Evaluating Multimodal Behavior Schemas With VoxWorld*

Christopher Tam[0000−0002−9126−4276], Richard Brutti[0000−0003−0449−4418], Kenneth Lai[0000−0003−2870−7019], and James Pustejovsky[0000−0003−2233−9761]

Brandeis University, Waltham, MA, USA
{christophertam,brutti,klai12,jamesp}@brandeis.edu

Abstract. The ability to understand and model human-object interactions is becoming increasingly important in advancing the field of human-computer interaction (HCI). To maintain more effective dialogue, embodied agents must utilize situated reasoning - the ability to ground objects in a shared context and understand their roles in the conversation [35]. In this paper, we argue that developing a unified multimodal annotation schema for human actions, in addition to gesture and speech, is a crucial next step towards this goal. We develop a new approach for visualizing such schemas, such as Gesture AMR [5] and VoxML [33], by simulating their output with VoxWorld [21] in the context of a collaborative problem-solving task. We discuss the implications of this method, including proposing a novel testing paradigm using the generated simulation to validate these annotations for their accuracy and completeness.

Keywords: Multimodal dialogue · AMR · Gesture annotation · Nonverbal behavior · Meaning representation.

1 Introduction

Collaborative problem-solving tasks, used in educational contexts to facilitate the development of critical thinking and teamwork skills, are a rich source of multimodal behavior. We believe that efforts toward developing a unified multimodal annotation schema for these and similar tasks will yield practical insights for modeling multi-party dialogue, and inform a greater range of expression for multimodal interactive virtual agents (IVAs).

In this paper, we first address the challenge of annotating multiple modalities in videos of two subjects engaged in task-based interactions. Then, to test and validate both the completeness and the expressiveness of the ensemble annotations, we use them to generate animated simulations of the interactions in the Unity-based VoxWorld platform [21].

*This work was supported in part by NSF grant DRL 2019805, to Dr. Pustejovsky at Brandeis University. We would like to express our thanks to Nikhil Krishnaswamy for his comments on the multimodal framework motivating the development of the simulation platform, FibWorld. The views expressed herein are ours alone.

The VoxWorld platform enables the deployment of embodied agents with contextual awareness, allowing them to interact and communicate from their virtual environments through speech and gesture [35]. Our research proposes a new approach for validating multimodal annotation schemas, such as VoxML [34] and Gesture AMR [5], by simulating their output with VoxWorld.

Our approach focuses on the *Shared Weights Task* [3], a collaborative task that involves deducing the weights of physical blocks using a scale. We capture several videos of participants performing this task, and develop a simple multimodal annotation schema to encode the speech, gestures, and physical actions performed within them. We model the environment and equipment for the Shared Weights Task digitally in Unity3D using the VoxWorld platform, with IVAs (*Diana* and *Apollo*) standing in as proxies for the human participants. The video annotations are then imported into Unity as a series of timestamped events, which can be executed sequentially to generate a real-time simulation of the annotated video, with the IVAs generating the corresponding behaviors to complete the task. Because this method succeeds at replicating the speech, select non-verbal behaviors (gestures), and actions of the task videos, it provides an informative visualization of the range of behaviors captured by the annotation schema. We plan to apply this method to a variety of candidate annotation schemas involving additional nonverbal modalities (such as gaze, pose, and sentiment), and then collect crowdsourced judgments of the simulation results to achieve the most accurate representation of the ground truth. The resulting framework, we believe, will help evaluate annotation schemas for multimodal interactions, both for their accuracy and their completeness.

2 Related Work

As multimodal interactive systems continue to become more commonplace and more sophisticated, users expect that their interactions will resemble interactions with other humans. A major challenge of human-robot interaction (HRI) and HCI involves communicating intentions, goals, and attitudes through simultaneous modalities in addition to language, including gesture, gaze, facial expression, and situational awareness [6, 9, 20, 26, 39, 42]. This in turn brings a need for capturing, representing, and annotating the data that encodes these various non-verbal modalities.

There are very few meaning representations that have been designed for or deployed in the context of true situated multi-party interaction that are both adequately expressive of the multimodal content and compact enough for corpus development. This is partially due to the fact that empirical evidence is often lacking for arriving at a data-based understanding of the nature of multimodal constructions in conversation [46]. Many existing approaches to annotating meaning representation treat verbal and nonverbal components as distinct and autonomous, while those that do address the interaction between them generally focus on form rather than meaning [17, 20]. For example, the Behavior Markup Language (BML) [19] is an exchange language originally intended to

describe an agent's communicative actions along a range of temporally marked behaviors, including gesture, pose, facial expressions, head movements, and gaze, among others. However, BML is not designed to represent the intents that give rise to such actions.

Embodied HCI systems must necessarily have an understanding of the objects in their environment and their associated affordances to achieve physical goals in their interactions. This requires the annotation of *actions* that result in the manipulation or transformation of objects in the environment. There has been significant interest in how encoding affordances might be used to improve the accuracy of human-object interaction (HOI) recognition and scene understanding models [12]. Modern annotation schemes for such affordances and actions range from general classification of human-object interactions in movie clips [10] to affordance annotation with the Gibsonian/telic (conventional) distinction [14]. An additional, critical layer to this task is visual semantic role labeling (VSRL) [11, 45] - the identification of how entities in an image fulfill the thematic roles of the events they are involved in; work has been done to apply this to cooking clips [44] and movie scenes [38].

After developing a satisfactory annotation scheme, the process of annotating and reviewing the various tracks of the multimodal data can be challenging and time-consuming. It is worthwhile to investigate methods to efficiently inspect and validate them beyond the software initially used to create them. Visualizing symbolic multimodal data has different requirements from simulating numerical data, such as mapping points from motion capture or pose tracking. Though previous studies have attempted to translate these annotations into visual images, such as finite state machines [30], another interesting approach is to instead directly use them to drive the behavior of embodied conversational agents (ECAs) [17]. In addition to being used to annotate the general form of gestures, this *copy-synthesis* approach [27] was also used to interpret emotional annotations to inform facial expression and gesture activity. To our knowledge, however, there has been little research on using ECAs to model changes of environment state as a result of actions, as situated reasoning has only been a recent area of focus in the field.

We use the VoxWorld platform to experiment with this visualization problem: it is the joint product of the VoxML modeling language [34] and its real-time Unity interpreter VoxSim [24], resulting in an environment based on rigorously defined interaction semantics. This architecture easily lends itself to action annotation interpretation, as annotation descriptions converted to linguistic entities defined in VoxML have a explicit correspondence in the simulation. Within VoxWorld, the Diana agent has been developed as an interface to recognize user speech and gesture [22], distinguishing itself from other IVAs with its ability to reason and act on a variety of objects with a well-defined affordance structure. It is equipped with a rig that can perform basic operations on simulated objects, like pointing, grabbing, lifting, and moving, and has achieved success at collaborating with users on a paired block structure-building task [23]. We use this agent as a starting point to directly model human behavior.

3 Method

3.1 Source data: the Shared Weights Task

Collaborative problem-solving tasks are specifically designed to promote student intercommunication and mutual learning, and their participants are often in constant interaction with the materials at hand. These tasks can thus be a prime candidate for dense multimodal annotation. However, we note that the collaborative aspect of these tasks is greatly diminished in the presence of what is known as *social loafing* [16], where a participant is inclined to make fewer contributions if they perceive that the other participants can complete the task on their own.

To ensure that the human interactions in our data are productive and meaningful, we require that the observed task promote equal interaction and communication. To meet these requirements, we introduce the *Shared Weights Task*, based off the classroom task described in [3]. This is a two-person problem, in which the objective is to determine the relative weights of six colored blocks using a provided scale. The block weights follow the Fibonacci sequence pattern, which the participants must deduce during the task. The participants are seated opposite each other at a table, with three blocks placed immediately in front of each of them, and the scale placed between the two sets of blocks. The participants are told the weight of one of the unit blocks (e.g., "The red block weighs one unit."), and are instructed to determine the relative weights of the rest of the blocks. The task is considered completed when each participant can relay the relative weights of their own blocks.

The Shared Weights Task imposes the additional constraint that each participant may only interact with the subset of blocks on their respective side of the table. This restriction is useful for several reasons. First, both participants must necessarily take a roughly equal number of actions, as it is impossible for either participant to complete the task on their own. Second, it introduces imperfect information. In our formulation, participants cannot directly compare blocks of different sets against each other by weighing them with both hands; they must communicate using language to coordinate use of the provided scale.

3.2 Data collection

An initial set of data was collected at the Brandeis University Lab for Linguistics and Computation, with various lab members as participants. Video was captured with a Microsoft Azure Kinect mounted above and behind the left shoulder of one of the seated participants. Audio was captured with a Realtek conferencing microphone placed on the edge of the table. The video and audio were synchronized using the OBS Studio software.

The combined audio and video were imported into ELAN [4], and annotated by the authors. Each modality was annotated on a separate track for each participant. We refer to the collected array of tracks as an annotation *score*. An example of the ELAN annotation interface is shown in Fig. 1.

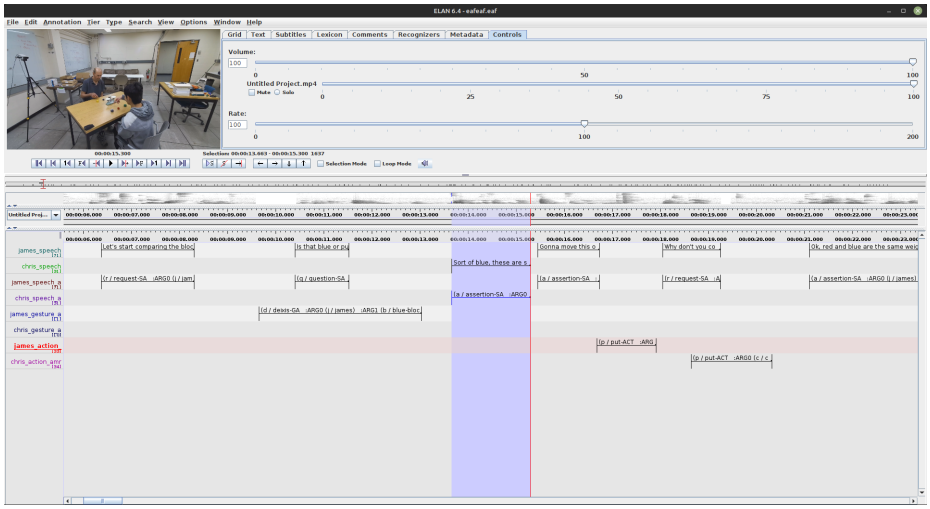

Fig. 1. ELAN annotation environment.

3.3 Annotation schema

To capture both the physical and communicative acts in these videos, we develop the following annotation schema:

Speech. First, the speech is transcribed, speaker-diarized, and segmented into utterances; this was done manually for our initial data set. We add the utterance strings to the appropriate tracks in ELAN, one track for each speaker. In addition to recording the form of each utterance, we also annotate its meaning using Abstract Meaning Representation (AMR) [1, 18]. AMR represents the predicate-argument structures of sentences in the form of graphs. Compared to other meaning representation systems such as Discourse Representation Theory [15] or Minimal Recursion Semantics [7], AMRs are designed for ease of annotation and parsing.

Specifically, we use Dialogue-AMR, an extension of AMR developed to represent meanings in task-based dialogues [2]. For example, an utterance "Is that blue or purple?", could be annotated as follows:

```
(q / question-SA
  :ARG0 (p1 / participant-1)
  :ARG1 (t / that
          :ARG1-of (a / amr-choice
                      :op1 (b / blue-01)
                      :op2 (p / purple-02)))
  :ARG2 (p2 / participant-2))
```

The above Dialogue-AMR is rooted by a speech act, here a question, and the `ARG0` and `ARG2` represent the speaker and addressee, respectively. The `ARG1` is

Fig. 2. Deictic gesture towards a blue block.

Fig. 3. Action of putting the blue block on the left side of the scale.

the semantic content of the utterance, similar to how it would be represented in standard AMR.

Gesture. We also use AMR to annotate gestures in our data; in this case, we use Gesture AMR [5]. Similarly to Dialogue-AMR, Gesture AMR classifies content-bearing gestures into four kinds of "gesture acts": deictic, iconic, emblematic, and metaphoric; and marks the gesturer, addressee, and semantic content of the gesture. For example, the situation in Fig. 2, involving a deictic, or pointing, gesture towards a blue block, can be represented like so:

```
(d / deixis-GA
  :ARG0 (p1 / participant-1)
  :ARG1 (b / blue-block)
  :ARG2 (p2 / participant-2))
```

Action. In addition to speech and gesture, we also annotate the actions performed by the participants, again using an AMR-style template. We create a taxonomy of actions by adapting relevant predicates from PropBank [29], that can be interpreted in VoxML. For instance, an action of putting a blue block on the left side of a scale, as shown in Fig. 3, is assigned the following annotation:

```
(p / put-ACT
  :ARG0 (p1 / participant-1)
  :ARG1 (b / blue-block)
  :ARG2 (o / on
          :op1 (l / left-scale)))
```

The argument structure of the action follows that of the corresponding PropBank predicate, in this case, `put-01`.[1]

[1] We are currently developing a much richer specification for action annotation (Action AMR), for both collaborative tasks as well as procedural texts and narratives.

Alignment Because agents can potentially use multiple modalities to communicate some piece of information, it is important to align the various modalities both temporally and semantically, so that we know (and can reconstruct) not only what is being communicated, but how it is done [25]. Temporal alignment is done within ELAN, as each annotation is stamped with its begin and end times. Semantic alignment, i.e., coreference relations between entities across tracks, is done using Uniform Meaning Representation (UMR) [41]. UMR is a meaning representation that adds document-level coreference annotation to AMR, along with aspect and scope relations, temporal and modal dependencies, and support for morphologically complex languages. This annotation is currently done outside of ELAN. As an example, the following annotation combines the previous three examples, and indicates that the object referred to as "that" in the speech, pointed to in the gesture, and moved in the action, are identical, as are the mentions of "participant-1" and "participant-2".

```
(s2q / question-SA
  :ARG0 (s2p1 / participant-1)
  :ARG1 (s2t / that
    :ARG1-of (s2a / amr-choice
      :op1 (s2b / blue-01)
      :op2 (s2p / purple-02)))
  :ARG2 (s2p2 / participant-2))

(g1d / deixis-GA
  :ARG0 (g1p1 / participant-1)
  :ARG1 (g1b / blue-block)
  :ARG2 (g1p2 / participant-2))

(a1p / put-ACT
  :ARG0 (a1p1 / participant-1)
  :ARG1 (a1b / blue-block)
  :ARG2 (a1o / on
    :op1 (a1l / left-scale)))
```

```
(g1 / gesture
  :coref ((g1p1 :same-entity s2p1)
          (g1b :same-entity s2t)
          (g1p2 :same-entity s2p2)))

(a1 / action
  :coref ((a1p1 :same-entity g1p1)
          (a1b :same-entity g1b)))
```

3.4 Environment modeling

To model the environment and participant actions in the Shared Weights Task, we use the VoxWorld platform to develop FibWorld, a virtual translation of the real-world task that embodied agents can fully interact with. The scene includes the original set of six colored blocks, arranged and assigned weights in the same way as in the physical setup. The scale is converted to a digital representation, with physics implemented so that when the combined weight of blocks on one side of the scale is heavier, a yellow indicator lights up on that side. This modified representation avoids the time-consuming "bouncing effect" of the physical scale when a weight is placed on either end.

To represent the participants, we modify the Diana paradigm from its user-centric model, introducing multiple IVAs into the simulation to function as proxies interacting with each other. To achieve this, we duplicate the original Diana

avatar, as well as its underlying cognitive architecture and event management systems, to serve as the second proxy. To differentiate the two avatars, we additionally replace the model of the second IVA with a male avatar designed in Reallusion Character Creator 4 [37], which we name "Apollo". The underlying keys in VoxWorld's blackboard architecture are adjusted to allow for separate streams of event messages to be assigned to each individual agent.

3.5 Annotation import and execution

This annotation score is exported from ELAN in tab-separated value (TSV) format, then converted to JSON as a list of annotation objects. The participant subject of each act is either marked as such in the AMR annotation, or implicitly suggested from the annotation track it is associated with. We assign one of the participants to the Diana avatar, and one to the Apollo avatar. The JSON objects are sent to the corresponding IVA in the simulation, sorted by starting timestamp, and read into annotation event queues, with each queue corresponding to a specific modality (speech, gesture, action). As the simulation runs, the annotation event at the head of each queue is automatically interpreted and executed when the current elapsed time exceeds its starting timestamp.

Speech utterance strings are fed directly into Diana's (or Apollo's) text-to-speech module, requiring no further interpretation. Gestures do not result in manipulation of the environment, and thus are handled separately from actions. For deixis, we simply trigger a pointing interaction for the duration of the timestamp, and utilize Diana's interruption-handling ability to undo previous actions to stop the interaction at the end of the timestamp. Other gesture types (icon, emblem, metaphor), did not appear in the data and were therefore not annotated, though future work will investigate how these gestures will be represented in VoxWorld.

For events, we rely on the internal mechanisms of VoxWorld to handle the recognition of object affordances, the situational habitats they require, as well as event satisfaction conditions. The composition and execution of events in VoxWorld are determined by their underlying specification in VoxML. Before interpretation, action AMR annotations must first be converted to VoxWorld event strings, which are predicate-argument structures of the following form:

```
PREDICATE(ARG1,ARG2,...)
```

For example, the action of putting a red block on the right side of the scale is converted to:

```
put(RedBlock,on(RightScale))
```

The main predicate corresponds to a VoxML program key, and the arguments to that predicate are listed out in sequence. As the event string is interpreted, it is broken down into compositions of primitive events, defined by a subevent structure in the main predicate's VoxML entry. For instance, the VoxML entry for the program *put* is:

Fig. 4. Simulation snapshot of Apollo putting a red block on the right side of the scale.

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

Here, the agent must first *grasp* the block, *move* it to the given location until it is at the given location, and finally *ungrasp* the block. Thus when we execute a given action annotation, it is interpreted in VoxML as a series of subevents that drive behavior in the Unity simulation. Fig. 4 shows the result of Apollo putting the red block on the right side of the scale.

4 Discussion and Future Work

The resulting simulation created upon execution of an annotation score is a faithful replay of the multimodal behavior in the source video, as defined by the combination of our chosen annotation schema and the behavioral generation capabilities of the VoxWorld platform. As a consequence of precisely modeling the locational transformations of objects with action annotations, the simulation

provides useful state tracking for the task itself - the positions of the blocks and the current value of the scale reading are accurate for any given point in time. Future modifications of FibWorld could conceivably keep track of problem-solving status, and provide predictions or suggestions for the next step towards the goal state. We anticipate this method being potentially useful in knowledge tracing and counterfactual (alternate outcome) simulations to better understand learning outcomes for the Shared Weights Task and other similar collaborative problem-solving tasks.

A visual inspection of the simulation provides valuable insight into each of its individual components: the annotation score; the annotation schema; and the IVA interpretation layer, as described below.

4.1 Annotation score validation

For researchers developing novel natural language annotation schemes, it is common practice to review exemplar documents in order to pilot early versions of the scheme. The typical early phase of annotation development takes place during the MAMA (Model-Annotate-Model-Annotate) sub-cycle [36], part of the larger MATTER annotation development methodology (Model, Annotate, Train, Test, Evaluate, Revise) [32].

Open source text annotation tools, such as BRAT [40], Doccano [28], and TreeAnnotator [13] all have clear visual interfaces for verifying and reviewing annotations. However, performing an "eyeball test" on a multimodal annotation scheme is much more difficult with video data, where the modalities are commonly separated by distinct tracks in the annotation environment and simple errors such as mislabeling a participant or an object under manipulation become difficult to detect.

Our simulation makes this analysis considerably more straightforward, especially for the assignment of semantic roles. Issues with under-generation can be identified by an evaluation of the alignment between the simulated speech, gesture and action. For instance, a missing action or deictic gesture often entails the associated object being underspecified in speech (e.g., saying "This block is heavier" without pointing to anything). On the other hand, missing speech utterances will be associated with objects being interacted with for no reason. Finally, any mislabeling of semantic roles in the action annotation will surface as clear state mismatches between the simulation and the video (e.g., a block being placed in an invalid location).

This validation process can easily be crowd-sourced given a user interface that displays both the simulation and the accompanying utterance content side-by-side. An annotation validator can highlight problems such as missing antecedents for demonstratives, excessive gestures and actions, and incorrect semantic role labeling. All feedback can then be used to adjust the annotation, and thus the simulation, resulting in a self-improving feedback loop.

4.2 Annotation schema

Natural language annotation schemes are commonly developed using an iterative approach [36], and are heavily dependent on machine learning goals or corpus objectives. Annotation schemes and annotator guidelines tend to evolve as corpora are annotated and as exceptions and edge cases are discovered in the data. Annotation schemes based on well-studied linguistic phenomena, such as named entities or semantic roles, will likely encounter fewer edge cases and therefore less modification to the original annotation specifications.

The primary goal of our pilot annotation schema, composed of speech utterances alongside verbal and nonverbal AMR, was to adequately capture the semantics of the Shared Weights Task to accurately model the problem state. It mostly succeeded in this aspect, allowing for a straightforward assignment of participant actions to agents and semantic role labels to VoxML arguments.

Beyond semantics, we note that AMR annotation does not capture the specific form of more abstract gestures, like icon, symbol, and beat gestures, as they do not encode quantitative pose data. As a result, the best pose estimations we can currently generate for icon gestures, for instance, would be predefined gestures based off the VoxML-defined geometry of the referenced object in question. Future work will investigate improvements and alternatives to our current annotation schema to handle these gesture types accurately: whether the AMR representations require additional expressive or timing markers, or if pose can be encoded and integrated alongside these annotations using markup languages focusing on form.

4.3 IVA expression

Even provided with an adequately expressive annotation schema, our method equally relies upon the subsequent interpretation and execution of these annotations by the IVA. Accurately representing nonverbal behavior is an especially crucial component in fostering positive human-computer interactions, and the field is currently moving towards a standardized evaluation of their quality [43].

A side-by-side comparison between the video and our simulation highlights distinctions between the ground truth and Diana's current capabilities. This contrasts with approaches that display trained model output alongside the participants' behaviors as in [25]. The Diana avatar treats all events as a strict interpretation of its VoxML definition, and does not currently consider the speed, handedness or manner of particular gestures or actions. The current implementation of the VoxML *put* program, for instance, takes a set amount of time to complete, and as a result the avatar struggles to relay drawn-out or rapid-fire actions. Additionally, we noticed that due to the mathematical nature of the task, gestures would often refer to multiple objects as a group. Though these gestures could easily be modeled in Gesture AMR, the particular manner of gesturing towards multiple objects at once could have had numerous potential interpretations with Diana (point averaging, back-and-forth gesturing, using two hands). These observations provide interesting avenues of exploration for

expanding Diana's expressiveness, involving iterative adjustments to the event primitive programming aimed at bridging the gap between video and simulation.

5 Conclusion

In this paper, we argue that, in the context of the evolving notion of embodied HCI [8, 31], there is a serious need to develop a unified multimodal annotation scheme, one that includes human-object interactions, in addition to speech and gesture. To this end, we introduced the Shared Weights Task, an collaborative problem-solving task involving deducing the weights of blocks, as a subject for multi-party dialogue investigation. To visualize the developing state of the interaction as captured by an annotation score on example Shared Weights Task videos, we used the VoxWorld platform to develop FibWorld, a virtual recreation of the task that allows embodied IVAs to stand in for human task participants, recreating their behavior across multiple modalities. Finally, we discussed how this simulation visualization can be used to intuitively correct and validate multimodal annotation schemas, as well as provide insight into expanding the ranges of expression of the associated embodied IVA agents.

References

1. Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., Schneider, N.: Abstract meaning representation (amr) 1.0 specification. In: Parsing on Freebase from Question-Answer Pairs. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL. pp. 1533–1544 (2012)
2. Bonial, C., Donatelli, L., Abrams, M., Lukin, S., Tratz, S., Marge, M., Artstein, R., Traum, D., Voss, C.: Dialogue-AMR: Abstract Meaning Representation for Dialogue. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 684–695 (2020)
3. Bradford, M., Hansen, P., Lai, K., Brutti, R., Dickler, R., Hirshfield, L.M., Pustejovsky, J., Blanchard, N., Krishnaswamy, N.: Challenges and opportunities in annotating a multimodal collaborative problem-solving task
4. Brugman, H., Russel, A.: Annotating multi-media/multi-modal resources with ELAN. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04). European Language Resources Association (ELRA), Lisbon, Portugal (May 2004), http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf
5. Brutti, R., Donatelli, L., Lai, K., Pustejovsky, J.: Abstract Meaning Representation for gesture. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1576–1583. European Language Resources Association, Marseille, France (Jun 2022), https://aclanthology.org/2022.lrec-1.169
6. Cassell, J., Sullivan, J., Churchill, E., Prevost, S.: Embodied conversational agents. MIT press (2000)
7. Copestake, A., Flickinger, D., Pollard, C., Sag, I.A.: Minimal recursion semantics: An introduction. Research on Language and Computation **3**(2-3), 281–332 (2005)

8. Evans, L., Rzeszewski, M.: Hermeneutic relations in vr: Immersion, embodiment, presence and hci in vr gaming. In: HCI in Games: Second International Conference, HCI-Games 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22. pp. 23–38. Springer (2020)
9. 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)
10. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6047–6056 (2018)
11. Gupta, S., Malik, J.: Visual semantic role labeling. arXiv preprint arXiv:1505.04474 (2015)
12. Hassanin, M., Khan, S., Tahtali, M.: Visual affordance and function understanding: A survey. ACM Computing Surveys (CSUR) **54**(3), 1–35 (2021)
13. Helfrich, P., Rieb, E., Abrami, G., Lücking, A., Mehler, A.: Treeannotator: versatile visual annotation of hierarchical text relations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018) (2018)
14. Henlein, A., Gopinath, A., Krishnaswamy, N., Mehler, A., Pustejovsky, J.: Grounding human-object interaction to affordance behavior in multimodal datasets. Frontiers in Artificial Intelligence **6**, 2
15. Kamp, H., Van Genabith, J., Reyle, U.: Discourse representation theory. In: Handbook of philosophical logic, pp. 125–394. Springer (2011)
16. Karau, S.J., Williams, K.D.: Social loafing: A meta-analytic review and theoretical integration. Journal of personality and social psychology **65**(4), 681 (1993)
17. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gestures: how to economically capture timing and form. Language Resources and Evaluation **41**, 325–339 (2007)
18. Knight, K., Baranescu, L., Bonial, C., Georgescu, M., Griffitt, K., Hermjakob, U., Marcu, D., Palmer, M., Schneifer, N.: Abstract meaning representation (AMR) annotation release 1.2.6. Web download (2019)
19. Kopp, S., Krenn, B., Marsella, S., Marshall, A.N., Pelachaud, C., Pirker, H., Thórisson, K.R., Vilhjálmsson, H.: Towards a common framework for multimodal generation: The behavior markup language. In: International workshop on intelligent virtual agents. pp. 205–217. Springer (2006)
20. Kopp, S., Wachsmuth, I.: Gesture in embodied communication and human-computer interaction, vol. 5934. Springer (2010)
21. Krishnaswamy, N., Beveridge, R., Pustejovsky, J., Patil, D., McNeely-White, D.G., Wang, H., Ortega, F.R.: Situational awareness in human computer interaction: Diana's world (2020)
22. Krishnaswamy, N., Narayana, P., Bangar, R., Rim, K., Patil, D., McNeely-White, D., Ruiz, J., Draper, B., Beveridge, R., Pustejovsky, J.: Diana's world: A situated multimodal interactive agent. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13618–13619 (2020)
23. Krishnaswamy, N., Pickard, W., Cates, B., Blanchard, N., Pustejovsky, J.: The voxworld platform for multimodal embodied agents. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. pp. 1529–1541 (2022)
24. Krishnaswamy, N., Pustejovsky, J.: Voxsim: A visual platform for modeling motion language. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations. pp. 54–58 (2016)

25. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The bielefeld speech and gesture alignment corpus (saga) (01 2010). https://doi.org/10.13140/2.1.4216.1922
26. Marshall, P., Hornecker, E.: Theories of embodiment in HCI. The SAGE handbook of digital technology research **1**, 144–158 (2013)
27. Martin, J.C., Niewiadomski, R., Devillers, L., Buisine, S., Pelachaud, C.: Multimodal complex emotions: Gesture expressivity and blended facial expressions. International Journal of Humanoid Robotics **3**(03), 269–291 (2006)
28. Nakayama, H., Kubo, T., Kamura, J., Taniguchi, Y., Liang, X.: doccano: Text annotation tool for human (2018), `https://github.com/doccano/doccano`, software available from https://github.com/doccano/doccano
29. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Computational Linguistics (2003)
30. Podlasov, A., Tan, S., O'Halloran, K.: Interactive state-transition diagrams for visualization of multimodal annotation. Intelligent Data Analysis **16**, 683–702 (07 2012). https://doi.org/10.3233/IDA-2012-0544
31. Pustejovsky, J., Krishnaswamy, N.: Embodied human computer interaction. Künstliche Intelligenz (2021)
32. Pustejovsky, J.: Unifying linguistic annotations: A timeml case study. In: Proceedings of Text, Speech, and Dialogue Conference (2006)
33. Pustejovsky, J., Krishnaswamy, N.: Voxml: A visualization modeling language. Proceedings of LREC (2016)
34. Pustejovsky, J., Krishnaswamy, N.: Voxml: A visualization modeling language. arXiv preprint arXiv:1610.01508 (2016)
35. Pustejovsky, J., Krishnaswamy, N.: Multimodal semantics for affordances and actions. In: Human-Computer Interaction. Theoretical Approaches and Design Methods: Thematic Area, HCI 2022, Held as Part of the 24th HCI International Conference, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings, Part I. pp. 137–160. Springer (2022)
36. Pustejovsky, J., Stubbs, A.: Natural Language Annotation for Machine Learning: A guide to corpus-building for applications. " O'Reilly Media, Inc." (2012)
37. Reallusion Inc.: Character Creator 4 (2022), `https://www.reallusion.com/character-creator/`
38. Sadhu, A., Gupta, T., Yatskar, M., Nevatia, R., Kembhavi, A.: Visual semantic role labeling for video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5589–5600 (2021)
39. 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)
40. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a web-based tool for nlp-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 102–107 (2012)
41. Van Gysel, J.E., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C.R., et al.: Designing a uniform meaning representation for natural language processing. KI-Künstliche Intelligenz pp. 1–18 (2021)
42. Wahlster, W.: Dialogue systems go multimodal: The smartkom experience. In: SmartKom: foundations of multimodal dialogue systems, pp. 3–27. Springer (2006)
43. Wolfert, P., Robinson, N., Belpaeme, T.: A review of evaluation practices of gesture generation in embodied conversational agents. IEEE Transactions on Human-Machine Systems (2022)

44. Yang, S., Gao, Q., Liu, C., Xiong, C., Zhu, S.C., Chai, J.: Grounded semantic role labeling. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 149–159 (2016)
45. 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)
46. Ziem, A.: Do we really need a multimodal construction grammar? Linguistics Vanguard **3**(s1) (2017)