# Situated UMR for Multimodal Interactions

Kenneth Lai[1], Richard Brutti[1], Lucia Donatelli[2], James Pustejovsky[1]
{klai12,brutti,jamesp}@brandeis.edu, donatelli@coli.uni-saarland.de

[1]Brandeis University, [2]Saarland University

## Introduction

- HCI & HRI involve communicating **intentions, goals, and attitudes through multiple modalities** beyond language, including gesture, gaze, and situational awareness.
- We outline desiderata for such a situated meaning representation and sketch a proposal based on **Abstract Meaning Representation (AMR)** (Banarescu et al., 2013).

## Background: AMR to UMR to SUMR

- AMR is a popular graph-based method to represent the logical meanings of sentences.

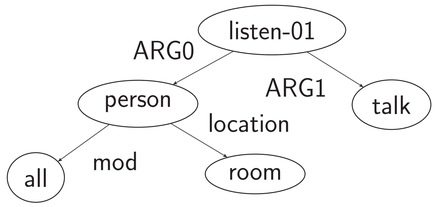


Figure 1:AMR for the English sentence "Everyone in the room listened to a talk."

- An extension of AMR, **Uniform Meaning Representation (UMR)** has been developed to be scalable, accomodate cross-linguistic diversity, and support lexical and logical inference (Van Gysel et al., 2021).
- UMR incorporates aspect, scope, temporal and modal dependencies, as well as inter-sentential coreference.

## Desiderata

- Accommodate the **structure** and **content** of the different modalities.
- Facilitate **alignment** and **binding** across modalities and to local environment (grounding).
- Possess basic facility for **situated grounding**; i.e., explicit mention of object and situational state in context.

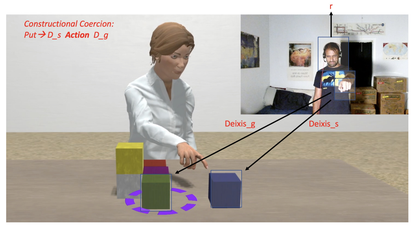


Figure 2:Multimodal interaction using language and gesture

## Common Ground in SUMR

- Components of a common ground structure (CGS) (Pustejovsky and Krishnaswamy, 2021; Krishnaswamy and Pustejovsky, 2021):
  - **A** the communicating agents;
  - **B**, the salient shared belief space;
  - **P**, the objects and relations that are jointly perceived in the environment; and
  - $\mathcal{E}$, the agents' joint embedding space.

## SUMR Example

*"that move there"*

```
(c / cgs
  :agent (a / agent)
  :agent (a2 / agent)
  :perception (b / block)
  :perception (l / location)
  :perception (l2 / location))

(s1c2 / command-00
  :ARG0 a1
  :ARG1 (c3 / communicative-act
          :gesture (g / gesture-unit
                     :op1 (d / deixis
                            :DIR (v / vector)
                            :OBJ b)
                     :op2 (a3 / action
                            :ACT (m / move-01)
                            :OBJ (i / implicit-role
                                   :op1 "moved")
                            :LOC (i2 / implicit-role
                                   :op1 "destination"))
                     :op3 (d2 / deixis
                            :DIR (v2 / vector))
                            :OBJ l))
          :speech (m2 / move-01
                    :mode imperative
                    :ARG0 (i3 / implicit-role
                            :op1 "mover")
                    :ARG1 (t / that)
                    :ARG2 (t2 / there))
  :ARG2 a2)

(s1 / sentence
  :coref ((a2 :same-entity i3)
          (b :same-entity i)
          (b :same-entity t)
          (l :same-entity i2)
          (l :same-entity t2)))
```

Figure 3:Example SUMR corresponding to the communicative act in Figure 4

## Intermodal Alignment

- The agents and perceived objects are listed in the CGS (**B** and $\mathcal{E}$ are omitted for brevity).
- For each communicative act, we have a sentence-level UMR representation with the gesture and speech modalities labeled. We assume the dialogue act annotation from Bonial et al. (2020).
- Document-level representation captures object coreference inherent in the discourse for all modalities (O'Gorman et al., 2018).

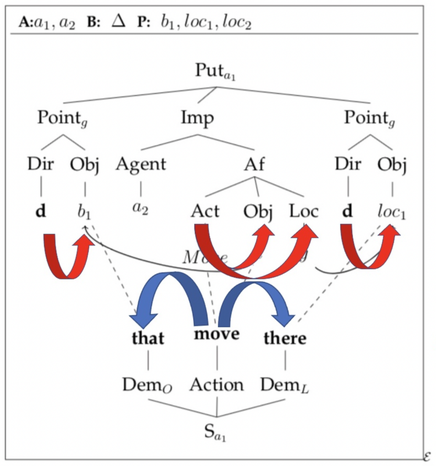


Figure 4:Intermodal alignment between linguistic and gesture dependency structures

## Discussion

- SUMR is a platform for multimodal situated dialogue annotation.
- SUMR is expressive at both utterance and dialogue levels, and easily accommodates dependency structures inherent in gestural expressions.
- Reentrancy facilitates the linking between modalities and situational grounding to contextual bindings.

## Open Questions and Future Work

- Is SUMR **expressive enough** to account for other modalities (e.g. gaze)? I.e., can we assume that structure and content across modalities is comparable?
- How applicable is SUMR outside of a **task-based** setting?
- How do we appropriately represent **alignment** between modalities and potential "emergent meaning" from such alignment?