Tutorial : Representation, Learning and Reasoning on Spatial Language for Downstream NLP Tasks
(Part 2)

Parisa Kordjamshidi, Michigan State University, USA, kordjams@msu.edu

Marie-Francine Moens, KU Leuven, Belgium, sien.moens@cs.kuleuven.be

James Pustejovsky, Brandeis University, USA, jamesp@cs.brandeis.edu

The 2020 Conference on Empirical Methods in Natural Language Processing

Nov 20th, 2020

Learning Reasoning

Heterogeneous

Table of Content

- Challenges and Motivating Applications
- Spatial Representation
- Spatial Information Extraction
- Spatial Reasoning
- Downstream tasks
  - Visual Question Answering
  - Navigation and Instruction Following
  - Dialogue Systems
  - Talking to Self-driving Cars

Dialogue, Context, and Situated Grounding

- Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication.
- Situated semantic grounding assumes shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.

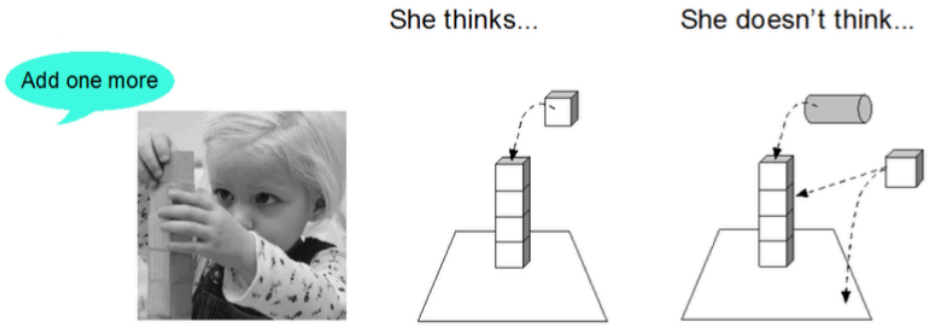
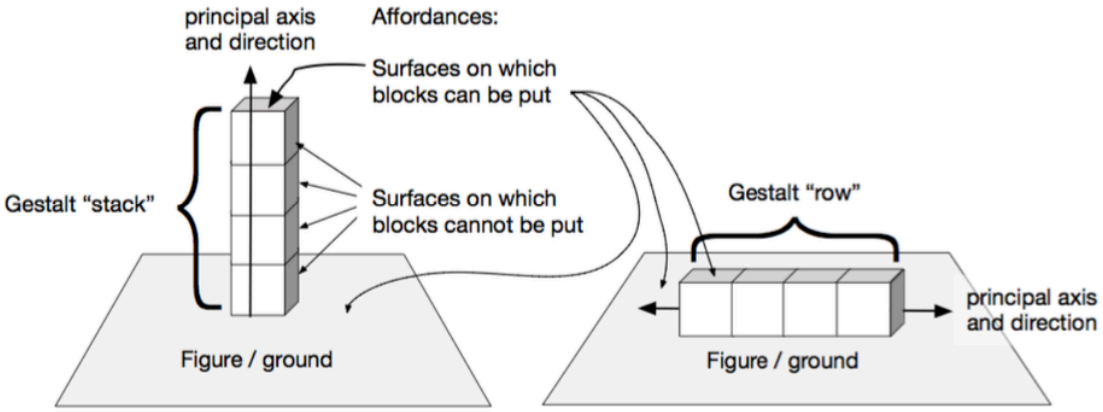


Figure 2. "Add one more" is ambiguous out of context, but given context it is remarkably precise.



DARPA Communicating with Computers BAA - Paul Cohen, PM (2015)

Semantics of ISO-Space

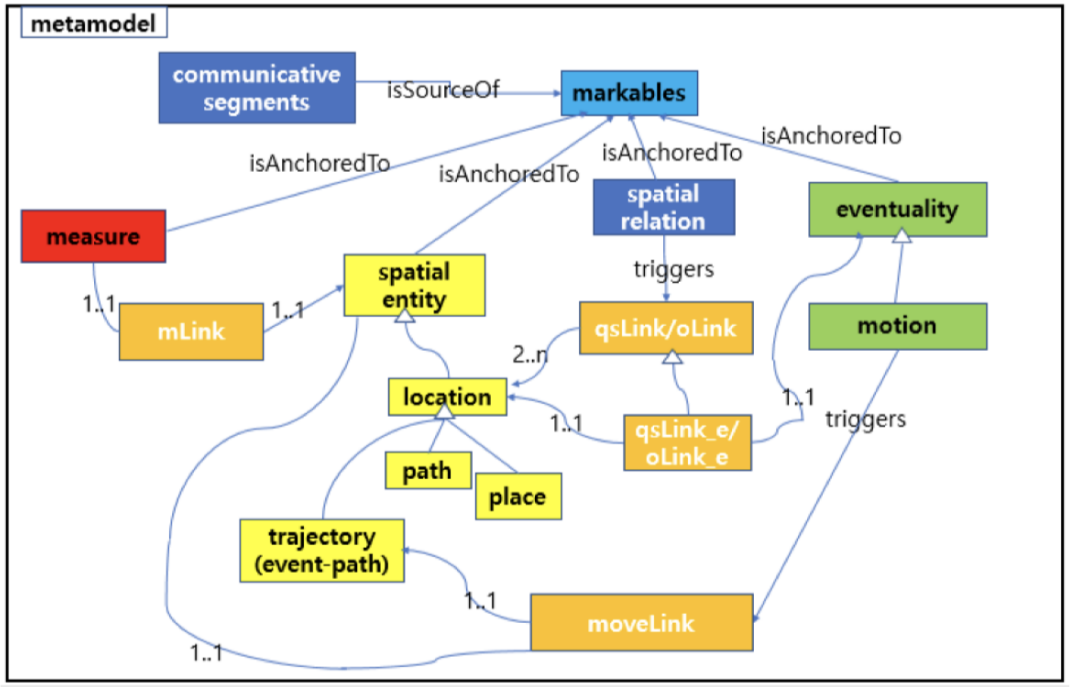


Figure: Metamodel of *ISO-Space*

Pustejovsky, J. (2017). ISO-Space: Annotating static and dynamic spatial information. In Handbook of linguistic annotation (pp. 989-1024). Springer, Dordrecht.

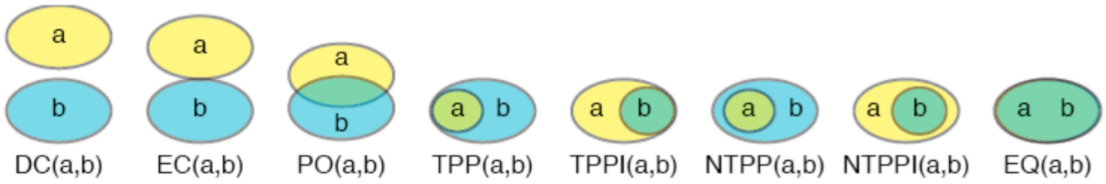
Unit Elements in ISO-Space

- Unit Elements
  - Spatial entities;
  - Eventualities;
  - Measures
- Relational Elements
  - QSLINK, qualitative spatial links;
  - OLINK, identifying orientation;
  - MOVELINK, specifying the figure and ground of a movement event;
  - MLINK, which identifies the metric of a region or distance between regions.

ISO-Space Qualitative Spatial Relations

Relation	Description
DC	Disconnected
EC	External Connection
PO	Partial Overlap
EQ	Equal
TPP	Tangential Proper Part
TPP_i	Inverse of TTP
NTTP	Non-Tangential Proper Part
$NTTP_i$	Inverse of NTTP

Table: RCC8 Relations.



Randell, D. A., Cui, Z., & Cohn, A. G. (1992). An interval logic for space based on "connection". In Proceedings of the 10th European conference on Artificial intelligence (pp. 394-398).

Randell, D. A., Cui, Z., & Cohn, A. G. (1992). A spatial logic based on regions and connection. KR, 92, 165-176.

Basic Types and Type Operations

We assume a model with the following basic types, corresponding generally to the elements in Figure 1.

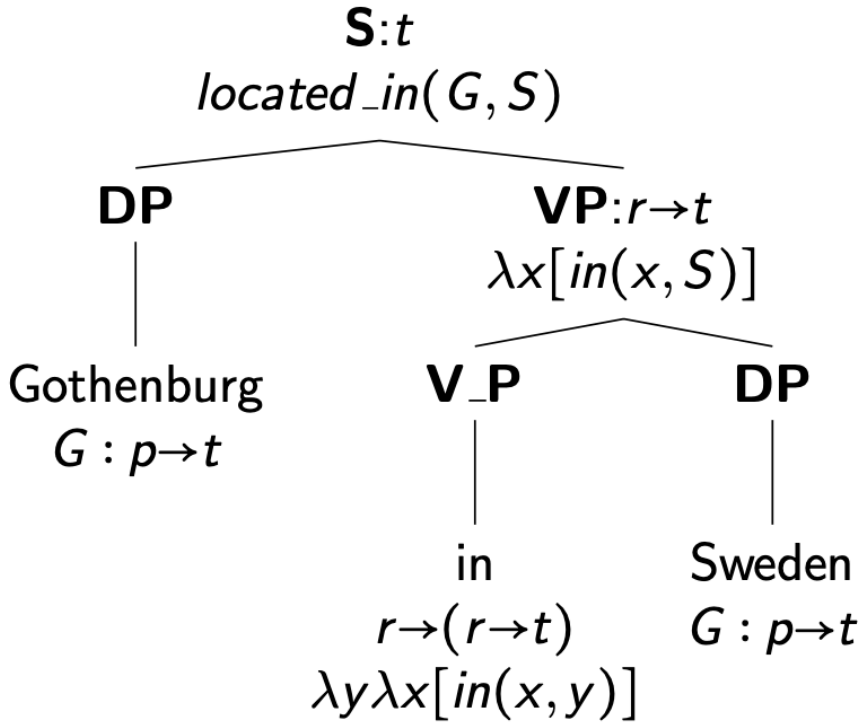
(2) a. e, the type of objects
b. i, the type of time points
c. p, the type of spatial points
d. ϵ, the type of events
e. m, the type of measures
f. t, the type of truth values.

Place and Spatial Entity

- The PLACE tag is used for annotating geolocations, such as *Germany* and *Boston*.
- In example (3), the qualitative spatial relation between the two entities is a relation between PLACEs. Both *Gothenburg* and *Sweden* are marked as PLACEs, which we will type as *regions*.
- A region, r, will be defined at a set of points, $p \rightarrow t$.[1]
- Further, a qualitative spatial mereotopological relation within RCC8 will be typed as a relation between regions: i.e., $QSLINK : r \rightarrow (r \rightarrow t)$.

(3) a. [**Gothenburg**$_{pl1}$] is [**in**$_{s1}$] [**Sweden**$_{pl2}$].
b. $[\![\text{Gothenburg}]\!] = G, \langle G{:}p \rightarrow t \rangle$
c. $[\![\text{Sweden}]\!] = S, \langle S{:}p \rightarrow t \rangle$
d. $[\![\text{in}]\!] = \lambda y \lambda x[in(x, y)], \langle \text{in}{:}r \rightarrow (r \rightarrow t) \rangle$
e. $\text{in}(G, S)$

Spatial Composition in ISO-Space



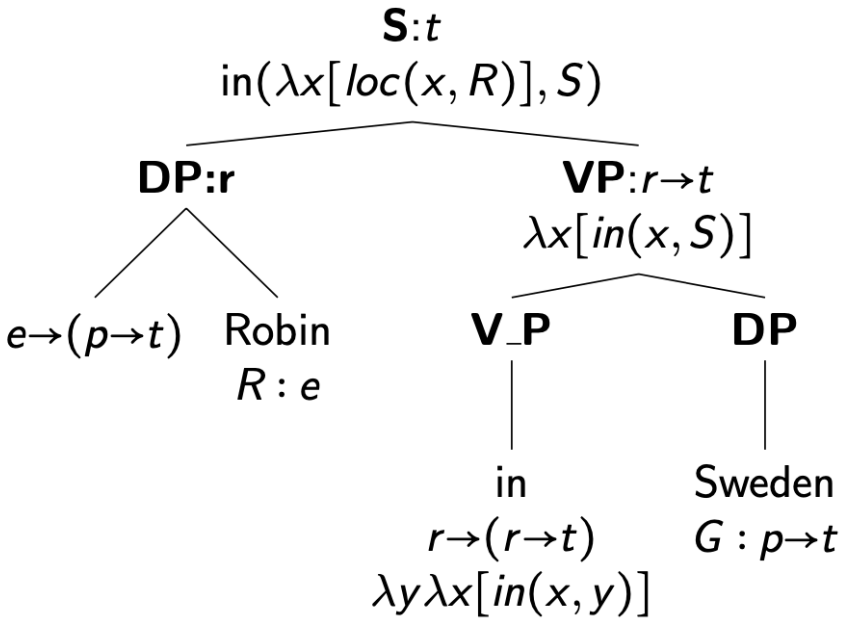
Kracht, M. (2002). On the semantics of locatives. *Linguistics and philosophy 25*(2), 157–232.

Eigenplace Function

- Humans and everyday objects carry a primary type of e, subtyped to SPATIAL_ENTITY.
- When they participate in spatial relations, there is a type coercion function, L, which returns the spatial region associated with that entity i.e., its location in space.
- This is the *eigenplace* for the entity (Klein, 1991): The type for this localization operator, L is: $e \rightarrow (p \rightarrow t)$.

(6) a. [**Robin**$_{sne1}$] is in [**Sweden**$_{pl1}$].
b. $[\![\text{Robin}]\!] = R, \langle R{:}e \rangle$
c. $[\![\text{Sweden}]\!] = S, \langle S{:}p \rightarrow t \rangle$
d. $[\![L(R)]\!] = \lambda x[loc(x, R)], \langle x{:}p, L{:}e \rightarrow (p \rightarrow t) \rangle$
e. $[\![\text{in}]\!] = \lambda y \lambda x[in(x, y)], \langle \text{in}{:}r \rightarrow (r \rightarrow t) \rangle$
f. $\text{in}(\lambda x[loc(x, R)], S)$

Eigenplace Coercion



Paths in ISO-Space

- We define a path as a subtype of locations (formally regions) that have the additional constraint of being directional, and are often construed as one-dimensional.
- Formally, paths have been analyzed as sequences of spaces Nam (1995) and sequences of vectors Zwarts and Winter (2000).
- Following Nam, let *int* be the type of the interval $[0, 1] \subset R$, and p be the type of a spatial point, as defined above.
- Then a *path*, π, will be that function $int \rightarrow p$, which indexes locations on the path to values from the interval [0,1].
- Similarly, if *vec* is the type of vectors, then a *vector-based path*, π_v, can be defined as the function $int \rightarrow vec$. That is, it indexes the vectors associated with the path to values from the interval [0,1].

Path Interpretaion

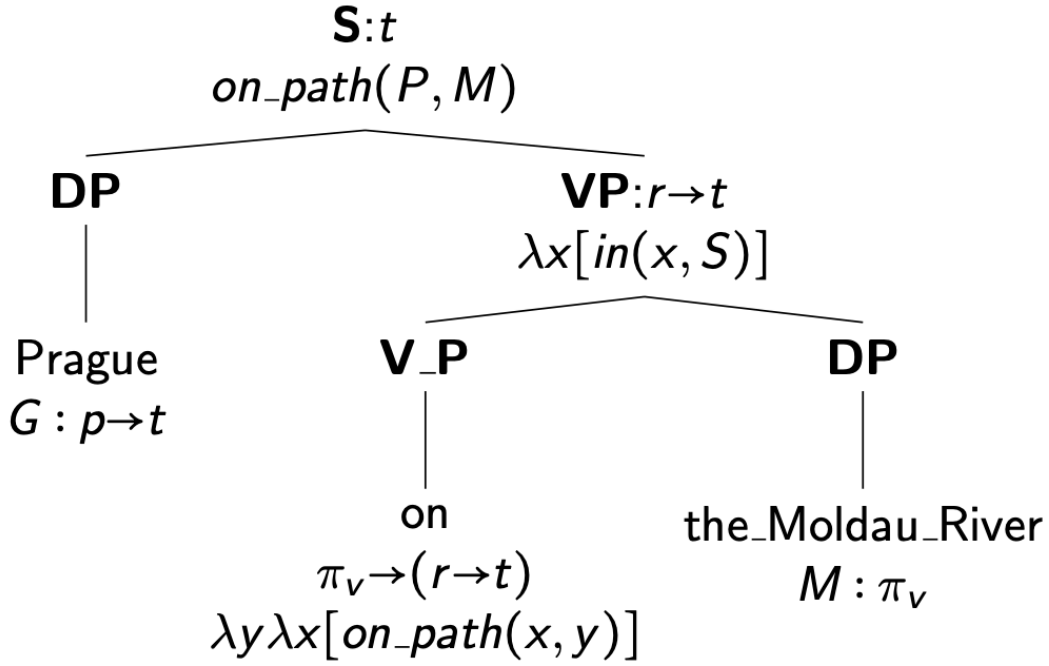
(7) a. [**Prague**$_{p/1}$] is on [**the Moldau River**$_{p1}$].
b. [**Boston**$_{p/1}$] is at the end of [**the Mass. Turnpike**$_{p1}$].

In these examples, the qualitative spatial relation introduced by the predication identifies a place as situated within (or on) a path. Hence, the preposition *on* which governs the path-PP, [$_{\mathrm{PP}}$ on [$_{\mathrm{NP}}$the Moldau River]], carries a more specific type than a general QSLINK relation, namely: $\pi_v \rightarrow (r \rightarrow t)$. The type derivation for (7a) is illustrated below.

(8) a. [**Prague**$_{p/1}$] is on [**the Moldau River**$_{p1}$].
b. $[\![\text{Prague}]\!] = P, \langle P{:}p \rightarrow t \rangle$
c. $[\![\text{the Moldau River}]\!] = M, \langle M{:}\pi_v \rangle$
d. $[\![\text{on}]\!] = \lambda y \lambda x [on\_path(x, y)], \langle \text{on\_path}{:}\pi_v \rightarrow (r \rightarrow t) \rangle$
e. $\text{on\_path}(P, M)$

Path Interpretation

(9)



Mani I., Pustejovsky J., Interpreting Motion: Grounded Representations for Spatial Language, Oxford University Press, 2012.

Path Interpretation

Formally, the expressions introducing end- and mid-point locations are acting as functions from paths to path positions: $\pi_v \rightarrow int$; e.g., given a path $\langle 3,4,5,2,1,8\rangle$, $end(\pi_v) = 8$.

(10) a. [**Boston**$_{pl1}$] is at the end of [**the Mass. Turnpike**$_{p1}$].
b. $[\![\text{Boston}]\!] = B$, $\langle B{:}p \rightarrow t\rangle$
c. $[\![\text{the Mass. Turnpike}]\!] = MT$, $\langle MT{:}\pi_v\rangle$
d. $[\![\text{end}]\!] = \lambda x[end\_of(x)]$, $\langle x{:}\pi_v, \text{end\_of}{:}\pi_v \rightarrow int\rangle$
e. $[\![\text{on}]\!] = \lambda y \lambda x[on\_path(x,y)]$, $\langle \text{on\_path}{:}\pi_v \rightarrow (r \rightarrow t)\rangle$
f. $\text{on\_path}(B, MT) \wedge end\_of(MT) = B$

As mentioned above, the eigenplace of a SPATIAL_ENTITY can be situated on a path by coercion: namely, L coerces *John* to his eigenplace, and then the spatial relation predication situates this region onto the path, π_v.

(12) a. [**John**$_{sne1}$] is on [**the road**$_{p1}$].
b. $[\![L(J)]\!] = \lambda x[loc(x,J)]$, $\langle x{:}p, L{:}e \rightarrow (p \rightarrow t)\rangle$

ISO-Space Annotation of Path

As sentence (7b) illustrates, end-points of paths can be explicitly mentioned in text. The ISO-Space annotated examples below demonstrate reference to both end-points and mid-points.

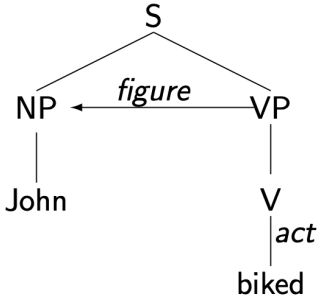
(12) a. . . . the [**railroad**$_{p1}$] between [**Boston**$_{pl1}$] and [**New York**$_{pl2}$] . . .
PATH (id=p1, beginID=pl1, endID=pl2, form=NOM)

b. John took the [**road**$_{p1}$] through [**Boston**$_{pl1}$].
PATH (id=p1, midIDs=pl1, form=NOM)

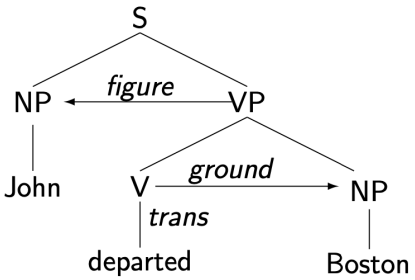
How Verbs Encode Paths

- Languages have two strategies to convey movement of an object through space
  - Path Motion: John arrived at home.
  - Manner Motion: John walked.

- Language encodes motion in Path and Manner constructions
  - Path: change with distinguished location
  - Manner: motion with no distinguished locations



· and paths may cc



How Verbs Make Paths

- In terms of their event structure, path-verbs are transitions while manner verbs are processes.
- In addition, path verbs are those predicates that *presuppose* a specific path for the moving object (the figure), along with a possible distinguished point or region on this path (the ground), which the figure is moving toward or away from.
- Manner verbs can be seen as *creating* a path as the motion event unfolds.
  - *Anchoring* relation @: indexes a proposition as holding at a specific event time: $\lambda p \lambda e[@(e, p)]$
  - Path-presupposing:
    $\lambda y \lambda x \lambda e \exists e_1, e_2, p[@(e, \text{arrive\_act}(x, p)) \wedge @(e_1, DC(x, y)) \wedge @(e_2, EC(x, y)) \wedge \text{end}(y, p) \wedge e = e_1 \circ e_2 \wedge e_1 \leq e_2 \wedge e_1 \preceq e \wedge e_2 \preceq e]$
  - Path-introducing:
    $\lambda p \lambda x \lambda e[\text{walk\_act}(e, x, p)]$

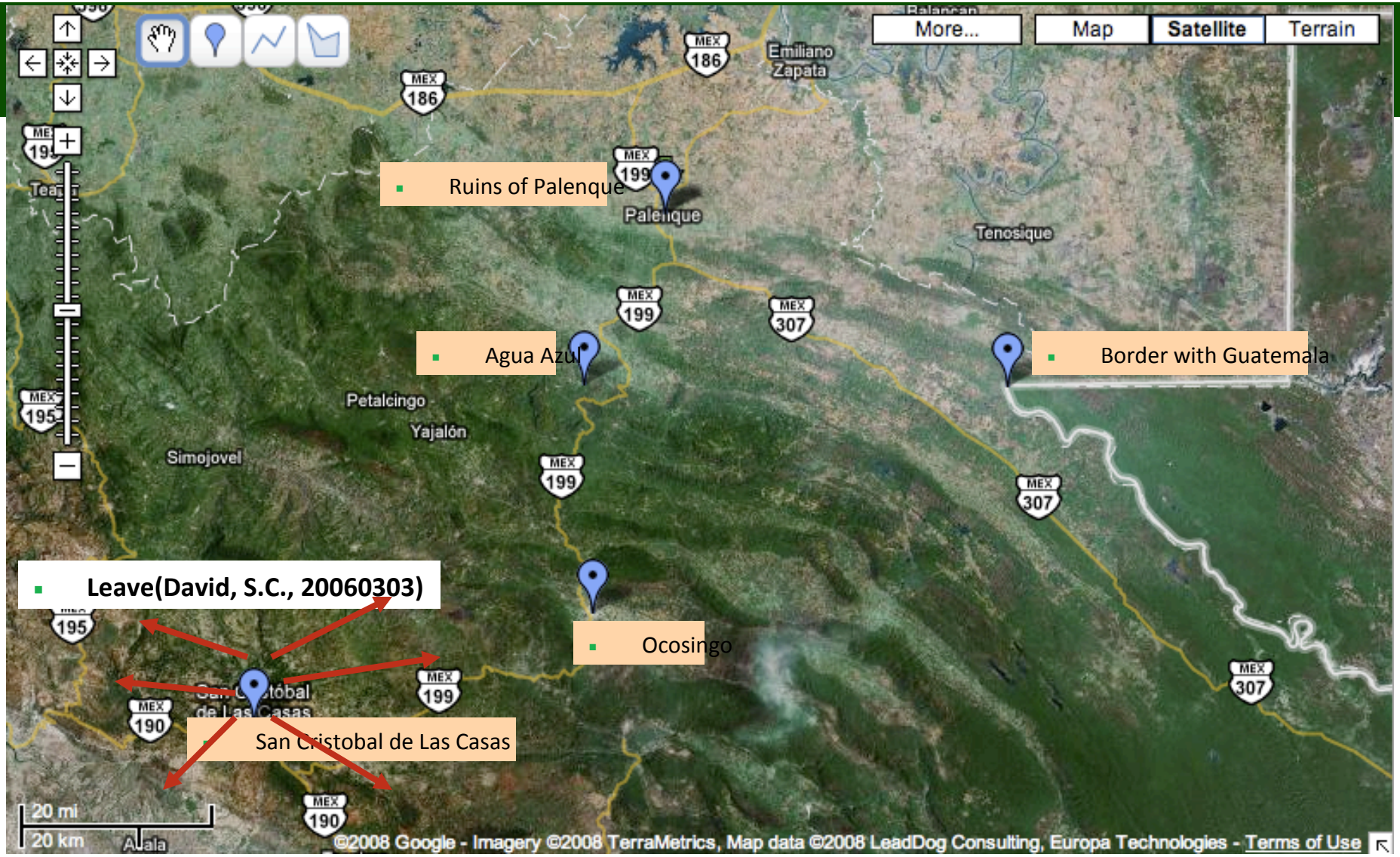
Reasoning with Paths

- The concrete syntax of ISO-Space, as deployed over a natural language example, receives an intermediate semantic interpretation, which can then be subsequently interpreted in a model.
- The semantics of ISO-Space validates each of the annotation structures by mapping it into a semantic form and then interpreting it model-theoretically.
- In an XML-based concrete syntax, the two elements `<eventPath>` and `<moveLink>` are implemented each with a list of attribute-value specifications.

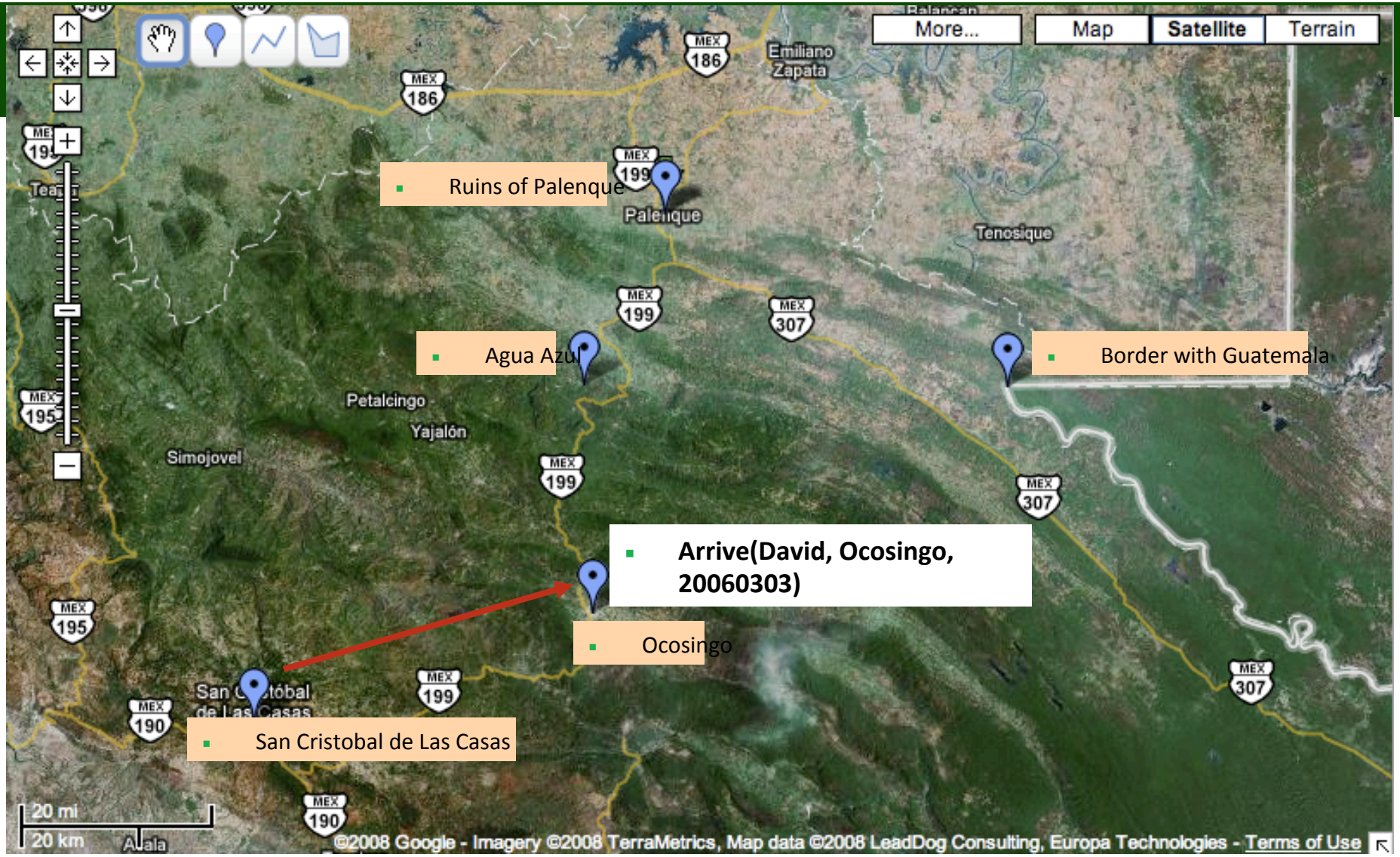
Temporally Anchored Paths in Text

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.

- Pustejovsky, J., & Moszkowicz, J. L. (2011). The qualitative spatial dynamics of motion in language. Spatial Cognition & Computation, 11(1), 15-44.
- Mani I., Pustejovsky J., Interpreting Motion: Grounded Representations for Spatial Language, Oxford University Press, 2012.
- Pustejovsky, J., Moszkowicz, J. L., & Verhagen, M. (2011, January). ISO-Space: The annotation of spatial information in language. In Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation (Vol. 6, pp. 1-9).



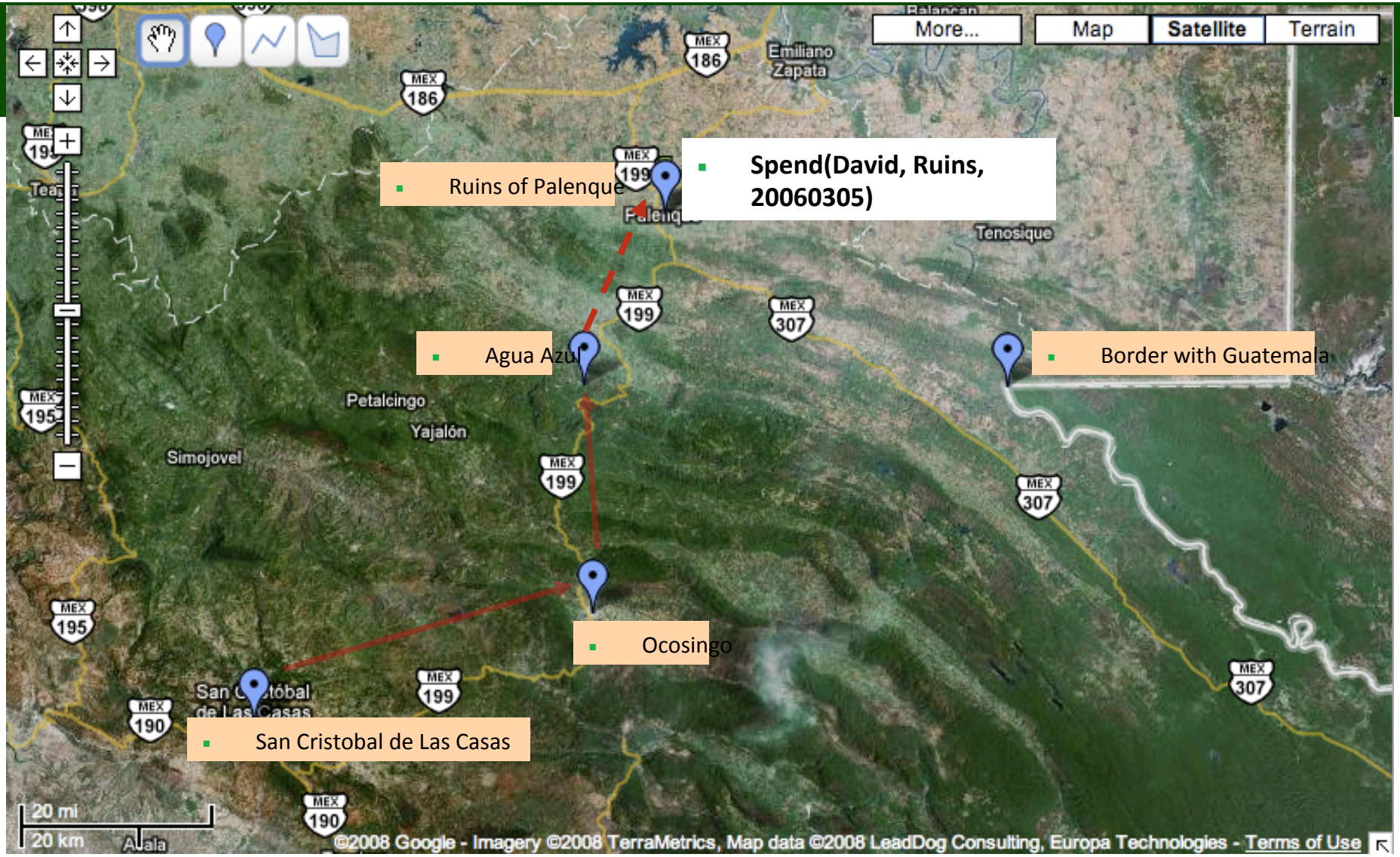
- David left San Cristobal de Las Casas 4 days ago.



- David arrived in Ocosingo that day.

- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.

- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.

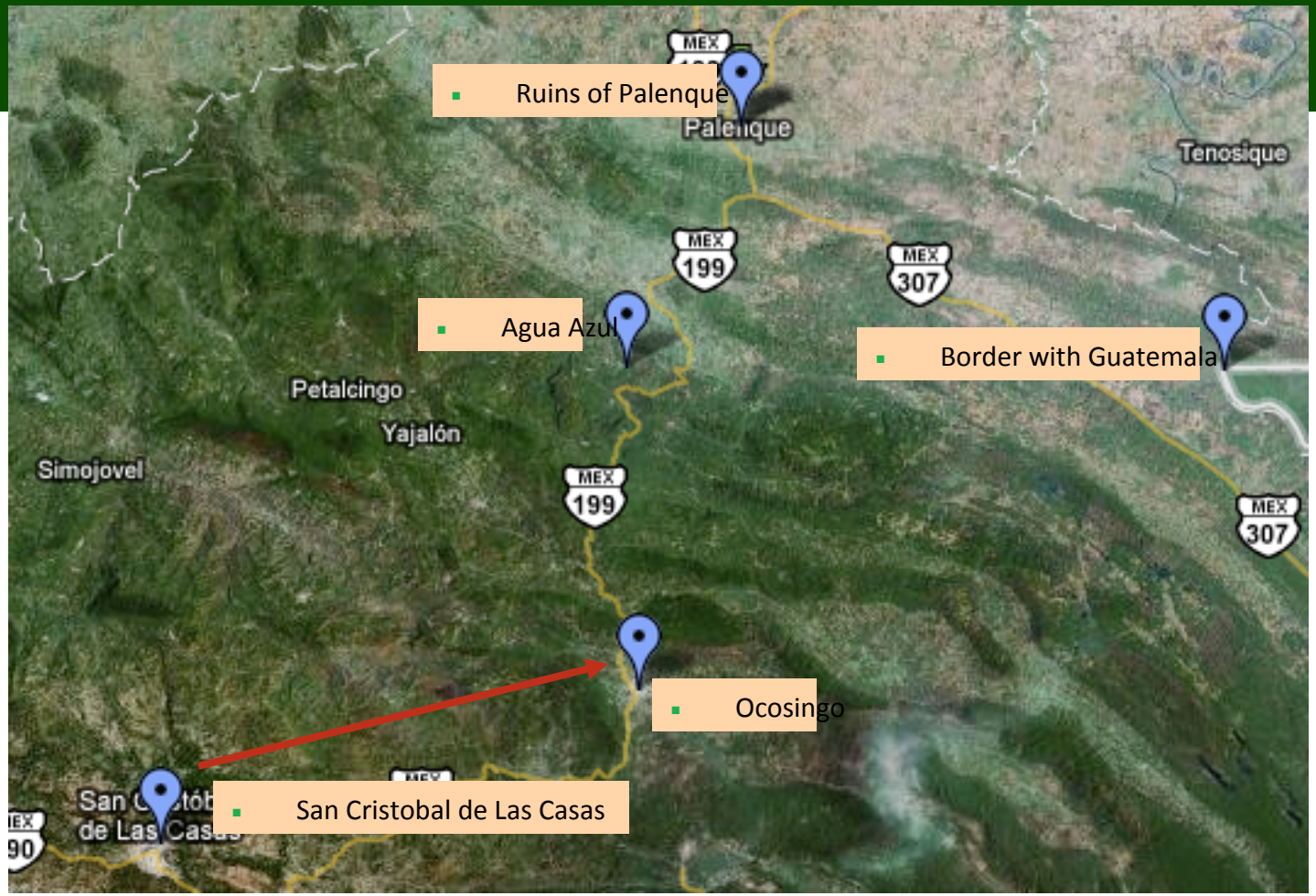


- David spent the next day at the ruins of Palenque.

- The following day, David drove to the border with Guatemala.

- March 3, 2006

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.



- March 3, 2006

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.

- March 4, 2006

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.

- 4 hours on
- March 4, 2006

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.

- March 5, 2006

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.

- March 6, 2006

- David left San Cristobal de Las Casas 4 days ago.
- David arrived in Ocosingo that day.
- The next day, David biked to Agua Azul and played in the waterfalls there for 4 hours.
- David spent the next day at the ruins of Palenque.
- The following day, David drove to the border with Guatemala.

Situated Grounding and Spatial Reasoning

- Multimodal Situated Grounding – co-perception and co-attention are necessary to understand deixis and relative spatial expressions
  - *Put the big one right here.*
  - *I want the cookie on the left behind the donut.*
  - *Show me a coffee shop around … here.*
- Understanding Events and their Results – actions change the spatial nature of the environment
  - *Mary opened the door and left the room.*
  - *Put the book in the bag. Take the bag to the car.*
  - *Remove the seeds and cut into thin strips, then brown in oil.*
- Appreciation of spatial properties of objects - intrinsic vs. relative Frame of Reference
  - *The tree behind the bench*
  - *The bench in front of the tree*

Situated Grounding – Foundational Work

- Cassell, J., Nakano, Y., Reinstein, G., & Stocky, T. (2003). Towards a model of face-to-face grounding. ACL.
- Holroyd, A., Rich, C., Sidner, C. L., & Ponsler, B. (2011). Generating connection events for human-robot collaboration. *2011 RO-MAN,* IEEE.
- Traum, D., and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. *Proc. Autonomous agents and multiagent systems*.
- Scheutz, M., Schermerhorn, P., Kramer, J., & Anderson, D. (2007). First steps toward natural human-like HRI. *Autonomous Robots*, *22*(4).
- Elliott, D., D. Kiela and A. Lazaridou (2016) Multimodal Learning and Reasoning, ACL Tutorial.
- Yatskar, M., Zettlemoyer, L., and Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. Proceedings of the IEEE CVPR.
- Hough, J., & Schlangen, D. (2017). A Model of Continuous Intention Grounding for HRI.
- Alikhani, M., and Stone, M. (2020). Achieving Common Ground in Multi-modal Dialogue. In Proceedings of the 58th Annual Meeting of ACL Tutorial

Situated Grounding and Spatial Reasoning

- Frames of Reference
  - *Absolute (coordinate system)*
  - *Relative (from an agent view)*
  - *Intrinsic (inherent property of object)*

Levinson, S. C. (2003). Space in language and culture: Explorations in cognitive diversity. Cambridge: Cambridge University Press

☺ On the left side of the picture is a big tree.

☹ or ☺ A tree is in the center of the scene.

☹ The tree's shadow is in the lower left corner.

☹ A bench is in front of a tree.

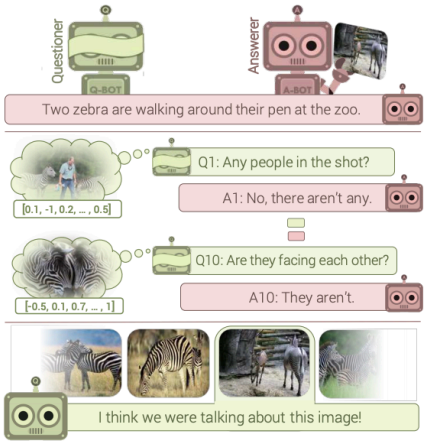
☹ On the left side of the picture is a big tree.

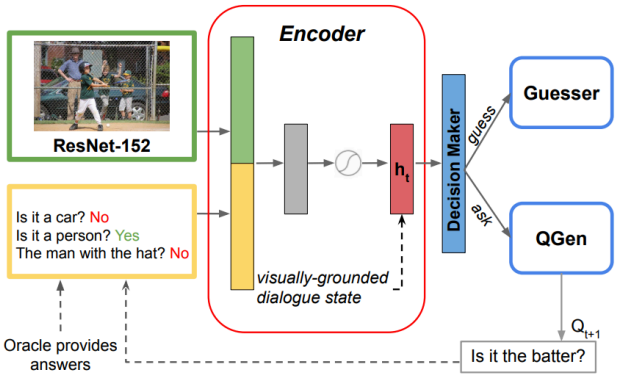
☺ A tree is in the center of the scene.

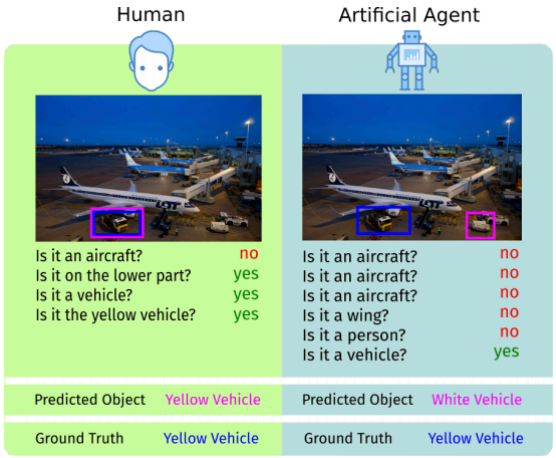
☺ The tree's shadow is in the lower left corner.

☺ A bench is in front of a tree.

Interactive Object Recognition in Dialogue







- Das, A., Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2951-2960).
- Shekhar, R., Testoni, A., Fernández, R., & Bernardi, R. (2019). Jointly Learning to See, Ask, Decide when to Stop, and then GuessWhat. In *CLiC-it*.
- Shukla, P., Elmadjian, C., Sharan, R., Kulkarni, V., Turk, M., & Wang, W. Y. (2019). What Should I Ask? Using Conversationally Informative Rewards for Goal-Oriented Visual Dialog. *arXiv preprint arXiv:1907.12021*.

Spatial Reasoning in Collaborative Tasks

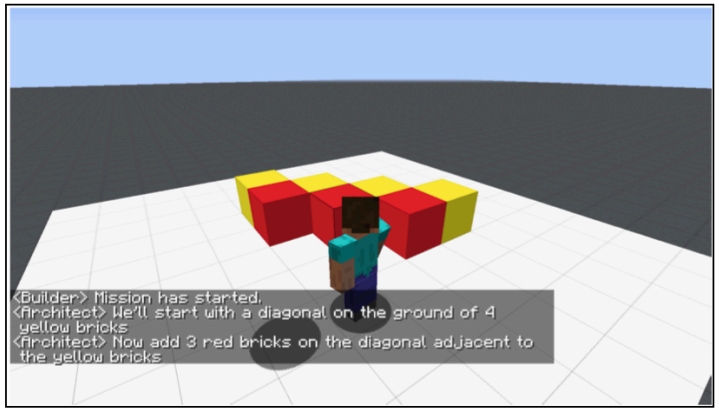


Figure 1: An instance of the collaborative building task. The last instruction was : Now add 3 red bricks on the diagonal adjacent to the yellow bricks.

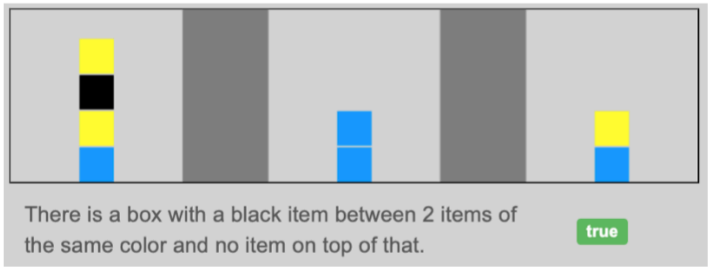


Figure 2: An example from the NLVR corpus that demonstrates *spatial focus shift* from the *black item* to the *yellow item.*

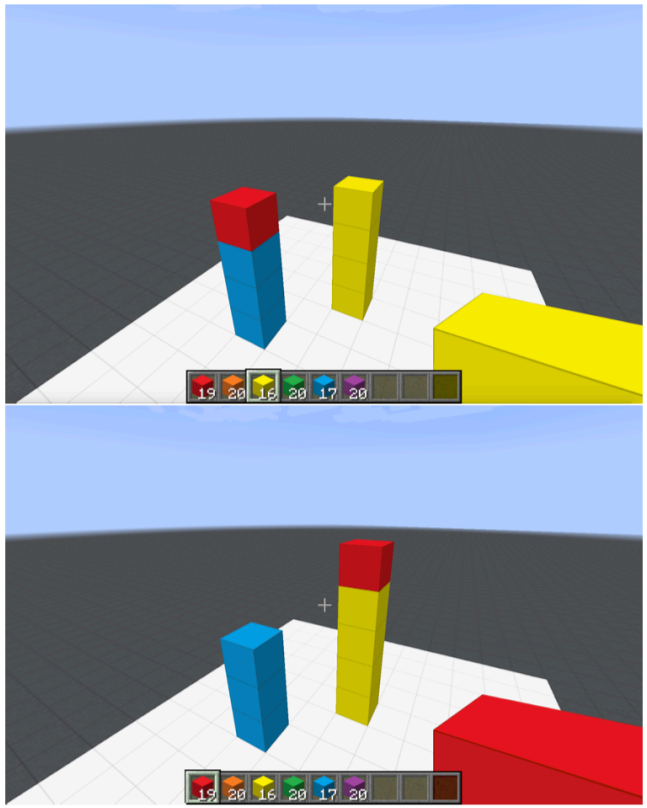


Figure 7: Move the large red block diagonally from the top of the blue column to the top of the yellow column ...

Dan, S., Kordjamshidi, P., Bonn, J., Bhatia, A., Cai, Z., Palmer, M., & Roth, D. (2020). From Spatial Relations to Spatial Configurations. *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5855-5864).

Spatial Reasoning and AMRs

	Configuration 1	Configuration 2
tr	$< t1, e1 >$	$< t2, e3 >$
lm	$< l1, e2 >, < l2, e3 >$	$< l3, e4 >$
sp	$<s1,from >$ $<s2,to >$	$<s1,from,$ $\{metric = 5spaces\}>$
m	$<m1,move, >$	NULL
path	$< l1, s1, begin >$ $< l2, s2, end >$ $\{orientation = diagonally\}$	NULL
FoR	$< l1, relative >$ $< l2, relative >$	$< l3, relative >$
v	first-person	first-person
QT	$<directional, relative>$	$<distal, quantitative>$ $<topological, DC>$

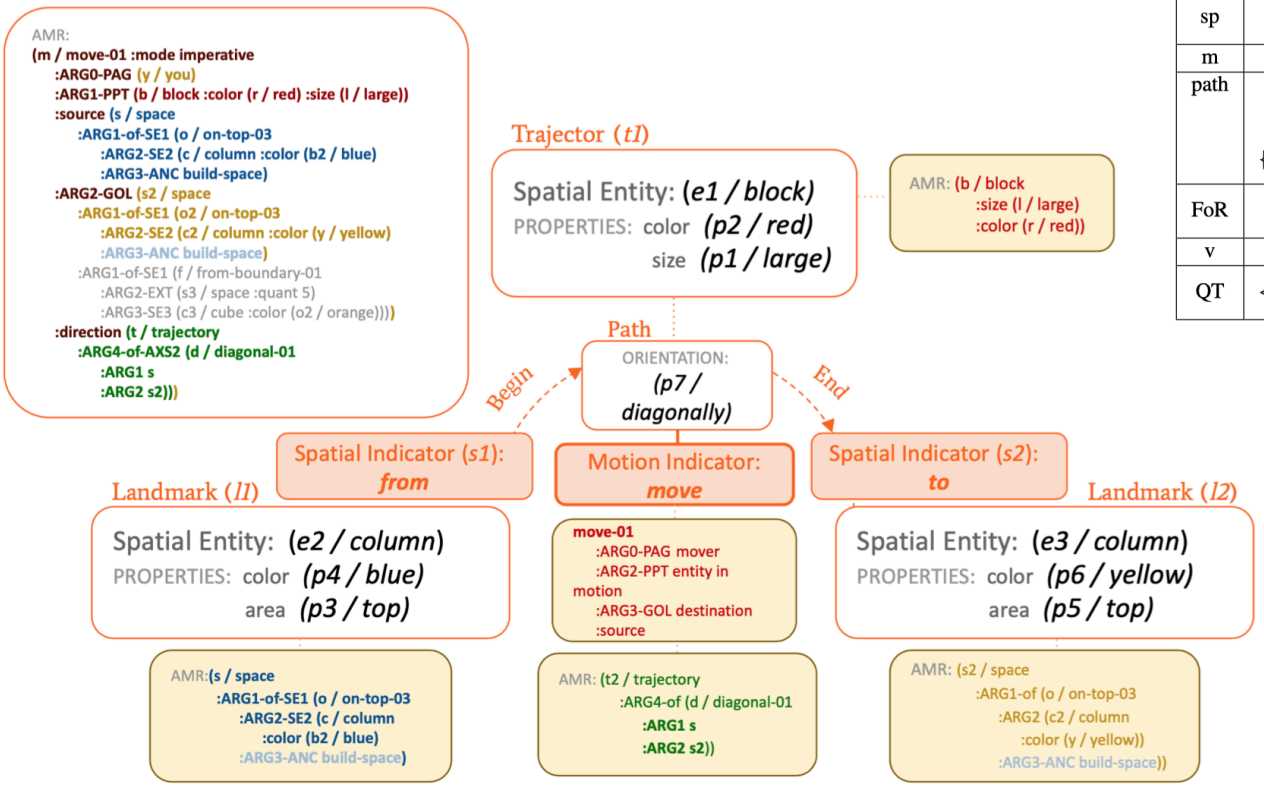


Figure 3: Graphical Representation of Configuration 1 of Table 3 with aligned AMR : ***Move the large red block diagonally from the top of the blue column to the top of the yellow column***, *which is 5 spaces from the orange cube.*

Dan, S., Kordjamshidi, P., Bonn, J., Bhatia, A., Cai, Z., Palmer, M., & Roth, D. (2020). From Spatial Relations to Spatial Configurations. *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 5855-5864).

Spatial Reasoning in Minecraft

Create models that generate spatial descriptions

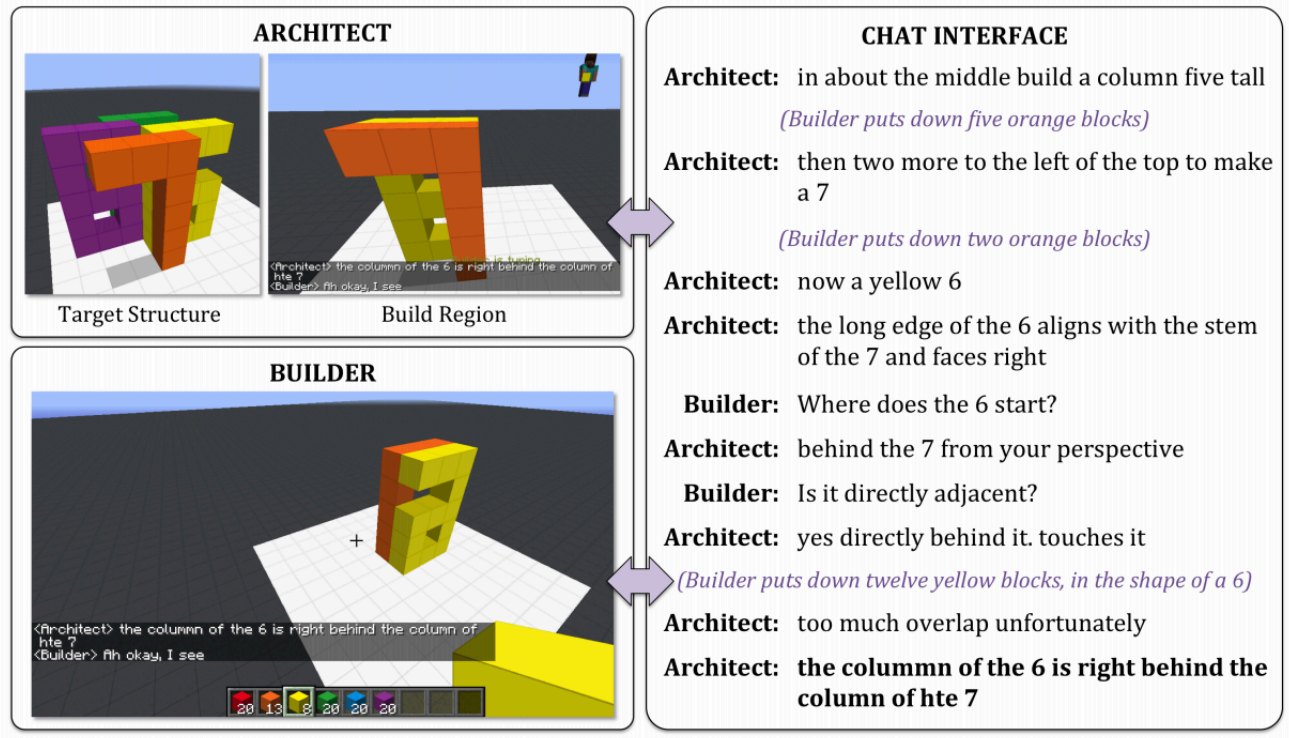


Figure 1: In the Minecraft Collaborative Building Task, the Architect (**A**) has to instruct a Builder (**B**) to build a target structure. **A** can observe **B**, but remains invisible to **B**. Both players communicate via a chat interface. (NB: We show **B**'s actions in the dialogue as a visual aid to the reader.)

Narayan-Chen, A., Jayannavar, P., & Hockenmaier, J. (2019). Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5405-5415).

Spatial Reasoning in Minecraft

Create models that execute spatial actions

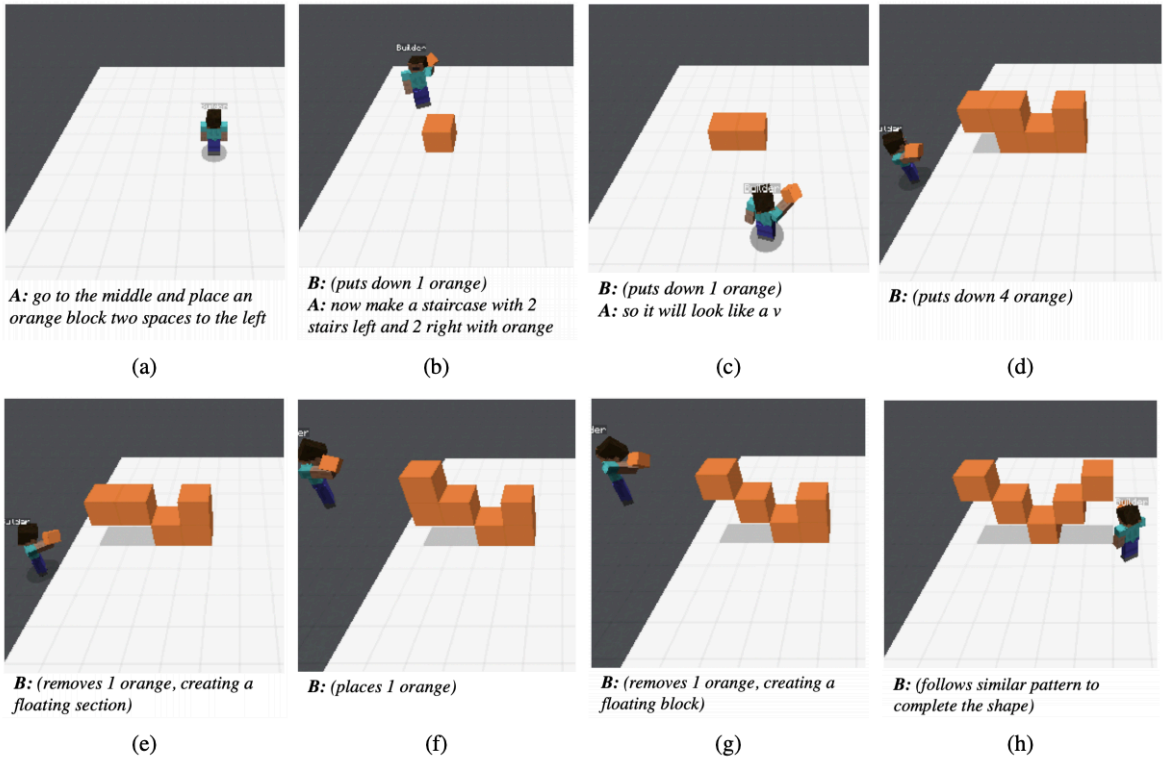
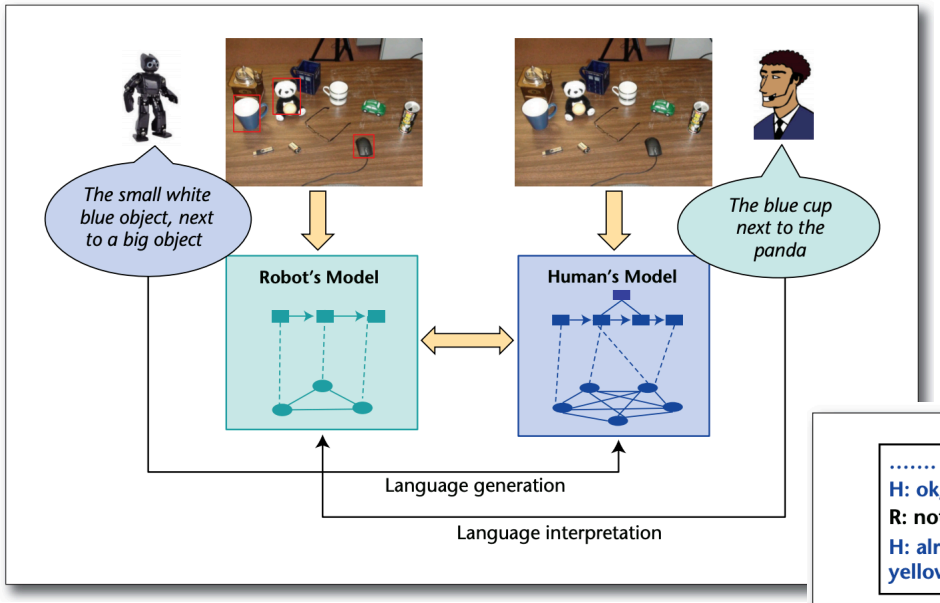


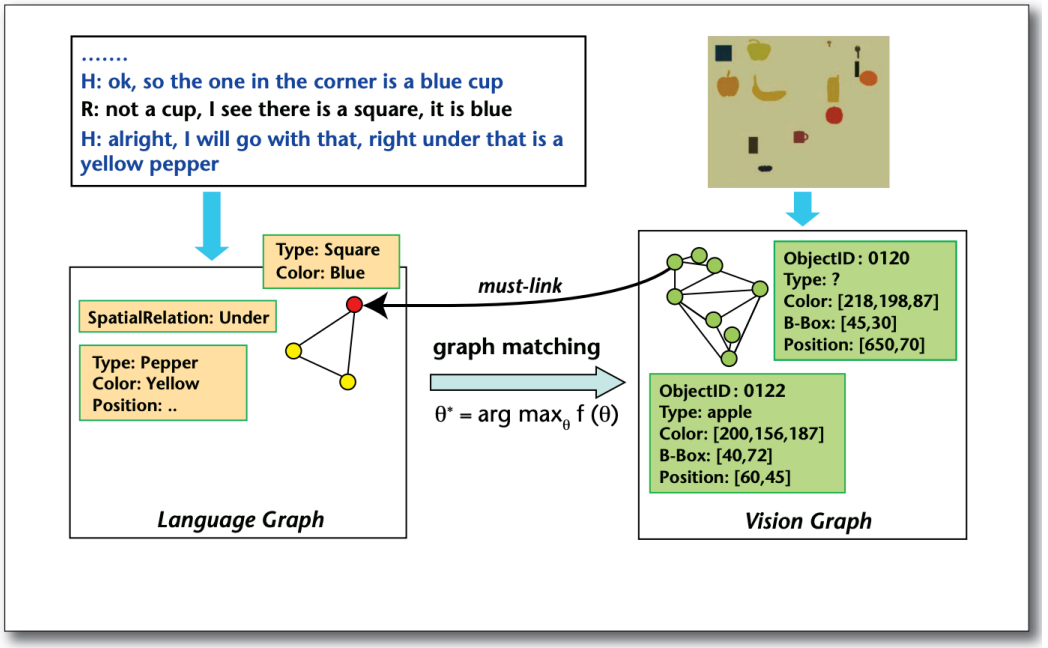
Figure 1: A sample sequence of human-human game states. The game starts with an empty grid and an initial **A** instruction (a), which **B** executes in the first action sequence (b) by placing a single block. In (c), **B** begins to execute the next **A** instruction given in (b). However, **A** interrupts **B** in (c), leading to two distinct **B** action sequences: (b)–(c) (single block placement), and (c)–(h) (multiple placements and removals).

Jayannavar, P., Narayan-Chen, A., & Hockenmaier, J. (2020). Learning to execute instructions in a Minecraft dialogue. In *Proceedings of the 58th Meeting of the Association for Computational Linguistics* (pp. 2589-2602).

Situated Grounding in Human Robot Dialogue



- Establish a Joint Perceptual Basis through language grounding



- Graph-Matching for Interpreting Referring Expressions

Chai, J. Y., Fang, R., Liu, C., & She, L. (2016). Collaborative language grounding toward situated human-robot dialogue. *AI Magazine*, *37*(4), 32-45.

Grounding - Multimodal Spatial Expressions

(1) Here$_{[deixis]}$ is the bus stop, a bit left of it$_{[deixis]}$ is a church and right in front of that$_{[deixis]}$ is the hotel.

Figure 1: Providing a multimodal description (*left*) of a scene (*right*).

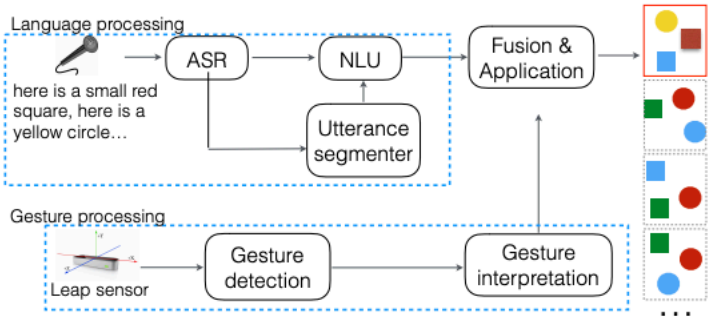


Figure 2: Multimodal system architecture.

- Interpreting multimodal spatial descriptions in route giving tasks.
- Gestures not only contribute information, but also help interpretations of speech incrementally, due to its parallel nature.

Han, T., Kennington, C., & Schlangen, D. (2018). Placing Objects in Gesture Space: Toward Real-Time Understanding of Spatial Descriptions. In *AAAI18.*

Situated Grounding and Pointing Actions

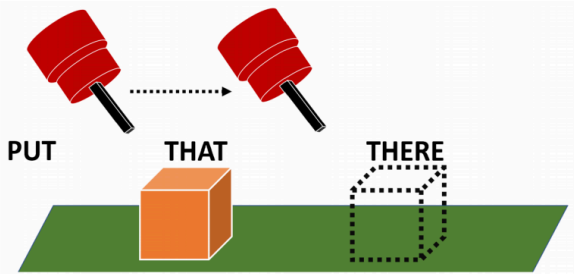


Figure 1: A pick-and-place task requires a *referential* pointing action to the object (orange cube) at the initial position, and a *locating* pointing action to a final placement position (dotted cube). Such an action by a robot (in red) can also be accompanied by verbal cues like *"Put that there."*

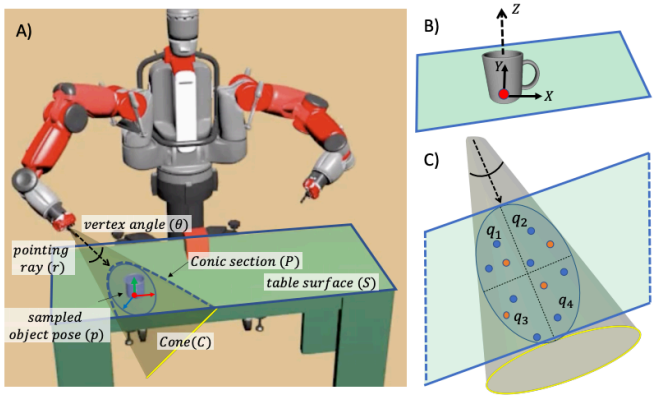


Figure 2: (A) Workspace setup showing the pointing cone and the corresponding conic section on the table. (B) The degrees-of-freedom considered for placement of the object on the table. (C) Sampling policy to sample object poses within the conic section.

- Pointing to something vs. somewhere
- Human subjects show greater flexibility in interpreting the intent of referential pointing compared to locating pointing, which needs to be more deliberate.

Alikhani, M., Khalid, B., Shome, R., Mitash, C., Bekris, K. E., & Stone, M. (2020). That and There: Judging the Intent of Pointing Actions with Robotic Arms. In AAAI (pp. 10343-10351).

Embodiment and Situated Grounding

- Task-oriented dialogues are embodied interactions between agents, where language, gesture, gaze, and actions are situated within a common ground shared by all agents in the communication.
- Situated semantic grounding assumes shared perception of agents with co-attention over objects in a situated context, with co-intention towards a common goal.
- VoxWorld : a multimodal simulation framework for modeling Embodied Human-Computer Interactions and communication between agents engaged in a shared goal or task.
- Embodied HCI and robot control in action.

Pustejovsky, J., & Krishnaswamy, N. (2020). Embodied Human-Computer Interactions through Situated Grounding. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents.

Spatial Semantics and Situated Grounding

- Identifying the *actions and consequences* associated with objects in the environment.
- Encoding a multimodal expression contextualized to the *dynamics of the discourse*
- *Situated grounding*: Capturing how multimodal expressions are anchored, contextualized, and situated in context

Spatial Reasoning and Situated Meaning

Situated Meaning in a Joint Activity

- Son: *Put it there (gesturing with co-attention)?*
- Mother: *Yes, go down for about two inches.*
- Mother: *OK, stop there. (co-attentional gaze)*
- Son: *Okay. (stops action)*
- Mother: *Now, start this one (pointing to another cupcake).*

Situated Meaning and Common Ground

Agents	mother, son
Shared goals	baking, icing
Beliefs, desires, intentions	Mother knows how to ice, bake, etc. Mother is teaching son
Objects	Mother, son, cupcakes, plate, knives, pastry bag, icing, gloves
Shared perception	the objects on the table
Shared Space	kitchen

Stalnaker R., “Common ground”, *Linguistics and philosophy*, vol. 25, no 5-6, p. 701-721, 2002

Clark H. H., Brennan S. E., “Grounding in communication”, *Perspectives on socially shared cognition*, vol. 13, p. 127-149, 1991.

Embodiment and Situated Grounding

- Pustejovsky, J., & Krishnaswamy, N. (2016). VoxML: A Visualization Modeling Language. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)..
- Krishnaswamy, N., & Pustejovsky, J. (2016). VoxSim: A visual platform for modeling motion language. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations.
- Pustejovsky, J., & Krishnaswamy, N. (2020). Embodied Human-Computer Interactions through Situated Grounding. In Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents.
- Krishnaswamy, N., & Pustejovsky, J. (2020). A Formal Analysis of Multimodal Referring Strategies Under Common Ground. LREC.
- Krishnaswamy, N., & Pustejovsky, J. (2017, August). Do you see what I see? effects of pov on spatial relation specifications. In Proc. 30th International Workshop on Qualitative Reasoning.

Spatial Properties of Objects

- Object size, shape, dimensionality, texture
- Orientation, frame of reference, facing (front/back)
- How we spatially interact with an object
- Space needed for Object function - affordance space
- Event space used for object function or purpose

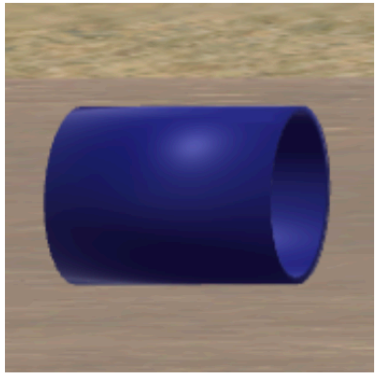
Pustejovsky, J., & Krishnaswamy, N. (2016). VoxML: A Visualization Modeling Language. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16).

Krishnaswamy, N., & Pustejovsky, J. (2016). VoxSim: A visual platform for modeling motion language. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations.

Spatial Properties of Objects

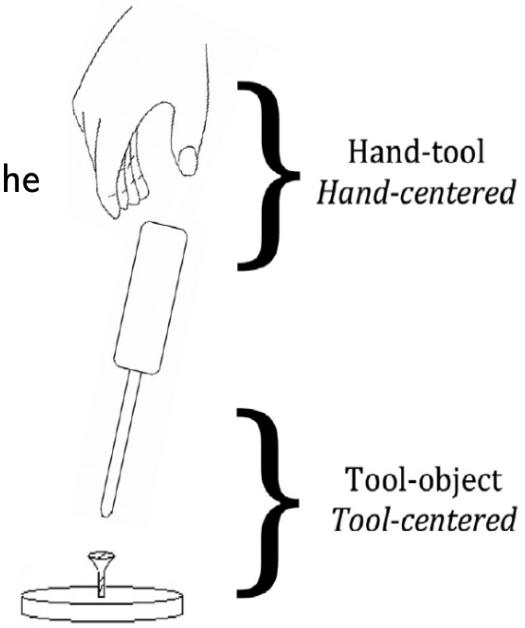
- Context of objects is described by their properties.
- Object properties cannot be decoupled from the events they facilitate.
  - *Affordances* (Gibson, 1979)
  - *Qualia* (Pustejovsky, 1995)

"He **slid** the cup across the table. Liquid spilled out."
"He **rolled** the cup across the table. Liquid spilled out."



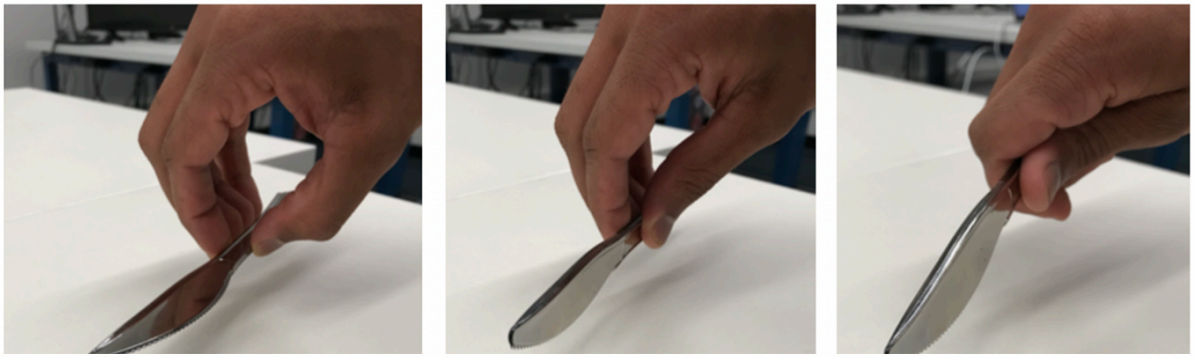
Reference Frames and Affordances

- Hand-tool Interface: based on the agent's biomechanical and morphological characteristics. For instance, a hammer is graspable by a human adult but not by a baby. Thus, the interface is centered on the agent.
- Tool-object Interface: independent of the agent's characteristics. The relationship is centered on objects external to the agent and the interaction is made possible because of the compatibility between the characteristics of the tool and the object.



Osiurak, F., Rossetti, Y., and Badets, A. (2017). What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, 77, 403-417.

Affordance Space and Grasp Poses



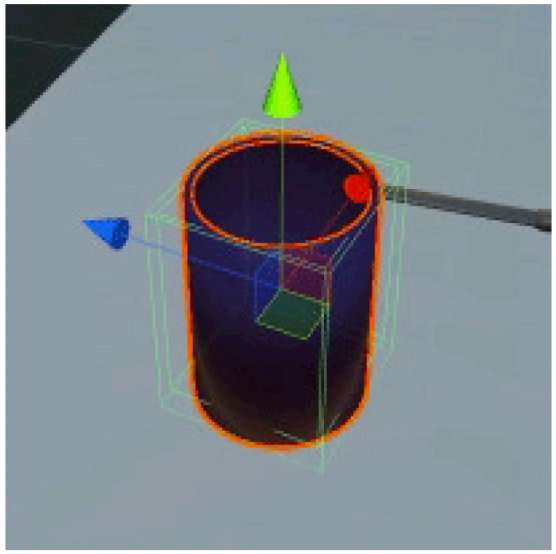
Pustejovsky, J., Krishnaswamy, N., and Do, T. (2017). Object embodiment in a multimodal simulation. In *AAAI Spring Symposium: Interactive Multisensory Object Perception for Embodied Agents*.

VoxML: Visual Object Concept Modeling Language

- Encodes afforded behaviors for each object
  - Gibsonian: afforded by object structure (Gibson,1977,1979)
    - grasp, move, lift, etc.
  - Telic: goal-directed, purpose-driven (Pustejovsky, 1995, 2013)
    - drink from, read, etc.
- Voxeme
  - Object Geometry: Formal object characteristics in R3 space
  - Habitat: Orientation, Situated context, Scaling
  - Affordance Structure:
    - What can one do to it
    - What can one do with it
    - What does it enable

VoxML - cup

$$
\left[\begin{array}{l}
\textbf{cup} \\
\text{LEX} = \left[\begin{array}{l} \text{PRED} = \textbf{cup} \\ \text{TYPE} = \textbf{physobj, artifact} \end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l} \text{HEAD} = \textbf{cylindroid[1]} \\ \text{COMPONENTS} = \textbf{surface, interior} \\ \text{CONCAVITY} = \textbf{concave} \\ \text{ROTATSYM} = \{Y\} \\ \text{REFLECTSYM} = \{XY, YZ\} \end{array}\right] \\
\text{HABITAT} = \left[\begin{array}{l} \text{INTR} = {}_{[2]}\left[\begin{array}{l} \text{CONSTR} = \{Y > X, Y > Z\} \\ \text{UP} = align(Y, \mathcal{E}_Y) \\ \text{TOP} = top(+Y) \end{array}\right] \\ \text{EXTR} = {}_{[3]}\left[\begin{array}{l} \text{UP} = align(Y, \mathcal{E}_{\perp Y}) \end{array}\right] \end{array}\right] \\
\text{AFFORD\_STR} = \left[\begin{array}{l} \text{A}_1 = H_{[2]} \rightarrow [put(x, on([1]))] support([1], x) \\ \text{A}_2 = H_{[2]} \rightarrow [put(x, in([1]))] contain([1], x) \\ \text{A}_3 = H_{[2]} \rightarrow [grasp(x, [1])] \\ \text{A}_4 = H_{[3]} \rightarrow [roll(x, [1])] \end{array}\right] \\
\text{EMBODIMENT} = \left[\begin{array}{l} \text{SCALE} = \textbf{<agent} \\ \text{MOVABLE} = \textbf{true} \end{array}\right]
\end{array}\right]
$$



VoxML for actions and relations

$$
\begin{bmatrix}
\textbf{put} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix}
\text{HEAD} = \textbf{transition} \\
\text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{bmatrix} \\
\text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while(hold(x,y), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)] \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

$$
\begin{bmatrix}
\textbf{on} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{on} \end{bmatrix} \\
\text{TYPE} = \begin{bmatrix}
\text{CLASS} = \textbf{config} \\
\text{VALUE} = \textbf{EC} \\
\text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:3D} \\ \text{A}_2 = \textbf{y:3D} \end{bmatrix} \\
\text{CONSTR} = \textbf{y}\rightarrow\text{HABITAT}\rightarrow\text{INTR}[align]
\end{bmatrix}
\end{bmatrix}
$$

VoxML - grasp

$$
\left[\begin{array}{l}
\textbf{grasp} \\
\text{LEX} = \left[\begin{array}{l}\text{PRED} = \textbf{grasp} \\ \text{TYPE} = \textbf{transition\_event}\end{array}\right] \\
\text{TYPE} = \left[\begin{array}{l}\text{HEAD} = \textbf{transition} \\ \text{ARGS} = \left[\begin{array}{l}\text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj}\end{array}\right] \\ \text{BODY} = \left[\text{E}_1 = grasp(x, y)\right]\end{array}\right]
\end{array}\right]
$$

VoxML – Composition [grasp + cup]

- Continuation-passing style semantics for composition
- Used within conventional sentence structures
- Used between sentences in discourse
- Used for gesture sequencing as well

Krishnaswamy, N., & Pustejovsky, J. (2019). Multimodal Continuation-style Architectures for Human-Robot Interaction. *arXiv preprint arXiv: 1909.08161*.

QSRLib

Table 1: Description of supported qualitative spatial relation families

qualitative spatial relation families	type	num of relations / variations	kind of entities
Qualitative Distance Calculus	distance	user specified	2D points
Probabilistic Qualitative Distance Calculus	distance	user specified	2D points
Cardinal Directions	direction	9	2D rectangles
Moving or Stationary	motion	2	2D points
Qualitative Trajectory Calculus	motion	B11: 9, C21: 81	2D points
Rectangle/Block Algebra	topology & direction	169/2197	2D/3D rectangles
Region Connection Calculus	topology	2, 4, 5, 8	2D rectangles
Ternary Point Configuration Calculus	direction	25	2D points

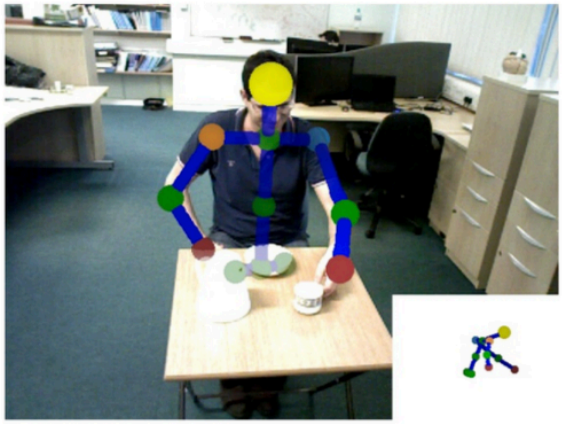


Figure 1: Activity recognition in a table top setting. Dyadic QSR relations between detected objects/skeleton points can be computed (bottom right inset).

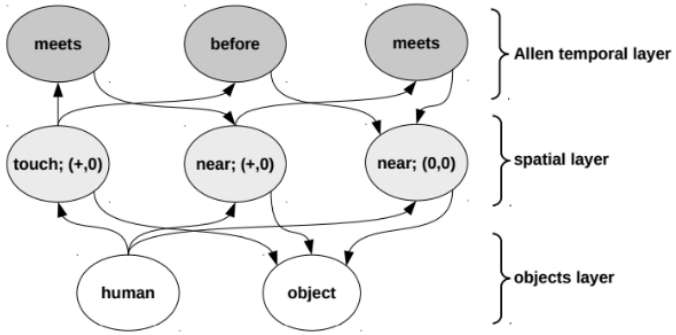
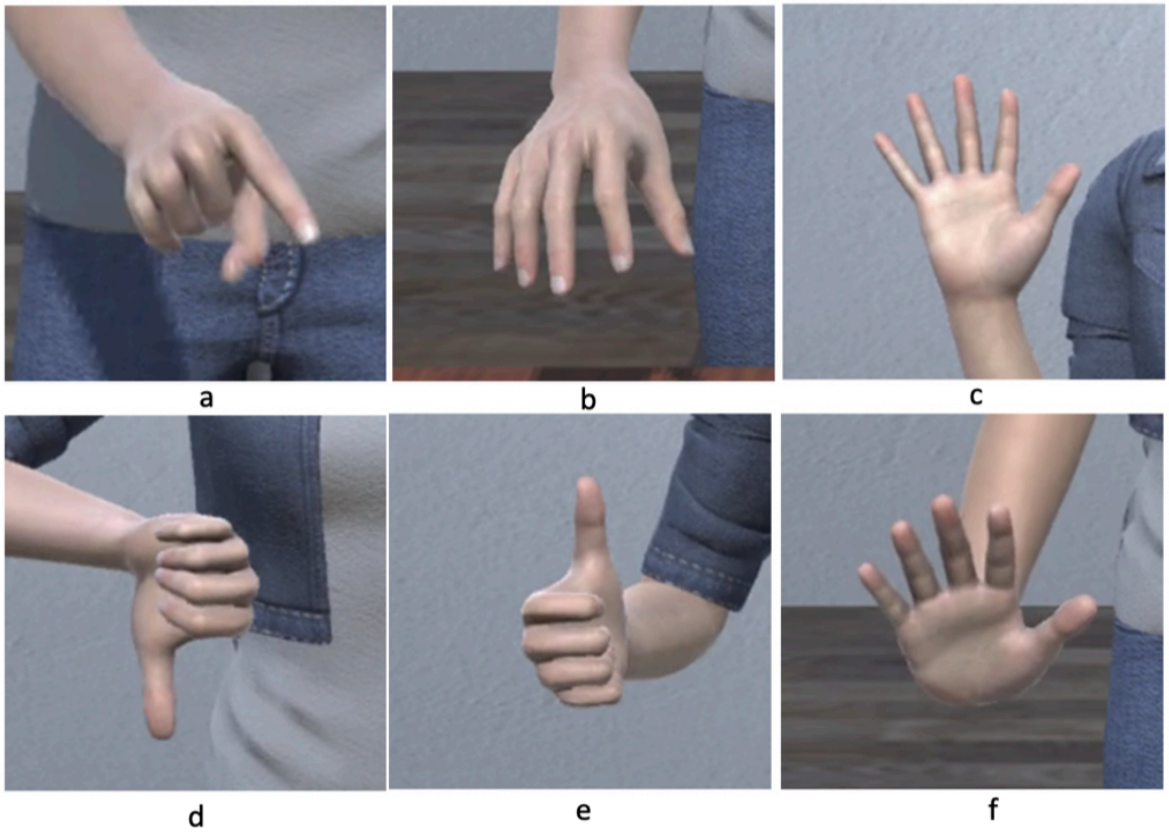


Figure 5: Example of a Qualitative Spatio-Temporal Activity Graph (QSTAG) between a human and an object; each spatial layer node encodes QSRs from two calculi: a QDC relation (touch/near) and a QTC_{B21} one ((+,0)/(0,0)).

- Gatsoulis, Yiannis, Muhannad Alomari, Chris Burbridge, Christian Dondrup, Paul Duckworth, Peter Lightbody, Marc Hanheide, Nick Hawes, D. C. Hogg, and A. G. Cohn. "Qsrlib: a software library for online acquisition of qualitative spatial relations from video." (2016).

Gestures Generated in VoxWorld

- VoxML encodes spatial configuration of gestures



Pustejovsky, J., Krishnaswamy, N., Beveridge, R., Ortega, F. R., Patil, D., Wang, H., & McNeely-White, D. G. Interpreting and Generating Gestures with Embodied Human Computer Interactions, GENEA Workshop, IVA20, 2020.

Situated Grounding in Dialogue

A non-verbal interaction between a human and IVA using gesture, gaze, and action.

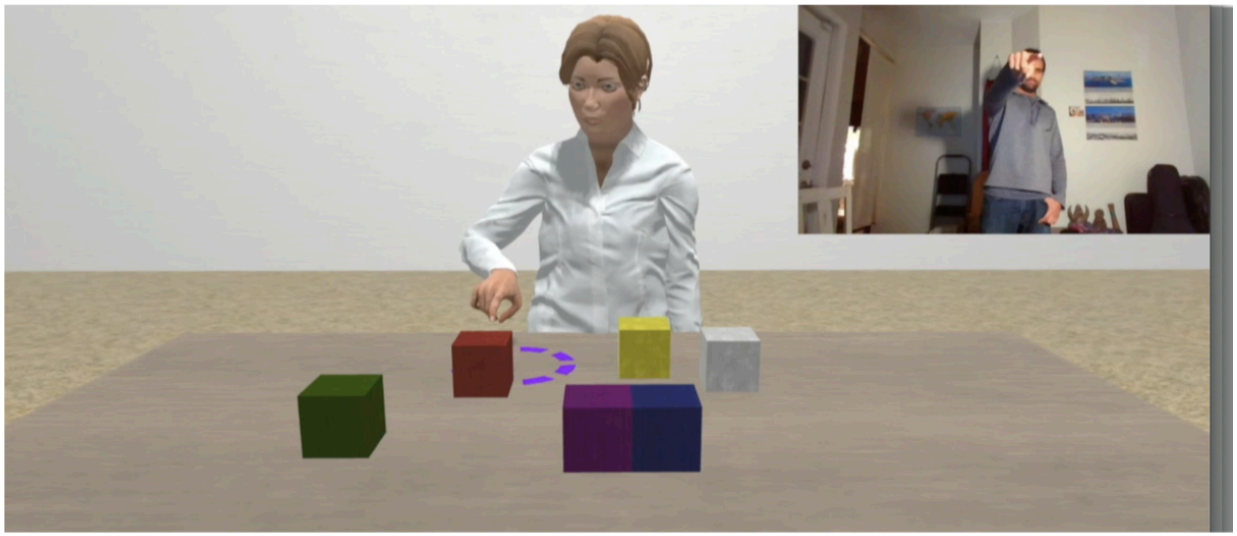
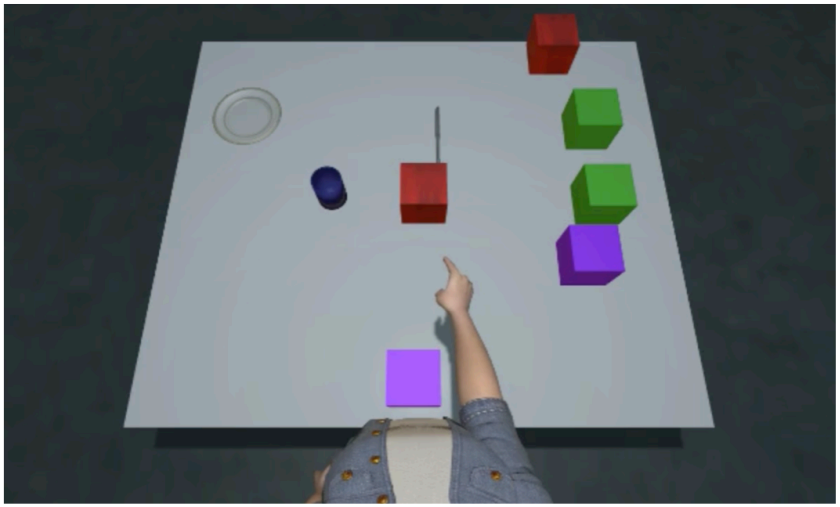


Figure: IVA Diana engaging in an embodied HCI with a human user.

Krishnaswamy, Nikhil, Pradyumna Narayana, Rahul Bangar, Kyeongmin Rim, Dhruva Patil, David McNeely-White, Jaime Ruiz, Bruce Draper, Ross Beveridge, and James Pustejovsky. "Diana's World: A Situated Multimodal Interactive Agent." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 09, pp. 13618-13619. 2020.

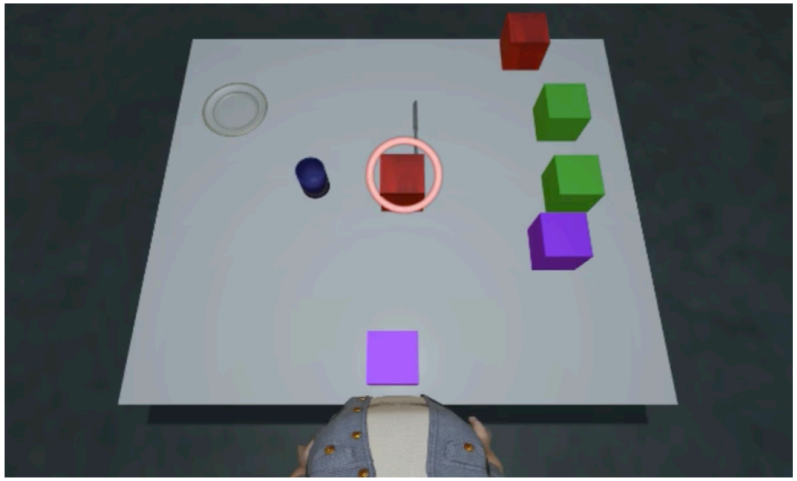
Multimodal Referring Expressions

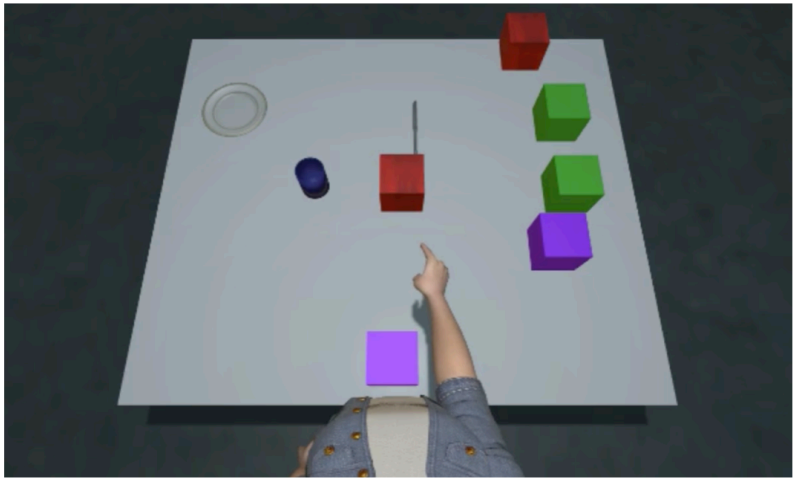
- Different modalities are better at expressing different information
- Humans make use of multiple modalities to refer to entities
- How do humans prefer to mix and match modalities?



Krishnaswamy, N., and Pustejovsky, J. (2020). A Formal Analysis of Multimodal Referring Strategies Under Common Ground. Proceedings of LREC.

Multimodal Referring Expressions

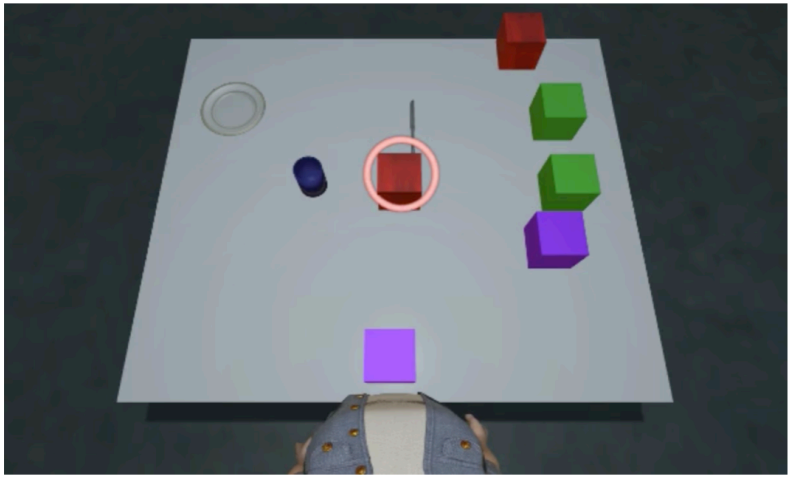


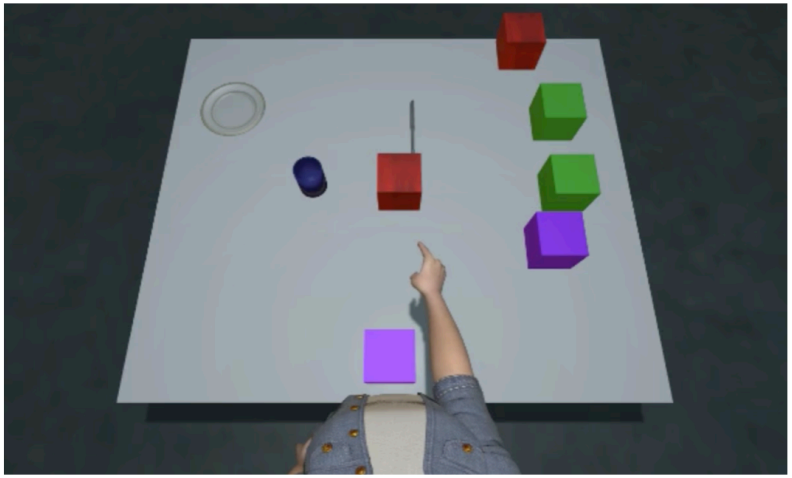


- Different modalities are better at expressing different information
- Humans make use of multiple modalities to refer to entities
- How do humans prefer to mix and match modalities?

Krishnaswamy, N., and Pustejovsky, J. (2020). A Formal Analysis of Multimodal Referring Strategies Under Common Ground. Proceedings of LREC.

Multimodal Referring Expressions

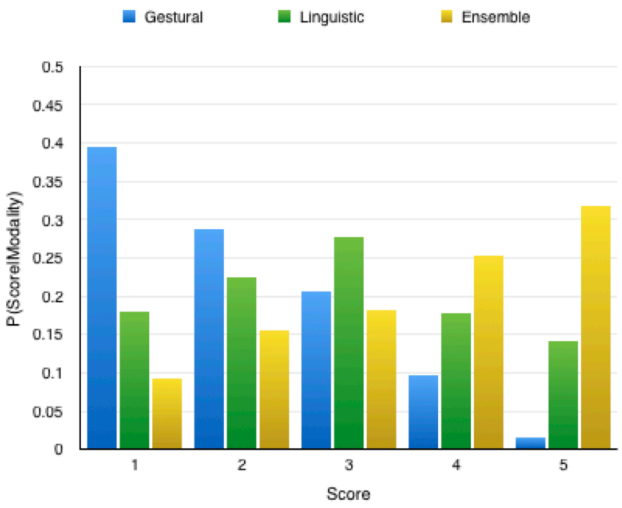


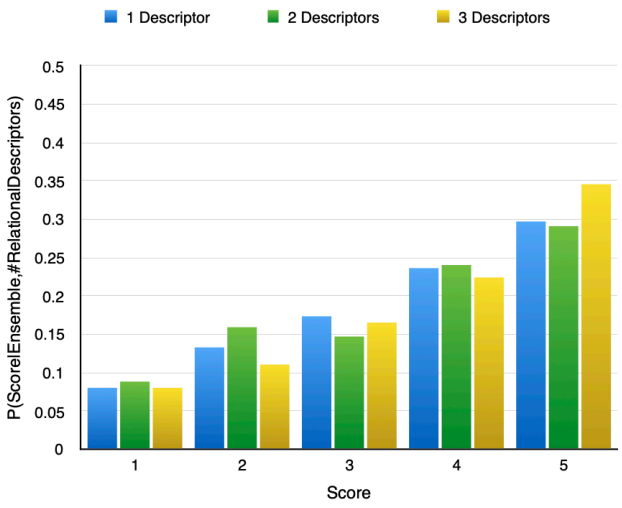


- EMRE dataset: 5 referencing strategies x 6 objects x 50 situations (1,500 videos)
- Parameters varied: modality, distance distinction, # relational descriptors, etc.
- MTurk Likert-type ranking (1-5): “How natural is the referring expression shown?”

Krishnaswamy, N., and Pustejovsky, J. (2019). Generating a Novel Dataset of Multimodal Referring Expressions. In Proceedings of the 13th International Conference on Computational Semantics.

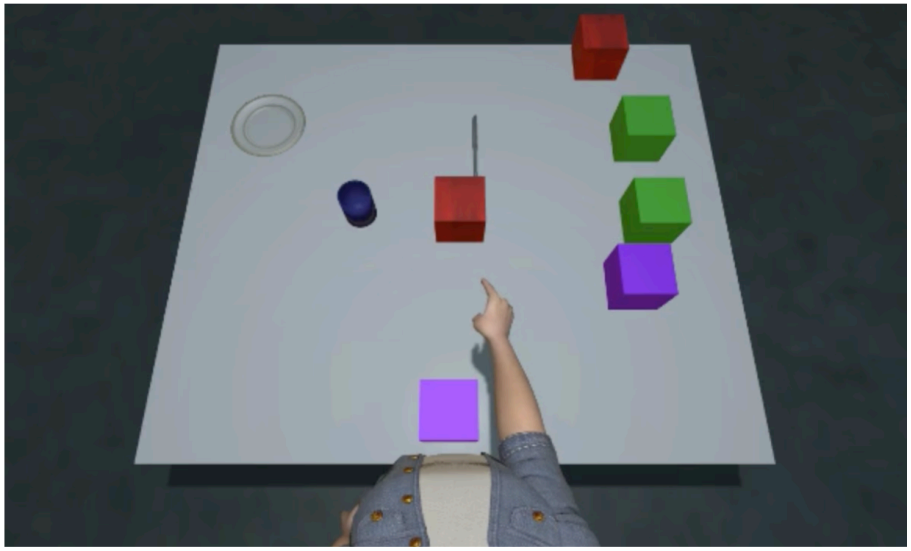
Multimodal Referring Expressions

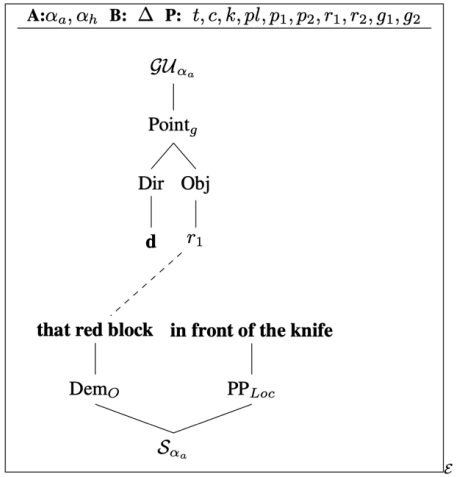




- Humans prefer mixed-modality referring strategies ✅
- Humans prefer more descriptive language ✅
- Not enough data to train a generation model 🚫

Multimodal Referring Expressions



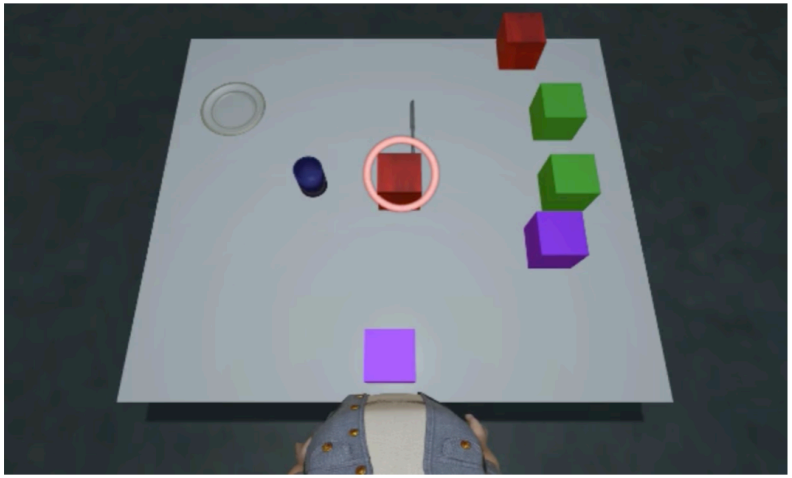


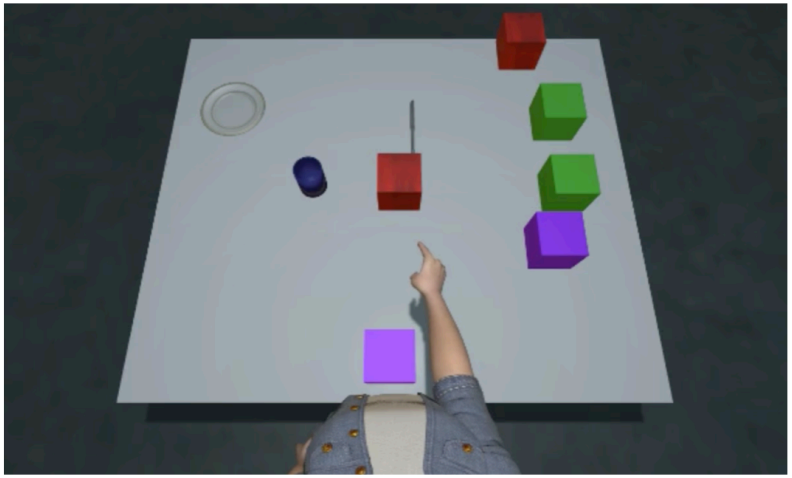
$\lambda k_s \otimes k_g(\textbf{that}(x)[\text{block}(x) \wedge \text{red}(x) \wedge \text{in\_front}(x, k, v)] \wedge k_s \otimes k_g(x)]$, where $v = \alpha_a$

"Common Ground Structure" (CGS) aligns modal information with continuation semantics

A: agents
B: belief space
P: perceived objects

Multimodal Referring Expressions



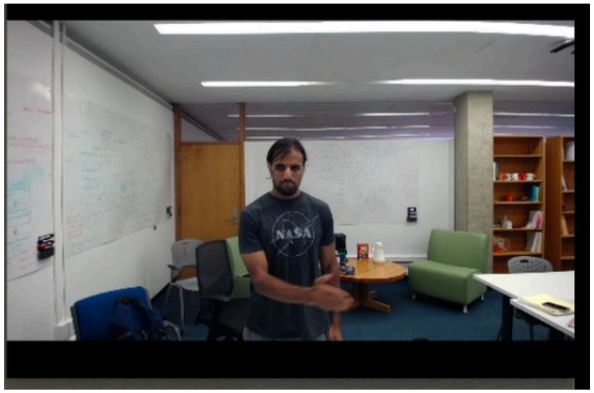


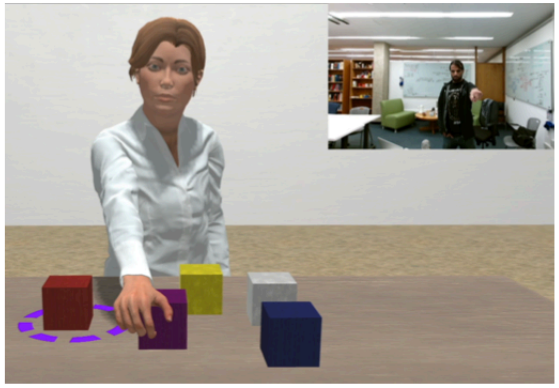
- EMRE dataset: 5 referencing strategies x 6 objects x 50 situations (1,500 videos)
- Parameters varied: modality, distance distinction, # relational descriptors, etc.
- MTurk Likert-type ranking (1-5): “How natural is the referring expression shown?”

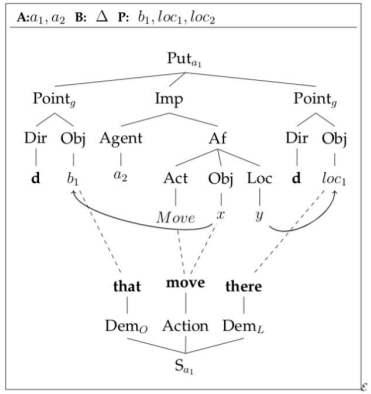
Krishnaswamy, N., and Pustejovsky, J. (2019). Generating a Novel Dataset of Multimodal Referring Expressions. In Proceedings of the 13th International Conference on Computational Semantics.

Multimodal Dialogue

- Language and Gesture determine Situated Grounding
- “That block, move it there.”







$\lambda k'_s \otimes k'_g.\overline{(\langle \textbf{that}, Point_1\rangle\langle \textbf{move}, Move\rangle)}(\lambda r_s \otimes r_g.\overline{\langle \textbf{that}, Point_2\rangle}$
$(\lambda k_s \otimes k_g.k'_s \otimes k'_g(k_s \otimes k_g r_s \otimes r_g)))$

Multimodal Dialogue

- Gesture sequence command

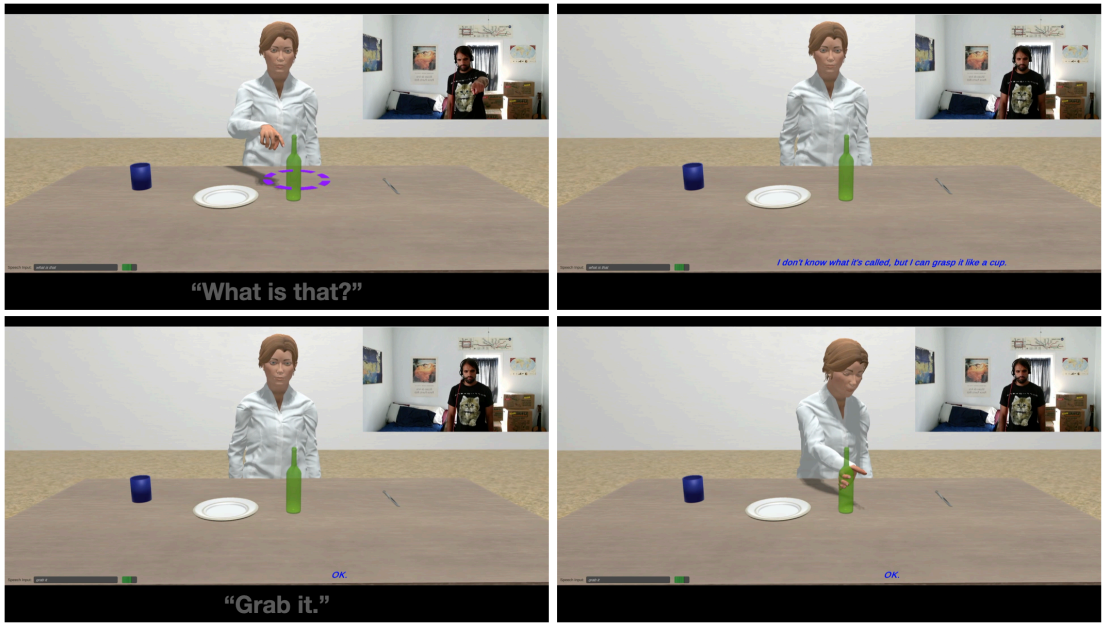
Single Modality (Gesture) Imperative

DIANA_1: $\mathcal{G} = [\textit{points to the purple block}]_{t1}$
DIANA_2: $\mathcal{G} = [\textit{makes move gesture}]_{t2}$
DIANA_3: $\mathcal{G} = [\textit{points to the blue block}]_{t3}$

Krishnaswamy, N., and Pustejovsky, J. (2018). Deictic Adaptation in a Virtual Environment. In *German Conference on Spatial Cognition* (pp. 180-196). Springer, Cham.

Spatial Reasoning and Affordance Learning

- Gibsonian/Telic affordances are associated with abstract properties:
  - spheres roll, sphere-like entities probably do too;
  - small cups are graspable, small cylindroid-shaped objects probably are too.
- Similar objects have similar habitats/affordances:
- This informs the way you can talk about items in context:
  - Q: "What am I pointing at?"
  - A: "I don't know, but it looks like {a ball/a container/etc.}



- Train over a sample of 17 different objects: blocks, KitchenWorld objects (apple, grape, banana, book, etc.)
- Trained 200 dimensional affordance and habitat embeddings using a Skip-Gram model, for 50,000 epochs with a window size of 3:
  - These embeddings serve as the inputs to the object prediction architectures
- Using the affordance embeddings in vector space, predict which object they belong to: using a 7-layer MLP; a 4-layer CNN with 1D convolutions