Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1576–1583
Marseille, 20-25 June 2022

Abstract Meaning Representation for Gesture

Richard Brutti[1], Lucia Donatelli[2], Kenneth Lai[1], James Pustejovsky[1]
[1]Brandeis University, [2]Saarland University
[1]Waltham, MA, USA, [2]Saarbrücken, Germany
{brutti, klai12, jamesp}@brandeis.edu, donatelli@coli.uni-saarland.de

Abstract

This paper presents Gesture AMR, an extension to Abstract Meaning Representation (AMR), that captures the meaning of gesture. In developing Gesture AMR, we consider how gesture form and meaning relate; how gesture packages meaning both independently and in interaction with speech; and how the meaning of gesture is temporally and contextually determined. Our case study for developing Gesture AMR is a focused human-human shared task to build block structures. We develop an initial taxonomy of gesture act relations that adheres to AMR's existing focus on predicate-argument structure while integrating meaningful elements unique to gesture. Pilot annotation shows Gesture AMR to be more challenging than standard AMR, and illustrates the need for more work on representation of dialogue and multimodal meaning. We discuss challenges of adapting an existing meaning representation to non-speech-based modalities and outline several avenues for expanding Gesture AMR.

Keywords: Dialogue, Gesture, Multimodal Interaction, AMR

1. Introduction

Meaning representations in the form of annotated graphbanks have become a popular tool to represent the semantics of language for various NLP tasks. Abstract Meaning Representation (AMR) is one such graph-based meaning representation that expresses the meaning of a sentence in terms of its predicate-argument structure (Banarescu et al., 2013). AMRs were designed to be easy for humans to annotate (supporting the creation of corpora/sembanks) and easy for computers to parse. An example AMR for the English language sentence "Put that block there", is shown in PENMAN (Matthiessen and Bateman, 1991) notation in Example (1) below:

(1) Put that block there.

```
(p / put-01
  :mode imperative
  :ARG0 (y / you)
  :ARG1 (b / block
          :mod (t / that))
  :ARG2 (t2 / there))
```

Though AMR was designed to represent meaning in English language, its design omits many important elements necessary to understand linguistic meaning in context. The AMR in Example (1) illustrates this. While the AMR conveys basic information about the desired action and the object upon which that action operates, it lacks precise grounding and spatial information necessary to successfully interpret and execute the command in the environment. Such information is provided by a gesture like the one shown in Figure 1, which can also specify the manner of motion (whether the "put" action is slow or fast) and clarify whether the addressee has understood the instructions correctly with gestural backchanneling. Example (1) lacks additional situated information characteristic of speech such as intonation, pauses, or disfluencies – temporal

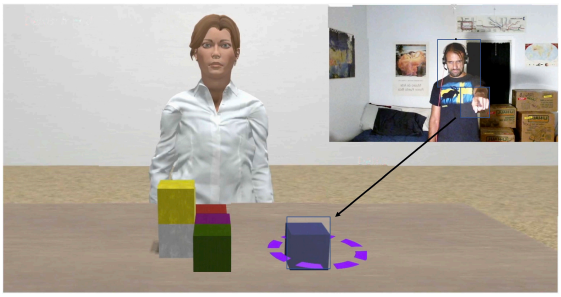

Figure 1: Person instructing an avatar through camera in a shared task of stacking blocks.

elements that interact with non-speech-based modalities such as gesture in meaningful ways. As a result, there is no clear way to align a representation of gesture (or another non-speech-based modality) to the language expressed in the AMR temporally, and much of the intended and interpreted meaning is lost.

In this paper, we address the challenge of extending AMR to additional modalities for greater expressivity and outline initial specifications for Gesture AMR. Our primary goals are to explore how the flexible design of AMR can accommodate the semantics of gesture, a non-verbal modality often crucial for full understanding of a speaker's meaning. Research on gesture has cast doubt on whether different modes of expression convey the same kinds of semantic content (Schlenker, 2018); as well as whether a coarse-grained, lexically-oriented meaning representation such as AMR can adequately capture a more underspecified morphology that can diverge from linguistic principles of compositionality (Cassell et al., 2007; McNeill, 2008). Extending AMR to gesture thus faces several challenges. While AMR aims to represent the salient semantic content of a linguistic utterance, the meaning of gesture requires integrating broader notions of context, in-

teractions with other communicative modalities, and speaker profiles. A question that arises from this analogy is whether a meaning representation should represent content independently of the mode of expression, or whether a "one-size-fits-all" approach to meaning representation is ill-suited to capture global meaning in context.

To develop the Gesture AMR schema, we are in the process of annotating gestures from the EGGNOG corpus (Wang et al., 2017b), a task-based corpus in which one participant instructs another to build a block structure using gesture and language. As these are initial specifications for Gesture AMR, we focus on content-bearing gesture that carries semantic weight independent of spoken language; gesture that carries expressive or ampliative meaning is saved for future work. We outline the specifications of Gesture AMR in such a way that leaves room for their expansion to non-task-based settings. Initial annotation results show that annotating both standard AMR and Gesture AMR in situated dialogue is challenging, and more work for both language and other modalities in such settings is needed.

The paper is structured as follows. Section 2 describes related work on gesture and other extensions to AMR. Section 3 grounds our work, and outlines what it aims to capture. We provide the specification of Gesture AMR in Section 4, and discuss our future work in Section 5.

2. Background

Gesture refers to the way speakers move their hands when they speak and communicate information. Several existing annotation schemes for gesture focus on its descriptive characteristics, such as hand shape, trajectory shape, location with respect to the body, palm orientation, trajectory movement, and trajectory relative size (Rohrer et al., 2020; Kopp and Wachsmuth, 2010; Kong et al., 2015). Additional work on gesture focuses on its integral link with spoken language in terms of timing and function (Kendon, 1997; Kendon, 2004; McNeill, 2008).

Our focus in this paper is gesture that is directly tied to speech and carries the same intentionality attributed to speech; such gesture carries meaning on its own or enhances the meaning provided by the verbal modality (Goldin-Meadow, 2003). To this end, we draw on annotation schemes that differentiate the following four referential forms of gestures (Ekman and Friesen, 1969; Mather, 2005; McNeill, 2011): iconic, deictic, metaphoric, and emblematic. We do not currently consider beat or rhythmic gesture that may control the flow of speech or emphasize certain words or phrases, though future work will investigate the meaningful properties of such gesture. We also draw inspiration from annotation schemes that focus specifically on the alignment and interaction of gesture and speech; nevertheless, such schema are primarily descriptive and do not encode meaning (Kopp and Wachsmuth, 2010; Kipp et al., 2007).

Beyond gesture, there are few meaning representations for situated (dialogue) interactions that are both adequately expressive of the content and compact enough for corpus development. Indeed, it has been noted that there is a lack of substantial empirical support to arrive at a detailed and data-based understanding of the nature of multimodal constructions in conversation (Ziem, 2017). With regards to extending AMR, AMR has been applied to multi-sentence settings (O'Gorman et al., 2018), to spatial information (Bonn et al., 2020), and to task-oriented dialogues to include the *how* as well as the *what* of what is said (Bonial et al., 2020). More recently, an extension of AMR, Uniform Meaning Representation (UMR), has been developed to be scalable, accommodate cross-linguistic diversity, and support lexical and logical inference (Van Gysel et al., 2021). To this end, UMR incorporates aspect, scope, temporal and modal dependencies, as well as inter-sentential coreference.

Other systems have been proposed for describing gesture in the context of embodied conversational agents, specifically the Behavior Markup Language (BML) (Kopp et al., 2006). BML is an exchange language originally intended to describe an agent's actions along a range of axes (e.g., gesture, gaze). BML descriptions subsequently require an interpretation layer, so that the agent can execute an action. In comparison, AMR is annotated only for textual data, and it is thus currently unable to capture layers of modality, such as gesture. These layers contribute to situated meaning, or meaning grounded in a physical environment and negotiated through the (often multimodal) dynamics of discourse (Pustejovsky and Krishnaswamy, 2021).

We consider these properties in designing Gesture AMR, as well as UMR's inter-sentential coreference, which is necessary in situated grounding to keep track of entities and events over time.

3. Approach

One of the most persistent and challenging problems facing the area of human-robot interaction involves communicating intentions, goals, and attitudes through multiple modalities beyond language, including gesture, gaze, facial expressions, and situational awareness (Cassell et al., 2000; Foster, 2007; Kopp and Wachsmuth, 2010; Marshall and Hornecker, 2013; Schaffer and Reithinger, 2019; Wahlster, 2006). In the context of task-oriented dialogues (Tellex et al., 2020) with robots, this introduces the problem of identifying and modifying the *common ground* between humans or human and robot (Clark and Brennan, 1991; Stalnaker, 2002; Tomasello and Carpenter, 2007).

In particular, in order to represent *situated meaning* (Pustejovsky and Krishnaswamy, 2021), that is, to ground actions and objects in their environment, we need to represent not only *what* is being communicated,

but also *how* it is being communicated, i.e., what the agents (e.g., the speaker/gesturer and the addressee(s)) are doing. In addition to the content of the utterance, our meaning representation must therefore also represent the agents involved, as well as indicate the mode of communication.

On the other hand, unlike previous work that describes gesture in terms of its physical attributes, we abstract away from physical descriptions, in the same way that AMR was designed to abstract away from syntax and individual lexical items. That being said, we aim to keep our proposal compatible with a future alignment between our semantic representation and a physical description of a gesture.

4. Gesture AMR

As noted, the initial Gesture AMR specification is based on the EGGNOG corpus (Wang et al., 2017a). EGGNOG is comprised of 8 hours of video across 40 participants, working in pairs on a shared task. The participants are located in different rooms connected by video and/or audio. One person (*actor*) has a set of wooden blocks, and the other (*signaler*) has a picture of a specific block arrangement. The signaler must get the actor to arrange the blocks as in the specific arrangement. EGGNOG videos are typically around 1 minute long, and feature natural continuous communication. We acknowledge that gesture can be culture- and individual-specific. While there is some variation in age (from 19 to 64 years), EGGNOG participants were largely recruited from a university setting, and are all English speakers. The somewhat homogeneous group of gesturers, in combination with the task-based premise, likely limits the range of gestures exhibited in the corpus.

Our student annotators are tasked with writing Gesture AMRs for each discrete content-bearing gesture, as per our guidelines[1]. The original release of EGGNOG contains time-stamped annotations of participant gestures, for both the gesturer's inferred intent, as well a physical description of the movement.

We use ELAN (Brugman and Russel, 2004) to carry out the annotations, due to its ability to annotate multiple "tracks" of information simultaneously, while viewing the existing EGGNOG labels. Annotators first create speech AMRs in one track, then the Gesture AMR in another. Each video is labeled by multiple annotators, and then adjudicated.

We propose the following general form for a Gesture AMR:

```
(g / [gesture]-GA
  :ARG0 [gesturer]
  :ARG1 [content]
  :ARG2 [addressee])
```

By analogy with Dialogue-AMR (Bonial et al., 2020), each Gesture AMR is rooted by one of four gesture act (GA) relations, described below. `ARG0` and `ARG2` are the gesturer and addressee, respectively, while `ARG1` contains the semantic content of the gesture, which varies by gesture type; this is also similar to Dialogue-AMR.

As noted above, our gesture act taxonomy reflects that of Ekman and Friesen (1969), Mather (2005), McNeill (2011), and Kong et al. (2015), among others:

- **deixis-GA**: A deictic gesture refers to an object or location, by pointing to it. The semantic content is then simply the pointed-to object or location itself, which can be represented in the same way as they are in a standard AMR, e.g., `(b / block)`. While in some cases it may be desirable to include additional detail, e.g., specific coordinates of locations, this may not be necessary (or feasible) in every case.

  For example, a pointing gesture aimed at a block will be annotated as:

```
(d / deixis-GA
  :ARG0 (g / gesturer)
  :ARG1 (b / block)
  :ARG2 (a / addressee))
```

- **icon-GA**: An iconic gesture refers to an object or action, by depicting some concrete property of it, such as an object's shape or an action's manner. These can also be represented as they are in a standard AMR. While English AMRs draw their inventory of verbs from PropBank (Palmer et al., 2005), specific applications can define their own lists of standardized concepts, in the same way that Dialogue-AMR maps several PropBank frames to a single robot action (Bonial et al., 2020).

  If a gesturer makes a "pushing" motion with their hands away from their body, indicating that they want the addressee to slide an object forward, that gesture will be annotated as:

```
(i / icon-GA
  :ARG0 (g / gesturer)
  :ARG1 (s / slide-01
          :direction (f / forward))
  :ARG2 (a / addressee))
```

- **metaphor-GA**: In contrast to iconic gestures, a metaphoric gesture depicts an abstract property of a concept or idea.

  As an example (as described by Kong et al. (2015)), a gesturer can, while saying "I want to show you something", trace a circle with their finger, the circle metaphorically denoting "something". In this case, the gesture will be annotated as:

```
(m / metaphor-GA
  :ARG0 (g / gesturer)
  :ARG1 (s / something)
  :ARG2 (a / addressee))
```

[1] Guidelines available at https://github.com/klai12/multimodal-amr-annotation-project

- **emblem-GA**: An emblematic gesture has a meaning that is set by convention, rather than by any physical or metaphorical similarity between the gesture and its semantic content.

  For example, a "thumbs up" gesture will be annotated as:

```
(e / emblem-GA
  :ARG0 (g / gesturer)
  :ARG1 (y / yes)
  :ARG2 (a / addressee))
```

In many cases, a single gesture can have multiple components to its meaning (Kendon, 2004). For example, if a gesturer arranges their hands in the shape of a square (denoting a block), while simultaneously moving them downwards towards a table (denoting a location on the table), that gesture has both iconic and deictic components to its meaning. In these cases, we annotate both components, and include them as subparts of a `(g / gesture-unit)`, as follows:

```
(g / gesture-unit
  :op1 (i / icon-GA
         :ARG0 (g2 / gesturer)
         :ARG1 (b / block)
         :ARG2 (a / addressee))
  :op2 (d / deixis-GA
         :ARG0 g2
         :ARG1 (l / location)
         :ARG2 a))
```

We utilize `(a / and)` to connect Gesture AMRs for coordinated gestures. Coordinated gestures are made simultaneously but are independent from each other (e.g., a gesturer pointing to two separate locations, one with each hand). In coordinated gestures, the meaning of each component gesture is separable from the other, unlike in a `(g / gesture-unit)`.

4.1. Alignment in Gesture AMR

In multimodal communication, the same objects can be referenced in both the speech and gesture modalities. When that occurs, we must have some way to mark coreference across the different modalities. Furthermore, the design of AMR abstracts away from the string form of a spoken or written sentence; this extends to Gesture AMR, in which we represent the semantic content of the gesture and not its temporal sequencing. As a result, temporal alignment between gesture, language, and other modalities is quite challenging if not impossible in AMR alone. We thus introduce an additional layer of representation to capture semantic and temporal alignment between modalities, whose formalization we sketch here.

A multimodal communicative act, C, as in Figure 2, consists of a sequence of gesture-language ensembles, (g_i, s_i), where an ensemble is temporally aligned in the common ground (Pustejovsky and Krishnaswamy, 2021). Let us assume that a linguistic subexpression, s, is either a word or full phrase in the utterance, while a gesture, g, comports with the Gesture AMR described above.

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

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

CO-GESTURAL SPEECH

HUMAN: s_1 = Put
g_1 = $\varnothing$
HUMAN: s_2 = [that block]
g_2 = [*points to the blue block*]
HUMAN: s_3 = there.
g_3 = [*points to the purple block*]

Figure 2: Communicative act with speech and gesture.

(a)
```
(s1c / command-00
  :ARG0 (g / gesturer)
  :ARG1 (c2 / communicative-act
          :gesture (d / deixis-GA
                     :ARG0 g
                     :ARG1 (b / block
                             :ARG1-of (b2 / blue-01))
                     :ARG2 (a / addressee))
          :gesture (d2 / deixis-GA
                     :ARG0 g
                     :ARG1 (b3 / block
                             :ARG1-of (p / purple-02))
                     :ARG2 a)
          :speech (p2 / put-01
                    :mode imperative
                    :ARG0 (y / you)
                    :ARG1 (b4 / block
                            :mod (t / that))
                    :ARG2 (t2 / there))
  :ARG2 a))
```

(b)
```
(s1 / sentence
  :coref ((b :same-entity b4)
          (a :same-entity y))
  :alignment ((d :overlap t)
              (d2 :overlap t2)
              (d :before d2)))
```

Figure 3: Meaning representation corresponding to the communicative act in Figure 2.

The example in Figure 2 combines the spoken command "Put that block there" as in Example (1) with the deictic gesture shown in Figure 1. In Figure 3(a), we represent the communicative act, with two Gesture AMRs as arguments of `:gesture` and an AMR for the speech as the argument of `:speech`. We additionally enclose the communicative act within a Dialogue-AMR (Bonial et al., 2020) "speech act" (that, in this example, is not limited to the speech modality) that marks its illocutionary force, namely, as a command.

Then, in Figure 3(b), we present the semantic and temporal alignments between the two modalities. For encoding the temporal relations between expressions in both modalities, we follow (Pustejovsky et al., 2010), adopting TimeML's encoding of Interval Temporal Logic (Allen, 1983). For the present discussion, we adopt a reduced subset of the 13 Allen relations, where, most significantly, `overlap` is a disjunction

of ITL's **overlap**, **overlap_inverse**, **during**, and **during_inverse**. Our enriched formalism is then based on that of UMR (Van Gysel et al., 2021), which in turn bases its approaches to inter-sentential coreference and temporal markup on Multi-sentence AMR (O'Gorman et al., 2018) and the TimeML-based temporal dependency structures (Zhang and Xue, 2018), respectively. Following the basic strategy employed in MultiML (Giuliani and Knoll, 2008) for aligning multiple modalities, we use an AMR-native device to capture the hybrid logic reentrancy binding from (Baldridge and Kruijff, 2002). We mark that the blue block being pointed to and "that block" mentioned in the speech are the same entity, as are the addressee of the gesture and "you", an implicit argument of the speech. We also mark that the first deictic gesture `d` temporally overlaps with the word "that", the second gesture `d2` overlaps with "there", and that `d` occurs before `d2`.

4.2. Spatial Description in Gesture AMR

Gesture provides additional meaning in the form of spatial information in ways more concisely than language. For example, in Figure 1, a linguistic description of exactly where the speaker is pointing would need to be quite long to capture the specific location indicated with a simple pointing gesture. The spatial coordinates of such a location should be documented for a complete grounding of the Gesture AMR to the environment. In addition to more precisely specifying locations, gesture also specifies spatial elements such as start point, end point, manner, and duration of motion; size of objects; and relative position of events and objects to a speaker and to each other. Such depiction is important for conveying and interpreting meaning and grounding language to environment (Capirci et al., 2022).

After our initial annotation on the EGGNOG dataset focusing on content-bearing gestures, we plan to augment our annotations with spatial information and Spatial AMR (Bonn et al., 2020). Spatial AMR adds spatial rolesets to the PropBank lexicon and is conceptualized around events and relations as construed in language; this includes whether events are static or dynamic, and it describes characteristics of relations related to location, orientation, configuration, and extent. Spatial AMR also incorporates Cartesian coordinates for mapping to physical space.

4.3. Gesture AMR Annotation

At publication, annotation for standard, speech AMR and Gesture AMR is in progress on the EGGNOG corpus. For our initial annotation, we compiled a small subset of the EGGNOG corpus, containing 21 videos, with a total length of 23 minutes. The ELAN environment allows a single annotator to complete both tasks within the same working session. We use the Smatch metric to measure inter-annotator agreement on both the speech and Gesture AMR. (Cai and Knight, 2013). We additionally perform manual agreement analysis

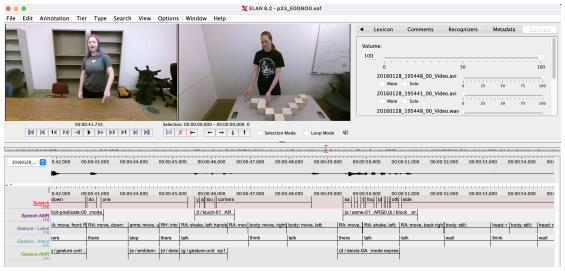

Figure 4: ELAN annotation environment for Gesture AMR, with EGGNOG videos.

for subgraphs of individual Gesture AMRs to highlight particular areas of agreement and disagreement. We also record the average time spent for annotating entire EGGNOG video sequences. Initial results show that individual Gesture AMRs can take up to several minutes each for annotation. A one-minute EGGNOG video clip with roughly 10 to 15 gesture units can take an expert annotator approximately 20 minutes to annotate. Annotator velocity is improving as annotation guidelines are clarified, and annotators have more practice with both the schema and tool.

As noted in the Introduction, AMR was developed to capture the semantics of sentences (Banarescu et al., 2013), which are an artifact of text as the mode of communication. AMR has been extended to specific speech-based use cases, such as with Dialogue-AMR (Bonial et al., 2020). Input utterances for Dialogue-AMR are from human-robot interactions, and are predominantly imperatives geared towards the robot's known constrained behaviors. However, the speech in EGGNOG tends towards monologues, with the signalers describing the block structures via largely one-way communication. The EGGNOG speech does not neatly arrange into utterances, and does not follow typical conversational turn-taking patterns (Sacks et al., 1974). As such, our annotators are implicitly tasked with providing endpoints to utterances, so that they can create AMRs for the speech. The lack of clear distinctions between utterances can lead to misalignment of the annotations, which presents challenges for adjudication and measuring agreement on a video level. Similar challenges are present with Gesture AMR. The distinction between a `(g / gesture-unit)` and two consecutive gestures may be difficult to discern, leading to similar alignment challenges.

4.4. Towards a Lexicon of Gesture AMR

In developing Gesture AMR, we are asking the question of whether or not there are additional "lexical" items, concepts, or relations unique to meaning conveyed in gesture that are absent in language and English AMR. Gesture AMR currently makes use of the framesets from PropBank (Kingsbury and Palmer, 2002; Palmer et al., 2005) as used in English AMR for natural language. In addition to this, the four gestural act (GA) frames explained above are used to signal the intended

interpretation of the gesture; these GAs are similar in function to the dialogue acts of Dialogue-AMR (Bonial et al., 2020).

As noted in Section 2, future work will extend Gesture AMR to additional gesture acts whose meaning is not strictly independent from language or compositional in a manner similar to language (Kendon, 1990). This work will address several aspects of gesture meaning, specifically: that it is *global*, such that the meanings of the parts are determined by the whole in a top-down (versus bottom-up) manner; that it is *synthetic*, in that a single gesture can accompany an entire sentence; that it is *instantaneous*, such that gesture meaning is not accumulated over time as in spoken language; and that it is *dynamic* and shaped by discourse context, speaker intention and memory, as well as the co-expressive, synchronous speech (McNeill, 2008). Such work necessarily incorporates more formalization of temporal and semantic alignment between speech and gesture. Focused, task-based datasets such as EGGNOG can provide initial data for this work; additional datasets that provide a broader range of gesticulation, pantomime, and other non-speech-linked gestures will allow more comprehensive analysis of gesture's meaning in other situated and non-task-based settings.

A related consideration in refining Gesture AMR is how to represent meaning in gesture that is non-propositional, such as beat and rhythmic gestures. We inherit a limited vocabulary of concepts related to English syntactic modal expressions from standard AMR. However, much meaning in gesture is performative and *expressive*, revealing subtleties about the speaker's persona and perspective that can impact how current and future gestures are interpreted (Cruse, 1986; Potts, 2007). AMR's `:mode expressive`, though a viable placeholder for now, will need to be expanded to capture the range of expressive meaning in gesture; this concern inevitably extends to other non-speech-based modalities.

5. Discussion and Future Work

This paper presents a specification for Gesture AMR, intended to capture semantics of gesture. We recognize that gesture can play a part of either the direct content of the utterance (Stojnić et al., 2020) or the cosuppositional content (Kendon, 1990; Schlenker, 2020). Hence, we must assume that natural interactions with computers and robots have to account for interpreting and generating language and gesture. As presented here, Gesture AMR focuses on content-bearing gesture, and is independent from the meanings represented in speech. Beginning with with the task-based EGGNOG corpus, Gesture AMR is designed to be extensible to other types of gesture. Incorporating gesture into AMR and its various extensions will allow for a fuller representation of the communicative act.

As introduced, each Gesture AMR is based on one of four gesture act relations, and contains reference to the gesturer, addressee, and the semantic content of the gesture. We introduce an additional layer of representation to capture the semantic and temporal alignment between the multiple modalities of speech and gesture (as well as to-be-described modalities). We look forward to completing our EGGNOG annotation, and further developing Gesture AMR using a wider range of corpora that include non-task-based contexts.

6. Acknowledgements

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 to James Pustejovsky, an NSF Student Grant funded by DRL 2019805, to Kenneth Lai, Richard Brutti, and Lucia Donatelli, as well as by the IIS Division of the NSF via Award No. 1763926 entitled "Building a Uniform Meaning Representation for Natural Language Processing". All views expressed in this paper are those of the authors and do not necessarily represent the views of the NSF.

7. Bibliographical References

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Baldridge, J. and Kruijff, G.-J. (2002). Coupling CCG and hybrid logic dependency semantics. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Bonial, C., Donatelli, L., Abrams, M., Lukin, S. M., Tratz, S., Marge, M., Artstein, R., Traum, D., and Voss, C. (2020). Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France, May. European Language Resources Association.

Bonn, J., Palmer, M., Cai, Z., and Wright-Bettner, K. (2020). Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France, May. European Language Resources Association.

Brugman, H. and Russel, A. (2004). Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*,

Lisbon, Portugal, May. European Language Resources Association (ELRA).

Cai, S. and Knight, K. (2013). Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria, August. Association for Computational Linguistics.

Capirci, O., Caselli, M. C., and Volterra, V. (2022). Interaction among modalities and within development.

Cassell, J., Sullivan, J., Churchill, E., and Prevost, S. (2000). *Embodied conversational agents*. MIT Press.

Cassell, J., Kopp, S., Tepper, P., Ferriman, K., and Striegnitz, K. (2007). Trading spaces: How humans and humanoids use speech and gesture to give directions.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition*, 13(1991):127–149.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Ekman, P. and Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1):49–98.

Foster, M. E. (2007). Enhancing human-computer interaction with embodied conversational agents. In *International Conference on Universal Access in Human-Computer Interaction*, pages 828–837. Springer.

Giuliani, M. and Knoll, A. (2008). MultiML: a general purpose representation language for multimodal human utterances. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 165–172.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Kendon, A. (1990). *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, Cambridge, U.K.

Kendon, A. (1997). Gesture. *Annual Review of Anthropology*, 26(1):109–128.

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

Kingsbury, P. and Palmer, M. (2002). From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain, May. European Language Resources Association (ELRA).

Kipp, M., Neff, M., and Albrecht, I. (2007). An annotation scheme for conversational gestures: how to economically capture timing and form. *Language Resources and Evaluation*, 41(3):325–339.

Kong, A. P.-H., Law, S.-P., Kwan, C. C.-Y., Lai, C., and Lam, V. (2015). A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a database of speech and gesture (DoSaGE). *Journal of Nonverbal Behavior*, 39(1):93–111.

Kopp, S. and Wachsmuth, I. (2010). *Gesture in embodied communication and human-computer interaction*, volume 5934. Springer.

Kopp, S., Krenn, B., Marsella, S., Marshall, A. N., Pelachaud, C., Pirker, H., Thórisson, K. R., and Vilhjálmsson, H. (2006). Towards a common framework for multimodal generation: The Behavior Markup Language. In *International Workshop on Intelligent Virtual Agents*, pages 205–217. Springer.

Marshall, P. and Hornecker, E. (2013). Theories of embodiment in HCI. *The SAGE Handbook of Digital Technology Research*, 1:144–158.

Mather, S. M. (2005). Ethnographic research on the use of visually based regulators for teachers and interpreters. *Attitudes, Innuendo, and Regulators*, pages 136–161.

Matthiessen, C. and Bateman, J. A. (1991). *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Burns & Oates.

McNeill, D. (2008). *Gesture and thought*. University of Chicago Press.

McNeill, D. (2011). *Hand and mind*. De Gruyter Mouton.

O'Gorman, T., Regan, M., Griffitt, K., Hermjakob, U., Knight, K., and Palmer, M. (2018). AMR beyond the sentence: the multi-sentence AMR corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3693–3702, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

Potts, C. (2007). The expressive dimension. *Theoretical Linguistics*, 33(2).

Pustejovsky, J. and Krishnaswamy, N. (2021). Embodied human computer interaction. *KI-Künstliche Intelligenz*, 35(3):307–327.

Pustejovsky, J., Lee, K., Bunt, H., and Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Rohrer, P. L., Vilà-Giménez, I., Florit-Pons, J., Esteve-Gibert, N., Ren, A., Shattuck-Hufnagel, S., and Prieto, P. (2020). The multimodal multidimensional (M3D) labelling scheme for the annotation of audiovisual corpora. In *7th Gesture and Speech in Interaction (GESPIN)*.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4 Part 1):696–735.

Schaffer, S. and Reithinger, N. (2019). Conversation is multimodal: thus conversational user interfaces

should be as well. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, pages 1–3.

Schlenker, P. (2018). Gesture projection and cosuppositions. *Linguistics and Philosophy*, 41(3):295–365.

Schlenker, P. (2020). Gestural grammar. *Natural Language & Linguistic Theory*, 38(3):887–936.

Stalnaker, R. (2002). Common ground. *Linguistics and Philosophy*, 25(5-6):701–721.

Stojnić, U., Stone, M., and Lepore, E. (2020). Pointing things out: in defense of attention and coherence. *Linguistics and Philosophy*, 43(2):139–148.

Tellex, S., Gopalan, N., Kress-Gazit, H., and Matuszek, C. (2020). Robots that use language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3:25–55.

Tomasello, M. and Carpenter, M. (2007). Shared intentionality. *Developmental Science*, 10(1):121–125.

Van Gysel, J. E. L., Vigus, M., Chun, J., Lai, K., Moeller, S., Yao, J., O'Gorman, T., Cowell, A., Croft, W., Huang, C.-R., Hajič, J., Martin, J. H., Oepen, S., Palmer, M., Pustejovsky, J., Vallejos, R., and Xue, N. (2021). Designing a uniform meaning representation for natural language processing. *KI-Künstliche Intelligenz*, 35(3):343–360.

Wahlster, W. (2006). Dialogue systems go multimodal: The SmartKom experience. In *SmartKom: foundations of multimodal dialogue systems*, pages 3–27. Springer.

Wang, I., Narayana, P., Patil, D., Mulay, G., Bangar, R., Draper, B., Beveridge, R., and Ruiz, J. (2017a). EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition*.

Wang, I., Narayana, P., Patil, D., Mulay, G., Bangar, R., Draper, B., Beveridge, R., and Ruiz, J. (2017b). Exploring the use of gesture in collaborative tasks. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17, pages 2990–2997, New York, NY, USA. ACM.

Zhang, Y. and Xue, N. (2018). Structured interpretation of temporal relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Ziem, A. (2017). Do we really need a multimodal construction grammar? *Linguistics Vanguard*, 3(s1).