The Role of Embodiment and Simulation in Evaluating HCI: Theory and Framework*

James Pustejovsky[1[0000-0003-2233-9761]] and
Nikhil Krishnaswamy[2[0000-0001-7878-7227]]

[1] Brandeis University, Waltham, MA 02453, USA
jamesp@brandeis.edu
[2] Colorado State University, Fort Collins, CO 80523, USA
nkrishna@colostate.edu

Abstract. In this paper, we argue that embodiment can play an important role in the evaluation of systems developed for Human Computer Interaction. To this end, we describe a simulation platform for building Embodied Human Computer Interactions (EHCI). This system, VoxWorld, enables multimodal dialogue systems that communicate through language, gesture, action, facial expressions, and gaze tracking, in the context of task-oriented interactions. A multimodal simulation is an embodied 3D virtual realization of both the situational environment and the co-situated agents, as well as the most salient content denoted by communicative acts in a discourse. It is built on the modeling language VoxML, which encodes objects with rich semantic typing and action affordances, and actions themselves as multimodal programs, enabling contextually salient inferences and decisions in the environment. Through simulation experiments in VoxWorld, we can begin to identify and then evaluate the diverse parameters involved in multimodal communication between agents. VoxWorld enables an embodied HCI by situating both human and computational agents within the same virtual simulation environment, where they share perceptual and epistemic common ground. In this first part of this paper series, we discuss the consequences of embodiment and common ground, and how they help evaluate parameters of the interaction between humans and agents, and demonstrate different behaviors and types of interactions on different classes of agents.

Keywords: Embodiment · HCI · Common ground · multimodal dialogue · VoxML.

*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, and Francisco Ortega for their comments and suggestions.

1 Introduction

As multimodal interactive systems become both more common and more sophisticated, naive users come to use them with increasing expectations that their interactions will approximate aspects of typical interactions with another human. With this increased interest in multimodal interaction comes a need to evaluate the performance of a multimodal system on the various levels with which it engages the user. Thus, system evaluations and metrics should be able to account for the communicative ability of the various modalities in use, as well as how the modalities interact with each other to facilitate communication. Such evaluation metrics should be modality-agnostic and assess the communication between human and computer based on the semantics of objects, events, and actions situated within the shared context created by the human-computer interaction.

One way to facilitate this type of evaluation is to position the human and the computational agent within a shared conceptual space, where the agent is able to sufficiently interpret multimodal behavior and communicative commands from the human. This suggests an *embodied* presence within a *simulated* environment. Here we argue that a simulation platform provides just such an environment for modeling communicative interactions, what we call *Embodied Human Computer Interaction*, one facilitated by a formal model of object and event semantics that renders the continuous quantitative search space of an open-world, real-time environment tractable. We provide examples for how a semantically-informed AI system can exploit the precise, numerical information provided by a game engine to perform qualitative reasoning about objects and events, facilitate learning novel concepts from data, and communicate with a human to improve its models and demonstrate its understanding.

As a case in point, consider the two interactions in Figure 1. On the left, we see a human-human interaction engaged in a joint task. On the right, the same task is being carried out between a human and an intelligent virtual agent (IVA), who is embodied in a simulation environment with the user.

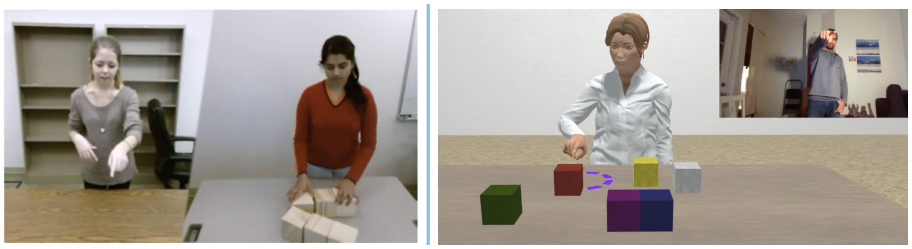


Figure 1: *Left:* Human-human collaborative interaction; *Right:* Human-avatar interaction.

The notion of embodiment has many diverse interpretations, depending on the discipline and field of study [1,12,43,52,68]. When discussing its role in HCI, we can identify (at least) three major factors of embodiment that contribute to how an artificial agent interacts effectively with its human partners:

- The artificial agent has some identifiable degree of *self-embodiment*: this is the "spatial presence" associated with the agent relative to the human partner, within the domain or space of the interaction (the embedding space). This might include a virtual presence on a screen, with face or even skeletal form; actual effectors for action and manipulation; and explicit sensors for audio and visual input.
- The agent is aware of the *human's embodiment*; that is, it has recognition of the human partner's linguistic and gestural expressions, facial expressions, and actions. The artificial agent continously receives inputs through which it constructs and maintains a representation of its human partner's embodiment.
- The interaction enables *situated meaning* for the objects and actions in the environment; an elementary understanding of how objects behave relative to each other and as a consequence of the agent's actions (affordances, action dynamics, etc). This also includes recognition of speaker intent and epistemic state.

While embodiment is a relatively recent theoretical development, the concept of simulation has played an important role in both AI and cognitive science for over 40 years. There are two distinct uses for the term *simulation*, particularly as used in computer science and AI. First, simulation can be used as a description for *testing a computational model*. That is, variables in a model are set and the model is run, such that the consequences of all possible computable configurations become known. Examples of such simulations include models of climate change, the tensile strength of materials, models of biological pathways, and so on. We refer to this as *computational simulation modeling*, where the goal is to arrive at the best model by using simulation techniques.

Simulation can also refer to an environment which 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: in these contexts, the software and gaming world assume an embodiment of the agent in the environment, either as a first-person restricted POV (such as a first-person shooter or RPG), or an omniscient movable embodied perspective (e.g., real-time or turn-based strategy). We refer to such approaches as *situated embodied simulations*. The goal is to simulate an agent within a situation.

Simulation has yet another meaning, however. Starting with Craik [18], we encounter the notion that agents carry a mental model of external reality in their heads. Johnson-Laird [40] 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 [39]. 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 that "mind reading" plays in modeling the mental representations of other agents and the content of their communicative acts [34,35,36,38]. Simulation semantics (as adopted within cognitive linguistics and practiced by Feldman [26], Narayanan [55], Bergen [7], and Evans [25]) 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 [37,6,75,74]. These simulation approaches can be referred to as *embodied theories of mind*. Their goal is to view the semantic interpretation of an expression by means of a simulation, which is either mental (a la Bergen and Evans) or interpreted graphs such as Petri Nets (a la Narayanan and Feldman).

We describe a simulation framework, VoxWorld, that integrates the functionality and the goals of all three approaches above. Namely, we situate 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. This platform provides an environment for experimenation with multimodal interactions between humans and avatars or robots.

In [63], we discuss the challenges involved in creating an embodied agent for HCI and HRI. Two issues present themselves, to this end. First, it will be important to identify an operational definition of embodiment for this domain; and secondly, we should acknowledge that an agent cannot simply be embodied without also embodying the interaction within which the agent is acting.

2 Prior Work

There is a long and established tradition of multimodal interfaces that combine language and gesture, starting with [8], which anticipated some of the issues discussed herein, including the use of deixis to disambiguate references, and also inspired a community surrounding multimodal integration (e.g., [22,42,72]).

The psychological motivation for multimodal interfaces, as epitomized by [66], holds that speech and gesture are coexpressive and processed partially independently, and therefore complement each other. Using both modalities increases human working memory and decreases cognitive load [22], allowing people to retain more information and learn faster.

Visual information has been shown to be particularly useful in establishing common ground [14,16,20,23,24], or mutual understanding that enables further communication. Other research in HCI additionally emphasizes the importance of shared visual workspaces in computer-mediated communication [28,29,30,44], highlighting the usefulness of non-verbal communication in coordination between humans [10,11].

[9] shows that allowing for shared gaze increased performance in spatial tasks in paired collaborations. Multimodal systems of gaze and speech have also been studied in interaction with robots and virtual avatars [2,54,69]. However, few

systems have centered the use of language and gesture in collaborative and communicative scenarios.

Communicating with computers becomes even more interesting in the context of shared physical tasks. When people work together, their conversation consists of more than just words. They gesture and they share a common workspace [13,47,49,53]. Their shared perception of this workspace is the context for their conversation, and it is this shared space that gives many gestures, such as pointing, their meaning [45]. The dynamic computation of discourse [4], furthermore, becomes more complex when multiple modalities are at play. Fortunately, embodied actions (such as coverbal gestures) do not seem to violate coherence relations [48].

As shown in Part 2 of this paper series, we approach evaluation from a semantics-centered perspective, and use distinct semantic properties of specific elements in the interaction to determine what about the interaction enabled or hindered “shared understanding.” This is typically referred to as the “common ground” in the literature, both in psychology and semantics [3,15,32,60,70,71].

Within HCI and human device interaction (HDI) design, a related area involves the evaluation of gesture fields [41] for the expression of image schemas and how they map to interactions with the computer. The results in [50] on the FIGURE corpus are relevant for design decisions, raising evaluation criteria distinct from the hallmarks mentioned above. Similar concerns and suggestions are discussed in [67], for how gestures can improve the behavior of embodied conversational agents (ECAs).

3 Common Ground in VoxWorld

3.1 VoxML: Encoding Actions and Objects

In order to characterize the many dimensions of human-computer interactions, we will introduce an approach to evaluating interactions drawing on the most relevant parameters in a co-situated communicative interactions. By introducing a formal model of shared context, we are able to to track the intentions and utterances, as well as the perceptions and actions of the agents involved in a dialogue. Our model, VoxWorld, integrates all three aspects of simulation discussed above into 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.

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. VoxWorld and VoxML [61], provide the following characteristics: object encoding with rich semantic typing and action affordances; action encoding as multimodal programs; it reveals the elements of the common ground in interaction between parties, be they humans or artificially intelligent agents. VoxWorld supports embodied HCI wherein artificial agents consume different sensor inputs for awareness of not only their own virtual space but also the surrounding physical space. It brings together the three definitions of simulation introduced above.

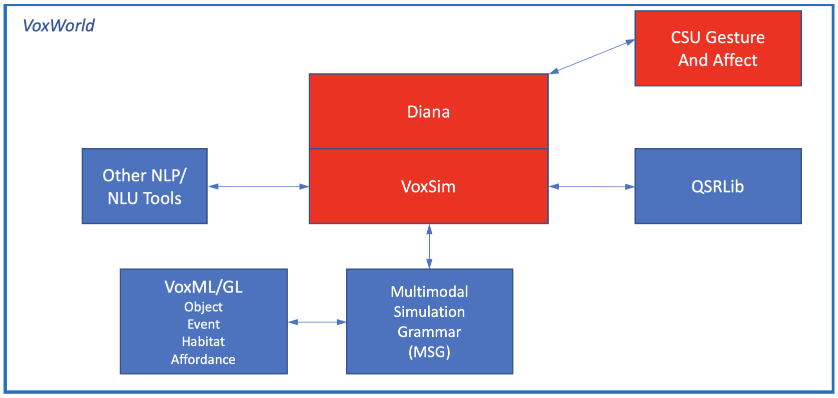


Figure 2: VoxWorld architecture schematic.

Within the computational context of VoxWorld, common ground relies on implementations of the following aspects of the interaction:

1. *Co-situatedness* and *co-perception* of the agents, such that they can interpret the same situation from their respective frames of reference, such as a human and an avatar perceiving the same virtual scene from different perspectives;
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 be able to refer to objects on the table in multiple modalities with a common model of differences in perspective-relative references;
3. *Co-intent* of a common goal, such that adversarial relationships between agents reflect a breakdown in the common ground. Here, human and agent are collaborating to achieve a common goal, each sharing their knowledge with the other.

VoxML (Visual Object Concept Markup Language) is the representation language used to encode knowledge about objects, events, attributes, and functions by linking lexemes to their visual instantiations, termed the "visual object concept" or *voxeme*. 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. For example, a *cup* can be typed as a cylindroid with concavity, as shown below:

$$\left[\begin{array}{l} \textbf{cup} \\ \text{LEXICAL} = \left[\begin{array}{l} \text{PREDICATE} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj} \bullet \textbf{artifact} \end{array}\right] \\ \text{TYPE} = \left[\begin{array}{l} \text{HEAD} = \textbf{cylindroid}[\textbf{1}] \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATIONAL\_SYMMETRY} = \{Y\} \\ \text{REFLECTION\_SYMETRY} = \{XY, YZ\} \end{array}\right] \end{array}\right]$$

An OBJECT voxeme's semantic structure also provides *habitats*, which are situational contexts or environments conditioning the object's *affordances*, which may be either "Gibsonian" affordances [31] or "Telic" affordances [57,58]. 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.

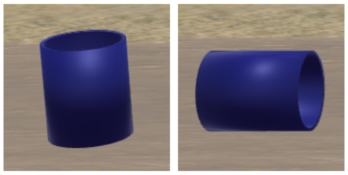
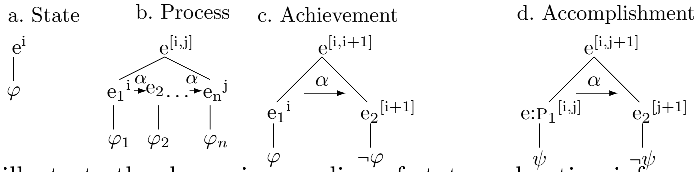


Figure 3: Cup in different habitats. Both allow **holding**, while the left allows **sliding** and the right allows **rolling**.

$$\left[\begin{array}{l} \textbf{cup} \\ \text{LEXICAL} = \left[\begin{array}{l} \text{PREDICATE} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj} \bullet \textbf{artifact} \end{array}\right] \\ \text{TYPE} = \left[\begin{array}{l} \text{HEAD} = \textbf{cylindroid}[\textbf{1}] \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATIONAL\_SYMMETRY} = \{Y\} \\ \text{REFLECTION\_SYMETRY} = \{XY, YZ\} \end{array}\right] \\ \text{HABITAT} = \left[\begin{array}{l} \text{INTRINSIC} = {\scriptstyle [2]}\left[\begin{array}{l} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{array}\right] \\ \text{EXTRINSIC} = {\scriptstyle [3]}\left[\text{UP} = align(Y, \mathcal{E}_{\perp Y})\right] \end{array}\right] \\ \text{AFF\_STR} = \left[\begin{array}{l} \mathrm{A}_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x) \\ \mathrm{A}_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x) \\ \mathrm{A}_3 = H_{[2]} \rightarrow [grasp(x, [1])]hold(x, [1]) \\ \mathrm{A}_4 = H_{[3]} \rightarrow [roll(x, [1])]\mathcal{R} \end{array}\right] \\ \text{EMBOD} = \left[\begin{array}{l} \text{SCALE} = <\textbf{agent} \\ \text{MOVABLE} = \textbf{true} \end{array}\right] \end{array}\right]$$

Activities and events are interpreted in VoxML as programs, π, in terms of a dynamic event semantics, Dynamic Interval Temporal Logic (DITL) [65]. The advantage of adopting a dynamic interpretation of events is that linguistic expressions map directly into simulations through an operational semantics. A formula is interpreted as a propositional expression, with assignment of a truth value in a specific state in the model. For our purposes, a state is a set of propositions with assignments to variables at a specific time index. Atomic programs are relations from states to states, and hence interpreted over an input/output state-state pairing (cf. also [27,56]).

The structure in (a) below represents a **state**, e^i, at time i, with the propositional content, φ. The event structure in (c) illustrates how program α takes the world from e^i with content φ, to the adjacent state, e_2^{i+1}, where the propositional content has been negated, $\neg\varphi$. This corresponds directly to **achievements**. From these two types, the other two Vendlerian classes can be generated. **Processes** can be modeled as an iteration of simple transitions, where two conditions hold: the transition is a change in the value of an identifiable attribute of the object; every iterated transition shares the same attribute being changed. This is illustrated in (b) below. Finally, **accomplishments** are built up by taking an underlying process event, e:P, denoting some change in an object's attribute, and synchronizing it with an achievement (simple transition): that is, e:P is unfolding while ψ is true, until one last step of the program α makes it the case that $\neg\psi$ is now true.



a. State b. Process c. Achievement d. Accomplishment

To illustrate the dynamic encoding of state and action information in VoxML, consider the voxeme for the accomplishment verb *put*, shown below.

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

In this way, simulation becomes a way of tracing the consequences of linguistic spatial cues through the narrative structure of an event and presenting the computer system's understanding of it.

VoxWorld 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., a pantomime of an embodied agent moving an object without the object itself), the system can present 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 that 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.

3.2 Multimodal Semantics for Common Ground

In the previous section, we illustrated how the objects in the embedding space shared by participants in an interaction are encoded as VoxML multimodal representation. In this section, we describe the module within VoxWorld that is responsible for encoding and tracking the shared situational elements in a dialogue. Given the emphasis on evaluation of multimodal communication, we will pay particular attention to the semantics of integrated multimodal expressions in the context of task oriented dialogues. This will include the co-situated space the conversational agents share, beliefs about and perception of the objects in the environment, and the goals and intentions associated with both the task and the users, respectively, but also affordances associated with the objects present in the environment.

Following [59,62], we model this *common ground structure* (CGS), the information associated with a state in a dialogue, as a state monad, $\mathbf{M}\alpha = State \rightarrow (\alpha \times State)$ [73]. We adopt a continuation-based semantics for both communicative acts in discourse, as outlined in [5,19]. The dialogue monad corresponds to computations that read and modify a particular state. The values returned by querying the monad include the following elements of the dialogue state:

- The communicative act, C_a, performed by an agent, a: a tuple of expressions from the diverse modalities involved. Broadly, this includes the modalities of a linguistic utterance, S (speech), gesture, G, facial expression, F, gaze, Z, and an explicit action, A: $C_a = \langle S, G, F, Z, A \rangle$.
- **A**: The agents engaged in communication;
- **B**: The salient shared belief space;
- **P**: The objects and relations that are jointly perceived in the environment;
- $\mathcal{E}$: The embedding space that both agents occupy in the communication.

Here we focus on a speech-gesture multimodal interaction, to illustrate how the common ground is computed. We first initialize the common ground based on shared beliefs and dynamic perceptual content for each of the agents. This can be represented graphically as below, where an agent, a_i, makes a communicative act either through gesture, $\mathcal{G}$ in (1a), or linguistically, as in (1b.)[3]

(1) a. $\boxed{\begin{array}{l}\mathbf{A}{:}a_1, a_2\ \ \mathbf{B}{:}\Delta\ \ \mathbf{P}{:}b\ \ \mathcal{E}:E \\ \hline \\ \mathcal{G}_{a_1}\end{array}}$ b. $\boxed{\begin{array}{l}\mathbf{A}{:}a_1, a_2\ \ \mathbf{B}{:}\Delta\ \ \mathbf{P}{:}b\ \ \mathcal{E}:E \\ \mathcal{S}_{a_1} = \text{“You}_{a_2}\text{ see it}_b\text{”}\end{array}}$

[3] This is similar in many respects to the representations introduced in [17,33] and [21] for modeling action and control with robots.

In order to see how embodiment and common ground structures contribute to the interpretation of multimodal expressions, let us review our assumptions regarding our model. Typically, a linguistic expression, S, is computed relative to a model, $\mathcal{M}$, and the relevant assignment functions, e.g., g: $[\![S]\!]^{\mathcal{M},g}$. For every expression made in the dialogue, the information state (our monad) is updated through continuation-passing, as in [5]. For example, given the current discourse, D, and the new utterance, S, S integrates into D as follows:

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

This states that the current discourse has two arguments, its left context i (where we are), and what is expected later in the discourse, k.

When we move into multimodal dialogues, however, this model is not expressive enough, since it does not capture the interpretation of other modalities in the communication that convey denotative information (such as gesture), nor does it provide a situated grounding for the expressions within the dialogue state of the current context.

In order to enable reference to other modalities and their situational denotations, we introduce a *simulation* within which communicative expressions are interpreted. A simulation, $\mathcal{S}$, is defined as a triple, $\langle \mathcal{M}, \mathcal{E}, \mathcal{CG} \rangle$, consisting of a conventional model, $\mathcal{M}$, an embedding space, $\mathcal{E}$, together with a common ground structure, $\mathcal{CG}$. This definition brings together the three types of simulation discussed above. Now we can refer to an interpretation of an expression, α, within a simulation, as $[\![\alpha]\!]^{\mathcal{S}}$.

Given a model within which we can potentially interpret additional modalities, let us briefly outline how one modality, gesture, can be modeled compositionally, and interpreted within a simulation, alone and when used in aligned co-gestural speech acts. We assume a dynamic interpretation for gesture that references the common ground structure in discourse. Extending the approach taken in [41] and [49], a 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*. In a multimodal dialogue, these gestures have two features; (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. Hence, we model two kinds of gestures in our dialogues: (a) establishing a reference; and (b) depicting an action-object pair.

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

A gesture is directly interpretable by the agents 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 [49,51,59].

(4) $[\![\mathbf{D}_{obj}]\!]^{\mathcal{S}} = End([\![ray]\!]^{\mathcal{S}}([\![\mathbf{d}]\!]^{\mathcal{S}}))$

In multimodal dialogue, language and gesture work together in a number of ways, where gesture might enhance an expression emotionally, or pick out a reference in context, or depict an action through iconic representation. For example, deictic gesture acts like a demonstrative in a referring expression, and embodied gesture, when enacted, becomes part of the embedding space. The embodied artificial agent can interpret and generate expressions like "this/that block," accompanied by deixis, and can do the same when referring to the embodiment of its interlocutor (e.g., "my/your arm"). While agents in the interaction are considered separately from objects in the model, the typing of embodied agents show they they have properties of physical objects (e.g., a convex hull, an interactiom with the physics of the world, etc.), and so can be discussed in similar terms.

In our theory, 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.

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

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

We assume an aligned language-gesture syntactic structure, for which we have provided a continuized semantic interpretation [46,64]. Both of these are contained in the common ground state monad introduced above. 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 the alignment of these two continuations as $k_s \otimes k_g$, seen in (6).

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

Each of these modalities will contribute information if it is present. We bind co-gestural speech to specific gestures in the communicative act, within a common ground, CGS. A dashed line in an ensemble expression indicates that a co-gestural speech element, $\mathcal{S}$, is aligned with a particular gesture, $\mathcal{G}$. For example, the CG structure for the expression in Figure 4 illustrates the alignment of the spoken demonstrative *that* with the denotation of the deictic gesture, following the computation in (4). This then takes the continuized right context of the gesture sequence, and binds this referent into the parameter structure for *grab*, resulting in the interpretation below.

(7) $Grab(a_2, b_1)$

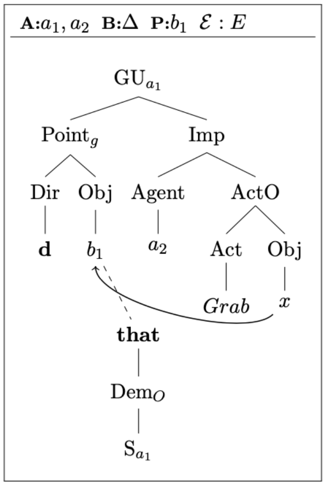


Figure 4: Common-ground structure for “that” (ensemble) + “grab” (speech).

4 Conclusion

Across different fields and in the existing AI, cognition, and game development literature, there exist many different definitions of “simulation.” Nonetheless, we believe the common thread between them is that simulations as a framework facilitate both qualitative and quantitative reasoning by providing quantitative data (for example, exact coordinates or rotations) that can be easily converted into qualitative representations. This makes simulation an effective platform for both producing and learning from datasets.

When combined with formal encodings of object and event semantics, at a level higher than treating objects as collections of geometries, or events as sequences of motions or object relations, 3D environments provide a powerful platform for exploring “computational embodied cognition.” Recent developments in the AI field have shown that common-sense understanding in a general domain requires either orders of magnitude more training data than traditional deep learning models, or more easily decidable representations, involving context, differences in perspective, and grounded concepts, to name a few.

Technologies in use in the gaming industry are proving to be effective platforms on which to develop systems that afford gathering both traditional data for deep learning and representations of common sense, situated, or embodied understanding. In addition, game engines perform a lot of “heavy lifting,” providing APIs for UI and physics, among others, which allows researchers to focus on implementing truly novel functionality and develop tools to deploy and examine the role of embodiment in human-computer interaction both quantitatively and qualitatively. In Part 2, we will describe such a system and experimental evaluations on it.

References

1. Anderson, M.L.: Embodied cognition: A field guide. Artificial intelligence **149**(1), 91–130 (2003)
2. Andrist, S., Gleicher, M., Mutlu, B.: Looking Coordinated: Bidirectional Gaze Mechanisms for Collaborative Interaction with Virtual Characters. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. pp. 2571–2582. CHI '17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3026033, `http://doi.acm.org/10.1145/3025453.3026033`
3. Asher, N.: Common ground, corrections and coordination. Journal of Semantics (1998)
4. Asher, N., Lascarides, A.: Logics of conversation. Cambridge University Press (2003)
5. Asher, N., Pogodalla, S.: Sdrt and continuation semantics. In: JSAI International Symposium on Artificial Intelligence. pp. 3–15. Springer (2010)
6. Barsalou, L.W.: Perceptions of perceptual symbols. Behavioral and brain sciences **22**(4), 637–660 (1999)
7. Bergen, B.K.: Louder than words: The new science of how the mind makes meaning. Basic Books (2012)
8. Bolt, R.A.: "Put-that-there": Voice and gesture at the graphics interface, vol. 14. ACM (1980)
9. Brennan, S.E., Chen, X., Dickinson, C.A., Neider, M.B., Zelinsky, G.J.: Coordinating cognition: The costs and benefits of shared gaze during collaborative search. Cognition **106**(3), 1465–1477 (Mar 2008). https://doi.org/10.1016/j.cognition.2007.05.012, `http://www.sciencedirect.com/science/article/pii/S0010027707001448`
10. Cassell, J.: Embodied conversational agents. MIT press (2000)
11. Cassell, J., Stone, M., Yan, H.: Coordination and context-dependence in the generation of embodied conversation. In: Proceedings of the first international conference on Natural language generation-Volume 14. pp. 171–178. Association for Computational Linguistics (2000)
12. Chrisley, R.: Embodied artificial intelligence. Artificial intelligence **149**(1), 131–150 (2003)
13. Clair, A.S., Mead, R., Matarić, M.J., et al.: Monitoring and guiding user attention and intention in human-robot interaction. In: ICRA-ICAIR Workshop, Anchorage, AK, USA. vol. 1025 (2010)
14. Clark, H.H., Brennan, S.E.: Grounding in communication. In: Resnick, L., B., L., John, M., Teasley, S., D (eds.) Perspectives on Socially Shared Cognition, pp. 13–1991. American Psychological Association (1991)
15. Clark, H.H., Brennan, S.E.: Grounding in communication. Perspectives on socially shared cognition **13**(1991), 127–149 (1991)
16. Clark, H.H., Wilkes-Gibbs, D.: Referring as a collaborative process. Cognition **22**(1), 1–39 (Feb 1986). https://doi.org/10.1016/0010-0277(86)90010-7, `http://www.sciencedirect.com/science/article/pii/0010027786900107`
17. Cooper, R., Ginzburg, J.: Type theory with records for natural language semantics. The handbook of contemporary semantic theory p. 375 (2015)
18. Craik, K.J.W.: The nature of explanation. Cambridge University, Cambridge UK (1943)

19. De Groote, P.: Type raising, continuations, and classical logic. In: Proceedings of the thirteenth Amsterdam Colloquium. pp. 97–101 (2001)
20. Dillenbourg, P., Traum, D.: Sharing solutions: Persistence and grounding in multimodal collaborative problem solving. The Journal of the Learning Sciences **15**(1), 121–151 (2006)
21. Dobnik, S., Cooper, R., Larsson, S.: Modelling language, action, and perception in type theory with records. In: Constraint Solving and Language Processing, pp. 70–91. Springer (2013)
22. Dumas, B., Lalanne, D., Oviatt, S.: Multimodal interfaces: A survey of principles, models and frameworks. In: Human machine interaction, pp. 3–26. Springer (2009)
23. Eisenstein, J., Barzilay, R., Davis, R.: Discourse topic and gestural form. In: AAAI. pp. 836–841 (2008)
24. Eisenstein, J., Barzilay, R., Davis, R.: Gesture salience as a hidden variable for coreference resolution and keyframe extraction. Journal of Artificial Intelligence Research **31**, 353–398 (2008)
25. Evans, V.: Language and time: A cognitive linguistics approach. Cambridge University Press (2013)
26. Feldman, J.: Embodied language, best-fit analysis, and formal compositionality. Physics of life reviews **7**(4), 385–410 (2010)
27. Fernando, T.: Situations in ltl as strings. Information and Computation **207**(10), 980–999 (2009)
28. Fussell, S.R., Kraut, R.E., Siegel, J.: Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work. pp. 21–30. CSCW '00, ACM, New York, NY, USA (2000). https://doi.org/10.1145/358916.358947, `http://doi.acm.org/10.1145/358916.358947`
29. Fussell, S.R., Setlock, L.D., Yang, J., Ou, J., Mauer, E., Kramer, A.D.I.: Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. Hum.-Comput. Interact. **19**(3), 273–309 (Sep 2004). https://doi.org/10.1207/s15327051hci1903$_3$, `http://dx.doi.org/10.1207/s15327051hci1903_3`
30. Gergle, D., Kraut, R.E., Fussell, S.R.: Action As Language in a Shared Visual Space. In: Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work. pp. 487–496. CSCW '04, ACM, New York, NY, USA (2004). https://doi.org/10.1145/1031607.1031687, `http://doi.acm.org/10.1145/1031607.1031687`
31. Gibson, J.J., Reed, E.S., Jones, R.: Reasons for realism: Selected essays of James J. Gibson. Lawrence Erlbaum Associates (1982)
32. Gilbert, M.: On social facts. Princeton University Press (1992)
33. Ginzburg, J., Fernández, R.: Computational models of dialogue. The handbook of computational linguistics and natural language processing **57**, 1 (2010)
34. Goldman, A.I.: Interpretation psychologized*. Mind & Language **4**(3), 161–185 (1989)
35. Goldman, A.I.: Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press (2006)
36. Gordon, R.M.: Folk psychology as simulation. Mind & Language **1**(2), 158–171 (1986)
37. Graesser, A.C., Singer, M., Trabasso, T.: Constructing inferences during narrative text comprehension. Psychological review **101**(3), 371 (1994)
38. Heal, J.: Simulation, theory, and content. Theories of theories of mind pp. 75–89 (1996)

39. Johnson-Laird, P.N., Byrne, R.M.: Conditionals: a theory of meaning, pragmatics, and inference. Psychological review **109**(4), 646 (2002)
40. Johnson-Laird, P.: How could consciousness arise from the computations of the brain. Mindwaves. Oxford: Basil Blackwell pp. 247–257 (1987)
41. Kendon, A.: Gesture: Visible action as utterance. Cambridge University Press (2004)
42. 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)
43. Kiela, D., Bulat, L., Vero, A.L., Clark, S.: Virtual embodiment: A scalable long-term strategy for artificial intelligence research. arXiv preprint arXiv:1610.07432 (2016)
44. Kraut, R.E., Fussell, S.R., Siegel, J.: Visual Information As a Conversational Resource in Collaborative Physical Tasks. Hum.-Comput. Interact. **18**(1), 13–49 (Jun 2003). https://doi.org/10.1207/S15327051HCI1812$_2$, http://dx.doi.org/10.1207/S15327051HCI1812_2
45. Krishnaswamy, N., Pustejovsky, J.: Multimodal semantic simulations of linguistically underspecified motion events. In: Spatial Cognition X: International Conference on Spatial Cognition. Springer (2016)
46. Krishnaswamy, N., Pustejovsky, J.: Multimodal continuation-style architectures for human-robot interaction. arXiv preprint arXiv:1909.08161 (2019)
47. Lascarides, A., Stone, M.: Formal semantics for iconic gesture. In: Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (BRANDIAL). pp. 64–71 (2006)
48. Lascarides, A., Stone, M.: Discourse coherence and gesture interpretation. Gesture **9**(2), 147–180 (2009). https://doi.org/10.1075/gest.9.2.01las, http://www.jbe-platform.com/content/journals/10.1075/gest.9.2.01las
49. Lascarides, A., Stone, M.: A formal semantic analysis of gesture. Journal of Semantics p. ffp004 (2009)
50. Lücking, A., Mehler, A., Walther, D., Mauri, M., Kurfürst, D.: Finding recurrent features of image schema gestures: the figure corpus. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 1426–1431 (2016)
51. Lücking, A., Pfeiffer, T., Rieser, H.: Pointing and reference reconsidered. J. of Pragmatics **77**, 56–79 (2015)
52. Marshall, P., Hornecker, E.: Theories of embodiment in hci. The SAGE handbook of digital technology research **1**, 144–158 (2013)
53. Matuszek, C., Bo, L., Zettlemoyer, L., Fox, D.: Learning from unscripted deictic gesture and language for human-robot interactions. In: AAAI. pp. 2556–2563 (2014)
54. Mehlmann, G., Häring, M., Janowski, K., Baur, T., Gebhard, P., André, E.: Exploring a Model of Gaze for Grounding in Multimodal HRI. In: Proceedings of the 16th International Conference on Multimodal Interaction. pp. 247–254. ICMI '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2663204.2663275, http://doi.acm.org/10.1145/2663204.2663275
55. Narayanan, S.: Mind changes: A simulation semantics account of counterfactuals. Cognitive Science (2010)
56. Naumann, R.: Aspects of changes: a dynamic event semantics. Journal of semantics **18**, 27–81 (2001)
57. Pustejovsky, J.: The Generative Lexicon. MIT Press, Cambridge, MA (1995)

58. Pustejovsky, J.: Dynamic event structure and habitat theory. In: Proceedings of the 6th International Conference on Generative Approaches to the Lexicon (GL2013). pp. 1–10. ACL (2013)
59. Pustejovsky, J.: From actions to events: Communicating through language and gesture. Interaction Studies **19**(1-2), 289–317 (2018)
60. Pustejovsky, J.: From experiencing events in the action-perception cycle to representing events in language. Interaction Studies **19** (2018)
61. Pustejovsky, J., Krishnaswamy, N.: VoxML: A visualization modeling language. In: Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA), Paris, France (May 2016)
62. Pustejovsky, J., Krishnaswamy, N.: Embodied human-computer interactions through situated grounding. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–3 (2020)
63. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. Künstliche Intelligenz (2021)
64. Pustejovsky, J., Krishnaswamy, N.: Situated meaning in multimodal dialogue: Human-robot and human-computer interactions. Traitement Automatique des Langues **62**(1) (2021)
65. Pustejovsky, J., Moszkowicz, J.: The qualitative spatial dynamics of motion. The Journal of Spatial Cognition and Computation (2011)
66. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: gesture and speech. ACM Transactions on Computer-Human Interaction (TOCHI) **9**(3), 171–193 (2002)
67. Ravenet, B., Pelachaud, C., Clavel, C., Marsella, S.: Automating the production of communicative gestures in embodied characters. Frontiers in psychology **9**, 1144 (2018)
68. Shapiro, L.: The Routledge handbook of embodied cognition. Routledge (2014)
69. Skantze, G., Hjalmarsson, A., Oertel, C.: Turn-taking, feedback and joint attention in situated human-robot interaction. Speech Communication **65**, 50–66 (Nov 2014). https://doi.org/10.1016/j.specom.2014.05.005, `http://www.sciencedirect.com/science/article/pii/S016763931400051X`
70. Stalnaker, R.: Common ground. Linguistics and philosophy **25**(5-6), 701–721 (2002)
71. Tomasello, M., Carpenter, M.: Shared intentionality. Developmental science **10**(1), 121–125 (2007)
72. Turk, M.: Multimodal interaction: A review. Pattern Recognition Letters **36**, 189–195 (2014)
73. Unger, C.: Dynamic semantics as monadic computation. In: JSAI International Symposium on Artificial Intelligence. pp. 68–81. Springer (2011)
74. Zwaan, R.A., Pecher, D.: Revisiting mental simulation in language comprehension: Six replication attempts. PloS one **7**(12), e51382 (2012)
75. Zwaan, R.A., Radvansky, G.A.: Situation models in language comprehension and memory. Psychological bulletin **123**(2), 162 (1998)