The Role of Embodiment and Simulation in Evaluating HCI: Experiments and Evaluation

Nikhil Krishnaswamy[1(✉)] and James Pustejovsky[2]

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

Abstract. In this paper series, we argue for the role embodiment plays in the evaluation of systems developed for Human Computer Interaction. We use a simulation platform, VoxWorld, for building Embodied Human Computer Interactions (EHCI). 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. In this second 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 compare and contrast evaluation schemes enabled by different levels of embodied interaction.

Keywords: Embodiment · HCI · Common ground · Multimodal dialogue · VoxML

1 Introduction

In Part 1, we described the theory of computational common ground and its underlying semantics. We focused on the role of an agent's *embodiment* in cre-

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.

V. G. Duffy (Ed.): HCII 2021, LNCS 12777, pp. 220–232, 2021.
https://doi.org/10.1007/978-3-030-77817-0_17

ating mechanisms through which to compute the parameter values that go into a common ground structure, such as the target of a pointing gesture.

This is crucial to evaluating human-computer interactions because it provides for bidirectional content: that is, each interlocutor has available all communicative modalities and can use them with reference to the current situation rather than having to communicate solely in abstractions, for lack of either a situated context or an ability to interact with it. Put simply, an agent needs to have at minimum the *notion* of a body and how it exists in an environment in order to reference said environment with any specificity. Figure 1 shows an example of this, with a human and an avatar making the smae gesture, that both of them can recognize and interpret.

Fig. 1. Bidirectional gesture recognition and generation.

Visual gesture recognition has long been a challenge [10,22]. Gesture recognition in our VoxWorld-based embodied HCI system is facilitated by Microsoft Kinect depth sensing [27] and ResNet-style deep convolutional neural networks (DCNNs) [7] implemented in TensorFlow [1]. As our goal in developing multimodal interactions is to achieve naturalistic communication, we must first examine what we mean by and desire of an interaction such as that illustrated in Sect. 2.

We take the view that a "meaningful" interaction with a computer system should model certain aspects of a similar interaction between two humans. Namely, it is one where each interlocutor has something "interesting" to say, and one that enables them to work together to achieve common goals and build off each other's contributions, thereby conveying the impression to the user that the computer system is experiencing the same events. We therefore build the evaluation scheme off of the following qualitative metrics:

1. Interaction has mechanisms to move the conversation forward [4,11]
2. System makes appropriate use of multiple modalities [2,3]
3. Each interlocutor can steer the course of the interaction [8]
4. Both parties can clearly reference items in the interaction based on their respective frames of reference [21,26,29]
5. Both parties can demonstrate knowledge of the changing situation [28]

In [18] we introduced a surface-level evaluation scheme that satisfies the above requirements. In this scheme, we took the view that a "meaningful" interaction

with a computer system should model certain aspects of similar interactions between two humans, namely that each interlocutor should have something to contribute that enables them to work together toward common goals, building off each other's contributions. This is discussed in Sect. 2.1.

This multimodal evaluation scheme was an attempt to quantify qualitative metrics based on the body of work underlying "common ground" in communication and multimodality in human-computer interaction.

These metrics, or "hallmarks" of communication, come from a rubric initially developed by the MITRE Corporation to evaluate peer-to-peer communication with computers on collaborative tasks. It was recently published as a technical report [13]. The hallmarks are intended to evaluate collaborative computer systems engaged in tasks of various complexities where there is not a single ground truth or "right answer" to compare to.

The results of the surface-level evaluation made it clear that the single dimension of response time as a proxy for communicativity of the preceding utterances or actions was not exposing the deep semantics or information content of the multimodal utterance, even when conditioned on context. Therefore a finer-grained evaluation scheme was needed—one that took into fuller consideration the parameters of common ground. This will be discussed in Sect. 2.2. We will then conduct a novel comparison of the two scenarios and evaluation methods with regard to the parameters of the common ground, in Sect. 2.3.

2 Evaluation Schemes

We have conducted a variety of studies on multimodal interactions using *Diana*, an embodied agent capable of interpreting linguistic and gestural inputs. Diana is one of many kinds of agents that can be implemented within the VoxWorld platform; she is designed to communicate with a human in the context of collaborative tasks. VoxSim (discussed in Part 1) handles Diana's language interpretation using inputs from 3rd-party or custom speech recognition, while gestures are recognized using custom 11-layer deep convolution neural nets (DCNNs) trained over 2048-dimensional feature vectors extracted from RGBD video data.

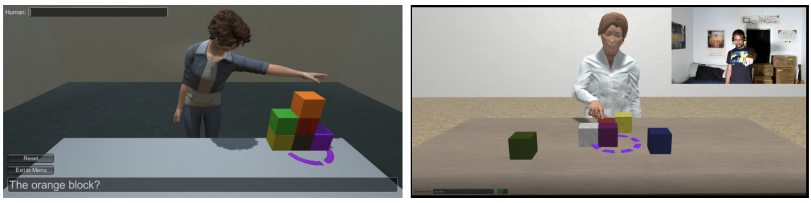


Fig. 2. L: Diana c. 2018; R: Diana c. 2020.

Diana has undergone numerous updates over time, from taking gestural inputs only [15] to word-spotting to recognizing complete utterances [23], and from a turn-taking interaction to one that is more asynchronous [14]. It is this specific embodied interactive agent that we conducted our evaluations against, as we detail subsequently.

2.1 Time-Based Evaluation

In [25], Wang et al. conducted user studies of two humans engaged in a collaborative building task wherein a "builder" with a target pattern of blocks had to instruct a "signaler" on how to build that pattern out of a physical set of blocks. Users were placed in one of three conditions:

1. *Video only*, where the signaler and builder can see but not hear each other and must rely on gesture to communicate;
2. *Audio only*, where the signaler can see the builder but the builder can only hear the signaler—the two can use only language to communicate bidirectionally;
3. *Both audio and video*, where both gestural and spoken communication are available.

These elicitation studies gave rise to the gesture set used by the Diana system, and also showed an interesting conclusion: the subjects could complete the task in all conditions, but when both linguistic and gestural modalities were available, the users could complete the task in significantly less time. Figure 3 shows these differences in trial time based on modality.

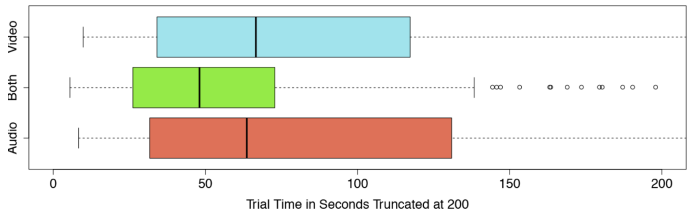


Fig. 3. In human-to-human collaborative studies, users complete tasks faster using both audio and visual channels. Figure credited to Dr. Jaime Ruiz of the University of Florida.

Therefore, in evaluating early versions of Diana, we adopted a similar assessment of communicative facility, where the time required to achieve a communicative goal was taken to be a surface-level indicator of the communicative content of the utterance, which was then assessed relative to the communicative modalities used.

This study was conducted using 20 graduate students placed in a live interaction with Diana (e.g., see Fig. 2L). They were tasked with building a 3-step

staircase out of six blocks and were told that Diana could understand gesture and language but were not given a specific vocabulary to use. We collected no identifying audio or video directly from the user but logged all instructions the computer recognized from the user, and Diana's responses.

Details are given in [18], but among other findings, we discovered discrepancies in the communicative facility of the handedness of the pointing (right-handed pointing prompted quicker responses than left-handed pointing), affirmative vs. negative acknowledgments (affirmatives prompted *slower* responses than negatives, particularly when spoken instead of gestured), and "push" gestures vs. "carry" gestures (pushing prompted quicker responses than carrying). These and other particulars can be ascribed to a number of factors, including variance in the gesture recognition, complexity of the gesture being made, and the use of positive acknowledgment as an explicit requirement for the conversation to proceed vs. negative acknowledgment as a contentful way of redirecting the discourse (cf. [12]).

These conclusions were useful in making improvements to the Diana agent, but given the coarse granularity of this high-level evaluation, it is clear that multiple dimensions are being masked; the important discriminative factor(s) in the communicativity of an utterance by an embodied interlocutor in a multimodal discourse might not be the time to receive a response, but rather how much and what information is being introduced via the multimodal utterance.

2.2 Common Ground-Based Evaluation

In [19], we presented the EMRE (Embodied Multimodal Referring Expressions) dataset. This dataset contains 1,500 individual videos of Diana generating multimodal references to 6 different objects in 50 different configurations using 5 different strategies (one gestural only, two linguistic only, and two multimodal) (Fig. 4).

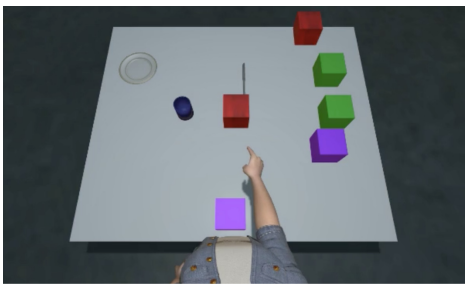


Fig. 4. Sample still from a video in the EMRE dataset. The accompanying utterance is "That red block on front of the knife". (Color figure online)

Each of these videos were then judged by 8 annotators on Amazon Mechanical Turk who indicated, on a Likert-type scale, how natural they thought the depicted referring expression was.

Initial analysis of the EMRE dataset provided similar surface-level conclusions to the time-based evaluation, and showed the evaluators preferred multimodal referring expressions and more descriptive language where language was used.

Following this, in [20], we conducted a further detailed evaluation of these object referencing strategies, using the common ground structure (CGS) as a feature generation strategy, and assessed how well individual classes of features predicted the Likert-type rating of the naturalness of a referring expression in the dataset. See Part 1 for an explanation of the CGS parameters that will be referenced below.

We extracted formal and propositional values as features from the EMRE data based on the information each feature introduces into the common ground. If the gesture ($\mathcal{G}$) or speech ($\mathcal{S}$) content in the referring expression demonstrates that either agent α (being either α_a the artificial agent or α_h the human) either *knows* or *perceives* some propositional content p that pertains to either the jointly perceived entities in **P** or the agents' beliefs **B** about what each other know or understand, this prompts an update to the common ground, and therefore new features for possible examination. This allows us to evaluate the behavior of annotators as a proxy for interlocutors, by examining what features are good predictors for naturalness judgments when interpreting referring expressions.

Details are given in [20]. We trained a multi-layer perceptron (MLP) classifier to predict the naturalness of a given referring strategy based on different combinations of input features. Possible input features included features taken directly from the EMRE dataset, sentence embedding features extracted from the linguistic portion of multimodal referring expressions, and features extracted from the CGS of each referring expression represented as individual one-hot vectors. For instance, if the communicative act $\mathcal{C}_a$ contains a speech component $\mathcal{S}$ that in turn contains the word "*other*" in conjunction with some attributive *Att* and an object type t, then this indicates that a knows 3 things: that there is more than one object of type t in the discourse, that they are distinct, and that *Att* predicates over both (or all) of them. Each of these knowledge elements $\mathcal{K}_a$ of the common ground is represented as a distinct one-hot vector.

This MLP classifier was then cross-validated on the EMRE data using 7 folds. We found that features that correlate formally with elements of the CGS improved the ability of the classifier to predict the annotator judgment on a referring expression by an average of 7–11%, when compared to the EMRE dataset features, with or without augmentation with sentence embeddings. When examining language-only referring expressions, addition of CGS-derived features improved classification accuracy by 5–16%, with using CGS-derived features *alone* providing the highest boost, bringing cross-validated classification accuracy up to ∼80% from a baseline of ∼64%.

CGS-derived features provided only a 1–5% boost in classifying ensemble (multimodal) referring expressions. However, we interpret this result as representing a redundancy in information communicated through the existence of a gesture as part of the referring expression, and the same content occurring

in the common ground structure: $Point_g \rightarrow Obj$ occurs in the CGS and its use indicates that the agent α_a knows what a pointing gesture is intended to communicate, but that information is already contained in the EMRE dataset features, which tracks the modality used by each RE.

2.3 Comparing Time-Based and Common Ground-Based Evaluation

Let us now examine a specific sample from each of the two evaluation schemes where an agent makes a reference to an object in the scene, and look at how much and what kind of information the time-based and CGS-based metrics provide, respectively.

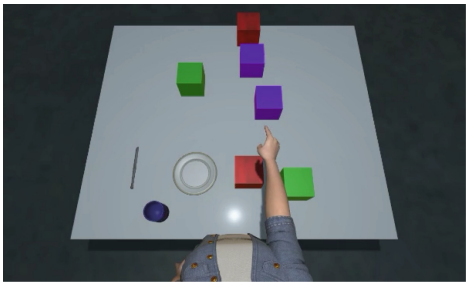
Figure 5 shows the log from the time-based evaluation. The first column is the index of the "move" in the interaction. The second column is a two-letter code representing the *A*vatar or the *H*uman plus the modality (*G*esture or *S*peech). The third column is the content of the move, which may be a utterance or gesture generated by the avatar, or a word recognized as being spoken by the human or a gesture recognized by the DCNN recognizer. For example, `right point low, X, Y` denotes a low-probability pointing gesture with the right hand at the 2D coordinates $<X, Y>$, which is then transformed in to 3D space. The fourth column is the timestamp (multiple moves can have the same timestamp).

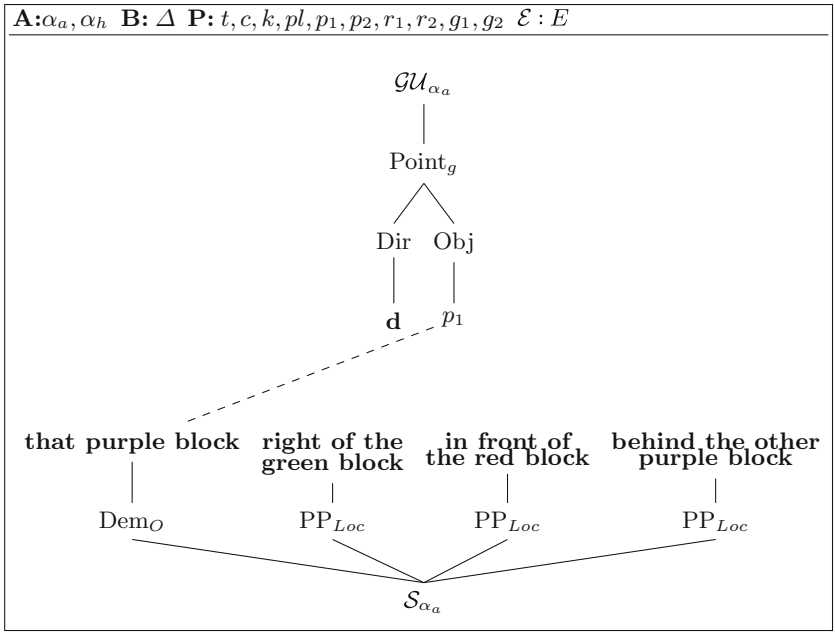
```
81 AS "Which object do you want?" 34.3762
82 HG right point low,0.24,1.59   37.148
83 HG right point stop            38.7285
84 HG right point low,0.24,1.60   38.99577
85 HS PURPLE                      39.1167
86 AS "OK, go on."                39.1167
87 AG reach(block7)               39.1167
```

Fig. 5. Sample of object referencing using time-based evaluation.

In this snippet, the avatar asked the human "Which object do you want?" (move 81) to which the human responded by starting to point (move 82). The tag `low` indicates that this gesture was not defind enough for the avatar to interpret. Eventually the human says "purple," (move 85) which the avatar is able to understand and respond to (move 86). From the point that the avatar requested input from the human to the point that the human supplied understandable input to the avatar, 4.7405 s elapsed. This number can thereafter be compared to similar blocks of moves where different gestures or different utterances are used to see how these sequence advances or slows the interaction. Examining this in isolation, we can also see that the most delay results from the difficulty the human has in pointing to a distinct, interpretable location (moves 82–84). The use of the linguistic modality is what allows the interaction to proceed here.

Figure 6 shows a still from a video in the EMRE dataset, with the corresponding common ground structure. The items in **P** are the non-agent items in





$\lambda k_s \otimes k_g(\textbf{that}(x)[\text{block}(x) \wedge \text{purple}(x) \wedge \text{right}(x, g_1, v) \wedge \text{in\_front}(x, r_1, v) \wedge \text{behind}(x, p_2, v)] \wedge k_s \otimes k_g(x)]$, where $v = \alpha_a$

Fig. 6. Sample from the EMRE dataset, with accompanying utterance "that purple block right of the green block, in front of the red block, and behind the other purple block," and corresponding common ground structure. The semantics of the RE includes a *continuation* (in the abstract representation sense in computer science, cf. Van Eijck and Unger [24]) for each modality, k_s and k_g, which will apply over the object in subsequent moves in the dialogue. v denotes the viewer, i.e., frame of reference.

the scene, where the non-uniquely colored blocks are denoted by subscripts $_1$ and $_2$. **B** is the belief space Δ which is populated by elements of the common ground. Items in this belief space are extracted as one-hot vector features in the evaluation described in Sect. 2.2.

This referring expression was presented to the 8 annotators mentioned in the EMRE study, alongside 3 other choices to refer to the same object:

1. Pointing only;
2. "The purple block in front of the red block" (language only);
3. "That purple block" (with pointing).

Of the 8 annotators, 6 judged this referring expression to be most natural (5 on the 1–5 Likert-type scale), while the remaining 2 judged it to be a 4. The MLP classifier also predicted that this RE would receive a score of 5. Examining the individual features introduced into the common ground by each candidate RE illuminates why:

- Pointing alone (1) is ambiguous. Performing the pointing gesture introduces into the common ground the fact that α_a knows how to point and what it means: $C_a = (\mathcal{G} \mid \mathcal{G} = Point_g \rightarrow Obj \rightarrow \_)$, but what fills that slot is unclear. From the camera angle shown the deixis could be indicating either of the purple blocks or even the red block at the back of the table.
- Adding the utterance "that purple block" (3) introduces the demonstrative "that" (in contrast to "this") and suggests that α_a has some knowledge of the near/far distance distinction on which the demonstrative distinction is based: $C_a = (\mathcal{S}, \mathcal{G} \mid \mathcal{G} = Point_g \wedge'' \, that'' \in \mathcal{S}) \rightarrow \mathcal{K}_a[[near(sfc)]] \neq [[far(sfc)]]_{\mathcal{M}}$. This is still ambiguous; the demonstrative is coupled with deixis, but both purple blocks are in the direction of the gesture.
- The language-only RE "the purple block in front of the red block" (2) introduces the color attribute as a distinction α_a uses, as well as the spatial term "in front of": $C_a = (\mathcal{S} \mid [''purple'','' red'', b_{1_s}, b_{2_s}] \in \mathcal{S}) \rightarrow \mathcal{K}_a[\![''purple''(b_1)]\!] \wedge [\![''red''(b_2)]\!] \wedge [\![''purple'']\!] \neq [\![''red'']\!]$ However, it is still ambiguous; there are red blocks on either side of the two purple blocks. Therefore, "in front of" could be interpreted as either (relative to the camera) "closer to me" or "away from me."
- Integrating deixis and the descriptive language of the RE given in Fig. 6 singles out three relations relative to the target object as well as an interpretation of "other" relative to the attribute it scopes over (in this case, "purple"): $C_a = (\mathcal{S} \mid [''other'', b_{1_s}, b_{2_s}] \in \mathcal{S} \wedge b_{1_s} = b_{2_s}) \rightarrow \mathcal{K}_a[\![Att(b_1 \wedge b_2)]\!]_{\mathcal{M}} \wedge \mathcal{K}_a b_1 \neq b_2$, such that if $Att =$ "purple," α_a knows what that means, knows that it applies to both blocks b_1 and b_2, and knows that the two objects are distinct. Only one of these blocks is right of a green block, in front of a red block, in the direction of deixis, and far enough from the agent to use "that" as a demonstrative, and so resolve to the correct target object (see Fig. 7).

Compared to the time-based evaluation, using common ground structures as a data structure from which to extract evaluation-relevant features allows the examination of specific features relative to the information they introduce into the interaction. Many of the most informative features, such as what an agent α knows about distance and spatial relations between objects, are dependent upon how the agent is situated or embodied in the world.

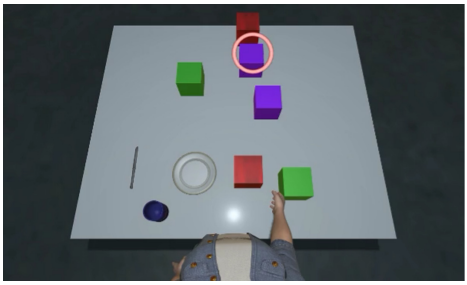


Fig. 7. The target object of the RE depicted in Fig. 6, shown highlighted with a circle.

2.4 How Embodiment Enables Evaluation

By embedding the communicative interaction within an embodied simulation environment, we are able to vary the parameters involved in the interaction the human has with a computational agent, and thereby measure the consequences these changes have on the effectiveness of the specific components of the interaction. VoxML provides a dynamic, interpretable model of objects, events, and their properties. This allows us to create visualized simulations of events and scenarios that are rendered analogues to the "mental simulations" discussed in Part 1. VoxSim [16,17] serves as the event simulator within which these simulations are created and rendered in real time, serving as the computer's method of visually presenting its interpretation of a situation or event. Because modalities are modes of presentation, a multimodal simulation entails as many presentational modes as there are modalities being modeled. The visual modality of presentation (as in embodied gaming) necessitates "situatedness" of the agent, as do the other perceptual modalities. Therefore, when we speak of *multimodal simulations*, they are inherently situated. In a human-computer interaction using such a simulation, the simulation is a demonstration of the computational agent's "mind-reading" capabilities (an *agent simulation*). If the two are the same (where the agent is a proxy for the player or user, then the "mind-reading" is just a demonstration of the scenario) If, on the other hand, the two are separate (agent is *not* proxy for the user), then the simulation/demonstration communicates the agent's understanding of the user and the interaction. In this case, this demonstration entails the illustration of both epistemic and perceptual content of the agent.

We believe that simulation can play a crucial role in human-computer communication; it creates a shared epistemic model of the environment inhabited by a human and an artificial agent, and demonstrates the knowledge held by the agent publicly. Demonstrating knowledge is needed to ensure a shared understanding with its human interlocutor. If an agent is able to receive information from a human and interpret that relative to its current physical circumstances, it can create an epistemic representation of that same information. However, without a modality to express that representation independently, the human is

unable to verify or query what the agent is perceiving or how that perception is being interpreted. In a simulation environment the human and computer share an epistemic space, and any modality of communication that can be expressed within that space (e.g., linguistic, visual, gestural) enriches the number of ways that a human and a computer can communicate within object and situation-based tasks, such as those investigated by Hsiao et al. [9], Dzifcak et al. [6], and Cangelosi [5], among others.

VoxWorld, and the accompanying simulation environment provided by VoxSim, includes the perceptual domain of objects, properties, and events. In addition, propositional content in the model is accessible to the simulation. Placing even a simple scenario, such as a blocks world setup, in a rendered 3D environment opens the search space to the all the variation allowed by an open world, as objects will almost never be perfectly aligned to each other or to a grid, with slight offsets in rotation caused by variations in interpolation, the frame rate, or effects of the platform's physics. Nevertheless, when the rendering is presented to a user, the user can use their native visual faculty to quickly arrive at an interpretation of what is being depicted.

3 Conclusion

In this paper series, we have brought together a number of definitions of "simulation" from the AI, cognitive science, and game development literature, into a single platform that creates both a formal and operational definition of "embodiment" in the content of Human-Computer Interaction. This framework provides both quantitative and qualitative outputs that can be used to produce, evaluate, and learn 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.

In Part 1, we introduced the underlying theory of computational common ground and its relation to the associated semantic literature. Included in this was the introduction of embodiment to our formulation and platform, as well as the formal notion of the common-ground structure, and how embodiment facilitates the populating thereof.

In Part 2, we presented and compared particular experiments done under versions of this framework using the embodied agent Diana. We hope to have demonstrated how the combination of formal semantics with the technologies provided by modern gaming engines lead to systems that afford gathering both traditional data for deep learning and representations of common sense, situated, or embodied understanding, thereby opening new doors for researchers to

deploy and examine the role of embodiment in human-computer interaction both quantitatively and qualitatively.

References

1. Abadi, M., et al.: TensorFlow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), Savannah, Georgia, USA (2016)
2. Arbib, M., Rizzolatti, G.: Neural expectations: a possible evolutionary path from manual skills to language. Commun. Cogn. **29**, 393–424 (1996)
3. Arbib, M.A.: From grasp to language: embodied concepts and the challenge of abstraction. J. Physiol. Paris **102**(1), 4–20 (2008)
4. Asher, N., Gillies, A.: Common ground, corrections, and coordination. Argumentation **17**(4), 481–512 (2003)
5. Cangelosi, A.: Grounding language in action and perception: from cognitive agents to humanoid robots. Phys. Life Rev. **7**(2), 139–151 (2010)
6. Dzifcak, J., Scheutz, M., Baral, C., Schermerhorn, P.: What to do and how to do it: translating natural language directives into temporal and dynamic logic representation for goal management and action execution. In: IEEE International Conference on Robotics and Automation, ICRA 2009, pp. 4163–4168. IEEE (2009)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Hobbs, J.R., Evans, D.A.: Conversation as planned behavior. Cognit. Sci. **4**(4), 349–377 (1980)
9. Hsiao, K.Y., Tellex, S., Vosoughi, S., Kubat, R., Roy, D.: Object schemas for grounding language in a responsive robot. Connect. Sci. **20**(4), 253–276 (2008)
10. Jaimes, A., Sebe, N.: Multimodal human–computer interaction: a survey. Comput. Vis. Image Underst. **108**(1), 116–134 (2007)
11. Johnston, M.: Building multimodal applications with EMMA. In: Proceedings of the 2009 International Conference on Multimodal Interfaces, pp. 47–54. ACM (2009)
12. Koole, T.: Conversation analysis and education. In: The Encyclopedia of Applied Linguistics, pp. 977–982 (2013)
13. Kozierok, R., et al.: Hallmarks of human-machine collaboration: a framework for assessment in the darpa communicating with computers program. arXiv preprint arXiv:2102.04958 (2021)
14. Krishnaswamy, N., et al.: Diana's world: a situated multimodal interactive agent. In: AAAI Conference on Artificial Intelligence (AAAI): Demos Program. AAAI (2020)
15. Krishnaswamy, N., et al.: Communicating and acting: understanding gesture in simulation semantics. In: 12th International Workshop on Computational Semantics (2017)
16. Krishnaswamy, N., Pustejovsky, J.: Multimodal semantic simulations of linguistically underspecified motion events. In: Barkowsky, T., Burte, H., Hölscher, C., Schultheis, H. (eds.) Spatial Cognition/KogWis -2016. LNCS (LNAI), vol. 10523, pp. 177–197. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68189-4_11

17. Krishnaswamy, N., Pustejovsky, J.: VoxSim: a visual platform for modeling motion language. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. ACL (2016)
18. Krishnaswamy, N., Pustejovsky, J.: An evaluation framework for multimodal interaction. In: Proceedings of LREC (2018, forthcoming)
19. Krishnaswamy, N., Pustejovsky, J.: Generating a novel dataset of multimodal referring expressions. In: Proceedings of the 13th International Conference on Computational Semantics-Short Papers, pp. 44–51 (2019)
20. Krishnaswamy, N., Pustejovsky, J.: A formal analysis of multimodal referring strategies under common ground. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 5919–5927 (2020)
21. Ligozat, G.F.: Qualitative triangulation for spatial reasoning. In: Frank, A.U., Campari, I. (eds.) COSIT 1993. LNCS, vol. 716, pp. 54–68. Springer, Heidelberg (1993). https://doi.org/10.1007/3-540-57207-4_5
22. Madeo, R.C.B., Peres, S.M., de Moraes Lima, C.A.: Gesture phase segmentation using support vector machines. Expert Syst. Appl. **56**, 100–115 (2016)
23. Narayana, P., et al.: Cooperating with avatars through gesture, language and action. In: Intelligent Systems Conference (IntelliSys) (2018, forthcoming)
24. Van Eijck, J., Unger, C.: Computational Semantics with Functional Programming. Cambridge University, Cambridge (2010)
25. Wang, I., et al.: EGGNOG: a continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In: To appear in the Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (2017)
26. Wooldridge, M., Lomuscio, A.: Reasoning about visibility, perception, and knowledge. In: Jennings, N.R., Lespérance, Y. (eds.) ATAL 1999. LNCS (LNAI), vol. 1757, pp. 1–12. Springer, Heidelberg (2000). https://doi.org/10.1007/10719619_1
27. Zhang, Z.: Microsoft kinect sensor and its effect. IEEE MulitMedia **19**, 4–10 (2012)
28. Ziemke, T., Sharkey, N.E.: A stroll through the worlds of robots and animals: applying Jakob von Uexkull's theory of meaning to adaptive robots and artificial life. Semiotica-La Haye Then Berlin **134**(1/4), 701–746 (2001)
29. Zimmermann, K., Freksa, C.: Qualitative spatial reasoning using orientation, distance, and path knowledge. Appl. Intell. **6**(1), 49–58 (1996)