Encoding Gesture in Multimodal Dialogue: Creating a Corpus of Multimodal AMR

Kenneth Lai*, **Richard Brutti***, **Lucia Donatelli**†, **James Pustejovsky***
*Brandeis University, †Vrije Universiteit Amsterdam
*Waltham, MA, USA, †Amsterdam, Netherlands
{klai12, brutti, jamesp}@brandeis.edu, l.e.donatelli@vu.nl

Abstract

Abstract Meaning Representation (AMR) is a general-purpose meaning representation that has become popular for its clear structure, ease of annotation and available corpora, and overall expressiveness. While AMR was designed to represent sentence meaning in English text, recent research has explored its adaptation to broader domains, including documents, dialogues, spatial information, cross-lingual tasks, and gesture. In this paper, we present an annotated corpus of multimodal (speech and gesture) AMR in a task-based setting. Our corpus is multilayered, containing temporal alignments to both the speech signal and to descriptions of gesture morphology. We also capture coreference relationships across modalities, enabling fine-grained analysis of how the semantics of gesture and natural language interact. We discuss challenges that arise when identifying cross-modal coreference and anaphora, as well as in creating and evaluating multimodal corpora in general. Although we find AMR's abstraction away from surface form (in both language and gesture) occasionally too coarse-grained to capture certain cross-modal interactions, we believe its flexibility allows for future work to fill in these gaps. Our corpus and annotation guidelines are available at https://github.com/klai12/encoding-gesture-multimodal-dialogue.

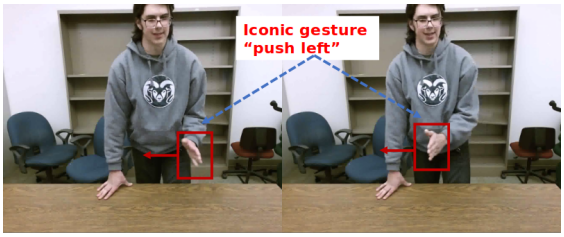
Keywords: Annotation, Dialogue, Gesture, Multimodal Interaction, AMR

1. Introduction

Determining the relationship between language and meaning has been an active research topic in linguistics, cognitive science, and artificial intelligence for decades. To this end, finding an adequate representation scheme to encode the meaning of linguistic expressions has been challenging. One representation that has become popular in recent years due to its clear structure and overall expressiveness is Abstract Meaning Representation (AMR) (Banarescu et al., 2013). As a general-purpose meaning representation for language, AMR bridges the gap between the subtleties of natural language and the explicit encoding required for computational understanding.

While AMR's design was primarily oriented towards capturing the meaning within English sentences in written text, AMR's architecture, ease of annotation, and the availability of comprehensive corpora have contributed to its widespread adoption in the research community. AMR's adaptability has also been tested across a range of domains beyond standalone sentences, encompassing broader linguistic structures such as full-length documents (O'Gorman et al., 2018), dialogues (Bonial et al., 2020), spatial information (Bonn et al., 2020), as well as cross-lingual tasks (Cai, 2022; Wein and Bonn, 2023) (Sec. 2).

Here, we explore further adaptation of AMR to a new domain: multimodal interaction through spoken language and gesture. Specifically, we create a corpus of speech and gesture meaning on top of the existing EGGNOG dataset (Wang et al., 2017), which annotates gesture morphology and intent in an English-instructed block-building task. We present a comprehensive annotation scheme with guidelines and accompanying annotated corpus of multimodal (speech and gesture) AMR (Sec. 3).



(1) "Push that block left."

```
(p / push-01
  :mode imperative
  :ARG0 (y / you)
  :ARG1 (b / block
          :mod (t / that))
  :ARG2 (l / left))
```

(2) Gesture for "push left"

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 (p / push-01
          :direction (l / left))
  :ARG2 (a / actor))
```

Figure 1: A multimodal communicative act, and its associated speech (1) and gesture (2) AMRs. Coreference relations are shown with colors.

LREC-COLING 2024, pages 5806–5818
20-25 May, 2024.

Our corpus consists of 21 one-minute long video segments and is multilayered: speech and gesture AMRs are temporally aligned with the speech signal and its transcription, as well as existing annotation for gesture morphology provided by EGGNOG (Sec. 4).

Figure 1 shows an example scenario in our data: a speaker says "Push that block left"; the AMR for that sentence is shown in PENMAN (Matthiessen and Bateman, 1991) notation. The speaker simultaneously moves his hand to the left; we represent the gesture's meaning using Gesture AMR (Brutti et al., 2022; Donatelli et al., 2022). Finally, we mark coreference relations across the two modalities; these are shown with colors (for simplicity, implicit roles in the gesture AMR are not shown). In the example, the word "push" in the speech denotes the same action as the pushing motion in the gesture; likewise, the word "left" and the direction of motion can be mapped to each other. Then, the implicit "you" in the imperative corresponds to the "actor" of the gesture.

Our work fills a gap in multimodal meaning representation, namely to link research on multimodal semantics (Sec. 2) to practical, application-oriented design as embodied in AMR. Additionally, our corpus allows for identification and analysis of cross-modal anaphora and coreference (Sec. 4, 5). Evaluation of our work both quantitatively and qualitatively shows our task to be both challenging yet useful: while multimodal AMR design is limited by the constraints of the AMR formalism and inter-annotator agreement is not as high as for other AMR corpora, this shows the need for continued exploration of this research space to better understand how modalities interact in conveying meaning. We discuss these aspects of our research, as well as potential avenues for integrating our work into broader representations of and systems for situated dialogue.

2. Related Work

2.1. AMR and Extensions

As previously mentioned, AMR is one of the more popular tools for representing the semantics of language, expressing the meaning of a sentence in terms of its predicate-argument structure (Banarescu et al., 2013). AMRs were designed to be both easy for humans to annotate and for computers to parse. AMR has been extended in several ways to represent additional aspects of language and communication.

For example, the Multi-sentence AMR (MS-AMR) corpus (O'Gorman et al., 2018) annotates existing sentential AMRs with information about coreference, implicit roles, and bridging relationships (Poesio et al., 1997). The scheme links individual AMRs using these relationships, presenting a representation of meaning across a document or discourse.

Dialogue-AMR (Bonial et al., 2020) introduces a detailed schema for representing illocutionary force in AMR; it is the most extensive of the annotation schemes that exist for AMR for spoken interaction (Bastianelli et al., 2014; Shen, 2018). Dialogue-AMR extends standard AMR with three extensions: (i) a taxonomy of speech acts (Searle, 1969; Bunt et al., 2012); (ii) annotations for tense and aspect to track task status and completion (Donatelli et al., 2018); and (iii) normalizing of propositional content to standard concepts for downstream operationalization.

Additionally, Uniform Meaning Representation (UMR), has been extended from AMR to accommodate cross-linguistic diversity, and support lexical and logical inference (Van Gysel et al., 2021) by incorporating aspect, scope, temporal and modal dependencies, as well as inter-sentential coreference, including co-reference with negation and quantification. As the amount of annotated UMR is currently relatively small, it has not been extended to gesture as of yet. There is work however, to extend UMR's schema for multimodal interactions (Lai et al., 2021).

2.2. Gesture and Multimodality

Gesture refers to the way people move their hands (and sometimes other body parts) when they speak and communicate information. Gestures can be classified according to their relation to speech, both in terms of their relative timing, as well as how gesture enhances or complements the speech content. *Co-speech* or *co-verbal* gestures, which co-occur with spoken words, are thought to contribute to meaning and discourse in the same way as lexical items (Kendon, 2004; McNeill, 2008), and can themselves project illocutionary propositions distinct from speech (Lücking and Ginzburg, 2020). Within the class of co-speech gestures, referential gestures (e.g., iconic, metaphoric, deictic) visually illustrate some aspect of the spoken utterance, while non-referential gestures (e.g., beat) align with important words and help structure the utterance and discourse. Meanwhile, *pro-speech* gestures (which fully replace spoken words) and *post-speech* gestures (which follow spoken expressions they modify) can also trigger various inferences (Schlenker, 2018). Finally, gesture can be analyzed for its contribution to dialogue structure: *interactive* gestures help manage turn-taking, indicate the next speaker, repair utterances, backchannel, and provide alignment between speakers (Bavelas et al., 2008; Lücking and Ginzburg, 2020; Lücking et al., 2021).

Formal work on integrating (co-speech) gesture into the semantics of discourse argues that the meaning of gesture is dependent on both its form and its links to accompanying linguistic context. For example, Lascarides and Stone (2009) argue that gesture and speech provide complementary information and together compose an integrated, overarching communicative act with a uniform force and consistent assignments of scope relationships. In such accounts, gesture is often thought to be underspecified or capable of carrying partial meaning in a way language cannot (Cassell et al., 2000). Language can thus serve to disambiguate gesture, especially iconic gesture (Lawler et al., 2017); this mirrors deictic gesture's disambiguating abilities with demonstratives in language (Lücking et al., 2006). Similarly, other approaches argue that utterance production is inherently multimodal, in which gesture complements speech in communicating linguistic and symbolic representations, retrieving words, and vying for optimal expressivity (Krauss et al., 2000; McNeill and Duncan, 2000; Kita and Özyürek, 2003; De Ruiter, 2004; Pustejovsky and Krishnaswamy, 2021).

A line of recent work has focused on automatically discretizing gestures into meaningful units similar to language tokens or word embeddings. Abzaliev et al. (2022) use contrastive pre-training to learn a joint embedding space that aligns language and gesture; the authors use this association to predict speaker native language and leverage gesture embeddings themselves to predict linguistic content. Inspired by distributional semantics, Vogel et al. (2023) adapt the idea of linguistic context vectors to gesture: gestures classified in the same semiotic types as those used in this paper (Sec. 3) are used as target items, and the authors explore the relative frequency of words in the contexts of gestures. Results show non-random interaction between gesture vectors and gesture type vectors. Finally, very recent work has fine-tuned large language models to take in discrete atomic motion elements represented as novel language tokens; this causal model can then produce real-time, semantically meaningful listener responses in the form of gesture (Ng et al., 2023).

There are several annotation schemes for gesture, some of which focus on its descriptive characteristics, such as hand shape, trajectory information, and location with respect to the body (Kong et al., 2015; Rohrer et al., 2020). Others focus on classifying gestures according to how they acquire their meaning; for example, referential gestures can be classified as deictic, iconic, metaphoric, or emblematic (Ekman and Friesen, 1969; McNeill, 1992; Mather, 2005). Some focus on the alignment of gesture and speech; however, typically these schema are primarily descriptive and do not encode gesture meaning independently (Kipp, 2001; Allwood et al., 2005; Kipp et al., 2007). Finally, the Behavior Markup Language (BML) describes gesture with respect to embodied conversational agents (Kopp et al., 2006). While BML can describe multiple forms of behavior (e.g., gesture, gaze, etc.), executing such actions requires an additional interpretation layer.

3. AMR for Gesture

In this paper, we adopt Gesture AMR (Brutti et al., 2022; Donatelli et al., 2022) to represent the semantics of gesture. We find that the flexibility of AMR is able to easily accommodate the structures in gestural expressions. Furthermore, by combining it with MS-AMR, we are able to provide not only links between the contents of the gesture and speech modalities, but also potentially allow for situated grounding to context (Sec. 5). While Gesture AMR does not currently consider non-referential beat or rhythmic gestures, because we are interested in gestures that carry their own meaning or intention similarly to speech, we find the focus on referential or content-bearing gestures to be sufficient for our current study.

Gesture AMRs follow a canonical template:

```
(g / [gesture]-GA
  :ARG0 (s / signaler)
  :ARG1 [content]
  :ARG2 (a / actor))
```

A gesture AMR is anchored by a gesture act (`GA`), where `[gesture]` can be `icon`, `deixis`, `emblem`, or `metaphor`. The `ARG0` represents the gesturer, `ARG1` contains the semantic content of the gesture, and `ARG2` represents the addressee. The overall form of a gesture AMR parallels that of a Dialogue-AMR, in which the top node represents a speech act (Bonial et al., 2020). In our corpus, the gestures are instructions or other communication on the part of one participant (the *signaler*) directed towards the other (the *actor*); see Sec. 4.1 for more details. We therefore use `(s / signaler)` and `(a / actor)` as the `ARG0` and `ARG2`, respectively.

Although the Gesture AMR specification includes metaphoric gestures, which show abstract properties of the concepts or ideas they denote, we did not observe any such gestures in our corpus. Because we focused on gestures in a task-based setting, we found that depictions of entities and events reflect their concrete properties, such as the shape of an object or the manner of an action. Each of the other three types of gesture acts is associated with a corresponding kind of semantic content, as described below.

Figure 2: Two examples of iconic gestures, denoting a block (left) and the number 3 (right).

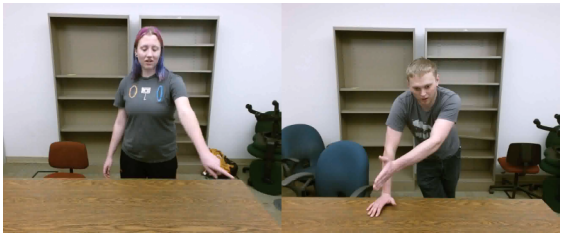


Figure 3: Two examples of deictic gestures, both denoting locations.

Iconic Gesture. Iconic gestures, in general, describe objects or actions by depicting concrete properties thereof. For example, figure 2 (left) shows a signaler making a block shape with his hand. This can be represented in Gesture AMR as follows:

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 (b / block)
  :ARG2 (a / actor))
```

Action predicates are drawn from PropBank (Palmer et al., 2005), while objects are described with words, as in an English language AMR.

Another example of an iconic gesture, shown in the right side of Figure 2, denotes the number 3. This is represented in Gesture AMR as follows:

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 3
  :ARG2 (a / actor))
```

Although numbers are neither concrete objects nor actions, and because the meaning of the gesture is directly derived from its physical form, namely, holding up three fingers to represent the number 3, it is considered iconic.

Deictic Gesture. Deictic gestures denote objects or locations through pointing. For example, figure 3 shows two examples of signalers pointing to locations on the table. In each of these cases, the corresponding gesture AMR is as follows:

```
(d / deixis-GA
  :ARG0 (s / signaler)
  :ARG1 (l / location)
  :ARG2 (a / actor))
```

Figure 4: Emblematic gesture, denoting positive acknowledgment.

Although the two gestures have different physical characteristics; for example, the two signalers are pointing in different directions, one with her finger, the other with his arm, etc.; their gesture AMRs are identical. To simplify the annotation process, we do not require annotators to include additional details, e.g., Cartesian coordinates of locations, as in Spatial AMR (Bonn et al., 2020).

Emblematic Gesture. Emblematic gestures have a conventional meaning agreed upon by members of some community, rather than one directly related to its form. Figure 4 shows a signaler making a "thumbs up" gesture, commonly used in English-speaking countries to express positive acknowledgment. In Gesture AMR, this is represented as:

```
(e / emblem-GA
  :ARG0 (s / signaler)
  :ARG1 (y / yes)
  :ARG2 (a / actor))
```

We note that the participants in our corpus (as well as our annotators) are English speakers in an American university setting. However, work is being done to explore how Gesture AMR can be applied in different cultural contexts, e.g., Arapaho speakers in conversation or monologues (Bonn et al., 2024).

Gesture with Multiple Meaning Components. Some gestures may contain multiple components to their meanings; these can be of the same gesture act type, or different ones. Figure 5 shows a signaler simultaneously outlining a square with his hands, and moving them slowly towards the table. The hand orientation is interpreted as an icon of a block; the motion is also interpreted as an icon of downward movement. Gesture AMR considers both components of the meaning, and incorporates them into a `(g / gesture-unit)` as follows:

Figure 5: Gesture with two iconic elements ("block" and "move down").

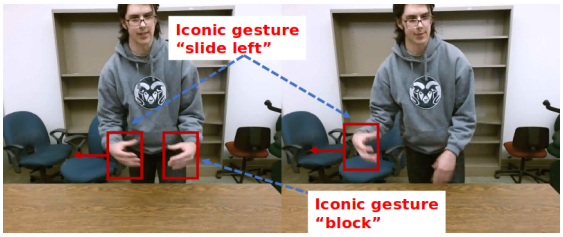


Figure 6: Two simultaneous gestures.

```
(g / gesture-unit
  :op1 (i / icon-GA
          :ARG0 (s / signaler)
          :ARG1 (b / block)
          :ARG2 (a / actor))
  :op2 (i2 / icon-GA
          :ARG0 s
          :ARG1 (m / move-01
                   :direction (d / down))
          :ARG2 a))
```

Coordinated or Simultaneous Gestures. In contrast to gesture units, which are considered single gestures, there are also instances of multiple independent gestures that occur simultaneously. An example is shown in Figure 6: a signaler makes a block shape with one hand, while moving the other hand leftward. Because the two gestures are separable (each gesture, done on its own with one hand, has a well-defined meaning), they are connected them under an `(a / and)` node as follows:

```
(a / and
  :op1 (i / icon-GA
          :ARG0 (s / signaler)
          :ARG1 (s2 / slide-01
                   :direction (l / left))
          :ARG2 (a2 / actor))
  :op2 (i2 / icon-GA
          :ARG0 s
          :ARG1 (b / block)
          :ARG2 a2))
```

4. Corpus of Multimodal AMR

4.1. Data Description

We used the EGGNOG corpus (Wang et al., 2017) as the base for our annotations. EGGNOG contains 360 videos (with a total length of eight hours) of pairs of participants working together on a shared task. One participant (the signaler) gives instructions to the other (the actor) for how to build a structure out of wooden blocks. Participants were English speakers between the ages of 19 and 64, recruited from a university setting.

For each video, the EGGNOG corpus includes time-stamped gesture labels, relevant part of the body, physical poses or motion (e.g., "body:still", "head:rotate"), and intent (e.g., "stack", "slide left").

We selected 21 videos, each around one minute long (23 minutes in total), in which the signaler was allowed to use both language and gesture to communicate. We then created speech transcripts for the videos; we used the Coqui speech-to-text toolkit with the English STT v1.0.0-huge-vocab model (Coqui, 2021), and manually corrected the output. The combination of EGGNOG's annotations, along our contribution of speech, gesture, and multi-sentence AMRs result in a rich, multilayered dataset for the exploration of in-context communication.

4.2. Annotation Methodology

Speech and Gesture AMR. First, for each video, given both types of EGGNOG gesture labels, the speech transcript, and the video itself, annotators were asked to create AMRs for each gesture and spoken utterance. In this step, annotators considered each modality (language and gesture) separately, and each individual gesture or utterance in isolation. We note that because the speech transcripts were created at the word level, annotators performed their own segmentation of the speech into utterances. Because the actors were mostly moving blocks, with little or no other communication, we annotated signaler gestures and utterances only. Annotation was done in ELAN (Wittenburg et al., 2006), with separate tracks for speech and gesture AMRs. An example of the ELAN annotation environment is shown in Figure 7.

We had a team of 5 annotators, made up of advanced undergraduate and master's students at an American university. We note that our annotators were drawn from a similar population as the participants in the EGGNOG corpus, who were also mostly students/young people at an American university. They were trained first on AMR, using the online guidelines[1], then on our Gesture AMR

[1] https://github.com/amrisi/amr-guidelines/blob/master/amr.md

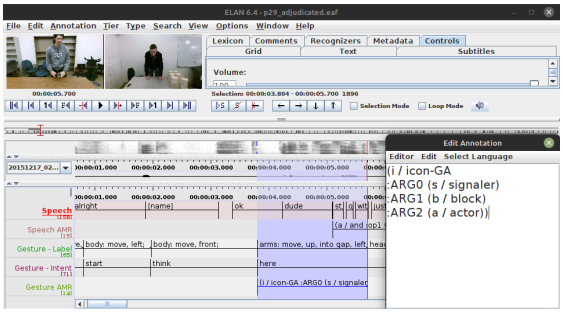


Figure 7: ELAN annotation environment.

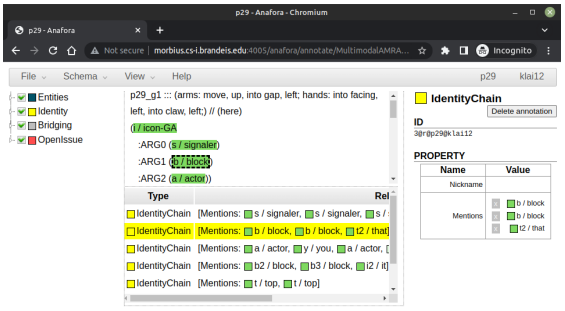


Figure 8: Anafora annotation tool.

schema. At least two annotators were assigned to each video: the first video was assigned to all annotators and used for training in our annotation scheme, while the other 20 were dually annotated. We held weekly meetings with our annotators, to answer questions, discuss issues with annotation, and refine our annotation guidelines. We (three of the authors, who are familiar with both standard and gesture AMR) then adjudicated the 21 videos (including the training video) to create a gold standard. Given the speech and gesture AMR annotations from the trained annotators, we decided which annotations were most appropriate given the data, correcting any annotations where necessary. While adjudication was done individually, we also held weekly meetings among ourselves, where we discussed and resolved any questions that arose during the adjudication process.

For each modality (speech, gesture) individually, we calculated inter-annotator agreement using SMATCH (Cai and Knight, 2013) and S^{2}MATCH (Opitz et al., 2020), across the 20 non-training videos. These scores measure the similarity between two annotators' AMRs: SMATCH measures the degree of overlap between two semantic feature structures; specifically, it measures the number of matching triples between two AMRs, given a mapping function m from the variables of one to those of the other. S^{2}MATCH, rather than requiring exact matches between instance triples, allows for soft matches. For the concepts in each triple, we extract their 100-dimensional GloVe vectors (Pennington et al., 2014) and compute their cosine similarity, with a minimum threshold of 0.5 below which the concepts are too dissimilar to match. In an equation, for two triples $t = (a, \text{:instance}, x)$ and $t' = (m(a), \text{:instance}, y)$, we compute a soft match score

$$f(t, t') = \begin{cases} \cos(x, y), & \text{if } \cos(x, y) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Because annotators may create different numbers of AMRs for the same video (due to, e.g., differences in utterance segmentation, or which body movements count as content-bearing gestures), we first manually aligned the two annotators' AMRs. Throughout the project, we found that annotating speech AMR required more effort than typical text AMR annotation, partially due to the questions of segmentation. Then we concatenated the AMRs (by embedding them under an `(a / and)` node, with `:op` numbers set by the manual alignment), creating, for each video, one speech and one gesture AMR for each annotator. We computed SMATCH and S^{2}MATCH using the SMATCH++ toolkit (Opitz, 2023), standardizing the AMRs by dereifying non-core relations and using integer linear programming to compute the optimal alignment between the AMRs. We report micro-averaged F1 scores and bootstrapped 95% confidence intervals.

Multimodal Multi-sentence AMR. Then, for each video, annotators marked coreference and certain bridging (specifically, *set-member* and *part-whole*) relations across the gold standard AMRs using MS-AMR. Relations could involve actions and objects within a single modality, or across both speech and gesture; some involved implicit roles not directly mentioned in the speech or depicted in a gesture. Annotation was done using Anafora (Chen and Styler, 2013), a web-based tool for coreference and temporal annotation. Figure 8 shows an example of the Anafora annotation tool. Again, at least two annotators were assigned to each video, with the first video used for training and the other 20 dually annotated. We continued to hold weekly meetings with our annotators to answer questions and provide guidance. Finally, we adjudicated the 21 videos and created a gold standard.

Similarly to O'Gorman et al. (2018), we calculated inter-annotator agreement by computing the CoNLL-2012 F1 score (Pradhan et al., 2014) on the coreference annotations in the 20 non-training videos. The CoNLL-2012 F1 score is the average of the MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and CEAF_e (Luo, 2005) metrics, which we computed using the reference implementation of Pradhan et al. (2014).

Gesture Act	Top-level	All
Icon	142	249
Deixis	63	129
Emblem	37	39
Gesture unit	50	-
Coordinated gestures	27	-

Table 1: Distribution of gesture act types.

Icon		Deixis		Emblem	
put-01	40	location	99	yes	19
block	31	block	24	ok	6
slide-01	17	left	2	no	5

Table 2: Most common semantic contents of gestures, by gesture act type.

4.3. Results

Our gold standard corpus contains 662 AMRs, of which 343 are speech and 319 are gesture AMRs. Among gesture AMRs, the distribution of gesture act types is shown in Table 1, both in terms of top-level nodes (including gesture units and coordinated/simultaneous gestures), and looking at all gesture act nodes (i.e., breaking gesture units and coordinated gestures into their individual components). Iconic gestures are the most common type overall, followed by deictic gestures. Interestingly, emblems were very rarely combined with other gesture components, almost always appearing alone.

The most common semantic contents (i.e., `ARG1`) for each gesture act type are shown in Table 2. Icons are roughly evenly divided between actions (e.g., put-01, slide-01, etc.) and objects (e.g., block, tower, etc.). The vast majority of deictic gestures denote locations, with some pointing to blocks, and very few with other content. Emblems were generally used for acknowledgement, both positive (yes, ok) and negative (no).

The EGGNOG task is very restricted in its domain, so most gesture patterns are grounded in the goal of the block arrangements. As discussed, the EGGNOG corpus contains labels describing gesture morphology (e.g., "RA: move, up"), which are continuous over the entire duration. After removing non-descriptive labels *still* and *unknown*, these physical descriptions cover approximately 60% of the corpus. We can therefore estimate that approximately 1/3 of the participant movements consist of beat or other non-referring gestures. When gesture AMRs overlapped with EGGNOG *still* and *unknown*, typically the signalers held gestures for extended durations while actors completed the actions indicated by the gestures.

We observe a wide variety of communication styles, both in terms of preferences for speech vs. gesture, as well as types of gestures performed. For example, the participant with the most spoken utterances (41) used the fewest gestures (6), while one of the participants with the fewest utterances (11) had the second-highest number of gestures (23, with 24 being the highest). One participant did not produce any purely iconic gestures at all, while another participant's gestures were exclusively iconic. This may be due to a range of factors, from individual personalities to native languages.

	SMATCH	S^2MATCH
Speech	48.9 (45.6-52.9)	64.8 (63.0-67.3)
Gesture	57.5 (47.5-65.2)	71.5 (61.5-77.4)

Table 3: Inter-annotator agreement (micro-averaged F1) scores for speech and gesture AMRs (95% confidence intervals in parentheses).

Our inter-annotator agreement scores for the speech and gesture AMRs are shown in Table 3. The SMATCH scores of 48.9 for speech and 57.5 for gesture are much lower than those for text-based AMRs, which are generally between 70-80 (Bonial et al., 2020). On the other hand, we see a considerable improvement when moving from SMATCH to S^2MATCH, with scores increasing to 64.8 for speech and 71.5 for gesture. These improvements are much larger than those observed by Opitz et al. (2020), who find S^2MATCH scores within 2% of SMATCH scores when testing parsers on a text-based corpus.

For gesture AMRs, we note that there is lexical variation in the words used to describe the semantic contents of gestures. In one example, a signaler moved their hands closer together; where one annotator used "closer", another used "together". Positive acknowledgement was variously annotated with "good", "great", or "yes". This is much more variation than would be expected for text-based AMR, since one can generally use the same words as are in the text.

Even when annotators agreed on particular lexemes, we occasionally observe variation in the specific inflected forms used. For example, some annotators used plural forms (e.g., "blocks", "corners", etc.) while others normalized them (following the general AMR guidelines) to their singular lemmas ("block", "corner"). We see this in both gesture and speech AMRs.

For some complex constructions, mainly in speech AMRs, we sometimes see annotators create "shortcuts" using hyphenated concepts. As an example, for "fourth row", one annotator wrote `(f / fourth-row)` instead of the complete AMR:

```
(r / row
  :ord (o / ordinal-entity
          :value 4))
```

In these cases, where annotators do not disagree on the basic meaning of the utterances or gestures, S^{2}MATCH allows for soft matches between the relevant concepts (e.g., "good" and "great", "block" and "blocks", "row" and "fourth-row", etc.). We therefore believe that S^{2}MATCH better reflects our annotators' agreement on the semantics of the speech and gesture in our corpus.

The multimodal multi-sentence part of our corpus contains 436 relations: 388 coreference (identity) chains, 28 set-member relations, and 20 part-whole relations. The 388 coreference chains have a total of 1,933 mentions, for an average length of 4.98; these cover approximately half of the 3,833 entities (including 699 implicit roles) in our 662 AMRs.

For inter-annotator agreement, we get a CoNLL-2012 F1 score of 60.46 (MUC: 71.58, B^3: 63.57, CEAF$_e$: 46.22). This is again lower than the 69.86 score O'Gorman et al. (2018) report for text-based MS-AMR, although not excessively so. The CEAF$_e$ score in particular is much lower than the others. Luo (2005) notes that CEAF$_e$ "reflects the percentage of correctly recognized entities" (in our case, the percentage of entities shared by the two annotators). Luo (2005) also defines a mention-based score, CEAF$_m$, that "reflects the percentage of mentions that are in the correct entities"; this score is much higher for our corpus (67.4). This indicates that there was much more agreement on longer chains (which tend to be more obvious examples of coreference) than shorter ones.

5. Discussion and Future Work

Facilitating the annotation process. As we have alluded to, despite AMR's ease of annotation compared to other meaning representations, our multimodal AMR annotation still proved challenging. In particular, while standard AMR is annotated on predetermined sentences, our annotators were required to segment speech transcripts into utterances themselves. Furthermore, while the EGGNOG gesture annotations are time-stamped, annotators were not bound to the time stamps, and could separate or combine gestures as they saw fit. As a result, annotators often did not agree on the number of utterances or gestures in a video, with downstream consequences when it came time to annotate the AMRs. Reaching an agreement on segmentation first, before AMR annotation, would alleviate these issues. Once the speech has been segmented into utterances, an additional idea to make the annotation process easier would be to automatically parse the utterances into AMRs, and have annotators make corrections, rather than writing their own AMRs from scratch.

Temporal patterns in gesture-language interactions. When thinking about alignment between gesture and speech, one can consider both semantic alignment (relations between entity and event mentions across the two modalities) and temporal alignment (the relative placement of gestures and utterances in time). Although this paper focuses on semantic alignment, leaving a more thorough analysis of temporal alignment to future work, we can note that gesture AMRs overall cover 41.2% of the total video duration. Individual participants gesture between 7% and 74% of the total video time, with the majority between 30% and 50%. In all cases, participants spoke for a longer duration than they gestured; participants spoke between 46% and 96% of the video, with the majority between 60% and 84%. Speech and gesture commonly co-occur in the data; the overlap ranges from 7% to 69% across the videos. These percentages track closely with the gesture AMR proportions in the videos; in other words, when people gesture, they are talking.

Gesture sense meanings. As seen in Sec. 4.3, the results of our annotation follow predictable patterns given the nature of the task we focused on. Table 2 shows an imbalance in iconic and deictic gestures towards certain semantic content: 35% of iconic gestures and 97% of deictic gestures are composed of three gesture concepts. Here we find both a limitation and a strength of our annotation scheme. On one hand, many spatiotemporal properties of gesture are not captured by the current annotation scheme. Though we find 40 instances of `put-01` as an iconic gesture, these gestures are not identical and vary in their speed, manner, and spatial location in relation to the speaker, addressee, and denoted objects. Similarly, deictic gestures vary in the perspective the speaker assumes in making the gesture (see Figure 3). How detailed morphological differences in gesture map onto semantic meaning is often challenging to determine (Lascarides and Stone, 2009), so we choose to leave this for future work. In addition to these spatiotemporal properties, gestures may also differ in whether the speaker uses one or two hands, which may also have subtle implications for the intended or received meaning. However, though the current version of our guidelines does not include these properties, they are easily integrated on top of the current annotation with non-core AMR roles such as `:manner` and `:mode` to more faithfully represent gesture meaning.

Situated grounding and AMR. In this paper, we have addressed the issue of alignment across channels in a multimodal (speech and gesture) dialogue, but we have not addressed the ques-

tion of *situated grounding* of an expression to the local environment, be it objects in a situated context, an image, or a formal registration in a database. Situated grounding involves identifying presupposed entities, relations, and events, that are co-perceived and co-present to the agents in the interaction during the dialogue (Pustejovsky and Krishnaswamy, 2021). Although we have not yet incorporated non-communicative context into the meaning representations employed in this paper, we believe that they have the basic facility for situated grounding; i.e., explicit mention of object and situational state in context (Lai et al., 2021). For example, researchers have started to integrate Gesture AMR with annotations of other kinds of multimodal interactions (e.g., actions) to track objects and beliefs in a situated task (Khebour et al., 2024a,b).

Challenges in multimodal coreference & anaphora. While annotating cross-modal coreference with MS-AMR, we observed several instances where determining whether a relationship was *coreferent*, *anaphoric*, or neither, was difficult. Largely, these challenging cases resulted from the situated and dialogic nature of our data. For example, one signaler instructed the actor to "**turn** [the block] so it's parallel" (potential coreferent or anaphoric tokens in bold) alongside a turn gesture; the actor then followed up two instructions later with "do the exact same **thing** there." Other examples often related to location descriptions: the instruction "put one **on** top...**in** the middle" was used by one signaler to indicate the precise location of a block in combination with a deictic gesture indicating the location.

The event anaphora examples resemble the so-called *sloppy identity* effect (Ross, 1967), in which the same verb phrase can be interpreted with different arguments (Partee, 1975; Webber, 1978; Carnie, 2021). Such anaphora may thus be better classified as *coreference under transformation* (Rim et al., 2023) than strict coreference or anaphora. For all such cases, the signaler gave instructions, the actor attempted these instructions, and the signaler referred to the attempt (the effect of the instructions) with a pronominal or deictic reference. Thus, while the events themselves were not coreferent in the sense of being identical given their situated nature, the effects of the actions involved in the event were identical. In regards to location, while the intentions of the linguistic expressions are not synonymous, the extensional semantics of the expressions denote the same place and can thus be coreferential. For both kinds of examples, we ultimately decided to mark coreference relations between the relevant tokens.

Limitations of AMR. The design of AMR and the particular interface for annotating MS-AMR does not allow us to capture certain cross-modal coreference relationships. We observed this in two particular cases. First, several participants used head nods or hand gestures to indicate negation. As negation is indicated in standard AMR with the attribute relation `:polarity -`, where `-` is considered a constant rather than a variable, it is impossible to link this to any concept in speech or gesture AMR. Second, we often observed speakers indicating quantity with iconic gestures of holding up the equivalent number of fingers, as for the utterance "seven blocks". As with negation, since this is denoted with the attribute relation and constant `:quant 7`, we cannot link it to the content of the iconic gesture.

In addition to these specific limitations, AMR's abstraction from surface syntax results in meaningful gaps in representing cross-modal interaction and semantics. As one of the first corpora to annotate AMR for speech instead of text, we faced challenges in annotating speech-specific phenomena such as pauses, disfluencies, and repetitions. Such phenomena play an important role in construing meaning in dialogue. Additionally, the lack of token alignment and inability of AMR to recreate surface structure is difficult when annotating alignment between modalities that is temporally determined and dependent. For example, we noticed variation in how long speakers held their gestures, as well as the specific timing between words contained in a single AMR and separate gestures. Though we attempted to capture this in our ELAN annotation interface, some subtle details could not be captured.

6. Conclusion

We present an annotated multi-layered corpus of speech and gesture AMR aligned temporally with speech signals and semantically through multi-sentence AMR to capture cross-modal coreference. Evaluation of our corpus both quantitatively and qualitatively shows multimodal meaning representation to be a challenging yet promising line of research to further both theoretical analysis of multimodal interaction and practical implementation of such representation schemes. Given the strictly defined domain of our corpus, we identify the need to develop even more fine-grained semantic representations of gesture, particularly as they are conveyed spatiotemporally. AMR proves to be a flexible representation scheme adaptable to this domain. Future work can continue to explore how to embed such representations in broader schema of situated grounding.

7. Acknowledgements

We would like to thank Dean Cahill, Gaby Dinh, Hayden McCormick, Ryan Partlan, Shiyi Shen, Christopher Tam, Alicia Tu, and Tali Tukachinsky for their assistance with this research. We would also like to thank the three anonymous reviewers for their detailed comments and suggestions.

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805 to James Pustejovsky, and in particular by an iSAT Trainee Grant funded by DRL 2019805 to Kenneth Lai, Richard Brutti, and Lucia Donatelli. The opinions expressed are those of the authors and do not represent views of the NSF.

8. Bibliographical References

Artem Abzaliev, Andrew Owens, and Rada Mihalcea. 2022. Towards understanding the relation between gestures and language. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5507–5520, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jens Allwood, Loredana Cerrato, Laila Dybkjaer, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. 2005. The MUMIN multimodal coding scheme. *NorFA yearbook*, 2005:129–157.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566, Granada.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Luca Iocchi, Roberto Basili, and Daniele Nardi. 2014. HuRIC: a human robot interaction corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4519–4526, Reykjavik, Iceland. European Language Resources Association (ELRA).

Janet Bavelas, Jennifer Gerwing, Chantelle Sutton, and Danielle Prevost. 2008. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2):495–520.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 684–695, Marseille, France. European Language Resources Association.

Julia Bonn, Matthew Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benét Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building an infrastructure for Uniform Meaning Representations. In *Proceedings of LREC-COLING 2024*.

Julia Bonn, Martha Palmer, Zheng Cai, and Kristin Wright-Bettner. 2020. Spatial AMR: Expanded spatial annotation in the context of a grounded Minecraft corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4883–4892, Marseille, France. European Language Resources Association.

Richard Brutti, Lucia Donatelli, Kenneth Lai, and James Pustejovsky. 2022. Abstract Meaning Representation for gesture. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1576–1583, Marseille, France. European Language Resources Association.

Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David Traum. 2012. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 430–437, Istanbul, Turkey. European Language Resources Association (ELRA).

Deng Cai. 2022. *Semantic Transformations across Natural Languages and Abstract Meaning Representation*. Ph.D. thesis, The Chinese University of Hong Kong. Copyright - Copyright ProQuest Dissertations Publishing 2022; Last updated - 2023-03-08.

Shu Cai and Kevin Knight. 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. Association for Computational Linguistics.

Andrew Carnie. 2021. *Syntax: A generative introduction*. John Wiley & Sons.

Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth Churchill. 2000. *Embodied Conversational Agents*. MIT Press.

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, pages 14–19, Atlanta, Georgia. Association for Computational Linguistics.

Coqui. 2021. English stt v1.0.0. Technical Report STT-EN-1.0.0, Coqui, https://coqui.ai/models.

Jan Peter De Ruiter. 2004. On the primacy of language in multimodal communication. In *LREC 2004 Workshop on Multimodal Corpora*, pages 38–41. ELRA-European Language Resources Association (CD-ROM).

Lucia Donatelli, Kenneth Lai, Richard Brutti, and James Pustejovsky. 2022. Towards situated AMR: Creating a corpus of gesture AMR. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Health, Operations Management, and Design*, pages 293–312, Cham. Springer International Publishing.

Lucia Donatelli, Michael Regan, William Croft, and Nathan Schneider. 2018. Annotation of tense and aspect semantics for sentential AMR. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 96–108, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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

Adam Kendon. 2004. *Gesture: Visible Action as Utterance*. Cambridge University Press.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data*.

Ibrahim Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard Brutti, Christopher Tam, Jingxuan Tu, Benjamin Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024b. Common ground tracking in multimodal dialogue. In *Proceedings of LREC-COLING 2024*.

Michael Kipp. 2001. ANVIL - a generic annotation tool for multimodal dialogue. In *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, pages 1367–1370.

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

Sotaro Kita and Asli Özyürek. 2003. What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, 48(1):16–32.

Anthony Pak-Hin Kong, Sam-Po Law, Connie Ching-Yin Kwan, Christy Lai, and Vivian Lam. 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.

Stefan Kopp, Brigitte Krenn, Stacy Marsella, Andrew N. Marshall, Catherine Pelachaud, Hannes Pirker, Kristinn R. Thórisson, and Hannes Vilhjálmsson. 2006. Towards a common framework for multimodal generation: The behavior markup language. In *Intelligent Virtual Agents*, pages 205–217, Berlin, Heidelberg. Springer Berlin Heidelberg.

Robert M. Krauss, Yihsiu Chen, and Rebecca F. Gottesman. 2000. *Lexical gestures and lexical access: a process model*, Language Culture and Cognition, page 261–283. Cambridge University Press.

Kenneth Lai, Richard Brutti, Lucia Donatelli, and James Pustejovsky. 2021. Situated UMR for multimodal interactions. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue - Poster Abstracts*, Potsdam, Germany. SEMDIAL.

Alex Lascarides and Matthew Stone. 2009. A Formal Semantic Analysis of Gesture. *Journal of Semantics*, 26(4):393–449.

Insa Lawler, Florian Hahn, and Hannes Rieser. 2017. Gesture meaning needs speech meaning to denote - A case of speech-gesture meaning interaction. In *Proceedings of the Workshop on Formal Approaches to the Dynamics of Linguistic Interaction 2017 co-located within the European Summer School on Logic, Language and Information (ESSLLI 2017), Toulouse, France, July 17-21, 2017*, volume 1863 of *CEUR Workshop Proceedings*, pages 42–46. CEUR-WS.org.

Andy Lücking and Jonathan Ginzburg. 2020. Towards the score of communication. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Virtually at Brandeis, Waltham, New Jersey. SEMDIAL.

Andy Lücking, Jonathan Ginzburg, and Robin Cooper. 2021. Grammar in dialogue. In Robert D. Borsley & Jean-Pierre Koenig Stefan Müller, Anne Abeillé, editor, *Head-Driven Phrase Structure Grammar: The handbook*, 1201-1250 26. Language Science Press, Berlin.

Andy Lücking, Hannes Rieser, and Marc Staudacher. 2006. Multi-modal integration for gesture and speech. In *Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Potsdam, Germany. SEMDIAL.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Susan M. Mather. 2005. Ethnographic research on the use of visually based regulators for teachers and interpreters. *Attitudes, innuendo, and regulators*, pages 136–161.

Christian Matthiessen and John A. Bateman. 1991. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Burns & Oates.

David McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

David McNeill. 2008. *Gesture and Thought*. University of Chicago Press.

David McNeill and Susan D. Duncan. 2000. *Growth points in thinking-for-speaking*, Language Culture and Cognition, page 141–161. Cambridge University Press.

Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, and Shiry Ginosar. 2023. Can language models learn to listen? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10083–10093.

Tim O'Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 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. Association for Computational Linguistics.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.

Juri Opitz, Letitia Parcalabescu, and Anette Frank. 2020. AMR similarity metrics from principles. *Transactions of the Association for Computational Linguistics*, 8:522–538.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Barbara Partee. 1975. Montague grammar and transformational grammar. *Linguistic Inquiry*, 6(2):203–300.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Massimo Poesio, Renata Vieira, and Simone Teufel. 1997. Resolving bridging references in unrestricted text. In *Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*

(Volume 2: Short Papers), pages 30–35, Baltimore, Maryland. Association for Computational Linguistics.

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

Kyeongmin Rim, Jingxuan Tu, Bingyang Ye, Marc Verhagen, Eben Holderness, and James Pustejovsky. 2023. The coreference under transformation labeling dataset: Entity tracking in procedural texts using event models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12448–12460, Toronto, Canada. Association for Computational Linguistics.

Patrick Rohrer, Ingrid Vilà-Giménez, Júlia Florit-Pons, Núria Esteve-Gibert, Ada Ren, Stefanie Shattuck-Hufnagel, and Pilar Prieto. 2020. The MultiModal MultiDimensional (M3D) labelling scheme for the annotation of audiovisual corpora. In *Gesture and Speech in Interaction Conference (GeSpIn 2020)*.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. Ph.D. thesis, MIT.

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

John Rogers Searle. 1969. *Speech acts: An essay in the philosophy of language*. Cambridge University Press.

Hongyuan Shen. 2018. Semantic parsing in spoken language understanding using abstract meaning representation. Master's thesis, Brandeis University.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35(3):343–360.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*.

Carl Vogel, Maria Koutsombogera, Anaïs Claire Murat, Zohreh Khosrobeigi, and Xiaona Ma. 2023. Gestural linguistic context vectors encode gesture meaning. In *Gesture and Speech in Interaction Conference (GeSpIn 2023)*.

Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J. Ross Beveridge, Bruce A. Draper, and Jaime Ruiz. 2017. EGGNOG: A continuous, multi-modal data set of naturally occurring gestures with ground truth labels. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 414–421.

Bonnie Lynn Webber. 1978. *A Formal Approach to Discourse Anaphora*. Routledge.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the 4th International Workshop on Designing Meaning Representations*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).