Multimodal Semantics for Affordances and Actions[*]

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, as HCI becomes more multi-modal with the integration of gesture, gaze, posture, and other nonverbal behavior, it is important to understand the role played by affordances and their associated actions in human-object interactions (HOI), so as to facilitate reasoning in HCI and HRI environments. We outline the requirements and challenges involved in developing a multimodal semantics for human-computer and human-robot interactions. Unlike unimodal interactive agents (e.g., text-based chatbots or voice-based personal digital assistants), multimodal HCI and HRI inherently require a notion of embodiment, or an understanding of the agent's placement within the environment and that of its interlocutor. We present a dynamic semantics of the language, VoxML, to model human-computer, human-robot, and human-human interactions by creating multimodal simulations of both the communicative content and the agents' common ground, and show the utility of VoxML information that is reified within the environment on computational understanding of objects for HOI.

Keywords: Affordances · HCI · habitats · Common ground · multi-modal dialogue · VoxML · Embodiment.

1 Introduction

In this paper, we argue that, as HCI becomes more multimodal with the integration of gesture, gaze, posture, and other nonverbal behavior [12,25,41,53,77,89], it is important to understand the role played by affordances and their associated actions in human-object interactions (HOI), so as to facilitate reasoning in HCI and HRI environments. We outline the requirements and challenges involved

in developing a multimodal semantics for human-computer and human-robot interactions. Unlike unimodal interactive agents (e.g., text-based chatbots or voice-based personal digital assistants), multimodal HCI and HRI inherently require a notion of embodiment [2,17,37,53,83], or an understanding of the agent's placement within the environment and that of its interlocutor [24,47,78,49,58].

As natural language technology becomes ever-present in everyday life, people will expect artificial agents to understand language use as humans do, in a situated context. Nevertheless, most advanced neural AI systems fail at some types of interactions that are trivial for humans. Certain problems in both human-human communication and HCI cannot be solved without *situated reasoning*, meaning they cannot be adequately addressed with ungrounded meaning representation or cross-modal linking of instances alone. Examples include grounding an object and then reasoning with it (*"Pick up this box. Put it there."*), referring to a previously-established concept or instance that was never explicitly introduced into the dialogue, underspecification of deixis, and in general, dynamic updating of context through perceptual, linguistic, action, or self-announcement [1,72]. Without *both* a representation framework and mechanism for grounding references and inferences to the environment, such problems may well remain out of reach for NLP.

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. This in turn impacts how we computationally model human-human communicative interactions, with particular relevance to the shared understanding of affordances [28] and actions over objects.

All multimodal human-to-human communicative acts are inherently embodied. Therefore, modeling similar capabilities with computational agents necessitates a notion of "embodied HCI" (EHCI). Agents in this framework are embodied and situated, which affords them the ability to affect the world they inhabit (either real or virtual), but also requires them to have accurate and robust interpretive capabilities for multiple input modalities, which must run in real time. In addition, an artificial agent must be able to communicate with its human interlocutors using all communicative modalities humans may use, including natural language, body language, gesture, demonstrated action, emotional cues, etc. This paper describes the semantics of actions and object affordances and the impact such knowledge has on embodied reasoning [72,45]. While the dynamic semantics of epistemic updating in discourse has been extensively modeled, there has been less development of integrated models of the dynamics of actions and affordances in cooperative or goal-directed discourse. We present a dynamic semantics of the language, VoxML [70], to model human-computer, human-robot, and human-human interactions by creating multimodal simulations of both the communicative content and the agents' common ground [4,19,85,87], which is formalized in a data structure known as a *common ground*

structure [71]. 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. VoxML provides a representation for the situated grounding of expressions between individuals involved in a communicative exchange. It does this by encoding objects with rich semantic typing and action affordances, and actions themselves as multimodal programs, enabling contextually salient inferences and decisions in the environment. Underlying this model is a dedicated platform, VoxWorld [44], that is used to create these 3D realizations and deploy embodied multimodal agents.

We believe that the major issues in HCI for situated reasoning involve the multimodal grounding of expressions, as well as contextual reasoning with this information. In particular, we address the question of how to encode the knowledge associated with Human Object Interactions (HOI): how is object-specific behavioral knowledge encoded in our everyday interactions with the entities we encounter?

2   Modeling Human Object Interactions

When humans engage in conversation, the objects under discussion can range from things and events present in their shared communicative space, to entities and situations removed from the present context, and potentially even hypothetical or irrealis in nature [30,50]. Because the focus here is on situated HCI and HRI, we restrict the domain of discourse between the agents to those objects and events that are either present or emergent in an environment shared by the interlocutors. Even with such a seemingly limited context, the objects in a dialogue, either between two humans or between human and computer, carry much more semantic information than conventionally assumed in planning research. This includes knowledge for how the objects can be manipulated and used by an agent in space and time, that is, their *physical and functional affordances* [28,65]. Such information also includes knowledge of how an object is situated in the environment relative to an agent for specific purposes and actions, that is, its *habitat* [66,54]. These two parameters constitute a kind of *teleological* knowledge [67], and in the discussion below, we describe this information and what role it plays in both reasoning and communication for HCI.

There is currently a disconnect between semantic models that support linguistic analysis and processing of narrative text, dialogue, and image captions, and the interpretation and grounding that is actually required to fully understand how an event is situated in a context. 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 [13,55,93]. Here 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. [36]).

Figure 1: "Woman drinking coffee."

Actual situated meaning is much more involved than aligning captions and bounding boxes in an image: e.g., [34] 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 [86] or cosuppositional content [80]. Hence, we must assume that natural interactions with computers and robots have to account for interpreting and generating language and gesture.

For example, consider the event depicted in Figure 1. We assume that conventional semantic composition results in a logical form such as that shown in (1b); for convenience, we will also employ a situated representation that takes advantage of contextual Skolemization; that is, "a woman" will be denoted by w, and "coffee" will be denoted by c.

(1) a. A woman drinks coffee.
     b. $\exists x \exists y [\text{woman}(x) \land coffee(y) \land \text{drink}(x,y)]$
     c. $drink(w,c)$

Such representations need to be grounded, hence the recent interest in linking text, and captions in particular, to image-based information (in the form of annotated bounding boxes, etc). As useful as such cross-modal (image-caption) linking can be for Question Answering tasks [3,51], it does not provide sufficient information to perform situated or embodied reasoning. That is, no true model of the underlying human-object interaction can be extracted from such alignments.

Let's examine just what kind of information would be necessary to have regarding an event and its participants, so that novel inferencing and reasoning can be performed. We begin by creating a verbose gloss or dense paraphrase for the caption in Figure 1.

(2) a. *A woman drinking coffee.*
     b. A upright seated woman is holding in her hand, a cup filled with coffee while she drinks it.
     c. The cup is upright so the container portion (inside) is able to hold coffee.
     d. She is holding the cup by an attached handle.
     e. The cup is tilted towards her and touches her partially open mouth, in order to allow drinking.

Similarly, the caption for Figure 2 is perfectly adequate as a description of the situation for a human to interpret. But for a computer to be able to understand the caption by itself or indeed even with the image provided, there needs to be an interpretation of how the human and the objects are interacting.

A similar "unpacking" of the situation would involve a dense paraphrase as shown below, where the semantic and pragmatic presuppositions in the caption are made explicit. As in the previous example in Figure 1, this spells out: the orientation and facing of the human to the object; touch points (hot spots) on the object, e.g., keyboard; pose and embodied actions, e.g., typing with both hands.

Figure 2: A man working at a desk.

(3) a. *A man working at a desk.*
     b. A upright man is seated in a chair, typing with both hands on the keyboard of a laptop, which is on the top surface of a table.
     c. The chair he is seated in is close enough to the table for him to reach the keyboard.
     d. The laptop is open, with the keyboard exposed flat and the screen facing the man.
     e. The man is facing the computer and the desk.

Where does this human-object interaction (HOI) information come from? Ideally, it can be learned through multimodal alignment of image and caption embeddings [92,31,74,14], but this is still a difficult problem within the knowledge acquisition community. Explicit representation has been disfavored in modern AI, but typical neural networks that learn implicit representations are treated as passive recipients of data, with the question of context- and situation-sensitive grounding treated as something of an inconvenience [16]. There is less attention paid to letting the current state of the world, as opposed to reams of pre-existing data, be "its own model", per Brooks [9]. In the present work, we start with an initial library of human-object interaction pairs, encoded as affordances in their habitats (cf. next section), and then discuss experiments where such HOI properties can be learned.

3   Modeling Habitats and Affordances

In everyday discourse, when referring to objects and events, humans expect each other to know more than what a word refers to: e.g., a cup is an artifact, coffee is a substance, and a toy is an inanimate object. Such categorical knowledge is typically represented as a type structure, such as that shown in Figure 3 below.

Such typing is useful in linguistic interpretations so as to ensure that predicates select the appropriate types of arguments in composition, as illustrated in the semantic type derivation in (4), for the caption from Figure 1, "Woman drinking coffee."
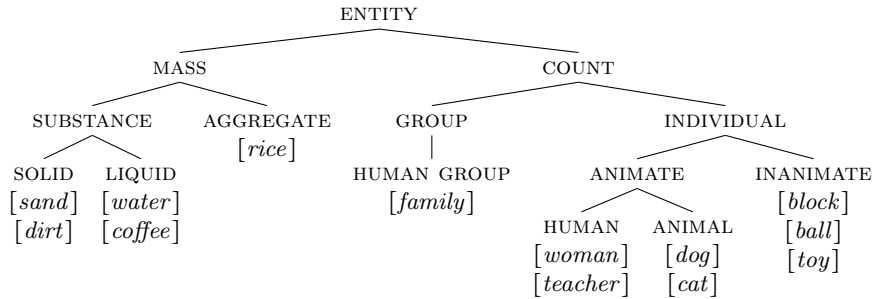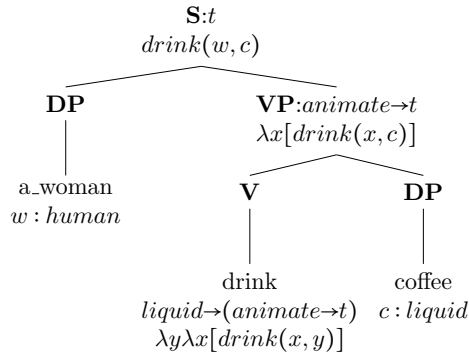
```
                                ENTITY
                    ┌──────────────┴──────────────┐
                  MASS                           COUNT
            ┌───────┴───────┐            ┌─────────┴─────────┐
        SUBSTANCE       AGGREGATE      GROUP             INDIVIDUAL
        ┌────┴────┐      [rice]          │          ┌────────┴────────┐
     SOLID     LIQUID              HUMAN GROUP   ANIMATE          INANIMATE
     [sand]    [water]              [family]   ┌────┴────┐         [block]
     [dirt]    [coffee]                      HUMAN    ANIMAL       [ball]
                                            [woman]   [dog]        [toy]
                                           [teacher]  [cat]
```

Figure 3: *Classical* ENTITY *subtyping*

(4)

```
                        S:t
                     drink(w, c)
              ┌──────────┴──────────┐
             DP                  VP:animate→t
              │                  λx[drink(x, c)]
              │               ┌───────┴───────┐
          a_woman             V              DP
          w : human           │               │
                              │               │
                            drink           coffee
                    liquid→(animate→t)    c : liquid
                      λyλx[drink(x, y)]
```

However, as the situated paraphrases from the previous section suggest, there is information missing from conventional semantic representations for deeper reasoning and inferencing about events and situations. In fact, we understand an entire set of object attributes as well as a network of relations concerning how the object appropriately participates in the situation under discussion. Many of these involve human-object interactions (HOIs), and our knowledge of things is predicated on are understanding of how to interact with them. Hence, just as a cup will conventionally have an ontological relationship of a handle to the whole structure, there is a conventional presupposition that the orientation of the cup exposes the concavity of the interior to enable the functioning of the cup. Notice that there are several things mentioned here that are partially spatial and partially teleological. This is what we will call a *teleotopological relationship* [68]. There are aspects of qualitative spatial reasoning that are implicated in this relationship such as orientation of the concavity [26].

In order to create a compositional interpretation from a sentences such as "the cup is on the table" and further "the coffee is in the cup," we must likely have some semantic encoding lexically associated with all these objects as well

as compositionally for how they are physically and spatially configured relative to each other.

Consider what some of the relevant parameters from this example are, that may be tied to the way specific objects inhabit their situation. Assuming that an object such as a cup, typed as a container is an asymmetric object, such as a cylinder, it would appear that orientation information is critical for enabling the use or function the object *qua* container. In fact, only when the cup's orientation facilitates containment can the function be "activated", as it were. This references two notions that are critical for reasoning about objects and HOI generally: we encode *what* the function associated with an object is (its affordance), but just as critically, we also identify *when* it is active (its habitat).

Similarly, consider the implicit knowledge we seem to exploit for the way we refer to and interact with an instrument such a spoon, knife, or fork. Each of these can be considered a tool for eating. Each is asymmetric in form with a handle, and "another" end that is associated with the function of eating, or a subevent of eating. Hence when asked to pick up a spoon and start eating from a bowl, we naturally grasp the handle and know where to put the spoon relative to the bowl, etc. This is a simple but telling example of the myriad actions which are afforded by and encoded with specific objects, in order to engage in specific activities.

Figure 4: Top: *Spoon* in different habitats allowing holding (left) and stirring (right). Bottom: *Knife* allowing spreading (left) and cutting (right).

In the images above, we see roughly the following condition-action pair:

(5) *If* **Habitat** *then* **Action**

For a spoon, we identify at least two functions, associated with distinct actions, each of which is enabled by distinct habitats in Figure 4 Top:

(6) a. If *spoon's concavity is vertical*, then it can *support containment of a sub-stance*;
     b. If *spoon's major axis is vertical*, then it can *support mixing.*

Similar remarks hold for the orientation of a knife, as illustrated above in Figure 4 (Bottom):

(7) a. If *knife's zero convexity (sheet) is horizontal*, then it can *support spreading of a substance*;
     b. If *knife's zero convexity (sheet) is vertical*, then it can *support cutting or separating.*

   Given the notion of affordance and how we interact with the objects in our environment, we can refactor the classic entity type ontology from Figure 3 in terms of how it is possible to interact with the objects in our environment. This is shown, in part, below in the modal definition of objects as possible behaviors in (8).

(8) Refactoring Entity Types as Modal Actions (Affordances)
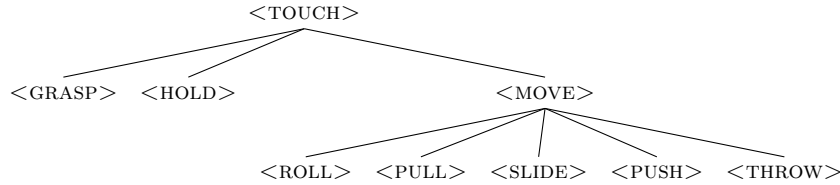
   In the next section, we outline a language that captures much of the information missing from conventional semantic models of events, as shown above: in particular, information encoding object geometry, as well as event-participant configuration and orientation constraints. This language, VoxML, encodes knowledge about objects, their attributes, events, and functions to their visual and spatial instantiations, called a "visual object concept", or *voxeme*.

4   VoxML

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". For example, a semantic model of perceived *object teleology*, as introduced by Generative Lexicon (GL) with the Qualia Structure, for example [65], as well as *object affordances* [28] is useful to help ground expression meaning to speaker intent. As an illustration, consider first how such information is encoded, and then exploited in reasoning. Knowledge of objects can be partially contextualized through their *qualia structure* [69], where each Qualia role can be seen as answering a specific question about the object it is bound to:

(9) a. *Formal*: encoding taxonomic information about the lexical item (the IS-A relation);

   b. *Constitutive*: encoding information on the parts and constitution of an object (PART-OF or MADE-OF relation);

   c. *Telic*: encoding information on purpose and function (the `used-for` or FUNCTIONS-AS relation);

   d. *Agentive*: encoding information about the origin of the object (the CREATED-BY relation).

In human-human communication, objects under discussion (cf. [30]) can be partially contextualized through their semantic type and their qualia structure: a food item has a TELIC value of *eat*, an instrument for writing, a TELIC of *write*, a cup, a TELIC of *hold*, and so forth. For example, the lexical semantics for the noun *chair* in (10), assuming a Generative Lexicon encoding, carries a TELIC value of *sit_in*, while the concept of *letter* carries a TELIC value of *read* and an AGENTIVE value of *write*. Such object-based information will need to be recognized by computational agents in HCI and HRI, as well, as it is so crucial for situational reasoning is dialogue and discourse.

(10) $\lambda x \begin{bmatrix} \textbf{chair} \\ \text{AS} = [\text{ARG1} = x : e] \\ \text{QS} = \begin{bmatrix} \text{F} = phys(x) \\ \text{T} = \lambda z, e[sit\_in(e, z, x)] \end{bmatrix} \end{bmatrix}$

Notice that, while an artifact may be designed for a specific purpose, this purpose can only be achieved under specific circumstances. To account for this context-dependence, [66,71] enrich 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 habitat (or 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. It is necessary to specify the precondition context $\mathcal{C}$, since this enables the local modality to be satisfied. An illustration of what the resulting knowledge structure for the habitat of a chair is shown in the QS entry below.

(11) $\lambda \mathcal{C} \lambda x \begin{bmatrix} \textbf{chair} \\ \text{F} = [phys(x), on(x, y_1), in(x, y_2), orient(x, up)] \\ \text{C} = [seat(x_1), back(x_2), legs(x_3), clear(x_1)] \\ \text{T} = \lambda z \lambda e[\mathcal{C} \rightarrow [sit(e, z, x)]\mathcal{R}_{sit}(x)] \\ \text{A} = [made(e', w, x)] \end{bmatrix}$

The habitat for an object is built by first placing it within an *embedding space* and then contextualizing it. For example, in order to use a table, the top has to be oriented upward, the surface must be accessible, and so on. A chair must also be oriented up, the seat must be free and accessible, it must be able to support the user, etc., [21,23].

The notion of habitat described above and the attached behaviors that are associated with an object are further developed in [70], where an explicit connection to Gibson's ecological psychology is made [29], along with a direct encoding of the *affordance structure* for the object [28]. 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.

Extending this notion, we define a habitat as a representation of an object situated within a simulation, a partial minimal model [8,40,43]; in this sense, it is a directed enhancement of the qualia structure. Multi-dimensional affordances determine how habitats are deployed and how they modify or augment the context, and compositional operations include procedural (simulation) and operational (selection, specification, refinement) knowledge.

The language used to construct this simulation is called VoxML (Visual Object Concept Modeling Language) [70]. VoxML 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 [44]. It adopts the basic semantic typing for objects and properties from Generative Lexicon and the dynamic interpretation of event structure developed in [73], along with a continuation-based dynamic interpretation for both sentence and discourse composition [6,5,22].

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*.

Entities modeled in VoxML can be OBJECTS, programs, or logical types. OBJECTS are logical constants; PROGRAMS are n-ary predicates that can take objects or other evaluated predicates as arguments; logical types can be divided into ATRIBUTES, RELATIONS, and FUNCTIONS, all predicates which take OBJECTS as arguments. ATTRIBUTES and RELATIONS evaluate to states, and FUNCTIONS evaluate to geometric regions. These entities can then compose into visualizations of natural language concepts and expressions. For example, the attributes associated with objects such as *cup*, *chair*, and *block*, include the following:

LEX	OBJECT's lexical information
TYPE	OBJECT's geometrical typing
HABITAT	OBJECT's habitat for actions
AFFORD_STR	OBJECT's affordance structure
EMBODIMENT	OBJECT's agent-relative embodiment

The LEX attribute contains the subcomponents PRED, the predicate lexeme denoting the object, and TYPE, the object's type according to Generative Lexicon.

Voxemes representing humans or IVAs are lexically typed as *agents*, but artificial 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]][3], 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.

An OBJECT's voxeme structure provides *habitats*, which are situational contexts or environments conditioning the object's *affordances*, which may be either "Gibsonian" affordances [28] or "Telic" affordances [65,66]. 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.

For example, the object geometry for the concept [[CUP]], along with the constraints on symmetry, is illustrated below.

$$(12) \quad \begin{bmatrix} \textbf{cup} \\ \text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{cylindroid[1]} \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATIONAL\_SYMMETRY} = \{Y\} \\ \text{REFLECTION\_SYMMETRY} = \{XY, YZ\} \end{bmatrix} \end{bmatrix}$$

Consider now the various habitats identified with [[CUP]].

$$(13) \quad \begin{bmatrix} \textbf{cup} \\ \text{HABITAT} = \begin{bmatrix} \text{INTRINSIC} = [2]\begin{bmatrix} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{bmatrix} \\ \text{EXTRINSIC} = [3]\begin{bmatrix} \text{UP} = align(Y, \mathcal{E}_{\perp Y}) \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Finally, given these habitats, we can identify the associated behaviors that are enabled (afforded) in such situations:

$$(14) \quad \begin{bmatrix} \textbf{cup} \\ \text{AFF\_STR} = \begin{bmatrix} A_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x) \\ A_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x) \\ A_3 = H_{[2]} \rightarrow [grasp(x, [1])]hold(x, [1]) \\ A_4 = H_{[3]} \rightarrow [roll(x, [1])]\mathcal{R} \end{bmatrix} \end{bmatrix}$$

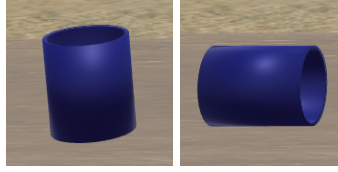[3]Beginning in [42], voxemes have been denoted [[VOXEME]].

Figure 5: Cup in different habitats allowing sliding and holding (left) and rolling (right).

Indeed, object properties and the events they facilitate are a primary component of situational context. In Fig. 12, 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. (15)). 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 conventions, where the Y-axis is vertical), and is smaller than and movable by the artificial agent. The remainder of VoxML typing structure is devoted to habitat and affordance structures, which we discuss below.

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

$$
\begin{bmatrix}
\textbf{cup} \\
\text{LEXICAL} = \begin{bmatrix} \text{PREDICATE} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj} \bullet \textbf{artifact} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix} \text{HEAD} = \textbf{cylindroid}[\textbf{1}] \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATIONAL\_SYMMETRY} = \{Y\} \\ \text{REFLECTION\_SYMMETRY} = \{XY, YZ\} \end{bmatrix} \\
\text{HABITAT} = \begin{bmatrix} \text{INTRINSIC} = [2]\begin{bmatrix} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{bmatrix} \\ \text{EXTRINSIC} = [3]\begin{bmatrix} \text{UP} = align(Y, \mathcal{E}_{\perp Y}) \end{bmatrix} \end{bmatrix} \\
\text{AFF\_STR} = \begin{bmatrix} A_1 = H_{[2]} \rightarrow [put(x, on([1]))]support([1], x) \\ A_2 = H_{[2]} \rightarrow [put(x, in([1]))]contain([1], x) \\ A_3 = H_{[2]} \rightarrow [grasp(x, [1])]hold(x, [1]) \\ A_4 = H_{[3]} \rightarrow [roll(x, [1])]\mathcal{R} \end{bmatrix} \\
\text{EMBOD} = \begin{bmatrix} \text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true} \end{bmatrix}
\end{bmatrix}
$$

(12-14) respectively show the typing, habitat, and affordance structure of [[CUP]], which are brought together in the complete VoxML encoding in (15). Bracketed numbers, (e.g., [1]) are reentrancy indices; terms annotated with the same number refer to the same entitiy. 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.

Now let us consider how this model informs the interactions available to agent in a simple environment of a child playing with blocks, as shown in Figure 6.

Figure 6: Girl stacking blocks.

Each of these blocks encodes the specific set of affordances associated with its class: namely, given the appropriate habitat, they can be grasped, and then moved (picked up, slid, pushed, puled, thrown, but not rolled!). In the current situation depicted in this image, however, only the top-most blocks are immediately available for these affordances.

As we show in the next section, object properties associated with how an agent can behave or interact with them can be the key towards classification and discrimination of objects in an otherwise homogeneous environment.

5   Reasoning with Affordances

To demonstrate the information that object affordances and situated grounding as encoded in VoxML provides, let us examine a simple object classification example. Humans are efficient at seeking out experiences that are maximally informative about their environment [52,59,61,76,81]. We explore the physical world to practice skills, test hypotheses, learn object affordances, etc. [11,32,33,60,63,64,84]. Young children, in particular, can rapidl generalize from previous to new experiences with few or even no examples [18,20,88].

Meanwhile, artificial neural networks require large numbers of samples to train. A single cortical neuron may take 5-8 layers of artificial neurons to approximate [7]. Common few-, one-, or zero-shot learning approaches in AI provide at best a rough simulacrum of human learning and generalization [39,62,94]. Recent success in few-shot learning in end-to-end deep neural systems still requires extensive pre-training and fine-tuning, often on special hardware [10] or specific task formulation [79].

In AI, object identification is usually approached as a computer vision task [46]. While convolutional neural nets, the most common method for object identification in modern computer vision, do appear to optimize for a form of invariance [35], and invariance is important for semantic interpretation [48], visual input alone is only one part of how humans learn to identify objects [90].

To examine the different affordances of objects, an artificial agent can interact with various objects and see how they behave differently under the same circumstances. To test this, we trained a TD3 reinforcement learning policy [27]

to learn to stack two cubes. We then used a successful cube-stacking policy to make the agent attempt to stack other spheres, cylinders, and capsules (Figure 7) on a block, forcing it to stack the other objects *as if they were cubes*. This control structure allowed us to identify differences in the behaviors of the different objects in the stacking task. Since these behaviors can be described in terms like "cubes stack successfully," "spheres roll off," etc., they can be described in terms of the object's *affordances*.
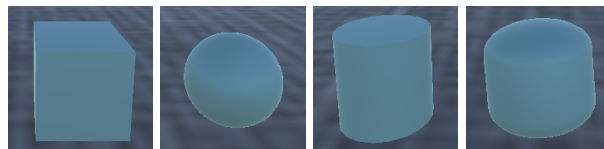

Figure 7: Test objects.

As the agent executes the trained policy over the various objects, we gather information about each stacking attempt from the VoxSim virtual environment. At each timestep we store the type of the theme object, its rotation in radians at episode start, radians between the world upright axis and the object upright (+Y) axis, the numerical action executed, the object rotation and offset from world upright after the action, the state observation after action completion, the reward for the attempt, the cumulative total reward over the episode, and the cumulative mean reward over the episode. At the end of the action, a small "jitter" is applied to the object, to simulate the small force exerted on an object when it is released from a grasp. We also store the vector representing the magnitude and direction of this small force. This jitter force is applied perpendicular to the major rotational axis of the theme object if one exists, or randomly if the object is symmetric around all axes. Therefore, the jitter applied to cylinder or capsule is applied perpendicularly to the object's Y-axis, while the jitter applied to a cube or sphere is random. Therefore the post-action jitter implicitly encodes information about the objects' habitats, while some other parameters gathered implicitly encode affordances. Compare a subset of parameters extracted from the environment from a single stacking attempt each with a cube and a sphere, and two attempts with a cylinder (Table 1).

Object	Jitter			θ after Action	Stack Height
cube	-1.472×10^{-4}	0	2.021×10^{-4}	0.02238165	2
sphere	8.165×10^{-5}	0	-2.363×10^{-5}	2.134116	1
cylinder	0	0	2.5×10^{-4}	0.01457105	2
cylinder	0	0	2.5×10^{-4}	1.570793	1

Table 1: Observations gathered during stacking task with multiple objects.

"Stack Height" indicates the number of stacked objects after the action was complete. 2 indicates the object was stacked successfully, while 1 indicates that it fell off when the action was complete, leaving only the bottom block in the stack. "θ after Action" is the net angle (in radians) of the object space upright ($+Y$) from the world space upright ($+Y$) axis. Values close to 0 indicate that the object is upright while values near $\pm\frac{\pi}{2}$ indicate that the object is lying on its side. A cube, which is flat on all sides, can rest stably at any multiple of $\frac{\pi}{2}$ radians. In Table 1, we see that the stacking attempt with a cube results in a stable stack (height 2) with the top object sitting upright. The sphere, which rolls off, does not stack (height 1), and comes to rest at an arbitrary angle.

The cylinder shares properties with cubes (flat ends) and others with spheres (round sides), and this behavior can be seen in the last two rows. In the second-to-last row, the cylinder stacks successfully (height 2), and is resting upright ($\theta \approx 0$). In the sample shown in the last row, the stacking attempt was made with the cylinder on its side ($\theta \approx \frac{\pi}{2}$). Since this places the cylinder's round surface in contact with its supporting surface, this habitat does not afford sustained support, and the cylinder rolls off (indicated by the height value of 1). We can also see that the direction of the jitter implicitly encodes the axis of symmetry of the object, and therefore the habitat.

We use all the parameters gathered above to train a model that will predict the object type from its behavior in the stacking task. We use a 1D convolutional neural net for this task. Since episodes can be variable length (depending on how successful the policy was at stacking the object in question), we pad out the length of each input to 10 timesteps, copying the last sample out to the padding length. Therefore an episode where the policy stacked the object successfully on the first try will consist of 10 identical timestep representations, while an episode where the agent tried and failed the stack the object 10 times will have 10 different timestep representations, so stackable objects like cubes and (upright) cylinders will have more consistent representations across the 10 timestep sample while less stackable objects will exhibit more variation across the 10 attempts.

The classifier consists of two convolutional layers (256 and 128 hidden units respectively). The filter size in the first layer is c, a variable equal to the number of parameters saved at each evaluation timestep during data gathering ($c = 19$ here) and a stride length of 8, and the second layer uses a filter size of 4 and a stride length of 2. This allows the convolutions to generate feature maps in the hidden layers that are approximately equal to the size of a single timestep sample, and convolving over this approximates observing each timestep of the episode in turn. The convolutional layers are followed by two 64-unit fully-connected layers and a softmax layer. We train for 500 epochs using the Adam optimizer [38], a batch size of 100 (= 10 episodes) and a learning rate of 0.001. Figure 8 shows the classification results over an unseen test set of 100 episodes worth of samples for each object.

Not only can the objects be predicted from their behavior, meaning the that affordances encode important information about object type, but the post-action jitter features, that specifically encode the dependency between habitat
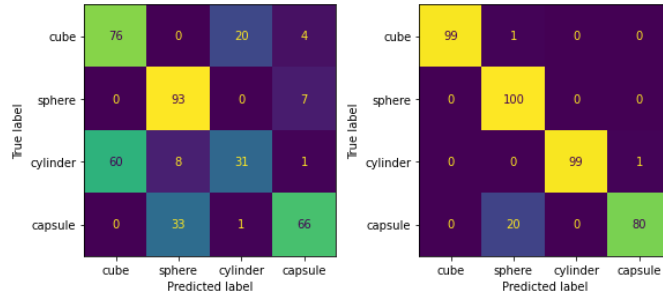
Figure 8: 1D CNN behavioral features classifier results. First chart shows results without the input of the implicit habitat and affordance information encoded in the post-action jitter. Second chart shows results with those input features.

and affordance by virtue of the object's axis of symmetry, increases classifier performance by *28%*, from **66.5%** without the jitter features to **94.5%**. Without the jitter features, most confusions are between the objects that share very broad stacking behavior: cube and cylinder (mostly stackable), and sphere and capsule (mostly unstackable). Implicit information about affordances and habitats makes the difference.

Of course, humans can readily tell that objects are different because they look different. Sometimes interaction is not needed. Therefore we compare the performance of the behavior-based classifier to a 2D CNN CIFAR-10 style object detector. We crop and downsample all images to 84×84 pixels, and use 3 2D convolutional layers with a filter size of 16, a 3×3 stride, and 2×2 average pooling, followed by a 64-unit fully-connected layer before the softmax layer. We train for 500 epochs using the Adam optimizer and a learning rate of 0.01, which are the same learning hyperparameters as the behavior-based 1D CNN.

This classifier achieves a validation accuracy of 97.5%, but when evaluated against an unseen test set of 140 novel images of the four object classes (35 images each), accuracy falls below the behavior-based classifier, to **90.7%** (Figure 9).
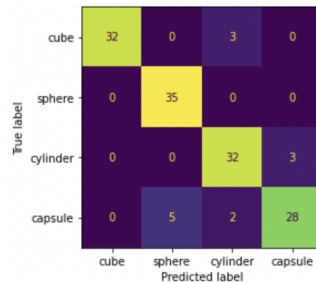

Figure 9: 2D CNN visual classifier results.

A look at the mistakes this classifier makes (Figure 10) reveals why. If the object is occluded, as in the top row, or distored due to perspective, the vision classifier naturally fails often.

Figure 10: Sample of misrecognized objects. From left: cylinder misrecognized as cube (1), and capsule (2), cylinder misrecognized as cube (3), sphere (4), and capsule (5), and capsule misrecognized as sphere (6).

Therefore while we can see that some objects are clearly visually distinct, like cube vs. sphere, other object classes are more difficult to distinguish visually. To confirm this, we go inside the 2D CNN and draw out the 64-dimensional embedding vectors from the final fully-connected layer. An embedding vector represents what a sample input is transformed into in the interior of a neural network after being multiplied by the optimized weight matrices in each layer in turn, until it reaches the softmax layer. These can be used to quantitatively assess the similarity of different input samples to each other. Figure 11 shows the cosine similarity of each of ther 140 test visual embeddings to each other embedding.
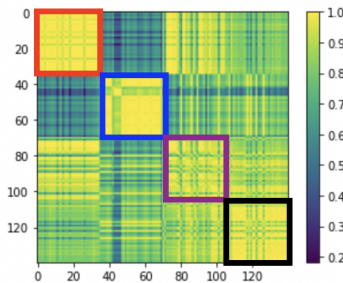

Figure 11: Cosine similarity matrix of visual embedding vectors from 2D CNN.

The red box indicates the cube embedding vectors, blue the sphere embeddings, purple the cylinder embeddings, and black the capsule embeddings. A brighter color indicates more similarity. We can see that cube embeddings are obviously most similar to each other, as are sphere embeddings, but when we look at the cylinder and capsule embeddings, the similarities are much less ob-

vious, and therefore the softmax layer of the networks makes its final prediction will less confidence in those cases.

Cube and sphere are also the most distinct objects in the behavior-based classifier, and the habitat and affordance information appears to be critical to making the more subtle and circumstantial distinctions in the behavior of cylinders and capsule. These implicit semantics are what VoxML is designed to explicitly encode.

We therefore propose that we can learn and populate VoxML encodings themselves through multiple inputs, such as interactive behavior, communication, and qualitative spatial relation calculi (e.g., [56,57,75]). For example, let A and B be two objects. If the Y-coordinate of A's position is above the Y-coordinate of B's position, and A and B are *externally connected* (touching), then it is likely that *B supports A* or *A is on top of B*. This can be learnable though interaction by learning a correlation between a embedding or vector representation of the resultant state and a symbol that denotes that state, such as a classification label, or ideally a word.

6   Conclusion and Future Directions

In this paper, we have outlined the requirements in developing a multimodal semantics for human-computer and human-robot interactions. We presented a model for how to encode, reason with, and learn object affordances in dynamic human-object interactions (HOI). Being able to identify and then perform inferencing from the modal possibilities inherent in the objects one encounters is an essential component of any natural HCI or HRI system. We introduced the language, VoxML, which provides a representation for the situated grounding of expressions between individuals involved in a communicative exchange. By providing a rich semantic typing and encoding of action affordances for objects, and for actions as multimodal programs, contextually salient inferences and decisions are made available in the environment as the interaction unfolds.

One current line of research is to examine how this model can inform the interpretation of student behavior in classroom interactions, as well as subsequent development of curriculum for use in middle school education. This is part of research conducted in the context of the NSF National AI Institute for Student-AI Teaming (iSAT)[4]. The goal is to imagine a range of increasingly sophisticated "AI partners" that a teacher could have as an assistant in a classroom. One particular task for this partner would be the interpretation of student behavior and attention to the lesson at hand. This is a challenging problem for AI but not for a seasoned teacher, since the latter can understand when a student is engaged in on-topic behavior or when they are acting out.

Following developments within the area of embodied cognition [15,83,91,82], it is interesting to see how this plays out. As a case in point, we have examined extensive videos of classroom lessons from an urban public school, while they

---

[4] https://www.colorado.edu/today/ai-education

are engaged in science immersion units. What is immediately clear is the need to distinguish between actions that are associated with an "embodied solution" to a task in the curriculum, from simply random actions performed off-topic. For example, consider the task of determining the characteristics of a *disk*, from a collection of objects. One way is to determine the relative height and diameter of each object. However, another means is to test the objects through experienced play, determining whether, for example, they can be thrown like a disk. In the images below, we see a girl testing a roll of tape, which satisfies the embodied action associated with a disk: *it can be grasped, flicked and then released, and it moves through the air.* That is, the girl is answering the question through actions on the object, using her body and reasoning from the consequences of the actions. This is embodied cognition.

Figure 12: Roll of tape can be thrown like a *disk*: it can be both "flicked" between thumb and fingers (left) and "released" (right).

This is an embodied cognitive solution utilizing the notion of affordances as developed in this paper, and is, we believe, an interesting direction for further research [45,72].

References

1. Alikhani, M., Khalid, B., Shome, R., Mitash, C., Bekris, K., Stone, M.: That and there: Judging the intent of pointing actions with robotic arms. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10343–10351 (2020)
2. Anderson, M.L.: Embodied cognition: A field guide. Artificial intelligence **149**(1), 91–130 (2003)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Asher, N.: Common ground, corrections and coordination. Journal of Semantics (1998)
5. Asher, N., Pogodalla, S.: Sdrt and continuation semantics. In: JSAI International Symposium on Artificial Intelligence. pp. 3–15. Springer (2010)
6. Barker, C., Shan, C.c.: Continuations and natural language, vol. 53. Oxford Studies in Theoretical Linguistics (2014)

7. Beniaguev, D., Segev, I., London, M.: Single cortical neurons as deep artificial neural networks. bioRxiv p. 613141 (2020)
8. Blackburn, P., Bos, J.: Computational semantics. Theoria: An International Journal for Theory, History and Foundations of Science pp. 27–45 (2003)
9. Brooks, R.A.: Intelligence without representation. Artificial intelligence **47**(1-3), 139–159 (1991)
10. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020)
11. Caligiore, D., Ferrauto, T., Parisi, D., Accornero, N., Capozza, M., Baldassarre, G.: Using motor babbling and hebb rules for modeling the development of reaching with obstacles and grasping. In: International Conference on Cognitive Systems. pp. E1–8 (2008)
12. Cassell, J., Sullivan, J., Churchill, E., Prevost, S.: Embodied conversational agents. MIT press (2000)
13. Chai, J.Y., Fang, R., Liu, C., She, L.: Collaborative language grounding toward situated human-robot dialogue. AI Magazine **37**(4), 32–45 (2016)
14. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: 2018 ieee winter conference on applications of computer vision (wacv). pp. 381–389. IEEE (2018)
15. Chemero, A.: Radical embodied cognitive science. MIT press (2011)
16. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: Proceedings of the IEEE international conference on computer vision. pp. 2722–2730 (2015)
17. Chrisley, R.: Embodied artificial intelligence. Artificial intelligence **149**(1), 131–150 (2003)
18. Clark, A.: Language, embodiment, and the cognitive niche. Trends in cognitive sciences **10**(8), 370–374 (2006)
19. Clark, H.H., Brennan, S.E.: Grounding in communication. Perspectives on socially shared cognition **13**(1991), 127–149 (1991)
20. Colung, E., Smith, L.B.: The emergence of abstract ideas: Evidence from networks and babies. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences **358**(1435), 1205–1214 (2003)
21. Coventry, K., Garrod, S.C.: Spatial prepositions and the functional geometric framework. towards a classification of extra-geometric influences (2005)
22. De Groote, P.: Type raising, continuations, and classical logic. In: Proceedings of the thirteenth Amsterdam Colloquium. pp. 97–101 (2001)
23. Dobnik, S., Cooper, R.: Interfacing language, spatial perception and cognition in type theory with records. Journal of Language Modelling **5**(2), 273–301 (2017)
24. Fischer, K.: How people talk with robots: Designing dialog to reduce user uncertainty. AI Magazine **32**(4), 31–38 (2011)
25. Foster, M.E.: Enhancing human-computer interaction with embodied conversational agents. In: International Conference on Universal Access in Human-Computer Interaction. pp. 828–837. Springer (2007)
26. Freksa, C.: Using orientation information for qualitative spatial reasoning. Springer (1992)
27. Fujimoto, S., Hoof, H., Meger, D.: Addressing function approximation error in actor-critic methods. In: International Conference on Machine Learning. pp. 1587–1596. PMLR (2018)
28. Gibson, J.J.: The theory of affordances. Perceiving, Acting, and Knowing: Toward an ecological psychology pp. 67–82 (1977)

29. Gibson, J.J.: The Ecological Approach to Visual Perception:. Psychology Press (1979)
30. Ginzburg, J.: Interrogatives: Questions, facts and dialogue. The handbook of contemporary semantic theory. Blackwell, Oxford pp. 359–423 (1996)
31. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8359–8367 (2018)
32. Gopnik, A.: How babies think. Scientific American **303**(1), 76–81 (2010)
33. Gottlieb, J., Oudeyer, P.Y.: Towards a neuroscience of active sampling and curiosity. Nature Reviews Neuroscience **19**(12), 758–770 (2018)
34. Hunter, J., Asher, N., Lascarides, A.: A formal semantics for situated conversation. Semantics and Pragmatics **11** (2018)
35. Kayhan, O.S., Gemert, J.C.v.: On translation invariance in cnns: Convolutional layers can exploit absolute spatial location. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14274–14285 (2020)
36. 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)
37. 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)
38. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
39. Knudsen, E.I.: Supervised learning in the brain. Journal of Neuroscience **14**(7), 3985–3997 (1994)
40. Konrad, K.: 4 minimal model generation. In: Model Generation for Natural Language Interpretation and Analysis, pp. 55–56. Springer (2004)
41. Kopp, S., Wachsmuth, I.: Gesture in embodied communication and human-computer interaction, vol. 5934. Springer (2010)
42. Krishnaswamy, N.: Monte-Carlo Simulation Generation Through Operationalization of Spatial Primitives. Ph.D. thesis, Brandeis University (2017)
43. Krishnaswamy, N., Pustejovsky, J.: Multimodal semantic simulations of linguistically underspecified motion events. In: Spatial Cognition X, pp. 177–197. Springer (2016)
44. Krishnaswamy, N., Pustejovsky, J.: VoxSim: A visual platform for modeling motion language. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics. ACL (2016)
45. Krishnaswamy, N., Pustejovsky, J.: The role of embodiment and simulation in evaluating hci: Experiments and evaluation. In: International Conference on Human-Computer Interaction. pp. 220–232. Springer (2021)
46. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
47. Kruijff, G.J.M., Lison, P., Benjamin, T., Jacobsson, H., Zender, H., Kruijff-Korbayová, I., Hawes, N.: Situated dialogue processing for human-robot interaction. In: Cognitive systems, pp. 311–364. Springer (2010)
48. Lakoff, G.: The invariance hypothesis: Is abstract reason based on image-schemas? (1990)
49. Landragin, F.: Visual perception, language and gesture: A model for their understanding in multimodal dialogue systems. Signal Processing **86**(12), 3578–3595 (2006)

50. Larsson, S., Ericsson, S.: Godis–issue-based dialogue management in a multi-domain, multi-language dialogue system. In: Demonstration Abstracts, ACL-02 (2002)
51. Lin, X., Parikh, D.: Leveraging visual question answering for image-caption ranking. In: European conference on computer vision. pp. 261–277. Springer (2016)
52. Markant, D.B., Gureckis, T.M.: Is it better to select or to receive? learning via active and passive hypothesis testing. Journal of Experimental Psychology: General **143**(1), 94 (2014)
53. Marshall, P., Hornecker, E.: Theories of embodiment in hci. The SAGE handbook of digital technology research **1**, 144–158 (2013)
54. McDonald, D., Pustejovsky, J.: On the representation of inferences and their lexicalization. In: Advances in Cognitive Systems. vol. 3 (2014)
55. Misra, D., Langford, J., Artzi, Y.: Mapping instructions and visual observations to actions with reinforcement learning. arXiv preprint arXiv:1704.08795 (2017)
56. Moratz, R., Nebel, B., Freksa, C.: Qualitative spatial reasoning about relative position. In: International Conference on Spatial Cognition. pp. 385–400. Springer (2002)
57. Moratz, R., Tenbrink, T.: Spatial reference in linguistic human-robot interaction: Iterative, empirically supported development of a model of projective relations. Spatial cognition and computation **6**(1), 63–107 (2006)
58. Muller, P., Prévot, L.: Grounding information in route explanation dialogues. (2009)
59. Najemnik, J., Geisler, W.S.: Eye movement statistics in humans are consistent with an optimal search strategy. Journal of Vision **8**(3), 4–4 (2008)
60. Neftci, E.O., Averbeck, B.B.: Reinforcement learning in artificial and biological systems. Nature Machine Intelligence **1**(3), 133–143 (2019)
61. Nelson, J.D., McKenzie, C.R., Cottrell, G.W., Sejnowski, T.J.: Experience matters: Information acquisition optimizes probability gain. Psychological science **21**(7), 960–969 (2010)
62. Niv, Y.: Reinforcement learning in the brain. Journal of Mathematical Psychology **53**(3), 139–154 (2009)
63. Piaget, J.: The attainment of invariants and reversible operations in the development of thinking. Social research pp. 283–299 (1963)
64. Piaget, J., Inhelder, B.: The psychology of the child. Basic books (1962)
65. Pustejovsky, J.: The Generative Lexicon. MIT Press (1995)
66. Pustejovsky, J.: Dynamic event structure and habitat theory. In: Proc. of the 6th International Conference on Generative Approaches to the Lexicon (GL2013). pp. 1–10. ACL (2013)
67. Pustejovsky, J.: Affordances and the functional characterization of space. In: Cognitive Processing. vol. 16, pp. S43–S43. Springer Heidelberg (2015)
68. Pustejovsky, J.: Computational models of events. ESSLLI Summer School, 2018, Sofia, Bulgaria (August 2018)
69. Pustejovsky, J., Boguraev, B.: Lexical knowledge representation and natural language processing. Artificial Intelligence **63**(1-2), 193–223 (1993)
70. Pustejovsky, J., Krishnaswamy, N.: Voxml: A visualization modeling language. Proceedings of LREC (2016)
71. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. KI-Künstliche Intelligenz **35**(3), 307–327 (2021)
72. Pustejovsky, J., Krishnaswamy, N.: The role of embodiment and simulation in evaluating hci: Theory and framework. In: International Conference on Human-Computer Interaction. pp. 288–303. Springer (2021)

73. Pustejovsky, J., Moszkowicz, J.L.: The qualitative spatial dynamics of motion in language. Spatial Cognition & Computation **11**(1), 15–44 (2011)
74. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 401–417 (2018)
75. Randell, D., Cui, Z., Cohn, A., Nebel, B., Rich, C., Swartout, W.: A spatial logic based on regions and connection. In: KR'92. Principles of Knowledge Representation and Reasoning: Proc. of the 3rd Int. Conference. pp. 165–176. Morgan Kaufmann, San Mateo (1992)
76. Renninger, L.W., Verghese, P., Coughlan, J.: Where to look next? eye movements reduce local uncertainty. Journal of vision **7**(3) (2007). https://doi.org/https://doi.org/10.1167/7.3.6
77. Schaffer, S., Reithinger, N.: Conversation is multimodal: thus conversational user interfaces should be as well. In: Proceedings of the 1st International Conference on Conversational User Interfaces. pp. 1–3 (2019)
78. Scheutz, M., Cantrell, R., Schermerhorn, P.: Toward humanlike task-based dialogue processing for human robot interaction. Ai Magazine **32**(4), 77–84 (2011)
79. Schick, T., Schütze, H.: It's not just size that matters: Small language models are also few-shot learners. arXiv preprint arXiv:2009.07118 (2020)
80. Schlenker, P.: Gesture projection and cosuppositions. Linguistics and Philosophy **41**(3), 295–365 (2018)
81. Schulz, L.E., Bonawitz, E.B.: Serious fun: preschoolers engage in more exploratory play when evidence is confounded. Developmental psychology **43**(4), 1045 (2007)
82. Shapiro, L.: Embodied cognition. Routledge (2010)
83. Shapiro, L.A.: The routledge handbook of embodied cognition (2014)
84. Son, L.K., Sethi, R.: Metacognitive control and optimal learning. Cognitive Science **30**(4), 759–774 (2006)
85. Stalnaker, R.: Common ground. Linguistics and philosophy **25**(5-6), 701–721 (2002)
86. Stojnić, U., Stone, M., Lepore, E.: Pointing things out: in defense of attention and coherence. Linguistics and Philosophy pp. 1–10 (2019)
87. Tomasello, M., Carpenter, M.: Shared intentionality. Developmental science **10**(1), 121–125 (2007)
88. Vlach, H., Sandhofer, C.M.: Fast mapping across time: Memory processes support children‚Äôs retention of learned words. Frontiers in psychology **3**, 46 (2012)
89. Wahlster, W.: Dialogue systems go multimodal: The smartkom experience. In: SmartKom: foundations of multimodal dialogue systems, pp. 3–27. Springer (2006)
90. Wallis, G., Bülthoff, H.: Learning to recognize objects. Trends in cognitive sciences **3**(1), 22–31 (1999)
91. Wilson, A.D., Golonka, S.: Embodied cognition is not what you think it is. Frontiers in psychology **4**, 58 (2013)
92. Xu, B., Wong, Y., Li, J., Zhao, Q., Kankanhalli, M.S.: Learning to detect human-object interactions with knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
93. Yatskar, M., Zettlemoyer, L., Farhadi, A.: Situation recognition: Visual semantic role labeling for image understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5534–5542 (2016)
94. Zador, A.M.: A critique of pure learning and what artificial neural networks can learn from animal brains. Nature communications **10**(1), 1–7 (2019)