Guidelines for the Annotation of Multimodal AMR

Last updated: March 2, 2022

In situated meaning, actions and objects are grounded in their environment, allowing participants in a dialogue to have a common understanding of the situation and how their actions affect it. Crucially, this grounding can occur through multiple modes of communication, not only through language, but also through gesture, gaze, etc. However, systems for representing meaning have thus far focused only on language, while frameworks that can account for multiple modalities do not describe their interpretation. In this project, we propose to create a corpus of multimodal Abstract Meaning Representation (AMR), incorporating aspects of meaning from both language and gesture.

For more details, one can read our grant proposal, located here.

1 Annotation Process

The annotation process is adapted from that used for the MUMIN multimodal coding scheme, used for annotation of multimodal communication in videos [1].

We envision a two-step annotation process.

1. For each video, given the list of gesture intents, the speech transcript, and the video itself, annotators will be asked to create AMRs for each gesture and spoken sentence.
2. Then, for each video, annotators will mark coreferent actions and objects across the gold standard AMRs using multi-sentence AMR (MS-AMR).

This version of the annotation guidelines only covers the first step.

1.1 Annotation Sessions

For each video, the annotation process can be divided into two "sessions":

1. Each annotator codes the video clip individually. The result is saved in a temporary file.
2. The annotators in each group get together and compare their annotations. Problems are noted. Adjustments to the codings are made to reduce differences, and results are saved in a second coding file.

1.2 Coding Passes

The following passes are recommended for an annotation session:

1. Watch entire video clip.
2. Create AMRs for each spoken sentence.
    - Annotators should refer to the AMR guidelines, located here, for guidance on how to annotate speech AMRs.
    - To create AMRs, an online AMR editor is available here. Alternatively, it may be easier to simply type them in a text editor and copy/paste them into our annotation environment.
        - A tutorial dataset is also available using the AMR editor: log in as guest → load workset at ISI → tutorial.
3. Create AMRs for each gesture.
    - Please see the next section for details of how to annotate gesture AMR.

Since understanding of phenomena and annotation tags usually changes as the coding proceeds, these passes should be gone through several times to ensure internal consistency.

1.3 Data

We will be annotating the EGGNOG (Elicited Giant Gallery of Naturally Occurring Gestures) data set [3]. The EGGNOG corpus contains 360 videos of participants tasked with directing another participant to build a structure of wooden blocks. Specifically, we will use the 192 videos in which the participants could use both language and gesture to communicate (in the others, participants could only use gesture). Gestures are labeled both in terms of physical pose or motion (e.g. body:still, head:rotate) and intent (e.g. stack, slide left). Each video also comes with a transcript of the speech used, if present. Of the 192 EGGNOG videos, we will annotate as many as funding will allow.

1.4 Annotation Environment

We will be using ELAN as our coding tool [4]. ELAN documentation, including the manual and a short how-to guide, is available here.

1.4.1 How to create an annotation in ELAN

To create an annotation, annotators should first define a selection, i.e., a time interval.

- Annotators are strongly encouraged to align their Gesture AMR annotations with the existing Gesture - Intent and/or Gesture - Label annotations. Clicking on an existing annotation will create a selection with the same begin and end time.
  - To create a selection spanning multiple annotations, one can Ctrl+Alt+click (or Command+Option+click) on the annotations.

Alternative ways of making selections are described on page 3 of the ELAN how-to guide here.

Given a selection, one can create an annotation by double clicking in the area where the selection and the Gesture AMR tier intersect. An AMR can then be typed or (preferably) copy/pasted into the annotation text editor.

- If typing an AMR directly in ELAN, the annotation editor can be detached/re-attached by Shift-Enter.

Changes should be committed by Ctrl+Enter.

For more details, please see pages 4-5 of the ELAN how-to guide.

2 Gesture AMR

We are interested in the following types of gestures. Descriptions are copied from [2].

1. Iconic gesture: refers to gestures that model the shape of an object or the motion of an action. For example, a speaker arcs his/her fist to form a cup and drink from it when saying 'I drink from a cup'.
2. Deictic gesture: refers to familiar pointing, indicating objects in conversational space. For example, when a speaker says 'I walked up the stairs', he/she points upward.

3. Emblem: gestures with standard properties and language-like features. For example, the OK sign has a culturally agreed upon meaning; and it is necessary to place the thumb and index finger together in order to form the sign.

Each type of gesture has a canonical AMR template:

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

Depending on the type of gesture, `[gesture]` should be `icon`, `deixis`, or `emblem`. `ARG0` and `ARG2` correspond to the gesturer and addressee; for this project, we will use `(s / signaler)` and `(a / actor)`, respectively. `ARG1`, the semantic `content` of the gesture, varies by gesture type, as described below.

2.1 Iconic Gesture

In an iconic gesture, the argument in `:ARG1` should represent the object or action being modeled. For example, the gesture in this video should be annotated as:

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 (s2 / slide-01
          :direction (f / forward))
  :ARG2 (a / actor))
```

Note that the above AMR is underspecified: there is no mention of the thing sliding (a block) or imperative mood. Only those aspects of meaning that can be explicitly seen in the gesture should be mentioned: "slide" and "forward". When in doubt, you can look at the EGGNOG intent annotation for clues; in this case, the intent annotation is "servo slide forward".

As another example, the gesture in this video should be annotated as:

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 1
  :ARG2 (a / actor))
```

Although "1" is neither an object nor an action, this is an example of a gesture that expresses "some semantic feature by similarity or homomorphism", namely, holding up one finger to represent

the number 1. These gestures should be annotated as iconic, as opposed to gestures "in which the relation between form and content is based on social convention", which should be annotated as emblems.

2.2 Deictic Gesture

In a deictic gesture, the argument in `:ARG1` should represent the object or location being pointed to. For example, the gesture in this video should be annotated as:

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

The gesture in this video should be annotated as:

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

You do not have to keep track of specific objects or mark specific locations at this time; simply writing `(b / block)` or `(l / location)` is fine. Differences in how one marks the location (e.g. by touching the table vs. pointing) also do not matter.

2.3 Emblem

In an emblem gesture, the argument in `:ARG1` should represent the conventional, culturally agreed-upon meaning of the gesture. For example, the gesture in this video should be annotated as:

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

2.4 Non-identifiable or Other

Some gestures may not fall clearly into one of the above categories; for example, metaphoric gestures that describe abstract ideas. Other gestures may seem to lack any discernable meaning at all.

(Often, but not always, these correspond to EGGNOG intent annotations such as "Unknown", "wait", "think", "out of context", etc.) If you are unsure whether a gesture should be annotated, or think it is meaningful and should be annotated despite not falling into one of the above categories, feel free to make a note of it. But otherwise, you are not required to annotate these.

2.5 Multiple Components in a Single Gesture

Some gestures may contain multiple components of meaning. For example, the gesture in this video contains both an icon of a block and a location being denoted. In this case, both components should be annotated using the above guidelines, and connected using `(g / gesture-unit)`:

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

Note that this is a single AMR. After the `signaler` and `actor` are introduced in the AMR once, you can use `s` and `a` for subsequent mentions within that AMR.

As another example, this video contains a sequence of three gestures. (If you are unsure whether something should count as a single gesture or multiple gestures, you can again look at the EGGNOG intent annotation for clues.) Here, all three gestures should be annotated with both iconic (block) and deictic (location) components:

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

(g / gesture-unit
  :op1 (i / icon-GA
         :ARG0 (s / signaler)
         :ARG1 (b / block)
         :ARG2 (a / actor))
  :op2 (d / deixis-GA
         :ARG0 s
         :ARG1 (l / location)
         :ARG2 a))

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

Note that this is three separate AMRs. The `signaler` and `actor` should be introduced once per AMR.

3 Alignment Between Language and Gesture

This section will be completed in a future version of the annotation guidelines.

4 General Tips

This section will contain frequently asked questions, and other issues that arise during the annotation process.

4.1 Changing .avi Files to .mp4 Files

By default, ELAN is set up to use .avi files. Some annotators (using Mac OS) have reported issues with the .avi files. In these cases, annotators can switch to using .mp4 files as follows:

1. Open the ELAN file (cancel the prompts to open the media file), then go to Edit > Linked Files.
2. Highlight the first file, click Update, then select the corresponding .mp4 file.
3. Highlight the second file, click Update, then select the corresponding .mp4 file.
4. Click Apply.

4.2 Speech AMR

4.2.1 Predicates

When creating a speech AMR, annotators are advised to first identify the "predicative core" (generally, the main verb) of the sentence, and select an appropriate PropBank predicate. A list of PropBank predicates is located here.

4.2.2 Implicit Predicates

In cases where a predicate is implicit in the speech (e.g., [put] "one block on each side"), you should annotate the implicit predicate in the speech if and only if it does *not* appear in the gesture. In these cases, annotators should use `(i / implicit-predicate-00)`. The `ARG0` of the predicate should be the subject/agent (typically implicit `(y / you)`), and the `ARG1` should be the direct object/patient. Other arguments should be named, as non-core roles, being as specific as possible (e.g., preferring `destination` over `location`).

```
(i / implicit-predicate-00
  :mode imperative
  :ARG0 (y / you)
  :ARG1 (b / block
          :quant 1)
  :destination (o / on
                 :op1 (s / side
                        :quant (e / each))))
```

4.2.3 Implicit Arguments

By contrast, in cases where an *argument* is implicit in the speech (e.g., "put [a block] a bit, a bit right"), you should in general not annotate the implicit argument.

```
(p / put-01
  :mode imperative
  :ARG0 (y / you)
  :ARG2 (r / right
          :degree (b / bit)))
```

Exceptions include:

- By convention, in imperatives, annotate `:mode imperative` and `:ARG0 (y / you)`.
- If you need to add a role to an implicit argument (e.g., "space two [blocks] out a little less than a block length"), use `(i / implicit-role)`.

```
(s / space-01
  :mode imperative
  :ARG0 (y / you)
  :ARG1 (i / implicit-role
          :quant 2)
  :ARG2 (q / distance-quantity
          :unit (b / block)
          :ARG1-of (h / have-quant-91
                     :ARG2 1
                     :ARG3 (l / less
                             :mod (l / little)))))
```

4.2.4 Conjunctions

In cases where two clauses are not connected by an overt conjunction (e.g., "Take one block; put it on top of those two in the middle"), but temporally and/or semantically act as a single utterance, annotators may choose to connect the two clauses with `(a / and)`. Annotators should use their best judgment whether this applies or not.

By contrast, in cases where two or more clauses are connected by overt conjunctions, but temporally and/or semantically act as multiple utterances (e.g., "okay, put one block wherever you want...and, uh, one block, uh, ahead of it...and one block left of the first block...and..."), annotators may choose to annotate them as separate AMRs. Annotators can use the timestamps (i.e., whether there is an extended pause between the clauses) to guide their decision.

4.2.5 Prepositions

The AMR guidelines note that "Most prepositions that signal semantic frame elements are dropped in AMR ... But time and location prepositions are kept if they carry additional information, using AMR's `:opN`". That being said, when in doubt, be as specific as possible, i.e., keep the prepositions.

4.2.6 Disfluencies

Disfluencies, such as fillers (e.g., "uh") or repeated words (e.g., "first block, first block") should not be annotated (repeated words should only be annotated once).

4.3 Gesture AMR

4.3.1 Coordinated Gestures

Coordinated gestures (independent gestures that occur simultaneously) should be connected using `(a / and)`. For example, if a signaler makes a "put here" gesture with each hand, meaning "put a block here and put another block there", this should be annotated as:

```
(a / and
  :op1 (g / gesture-unit
         :op1 (d / deixis-GA
                :ARG0 (s / signaler)
                :ARG1 (l / location)
                :ARG2 (a2 / actor))
         :op2 (i / icon-GA
                :ARG0 s
                :ARG1 (p / put-01)
                :ARG2 a2))
  :op2 (g2 / gesture-unit
         :op1 (d2 / deixis-GA
                :ARG0 s
                :ARG1 (l2 / location)
                :ARG2 a2)
         :op2 (i2 / icon-GA
                :ARG0 s
                :ARG1 (p2 / put-01)
                :ARG2 a2)))
```

Note that annotators should not connect the two gestures using `(g / gesture-unit)`, which is used to connect multiple components within a single gesture. As a diagnostic, annotators should

consider whether the gestures are *separable*. In the above example, each "put here" gesture, done on its own with one hand, has a well-defined meaning.

4.3.2 Iterated Gestures

Iterated gestures (gestures done multiple times, where the number of times the gesture is repeated is not semantically meaningful) should be annotated once, using `:mode expressive` to mark the iteration. For example, if a signaler moves their hands back and forth multiple times, meaning "move closer", this should be annotated as:

```
(i / icon-GA
  :mode expressive
  :ARG0 (s / signaler)
  :ARG1 (m / move-01
          :ARG2 (c / close-10))
  :ARG2 (a / actor))
```

4.4 Non-task-based Speech/Gesture (Front/Back Matter)

In cases where there is speech or gesture outside of the task context (e.g., before or after the main content), you are not required to annotate these utterances or gestures.

References

[1] Jens Allwood, Loredana Cerrato, Laila Dybkjaer, Kristiina Jokinen, Costanza Navarretta, and Patrizia Paggio. The mumin multimodal coding scheme.

[2] Anthony Pak-Hin Kong, Sam-Po Law, Connie Ching-Yin Kwan, Christy Lai, and Vivian Lam. 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, 2015.

[3] Isaac Wang, Mohtadi Ben Fraj, Pradyumna Narayana, Dhruva Patil, Gururaj Mulay, Rahul Bangar, J Ross Beveridge, Bruce A Draper, and Jaime Ruiz. 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. IEEE, 2017.

[4] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: A professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, 2006.