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

Kenneth Lai[1], Richard Brutti[1], Lucia Donatelli[2], James Pustejovsky[1]

[1]Brandeis University, [2]Vrije Universiteit Amsterdam

LREC-COLING 2024

May 20-25, 2024

Brandeis UNIVERSITY

VU VRIJE UNIVERSITEIT AMSTERDAM

# Contributions

- We create a corpus of speech and gesture meaning in a task-based setting
  - Temporal alignment of meaning representations with speech signal and gesture morphology
  - Semantic alignment: cross-modal coreference and anaphora

- Our corpus allows for structured analysis of multimodal semantic phenomena, enabling downstream applications for multimodal dialogue understanding

- Corpus and annotation guidelines available at https://github.com/klai12/encoding-gesture-multimodal-dialogue

# Outline

- Introduction and Background
- Gesture AMR and Multi-sentence AMR
- Data and Annotation
- Results
- Discussion and Conclusion

# Gesture

- Idea: Humans naturally communicate using multiple modalities beyond language

- Gesture: how people move their hands or other body parts when they speak or communicate information

# Gesture

- Gestures can be classified according to:
  - Relation with speech: co-speech, post-speech, pro-speech
  - Function within discourse: referential, non-referential, interactive

- In this project, we focus on co-speech, referential gestures
  - Other kinds of gestures are future work

# Abstract Meaning Representation (AMR)

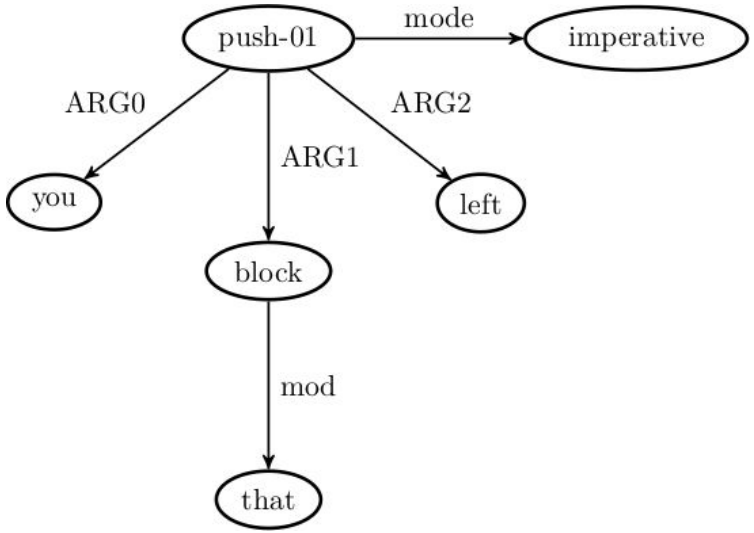
AMR is a graph-based meaning representation that expresses the meaning of a sentence in terms of its predicate-argument structure

- Relatively easy to annotate
- Readable by both humans and machines
- Existing community of researchers

# Abstract Meaning Representation (AMR)

- Push that block left.

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



# Gesture AMR

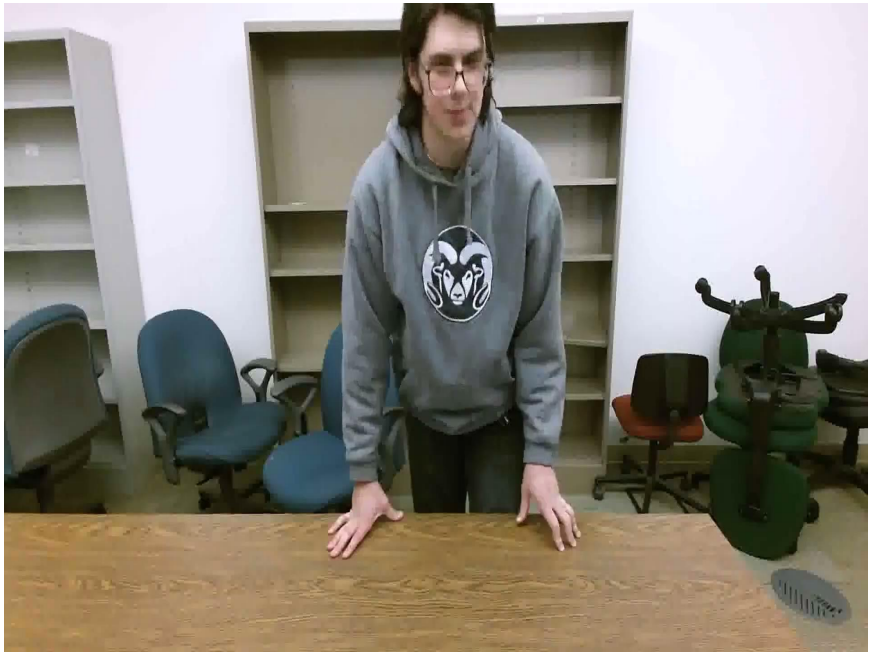
Gesture AMR can represent the dialogue participants, and objects and actions being referred to, while abstracting away from physical descriptions

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

- Root node: gesture act type
  - ARG0: initiator of the gesture
  - ARG1: semantic content of the gesture
  - ARG2: recipient of the gesture

# Gesture AMR

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



# Gesture AMR

- Basic gesture act relations
  - Iconic: describe concrete properties of objects or actions
  - Deictic: point to objects or locations
  - Emblematic: meanings set by convention
  - Metaphoric: describe abstract properties of concepts or ideas
    - No examples in our corpus
- Complex gestures
  - Gesture units: single gestures with multiple meaning components
  - Coordinated gestures: separate gestures that are simultaneous but independent

# Multi-sentence AMR

Multi-sentence AMR (MS-AMR) extends AMR with coreference, implicit roles, and bridging relations across sentences

- Push that block left.

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

- Gesture for “push left”

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 (p / push-01
    :ARG0 (i / implicit-role: pusher)
    :ARG1 (i2 / implicit-role: thing pushed)
    :direction (l / left))
  :ARG2 (a / actor))
```

# EGGNOG Corpus

We annotate on top of the EGGNOG (Elicited Giant Gallery of Naturally Occurring Gestures) corpus

- Shared task of arranging block structures
  - Signaler gives instructions
  - Actor interprets and builds structures

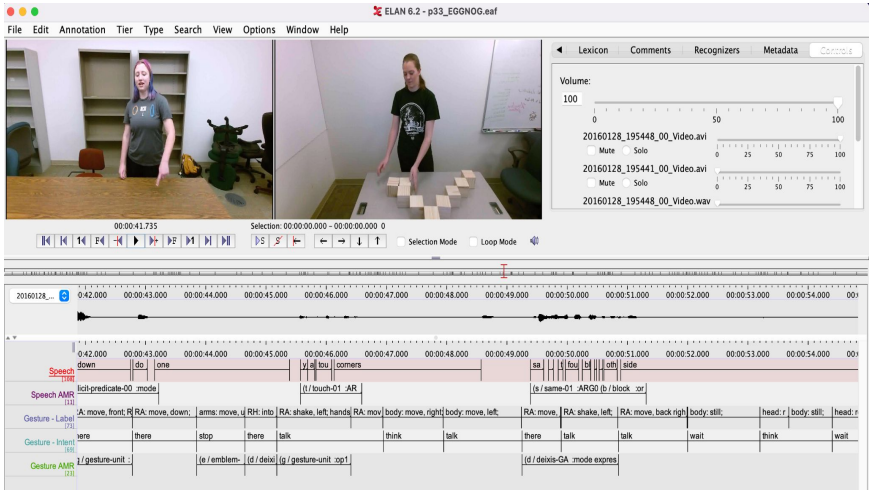
# EGGNOG Corpus

# EGGNOG Corpus

- Time-stamped annotations of participant gestures, for both the gesturer's inferred intent (such as "here"), as well a physical description of the movement (such as "arms: apart, left; hands: claw, down;")
- We created speech transcripts for the videos using the Coqui speech-to-text toolkit and manually corrected the output

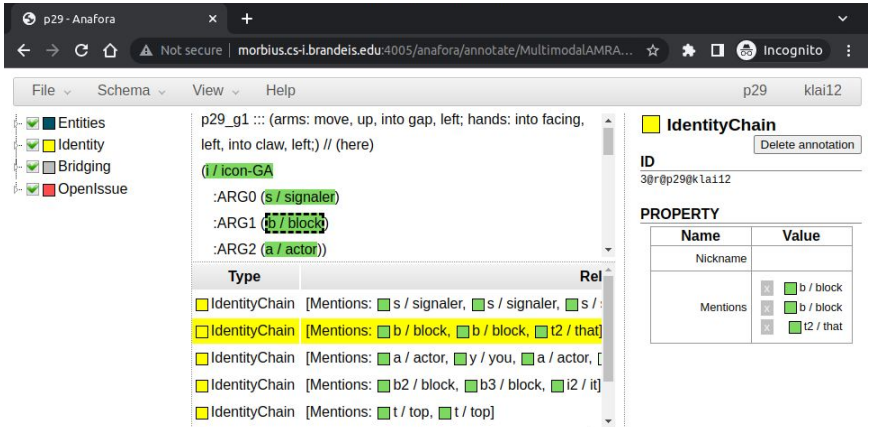
# Annotation Process

- We dually annotated 21 videos of task-based dialogues with speech and gesture AMRs
- We calculated inter-annotator agreement (IAA) using Smatch and $S^2$match



# Annotation Process

- We then marked coreference and bridging relations between actions and objects across modalities using MS-AMR
- We calculated IAA using CoNLL-2012 F1 score



# Results

Step 1: 343 speech AMRs and 319 gesture AMRs

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

# Results

Step 1: Most common semantic contents of gestures

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

# Results

Step 1: Inter-annotator agreement (micro-averaged F1) scores

| | Smatch | $S^2$match |
|---|---|---|
| Speech | 0.489 | 0.648 |
| Gesture | 0.575 | 0.715 |

# Results

Step 2: 436 relations: 388 coreference chains, 28 set-member relations, 20 part-whole relations

- ~5 mentions per coreference chain, covering ~50% of entities/events in the corpus

Inter-annotator agreement (CoNLL-2012 F1 score): 60.46

# Discussion

We observe a wide variety of communication styles

- Some participants spoke much more than they gestured; others, vice versa
- One participant's gestures were exclusively iconic, while another did not produce any purely iconic gestures at all

~⅓ of participant movements consist of beat or other non-referring gestures

# Discussion

Annotation is hard!

- Annotators were required to segment speech transcripts into utterances themselves, and could separate or combine gestures as they saw fit

Future work:

- Separate utterance/gesture segmentation from AMR annotation
- Automatic AMR parsing followed by manual correction

# Discussion

Our annotation scheme abstracts away from many spatiotemporal properties of gesture

- Iconic gestures vary in their speed, manner, spatial coordinates, etc.
- Deictic gestures vary in the perspective the speaker assumes, etc.

How detailed morphological differences in gesture map onto semantic meaning is often challenging to determine, so we choose to leave this for future work

- We can use roles such as `:manner` and `:mode` to more faithfully represent gesture morphology

# Conclusion

- We present an annotated corpus of gesture and speech AMR in a task-based setting, that enables analysis of how the semantics of gesture and natural language interact
- Although our current annotation scheme is rather coarse-grained, AMR proves to be a flexible representation scheme adaptable to this domain

# 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.

This research is supported by the NSF National AI Institute for Student-AI Teaming (iSAT), under grant DRL 2019805.

# Thank you!