies how an object typically occupies a space. When we are challenged with computing the embedding space for an event, the individual habitats associated with each participant in the event will both define and delineate the space required for the event to transpire. Affordances are used as attached behaviors, which the object either facilitates by its geometry (Gibsonian) or purposes for which it is intended to be used (Telic). For example, a Gibsonian affordance for [[CUP]] is "grasp," while a Telic affordance is "drink from." This allows procedural reasoning to be associated with habitats and affordances, executed in real time in the simulation, inferring the complete set of spatial relations between objects at each frame and tracking changes in the shared context between human and computer.

Indeed, object properties and the events they facilitate are a primary component of situational context. In Fig. 4, we understand that the cup in the orientation shown can be *rolled* by a human. Were it not in this orientation, it might be able to be only *slid* across its supporting surface (cf. (2)). This voxeme for [[CUP]] gives the object appropriate lexical predicate and typing (a *cup* is a PHYSICAL OBJECT and an ARTIFACT). It denotes that the cup is roughly cylindrical and concave, has a surface and an interior, is symmetrical around the Y-axis and across associated planes (VoxML

adopts 3D graphics convention where the Y-axis is vertical), and is smaller than and movable by the agent. The remainder of VoxML typing structure is devoted to habitat and affordance structures, which we discuss below.

(2) Objects encoding semantic type, habitat, and affordances:

$$\left[\begin{array}{l}\textbf{cup}\\ \textsc{lexical} = \left[\begin{array}{l}\textsc{predicate} = \textbf{cup}\\ \textsc{type} = \textbf{physobj, artifact}\end{array}\right]\\ \textsc{type} = \left[\begin{array}{l}\textsc{head} = \textbf{cylindroid[1]}\\ \textsc{components} = \textbf{surface, interior}\\ \textsc{concavity} = \textbf{concave}\\ \textsc{rotational\_symmetry} = \{Y\}\\ \textsc{reflection\_symetry} = \{XY, YZ\}\end{array}\right]\\ \textsc{habitat} = \left[\begin{array}{l}\textsc{intrinsic} = [2]\left[\begin{array}{l}\textsc{constr} = \{Y > X, Y > Z\}\\ \textsc{up} = align(Y, \mathcal{E}_Y)\\ \textsc{top} = top(+Y)\end{array}\right]\\ \textsc{extrinsic} = [3]\left[\textsc{up} = align(Y, \mathcal{E}_{\perp Y})\right]\end{array}\right]\\ \textsc{affordance\_structure} = \left[\begin{array}{l}\mathrm{A}_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x)\\ \mathrm{A}_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x)\\ \mathrm{A}_3 = H_{[2]} \rightarrow [grasp(x, [1])]hold(x, [1])\\ \mathrm{A}_4 = H_{[3]} \rightarrow [roll(x, [1])]\mathcal{R}\end{array}\right]\\ \textsc{embod} = \left[\begin{array}{l}\textsc{scale} = <\textbf{agent}\\ \textsc{movable} = \textbf{true}\end{array}\right]\end{array}\right]$$

In VoxML encodings like 2, bracketed numbers, e.g., [1] are reentrancy indices, such that terms annotated with the same number refer to the same entity. For instance, in habitat 2 ($H_{[2]}$), the intrinsic habitat where the cup has an upward orientation, if an agent puts some x inside the cup's cylindroid geometry ([1]), the cup contains x.

One of the major improvements to the notion of habitat developed in VoxML over that given originally in Pustejovsky (2013) is how the preconditions to actions are encoded and scoped. Notice how in the example in (1), the constraint on relative size of the chair to its user (along with all constraints) is specified outside the modal context in the TELIC, while the VoxML representation using Habitats in (3) provides a reentrant binding for the situational variables.

(3) Habitat and affordance structure for *chair*:

$$\left[\begin{array}{l}\textbf{chair}\\ \textsc{habitat} = \left[\begin{array}{l}\textsc{intr} = [2]\left[\begin{array}{l}\textsc{constr} = \{Y > X, Y > Z\}\\ \textsc{up} = align(Y, \mathcal{E}_Y)\\ \textsc{top} = top(+Y)\end{array}\right]\end{array}\right]\\ \textsc{afford\_str} = \left[\mathrm{A}_1 = H_{[2]} \rightarrow [sit(y, on([1]))]support([1], y)\right]\end{array}\right]$$

VoxML treats actions and events within a dynamic event semantics as programs (Pustejovsky and Moszkowicz, 2011; Mani and Pustejovsky, 2012). The advantage of adopting a dynamic interpretation of events is that one can map linguistic expressions directly into simulations through an operational semantics (Miller and Johnson-Laird, 1976). Models of processes using updating typically make reference to the notion of

a state transition (Harel, 1984). Each event, such as *put* in (4), can be seen as a traced structure over a Labeled Transition System. The approach is similar in many respects to that developed in both Fernando (2009) and Naumann (2001).

This also allows the system to reason about objects and actions independently. When simulating the objects alone, the simulation presents how the objects change in the world. By removing the objects and presenting only the actions that the viewer would interpret as *causing* the intended object motion (i.e., an embodied agent pantomiming the object motion), the system presents a "decoupled" interpretation of the action, for example, as an animated gesture that traces the intended path of motion. By composing the two, it demonstrates a particular instantiation of the complete event. This allows an embodied situated simulation approach to easily compose objects with actions by directly interpreting at runtime how the two interact.

For the simulation to run, all parameters (e.g., object location, agent motion, etc.) must have values assigned. The simulation environment itself facilitates the calculation of these values, including a common path that the object and agent's manipulator must follow while completing an action; adhering to these common paths and positional values keeps the two synchronized.

(4) Events as Programs:

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

The logic of event structure encodes only minimal temporal constraints on how the subevents interact or play out. The rendering engine itself maintains an internal clock and regulates frame rate, and therefore the time it takes to conduct movements, obviating the need to regularly model this temporal aspect in operationally defined events in VoxML, although scalara attributives like *faster* or *slower* can provide temporal modifiers.

4. VoxWorld: A Platform for Multimodal Simulations

In this section, we introduce a simulation framework, VoxWorld, that situates an embodied agent in a multimodal simulation, with the capability of understanding and generating language and gesture, and the ability to synthetically perceive an interlocutor human as well as objects in its virtual surroundings, and act on them through a limited inventory of actions.

4.1. *Modes of Simulation*

The concept of simulation has played an important role in both AI and cognitive science for over forty years. The two most common uses for the term *simulation* as used in computer science and AI include: (a) *computational simulation modeling*, where variables in a model are set, the model is run, and the consequences of all possible computable configurations become known; and (b) *situated embodied simulations*, where an environment allows a user to interact with objects in a "virtual or simulated world", where the agent is embodied as a dynamic point-of-view or avatar in a proxy situation. Such simulations are used for training humans in scripted scenarios, such as flight simulators, battle training, and of course, in video gaming, where the goal is to simulate an agent within a situation.

Simulation has yet another meaning, where starting with Craik (1943), we encounter the notion that agents carry a mental model of external reality in their heads. Johnson-Laird (1987) develops his own theory of a mental model, which represents a situational possibility, capturing what is common to all the different ways in which the situation may occur. This is used to drive inference and reasoning, both factual and counterfactual. Simulation Theory, as developed in philosophy of mind, has focused on the role "mind reading" plays in modeling the mental representations of other agents and the content of their communicative acts (Goldman, 2006). Simulation semantics (Feldman, 2010; Narayanan, 2010) argues that language comprehension is accomplished by means of such mind reading operations. Similarly, within psychology, there is an established body of work arguing for "mental simulations" of future or possible outcomes, as well as interpretations of perceptual input (Barsalou, 1999). These approaches we refer to as *embodied theories of mind*.

4.2. *VoxWorld*

VoxWorld integrates the functionality and the goals of all three approaches above. The platform situates an embodied agent in a multimodal simulation, with *mind-reading* interpretive capabilities, facilitated through assignment and evaluation of object and context parameters within the environment being modeled.

4.2.1. *Architecture*

VoxWorld is based on the semantic scaffold provided by the VoxML modeling language (Pustejovsky and Krishnaswamy, 2016), which provides a dynamic, interpretable model of objects, events, and their properties. This allows us to create visualized simulations of events and scenarios that are rendered analogues to the "mental simulations" discussed above. We can restrict mind-reading to events that are tangible and perceptually reflective or transparent. So, mental events (desires, beliefs by themselves, etc.) will not be modeled here as simulations themselves, but rather as modal signatures or propositional content of a common ground—that is an agent's desire for food may manifest as holding their stomach or opening the refrigerator, themselves modeled as distinct events stemming from that cause. VoxSim (Krishnaswamy and Pustejovsky, 2016) serves as the event simulator within which these simulations are created and rendered in real time, serving as the computer's method of visually presenting its interpretation of a situation or event. Because modalities are modes of

presentation, a multimodal simulation entails as many presentational modes as there are modalities being modeled. The visual modality of presentation (as in embodied gaming) necessitates "situatedness" of the agent, as do the other perceptual modalities. Therefore, when we speak of *multimodal simulations*, they are inherently situated. In a human-computer interaction using such a simulation, the simulation is a demonstration of the computational agent's "mind-reading" capabilities (an *agent simulation*). If the two are the same (where the agent is a proxy for the player or user), then the "mind-reading" is just a demonstration of the scenario. It observes what the user can. The user observes the agent act as if they share the same perspective. If, on the other hand, the two are separate (agent is *not* proxy for the user), then the simulation/demonstration communicates the agent's understanding of the user and the interaction. In this case, this demonstration entails the illustration of both epistemic and perceptual content of the agent. The agent's actions within the scene facilitate the human's "mind-reading" based on the agent's demonstrated interpretation of propositional content within the scene. We assume an agent has present epistemic knowledge and the relevant inferences reasonably associated with/derivable from these propositions. The agent may know that an object is graspable and can be held in a certain way. This also means that the agent "knows" that it is touchable and moveable, similarly for propositional knowledge associated with logical entailments, etc.

The current architecture of the VoxWorld system is shown in Fig. 5.

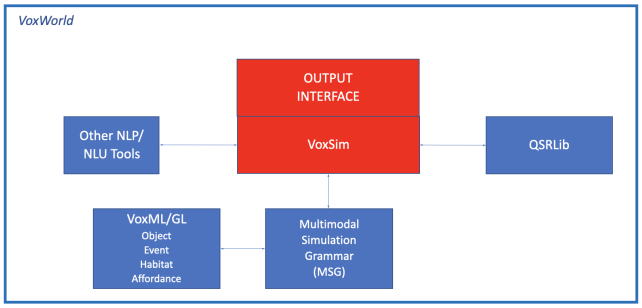


Figure 5. *VoxWorld Architecture schematic.*

At the center is VoxSim, the software that handles visual event simulation in three dimensions, written with the Unity game engine. VoxSim connects to a number of other default VoxWorld components, including some native natural language processing capabilities, VoxML encodings/GL knowledge as interpreted through the multimodal semantics discussed in Section 5, and 3rd-party libraries, e.g., QSRLib (Gatsoulis *et al.*, 2016). Individual agent, such as the interactive avatar Diana (discussed below), are arbitary output interfaces that can also connect to 3rd-party endpoints; in the case of Diana, this is custom gesture and affect recognition (Narayana *et al.*, 2018).

4.2.2. *Usage*

VoxSim contains scenes in a Blocks World domain, plus a set of more complicated or interesting everyday objects (e.g., cups, plates, books, etc.). In scenes without an avatar, the user can direct the computer to manipulate objects in space or create an avatar that can act upon objects and respond to the user's input. VoxWorld includes

other software, models, and interfaces, e.g., to consume input from CNN-based gesture recognizers (Narayana *et al.*, 2018), or to track the agent's epistemic state or knowledge of what its interlocutor knows.

It is straightforward to create new scenes with 3D geometries with packaged code that creates and instantiates voxemes, handles their interactions and performs basic spatial reasoning over them. VoxWorld contains a library of basic motion predicates and methods to compose them into more complex actions using VoxML.

4.2.3. *Situated Reasoning in VoxWorld*

Situational embodiment takes place in real time, so in a situation where there may be too many variables to predict the state of the world at time t from initial conditions at time 0, situational embodiment within the simulation allows the agent to reason forward about a specific subset of consequences of actions taken at time t, given the agent's current conditions and surroundings. Situatedness and embodiment is required to arrive at a complete, tractable interpretation given any element of non-determinism. E.g., an agent trying to navigate a maze from start to finish could easily do so with a map that provides complete or sufficient information about the scenario. However, if the scene is disrupted (e.g., the floor crumbles, or doors open and shut randomly), the agent would be unable to plot a course to the goal. It would have to start moving, assess circumstances at every timestep, and choose the next move(s) based on them. Situated embodiment allows the agent to assess the next move based on the current set of relations between itself and the environment (e.g., ability to move forward but not leftward at the current state). This allows reasoning that saves computational resources and performs more analogously to human reasoning.

Given the continuous tracking of object parameters such as position and orientation, facilitated by a game engine or simulation, and the knowledge of object, event, and functional semantics facilitated by a formal model, an entity's interpretation at runtime can be computed in conjunction with the other entities it is currently interacting with and their properties. One such canonical example would be placing an object [[SPOON]] in an [[IN]] relation with another object [[MUG]] (Fig. 6).

Figure 6. *[[*SPOON IN MUG*]].*

The mug has an intrinsic top, which is aligned with the upward Y-axis of the world or embedding space (denoted in VoxML as $\{align(Y, \mathcal{E}_Y)$, $top(+Y)\}$). The mug is a concave object, and the mug's geometry (the [[CUP]], excluding the handle) has reflectional symmetry across its inherent (object-relative) XY- and YZ-planes, and rotational symmetry around its inherent Y-axis such that when the object is situated in its inherent top habitat, its Y-axis is parallel to the world's. From this we can infer that the *opening* (e.g., access to the concavity) must be along the Y-axis. Encoding the object's concavity allows fast computation for physics and collisions using bounding boxes, while still facilitating reasoning over concave objects.

An embodied simulation model such as VoxWorld is an approach that integrates all three aspects of simulation: a situated embodied environment built on a game engine platform. The computer, either as an embodied agent distinct from the viewer,

or as the totality of the rendered environment itself, presents an interpretation (*mind-reading*) of its internal model, down to specific parameter values, which are often assigned for the purposes of testing that model. As such, it provides a rich environment within which to experiment with task-oriented dialogues, such as those explored in Section 6, because of the requirement that the agent have a situated embodiment in which it interprets its environment and its interlocutor. This in turn requires the creation of common ground (CG) between the human and the AI that allows them to communicate. The parameters within this CG structure can be varied and set according to various experimental configurations, allowing us to both qualitatively and quantitatively measure the effect of different CG structures on the communication. For example, we can experiment with variable settings for the composition of multimodal referring descriptions as well as action or event predicates; that is, what aspects of the content of the expression are conveyed through each modality, speech or gesture? Another variation involves the degree of alignment of information in each modal channel; that is, whether a linguistic expression and gesture are synchronous or asynchronous when generated. The interaction in Fig. 7 illustrates a person directing an avatar to pick up a block, using an asynchronous multimodal expression.

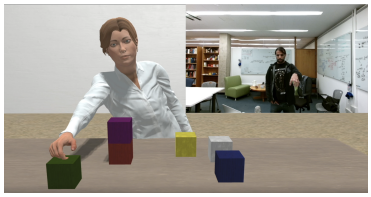


Figure 7. *Asynchronous ensemble dialogue: Human grasping gesture precedes his linguistic utterance, "Grab it".*

We assume that a simulation is 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 discourse between them. The encoding that VoxML provides for objects, with its rich semantic typing and action affordances, enables VoxWorld to describe agent actions as multimodal programs, as well as identifying and tracking the elements of the common ground that are revealed in the interaction between parties, be they humans or artificially intelligent agents.

5. Multimodal Semantics for Common Ground

The theory of common ground has a rich and diverse literature concerning what is shared or presupposed in human communication (Clark and Brennan, 1991; Stalnaker, 2002; Asher, 1998; Ginzburg and Fernández, 2010). With the presence of a common ground during shared experiences, embodied communication assumes agents can understand one another in a shared context, through the use of co-situational and co-perceptual anchors, and a means for identifying such anchors, such as gesture, gaze, intonation, and language. In this section, we develop a computational model of common ground for multimodal communication.

We assume generally a model of discourse semantics as proposed in Asher and Lascarides (2003), as it facilitates the adoption of a continuation-based semantics for our phrase-level compositional semantics (Barker and Shan, 2014), as well for discourse, as outlined in De Groote (2001) and Asher and Pogodalla (2010). For the present discussion, however, we will not refer to SDRT representations, but focus

instead on the semantics integrated multimodal expressions in the context of task oriented dialogue, as presented first in (Pustejovsky, 2018) and extended here.

Here, we introduce the notion of a *common ground structure*, the information associated with a state in a dialogue or discourse. We model this as a state monad (Unger, 2011), as illustrated in (5).

(5) State Monad: $\mathbf{M}\alpha = State \rightarrow (\alpha \times State)$

A state monad corresponds to computations that read and modify a particular state, in this case 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. This monad consists of the following state information:

(6) a. the communicative act, C_a, performed by an agent, a: a tuple of expressions from the modalities involved. For our present discussion, we restrict this to a linguistic utterance, S (speech) and a gesture, G. There are hence three possible configurations in performing a C: $C_a = \{(G), (S), (S, G)\}$;
b. **A**: the agents engaged in communication;
c. **B**: the shared belief space;
d. **P**: the objects and relations that are jointly perceived in the environment;
e. $\mathcal{E}$: the embedding space that both agents occupy in the communication.

The common ground structure (CGS) can be represented graphically as in (7), where an agent, a_i, makes a communicative act either through gesture, $\mathcal{G}$ in (7a), or linguistically, as in (7b.) [1]

(7) a. $\dfrac{\textbf{A:}a_1, a_2 \;\; \textbf{B:}\Delta \;\; \textbf{P:}b \;\; \mathcal{E} : E}{\mathcal{G}_{a_1}}$ b. $\textbf{A:}a_1, a_2 \;\; \textbf{B:}\Delta \;\; \textbf{P:}b \;\; \mathcal{E} : E$
$\mathcal{S}_{a_1}$ = "You$_{a_2}$ see it$_b$"

(7a) specifies that two agents, a_1 and a_2, co-inhabiting an embedding space, E, within which the experience is embodied, share a set of beliefs, Δ, where they can both see the object, b. Given this representation, the gesture is now situated to refer to objects and knowledge within the CG structure. In (7b), the linguistic expression, $\mathcal{S}_{a_1}$, is grounded relative to the parameters of common ground, where the indexical *you* will denote the agent, a_2, and the pronoun *it* will denote the object, b.

We have augmented and extended the approach taken in Kendon (2004) and Lascarides and Stone (2009), where gestures are simple schemas consisting of distinct sub-gestural phases, where **Stroke** is the content-bearing phase of the gesture.

(8) $G \rightarrow (Prep)\ (Pre\_stroke\ Hold)\ \mathbf{Stroke}\ Retract$

[1] This is similar in many respects to the representations introduced in Cooper and Ginzburg (2015), Ginzburg and Fernández (2010) and Dobnik *et al.* (2013) for modeling action and control with robots.

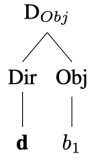
In the context of multimodal dialogues and interactions with computational agents and robots, gesture's **Stroke** will denote a range of primitive action types, $\mathcal{ACT}$, e.g., *grasp*, *pick up*, *move*, *throw*, *pull*, *push*, *separate*, and *put together*. There are many ways to convey intent to carry out these actions, but they all involve two characteristics: (a) the action's object is an embodied reference in the common ground; and (b) the gesture sequence must be interpreted dynamically, to correctly compute the end state of the event. To this end, we model two kinds of gestures in our dialogues: (a) establishing a reference; and (b) depicting an action-object pair.

(9) a. **Deixis**: $D_{obj} \rightarrow Dir \;\; Obj$
b. **Action**: $G_{Af} \rightarrow Act \;\; Obj$

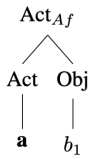
We introduce the notion of an interpreted gesture tree in (10a), which indicates that the gesture D_{obj} functionally consists of a deictic orientation, Dir, with the demonstratum, **d**, and the referenced or denoting entity, Obj, denoting b_1.

(10) Interpreted Gesture Tree:

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



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



As gesture is intended for visual interpretation, it is directly interpretable by the interlocutor in the context if and only if the value is clearly evident in the common ground, most likely through visual inspection. Directional or orientational information conveyed in a gesture identifies a distinct object or area of the embedding space, E, by directing attention to the End of the designated pointing ray (or cone) trace (Lascarides and Stone, 2009; Lücking *et al.*, 2015; Pustejovsky, 2018).

(11) $[\![\mathbf{D}_{obj}]\!] = [\![End(ray(\mathbf{d}))]\!]$

We model the interpretation function, $[\![.]\!]$, as fully determining the value of the deixis in the context, supplied by the common ground, which we discuss below. In (10b), the action gesture type, G_{Af}, consists of an action-object pairing, where the action, **a**, is applied to the object, b_1, in some prototypical manner. The strategies available are outlined in (12-14).

(12) a. ACTION-OBJECT: e.g., *grab* [**Object**]
b. $GvP_1 \rightarrow G_{Af} \;\; D_{obj}$ (Action Focus)
$\rightarrow \;\; D_{obj} \;\; G_{Af}$ (Object Focus)

(13) 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)

(14) a. ACTION-RESULT: e.g., *move* [**Object**] [**Direction**]
b. $GvP_3 \rightarrow G_{Af} \;\; D_{obj} \;\; D_{dir}$

As mentioned above, the deictic gesture in (9a) and (10a) actually serves to indicate both a location and objects within that location, suggesting that deixis denotes a *dot object*, viz., PHYSOBJ•LOCATION (Pustejovsky, 1995). Either of these type components may be exploited by the deictic reference, which is then interpreted in context, either as a selection (exploiting the PHYSOBJ) or as a destination (exploiting either). For example, should an object b_1 already be selected through a deixis $\mathbf{d}_a$, as in (10a), a subsequent deixis $\mathbf{d}_b$ may be interpreted as selecting a destination location in isolation (in which case the interpretation exploits the LOCATION of $\mathbf{d}_b$), or as selecting a location relative to another object (exploiting the PHYSOBJ type of $\mathbf{d}_b$). We discuss this further below.

With conventional treatments of continuation-style passing within the utterance, all linguistic expressions are continuized within the sentence. This has a distinct advantage in multimodal processing, because it allows for an *informational distribution* among the expressions being used in composition to form larger meanings.

By treating the common ground as a state monad, as described above, we can continuize the composition above the level of the sentence as well. Following De Groote (2001), Asher and Pogodalla (2010) and further developments in Van Eijck and Unger (2010), we represent a context as a stack of items and the type of left contexts to be lists of entities, $[e]$. Right contexts will be interpreted as continuations: a discourse that requires a left context to yield a truth value. The type of a right context is therefore $[e] \to t$. Hence, context transitions get the type $[e] \to [e] \to t$; they are characteristic functions of binary relations on contexts. The continuized semantics for gesture phrases is in (15).

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

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

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

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

The discourse updating operation is accomplished through continuation-passing as well, as in (Asher and Pogodalla, 2010). We apply a CPS transformation to arrive at the continuized type for each expression, notated as an overlined expression (Van Eijck and Unger, 2010). Given the current discourse, T, and the new utterance, C, we take the integration of C into T as follows:

(16) $[\![\overline{(\mathbf{T}.\mathbf{C})}]\!]^{M,cg} = \lambda k.[\![\overline{\mathbf{T}}]\!](\lambda n.[\![\overline{\mathbf{C}}]\!](\lambda m.k(m\ n)))$

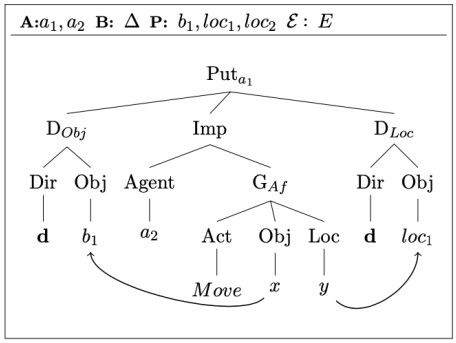
To illustrate how continuations help in the interpretation of gesture sequences, consider a single modality gesture imperative.

SINGLE MODALITY (GESTURE) IMPERATIVE

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

Through its own continuation, the referent identified in the first deixis, $\mathbf{D}_{Obj}$, is passed to the action ($\lambda k.k([\![\mathbf{Move}]\!])$), while the continuized interpretation of the action delays the computation of its argument until the appropriate binding has been identified. Finally, the goal location for the movement selected for by the *move* gesture is identified through the action of the continuized location deixis, $\mathbf{D}_{Loc}$. This is illustrated in (18), along with the common ground structure that is computed, shown in (17).

(17)



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

Given a description of the gesture grammar as used in our multimodal dialogues, let us explore a communicative act that exploits a combination of both speech and gesture, (S, G). We identify three configurations for how a language-gesture *ensemble* can be interpreted, depending on which modality carries the majority of semantic content: (a) language with *co-speech gesture*, where language conveys the bulk of the propositional content and gesture adds situated grounding, affect, effect, and presuppositional force (Cassell *et al.*, 2000; Lascarides and Stone, 2009; Schlenker, 2020); (b) *co-gestural speech*, where gesture plays this role (Pustejovsky, 2018); and (c) a truly mixed modal expression, where both language and gesture contribute equally to the meaning. In practice, while many of the interaction in our dialogues have this property, the discourse narrative is broadly guided by gesture. For this reason, we model the multimodal interactions as content-bearing gesture with *co-gestural speech*.

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. Let us assume that a linguistic subexpression, s, is either a word or full phrase in the utterance, while a gesture, g, comports with the gesture grammar described above.

(19) **Co-gestural Speech Ensemble**:
$$\begin{bmatrix} \mathcal{G} & g_1 & \dots & g_i & \dots & g_n \\ \mathcal{S} & s_1 & \dots & s_i & \dots & s_n \end{bmatrix}$$

We assume an aligned language-gesture syntactic structure, for which we provide a continuized semantic interpretation. Both

of these are contained in the common ground state monad introduced above (6). For each temporally indexed and aligned gesture-speech pair, (g, s), we have a continuized interpretation, as shown below. Each modal expresssion carries a continuation, k_g or k_s, and we denote alignment of these two continuations as $k_s \otimes k_g$, seen (20).

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

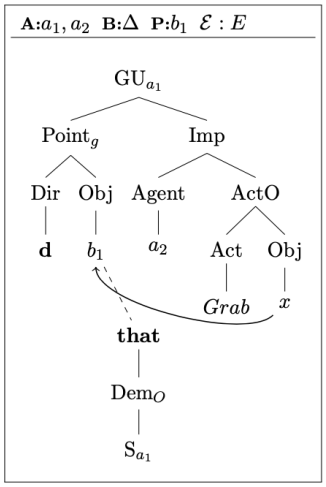
We bind co-gestural speech to specific gestures in the communicative act, within a common ground, CGS. A dashed line in (21) indicates that a co-gestural speech element, $\mathcal{S}$, is aligned with a particular gesture, $\mathcal{G}$. For example, consider the co-gestural speech expression below.

The CG structure for this expression, $\begin{bmatrix} \mathcal{G} & D_{Obj} & Grab_g \\ \mathcal{S} & \text{THAT} & \_\_\_ \end{bmatrix}$, is shown in (21).

SINGLE MODALITY (GESTURE) IMPERATIVE

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

(21) $[\![\langle \text{THAT}, \mathbf{D}_{Obj}\rangle.\langle \_, \mathbf{Grab}\rangle]\!] = \lambda k_s \otimes k_g.([\![\mathbf{D}_{Obj}]\!]; \lambda j_g.((([\![\mathbf{Grab}]\!]j_g)k_s \otimes k_g))$



Common ground updates will also include executing modal operations over the belief space **B**, where each new element from the discourse is introduced via a *public announcement logic* (PAL) formula, and each new perceived object or relation is introduced into **P** via an analogous *public perception logic* (PPL) formula (Plaza, 2007; Van Ditmarsch *et al.*, 2007; Van Benthem, 2011). We will use $[\alpha]\varphi$ to denote that an agent "α *knows* φ". Public announcements are implemented as: $[!\phi_1]\phi_2$. Any proposition, φ, in the common knowledge held by two agents, α and β, is computed as: $[(\alpha \cup \beta)^*]\varphi$.

Similarly, an agent α's perception is encoded as sets of accessibility relations, α, between situations. What is seen in a situation is encoded as either a proposition, φ, or existential statement of an object, x, $\hat{x}$. $[\alpha]_\sigma\varphi$ denotes that agent "α perceives that φ". $[\alpha]_\sigma\hat{x}$ denotes that agent "α perceives that there is an x."

(22) a. **block**: Pick me up!, Move me!
b. **cup**: Pick me up!, Drink what's in me!
c. **knife**: Pick me up!, Cut that with me!

Given the theory of two-level affordances proposed here (Gibsonian/Telic), we can naturally think of objects as *antecedents to the actions performable on them*. For each object in (22), we identify attached behaviors. This naturally suggests that affordances are a subclass of continuations. For example, both *cup* and *block* have similar Gibsonian affordance values, but quite distinct Telic affordance values. This can be distinguished by the nature of their respective Telic continuation sets as follows, where **sel** is a function that selects a suitable discourse antecedent inside the continuation set (Asher and Pogodalla, 2010): $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(cup)$, $grab \subseteq$ **sel** k_{Gib}, $drink \subseteq$ **sel** k_{Telic}, $\lambda k_{Gib} \otimes k_{Telic}.k_{Gib} \otimes k_{Telic}(block)$, $grab \subseteq$ **sel** k_{Gib}, $pick\_up \subseteq$ **sel** k_{Gib}, $move \subseteq$ **sel** k_{Gib}. This is the subject of ongoing research.

6. Experiments with Multimodal Dialogues

6.1. *Aspects of Multimodal Compositionality*

In this section, we provide additional formal analysis of experimental data gathered from multimodal dialogues between a human and a computational agent, represented as an avatar in VoxWorld. We examine extracts from dialogues between humans and computational agents in various tasks, in order to examine the nature of the communicative act in the context of the common ground structure. We illustrate how the situated meaning of the multimodal expression is constructed in each case. In particular, we look at three aspects of multimodal compositionality in these examples:

(23) a. generating referring expressions using different modalities;
b. generating and interpreting action and event expressions;
c. generating full action descriptions using both gesture and language.

Recall that 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. For the examples below, we annotate the dialogue with the contribution of both speech and gesture for each agent. Each dialogue turn encodes a multimodal ensemble, $\begin{bmatrix} \mathcal{S} \\ \mathcal{G} \end{bmatrix}$, which may or may not be realized in both modalities. In the annotation below, alignment between the modalities is indicated through a temporal indexing on the appropriate modal expression, e.g., t_i.

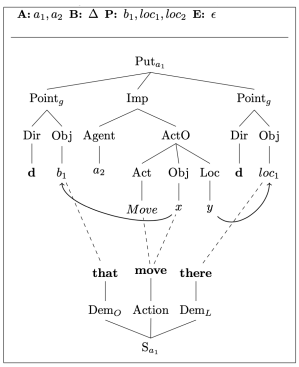
Figure 8. *Co-gestural speech imperative.*

Since we can use speech and gesture to indicate objects, location, and actions, we bias our speech recognition toward syntactic categories that represent partial information (e.g., NPs for objects, PPs for locations, VPs for actions), using incrememntal predictivity (cf. Hough *et al.* (2015)). We parse input in both directions, so we can take inputs like "put a block on the purple block" without resolving "a block" to the purple block, to prevent the agent from putting the purple block on itself.

(24)

MIXED MODALITY IMPERATIVE

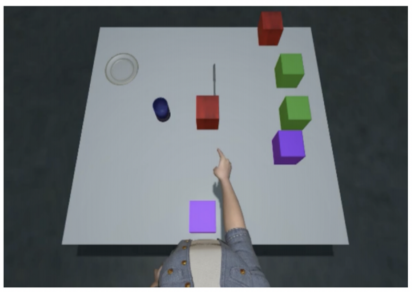
HUMAN$_1$: $\mathcal{S}$ = That$_{t1}$
$\mathcal{G}$ = [*points to the purple block*]$_{t1}$
HUMAN$_2$: $\mathcal{S}$ = Move$_{t2}$
$\mathcal{G}$ = [*makes move gesture*]$_{t2}$
HUMAN$_3$: $\mathcal{S}$ = There$_{t3}$
$\mathcal{G}$ = [*points to the red block*]$_{t3}$



$\lambda r_s \otimes r_g.([\![\langle \text{THERE}, \mathbf{D}_{Loc} \rangle]\!] \lambda k_s \otimes k_g.([\![\langle \text{THAT}, \mathbf{D}_{Obj} \rangle]\!]; \lambda j_s \otimes j_g.((([\![\langle \text{MOVE}, \mathbf{Move} \rangle]\!] j_s \otimes j_g) k_s \otimes k_g)) r_s \otimes r_g)$

6.2. *Multimodal Referring Expressions*

The *E*mbodied *M*ultimodal *R*eferring *E*xpressions (EMRE) dataset (Krishnaswamy and Pustejovsky, 2019) consists of 1,500 visual simulated situations showing an agent (Diana) indicating various object in a scene each accompanied by a definite referring expression. Referring expressions may take the form of deictic gesture only, a spoken description only with no demonstratives (e.g., "the red block in front of the knife and left of the green block"), or a mixed-modality referring expression as in Fig. 9 (right). Fig. 9 (left) shows a sample still that accompanies the utterance, with an equivalent common ground structure one the right.



$\lambda k_s \otimes k_g(\textbf{that}(x)[\mathsf{block}(x) \wedge \mathsf{red}(x) \wedge \mathsf{in\_front}(x, k, v)] \wedge k_s \otimes k_g(x)]$, where $\mathsf{v} = \alpha_a$

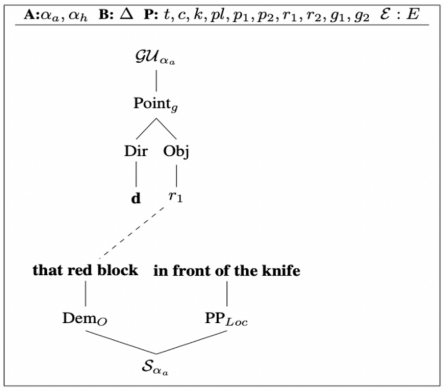


Figure 9. *Left: Sample still from the EMRE dataset (L), with CGS (R) and semantics of the RE (below), showing a continuation for each modality, k_s and k_g, which apply over the object subsequently in the dialogue.*

Amazon Mechanical Turk workers evaluted the EMRE dataset on a Likert-type scale for naturalness of the depicted referring expression for the indicated object. We found a clear preference for the multimodal referring expressions, suggesting that the redundancy provided by co-occurring language and gesture made for the clearest, most natural references to objects.

In Krishnaswamy and Pustejovsky (2020), we extracted formal features from the data as one-hot vectors representing elements of common ground structures. If in one of visualized REs, the avatar points to b, one of the jointly perceived objects $\in \mathbf{P}$, such that $\forall b\ (b \in \mathbf{P} \rightarrow \mathcal{K}_{\alpha_h}\mathcal{P}_{\alpha_a} b \wedge \mathcal{K}_{\alpha_a}\mathcal{P}_{\alpha_h})$. This demonstrates the avatar can point, and knows that b is the target $[C_{\alpha_a} = Point_g \rightarrow Dir\ b!]\mathcal{K}_{\alpha_h}\mathcal{K}_{\alpha_a}(Point_g \wedge target(b))$, which is encoded as a single feature. An agent may introduce a new object into the dicussion, making common the knowledge of its existence. Or an agent a uses a term t to make public the knowledge of a's interpretation of t.

We used these CGS-extracted features to train a neural net to predict the naturalness of a given referring expression, using the naturalness judgments from the EMRE dataset as ground truth. The EMRE dataset contains situational information about the specific configuration in which the referring expression was generated, and the linguistic referring expression itself, so we tested the effects of including formal, CGS-derived features by training classifiers on combinations of the symbolic situational features, embedding vectors of the linguistic RE, and the CGS-derived features.

We trained a multilayer preception, a simple, fast architecture that can distinguish dependencies in linearly-inseparable regions of data. This architecture consists of three fully-connected hidden layers of 32, 128, and 64, respectively, prior to a $softmax$ output layer. The layers use $tanh$, ELU, and $tanh$ activation, respectively, cross-entropy loss and Adam optimization, and is trained for 1,000 epochs with a batch size of 50. We perform 7-fold cross-validation in order to achieve a more balanced sample across all classes of annotator judgments. $k = 7$ is chosen here to approximate a leave-one-out cross-validation approach over the 8 annotator judgments on each visualized referring expression. The "most likely" annotator judgment in the EMRE dataset is a probability distribution so, we regard a "correct" prediction by the classifier as one that falls within the correct quintile of the distribution over all annotator judgments of that visualized referring expression.

	Raw features	**Raw feat. + SE**
μ **Acc. (1K)**	0.6757	0.6429
σ **Acc. (1K)**	0.0230	0.0111

	Raw + form.	**Raw + form. + SE**	**Formal only**
μ **Acc. (1K)**	0.7214	0.6671	0.7471
σ **Acc. (1K)**	0.0398	0.0243	0.0269

Figure 10. *Classification accuracy using formal features (mean and standard deviation).*

Fig. 10 shows that inclusion of formal features derived from the elements of common-ground structures improved classifier prediction accuracy by between 7% and 11% relative to baseline predictions that used the raw features of the EMRE dataset, plus sentence embedding representations of the referring expression itself. This suggests that common ground structures provide a dense, interpretable representation of the dialogue state, facilitating generation of natural, situation-appropriate referring expressions, and predicts the natural quality of a referring expression beyond other strong predictors of naturalness, e.g., modality.

6.3. *Interruptions and Corrections in Dialogue*

Figure 11. *Correcting and undoing an action.*

Establishing entities in a common ground structure so they can be recombined appropriately and interpreted in context allows us to build asynchronous agent behaviors capable of interruption and correction. Correction (Fig. 12) is currently implemented by performing three functions: (a) **Undo**, which recontinuizes an expression which has saturated its parameters, i.e., **undo** $k = \lambda k.k(grab)$; (b) **Rewind**, which reintroduces the previous monad; and (c) **Reassign**, which takes the corrected value and assigns it, resulting in $M, cg_2 \models grab(white)$.

In this manner, parameters can be unbound from either object or location argument, depending on the typing of the content communicated. Fig. 11 shows one such situation, where the replacement content "*on* the white one" is evaluated to a location. The state monad containing the location on the blue block is rewound, and the argument reassigned to the location on top of the white block. Had the utterance been "the

REFERENCE REPAIR

H: $\mathcal{G}$ = [*points to area around yellow and white blocks*]
D: $\mathcal{S}$ = Okay.$_{t1}$
$\mathcal{G}$ = [*picks up yellow block*]$_{t1}$
H: $\mathcal{S}$ = No, the white one.
D: $\mathcal{S}$ = Okay.$_{t2}$
$\mathcal{G}$ = [*picks up white block*]$_{t2}$

The user ambiguously points to yellow and white blocks. Diana chooses the yellow block ($\lambda k.k(grab) \Rightarrow M, cg_1 \models grab(yellow)$). The user corrects her, focus is unbound from the yellow block and assigned to the white block.

Figure 12. *Correcting deictic reference*

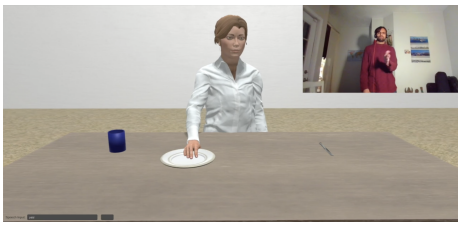
white one," the action would be reassigned with the white block as the theme, with the previously-existing target location, and Diana would put down the yellow block and put the *white* block on the blue block.

6.4. *Affordance Structure and Transfer Learning*

Diana may come across objects with different affordances from the typical Blocks World scenario. In these cases, the semantics of each object provided by VoxML allows Diana to learn new gestures associated with specific affordances of specific objects. Fig. 13 specifies such an interaction.

Using a random forest classifier, the gesture the human makes to associate with the specific affordance is situated in the search space defining the existing known gestures. Those learned grasp semantics can then be propagated down to any other event containing [[GRASP]] as a subevent, as shown in (25).

while(C, A) states that an activity, A, is performed only if a constraint, C, is satisfied at the same moment. Thus, if the agent encounters a [[SLIDE]] action with an outstanding variable ($\lambda y.slide(y, loc)$), and the human makes a gesture denoting $grasp(plate)$, the agent can directly lift $grasp(plate)$ to the slide action and apply the argument $plate$ to y: $\lambda y.slide(y, loc)@plate \Rightarrow \lambda y.slide(y, loc)$.



AFFORDANCE LEARNING IN KITCHENWORLD

HUMAN: $\mathcal{S}$ = The plate.
DIANA: $\mathcal{S}$ = Okay.$_{t1}$
$\mathcal{G}$ = [*points to the plate*]$_{t1}$
HUMAN: $\mathcal{G}$ = [*makes "claw down" gesture*]
DIANA: $\mathcal{S}$ = Should I grasp it like this$_{t2}$?
$\mathcal{G}$ = [*grasps plate from the side*]$_{t2}$
HUMAN: $\mathcal{S}$ = Yes.
DIANA: $\mathcal{S}$ = Is there a gesture for that?
HUMAN: $\mathcal{G}$ = [*makes "grasp plate" gesture*]

Figure 13. *Diana and human interacting.*

(25) $grasp(e_1, \text{AG}, y)$; **while**$(hold(\text{AG}, y) \wedge on(y, \text{SURF}) \wedge \neg at(y, \text{LOC}))$, $move\_to(e_2, \text{AG}, y, \text{LOC}))$; **if**$(at(y, \text{LOC}), ungrasp(e_3, \text{AG}, y))$

Model	% correct cluster
MLP (Habitats)	78.82
MLP (Affordances)	**84.71**
CNN (Habitats)	78.82
CNN (Affordances)	81.18

Figure 14. *Prediction accuracy w/ 6 means.*

Affordance properties can also be transferred between objects. Given that similar habitats serve as necessary (but not sufficient) preconditions to behaviors (e.g., to be rolled, an apple, cup, and bottle must all be turned on their sides), the ability to assess an unknown object relative to known ones allows an agent to transfer properties between them, to gain a handle on interacting with and discussing a novel object. Consider Fig. 15, where Diana has no semantics for what we recognize as a bottle.

Diana observes similarities in the cup's habitats and the bottle's (e.g., similar orientation, symmetry and size constraints), infers they may share behaviors, and so grasps one like the other. Links between habitats and affordances allows inference of similar objects and behaviors in the current situation.

Over 17 VoxML objects (e.g., Fig.2), we trained 200D *habitat and affordance embeddings* using a Skip-Gram model for 50,000 epochs with window size 3. Objects were represented as averaged habitat or affordance vectors. These embeddings were run through a 7-layer MLP and a 4-layer (1D) CNN, that chose the known object most similar to the unlabeled vector. E.g., a vector representing a *plate*'s affordances was predicted to be similar to a *cup* or *bottle* due to its containment affordance.

For each object, 8 annotators chose the 2 most similar objects in the vocabulary, in terms of their afforded behaviors, and we performed k-means clustering over these annotations. Our models trained on habitat or affordance embedding vectors successfully predicted an object in the correct cluster 80% of the time (Fig. 14). Diana then enacted known behaviors over novel objects (Fig. 15, top right). Further analysis of these models and their properties are ongoing but these early results show how affordances can be used to train useful models over small sample sizes.

TRANSFER OF OBJECT AFFORDANCES

HUMAN: $\mathcal{S}$ = What is that?$_{t1}$
$\mathcal{G}$ = [*points to the cup*]$_{t1}$
DIANA: $\mathcal{S}$ = That's a cup$_{t2}$
$\mathcal{G}$ = [*points to the cup*]$_{t2}$
HUMAN: $\mathcal{S}$ = What is that?$_{t3}$
$\mathcal{G}$ = [*points to the bottle*]$_{t3}$
DIANA: $\mathcal{S}$ = I don't know, but I could grasp it like a cup$_{t4}$.
$\mathcal{G}$ = [*points to the bottle*]$_{t4}$
HUMAN: $\mathcal{S}$ = Grab it.
DIANA: $\mathcal{G}$ = [*grasps bottle from the side*]$_{t5}$

Figure 15. *Transferring affordance properties through dialogue.*

7. Conclusion

Multimodal peer-to-peer interfaces require robust integration of conversational modalities in a naturalistic fashion. We have outlined the first steps toward such integration, based on the logic of our multimodal simulation semantics and 3D environment as the platform for shared common ground. We give our computational agent a framework for major faculties natively available to humans using computer vision techniques to recognize gesture and by laying the groundwork for a modal logic of synthetic vision. The result is a framework and platform that interweaves linguistic and non-linguistic modalities in the completion of a shared task by exploiting the relative strengths of linguistic and non-linguistic context to exchange information in a situated communication. We have also developed this framework into an interaction with a mobile robot mediated by a virtual rendition of the environment the robot sees as it explores. The human then gestures to objects and locations on the screen and gives the robot grounded instructions with spoken English and gesture.

We hope to have demonstrated that the notion of situatedness involves embedding linguistic expressions and their grounding within a multimodal semantics. This approach allows environmentally-aware models that can be validated; if one model of expression (e.g., gesture) is insufficiently communicative, another (e.g., language) can be used to examine where it went wrong. Each additional modality provides an avenue through which to validate models of other modalities.

Acknowledgements

This work was supported by Contract W911NF-15-C-0238 with the US Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO). Approved for Public Release, Distribution Unlimited. The views expressed herein are ours and do not reflect the official policy or position of the Department of Defense or the U.S. Government. We would like to thank Ken Lai, Bruce Draper, Ross Beveridge, Francisco Ortega, and Lucia Donatelli for their comments and suggestions.

8. References

Asher N., "Common ground, corrections and coordination", *Journal of Semantics*, 1998.

Asher N., Lascarides A., *Logics of conversation*, Cambridge University Press, 2003.

Asher N., Pogodalla S., "SDRT and continuation semantics", *JSAI International Symposium on Artificial Intelligence*, Springer, p. 3-15, 2010.

Barker C., Shan C.-c., *Continuations and natural language*, vol. 53, Oxford studies in theoretical linguistics, 2014.

Barsalou L. W., "Perceptions of perceptual symbols", *Behavioral and brain sciences*, vol. 22, n° 4, p. 637-660, 1999.

Cassell J., Stone M., Yan H., "Coordination and context-dependence in the generation of embodied conversation", *Proc. of 1st Int. Conf. on NLG*, ACL, p. 171-178, 2000.

Chai J. Y., Fang R., Liu C., She L., "Collaborative language grounding toward situated human-robot dialogue", *AI Magazine*, vol. 37, n° 4, p. 32-45, 2016.

Clark H. H., Brennan S. E., "Grounding in communication", *Perspectives on socially shared cognition*, vol. 13, n° 1991, p. 127-149, 1991.

Cooper R., Ginzburg J., "Type Theory with Records for Natural Language Semantics", *The handbook of contemporary semantic theory*p. 375, 2015.

Craik K. J. W., *The nature of explanation*, Cambridge University, Cambridge UK, 1943.

De Groote P., "Type raising, continuations, and classical logic", *Proceedings of the 13th Amsterdam Colloquium*, p. 97-101, 2001.

Dobnik S., Cooper R., Larsson S., "Modelling language, action, and perception in type theory with records", *Constraint Solving and Language Processing*, Springer, p. 70-91, 2013.

Feldman J., "Embodied language, best-fit analysis, and formal compositionality", *Physics of life reviews*, vol. 7, n° 4, p. 385-410, 2010.

Fernando T., "Situations in LTL as strings", *Information and Computation*, vol. 207, n° 10, p. 980-999, 2009.

Fischer K., "How people talk with robots: Designing dialog to reduce user uncertainty", *AI Magazine*, vol. 32, n° 4, p. 31-38, 2011.

Gatsoulis Y., Alomari M., Burbridge C., Dondrup C., Duckworth P., Lightbody P., Hanheide M., Hawes N., Hogg D., Cohn A. *et al.*, "QSRlib: a software library for online acquisition of Qualitative Spatial Relations from Video", 2016.

Gibson J. J., "The Theory of Affordances", *Perceiving, Acting, and Knowing: Toward an ecological psychology*p. 67-82, 1977.

Ginzburg J., "Interrogatives: Questions, facts and dialogue", *The handbook of contemporary semantic theory*p. 359-423, 1996.

Ginzburg J., Fernández R., "Computational Models of Dialogue", *The handbook of computational linguistics and natural language processing*, vol. 57, p. 1, 2010.

Goldman A. I., *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*, Oxford University Press, 2006.

Harel D., "Dynamic Logic", *in* M. Gabbay, F. Gunthner (eds), *Handbook of Philosophical Logic, Volume II: Extensions of Classical Logic*, Reidel, p. 497-604, 1984.

Hough J., Kennington C., Schlangen D., Ginzburg J., "Incremental semantics for dialogue processing: Requirements, and a comparison of two approaches", 2015.

Hunter J., Asher N., Lascarides A., "A formal semantics for situated conversation", *Semantics and Pragmatics*, 2018.

Johnson-Laird P., "How could consciousness arise from the computations of the brain", *Mindwaves. Oxford: Basil Blackwell*p. 247-257, 1987.

Kendon A., *Gesture: Visible action as utterance*, Cambridge University Press, 2004.

Kennington C., Kousidis S., Schlangen D., "Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information", *Proceedings of SigDial 2013*, 2013.

Krishnaswamy N., Pustejovsky J., "VoxSim: A Visual Platform for Modeling Motion Language", *Proceedings of COLING 2016*, ACL, 2016.

Krishnaswamy N., Pustejovsky J., "Generating a Novel Dataset of Multimodal Referring Expressions", *Proc. of 13th Int. Conference on Computational Semantics*, p. 44-51, 2019.

Krishnaswamy N., Pustejovsky J., "A Formal Analysis of Multimodal Referring Strategies Under Common Ground", *Proceedings of The 12th LREC*, p. 5919-5927, 2020.

Landragin F., "Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems", *Signal Processing*, vol. 86, n° 12, p. 3578-3595, 2006.

Lascarides A., Stone M., "A formal semantic analysis of gesture", *Journal of Semantics*p. ffp004, 2009.

Lücking A., Pfeiffer T., Rieser H., "Pointing and reference reconsidered", *Journal of Pragmatics*, vol. 77, p. 56-79, 2015.

Mani I., Pustejovsky J., *Interpreting Motion: Grounded Representations for Spatial Language*, Oxford University Press, 2012.

Marge M., Rudnicky A. I., "Towards evaluating recovery strategies for situated grounding problems in human-robot dialogue", *2013 IEEE RO-MAN*, IEEE, p. 340-341, 2013.

Miller G. A., Johnson-Laird P. N., *Language and perception.*, Belknap Press, 1976.

Narayana P., Krishnaswamy N., Wang I., Bangar R., Patil D., Mulay G., Rim K., Beveridge R., Ruiz J., Pustejovsky J., Draper B., "Cooperating with Avatars Through Gesture, Language and Action", *Intelligent Systems Conference (IntelliSys)*, 2018.

Narayanan S., "Mind changes: A simulation semantics account of counterfactuals", *Cognitive Science*, 2010.

Naumann R., "Aspects of changes: a dynamic event semantics", *Journal of semantics*, vol. 18, p. 27-81, 2001.

Plaza J., "Logics of public communications", *Synthese*, vol. 158, n° 2, p. 165-179, 2007.

Pustejovsky J., *The Generative Lexicon*, MIT Press, 1995.

Pustejovsky J., "Dynamic Event Structure and Habitat Theory", *Proc. of the 6th International Conference on Generative Approaches to the Lexicon (GL2013)*, ACL, p. 1-10, 2013.

Pustejovsky J., "From actions to events: Communicating through language and gesture", *Interaction Studies*, vol. 19, n° 1-2, p. 289-317, 2018.

Pustejovsky J., Krishnaswamy N., "VoxML: A Visualization Modeling Language", *Proceedings of LREC*, 2016.

Pustejovsky J., Moszkowicz J., "The qualitative spatial dynamics of motion", *The Journal of Spatial Cognition and Computation*, 2011.

Scheutz M., Cantrell R., Schermerhorn P., "Toward humanlike task-based dialogue processing for human robot interaction", *Ai Magazine*, vol. 32, n° 4, p. 77-84, 2011.

Schlenker P., "Gestural grammar", *Natural Language & Linguistic Theory*p. 1-50, 2020.

Stalnaker R., "Common ground", *Linguistics and philosophy*, vol. 25, n° 5-6, p. 701-721, 2002.

Stojnić U., Stone M., Lepore E., "Pointing things out: in defense of attention and coherence", *Linguistics and Philosophy*p. 1-10, 2019.

Unger C., "Dynamic semantics as monadic computation", *JSAI International Symposium on Artificial Intelligence*, Springer, p. 68-81, 2011.

Van Benthem J., *Logical dynamics of information and interaction*, Cambridge, 2011.

Van Ditmarsch H., van Der Hoek W., Kooi B., *Dynamic epistemic logic*, vol. 337, Springer Science & Business Media, 2007.

Van Eijck J., Unger C., *Computational semantics with functional programming*, Cambridge, 2010.

Williams T., Bussing M., Cabrol S., Boyle E., Tran N., "Mixed reality deictic gesture for multimodal robot communication", *IEEE Int'l Conf. on HRI*, IEEE, p. 191-201, 2019.