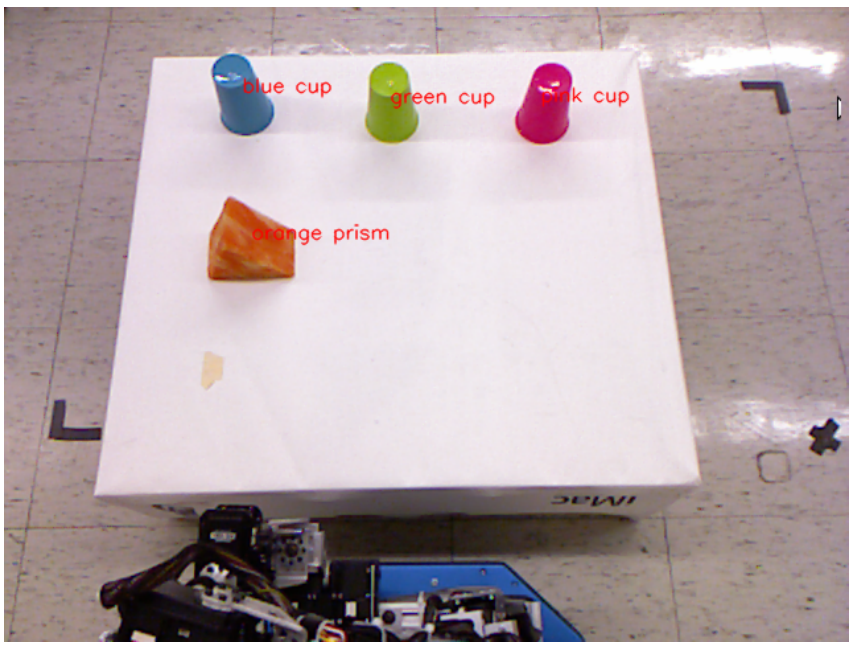
A generative probabilistic framework for learning spatial language

Colin Dawson, Jeremy Wright, Antons Rebguns, Marco Valenzuela Escárcega, Daniel Fried and Paul Cohen

University of Arizona
School of Information: Science, Technology, and Arts

August 20, 2013

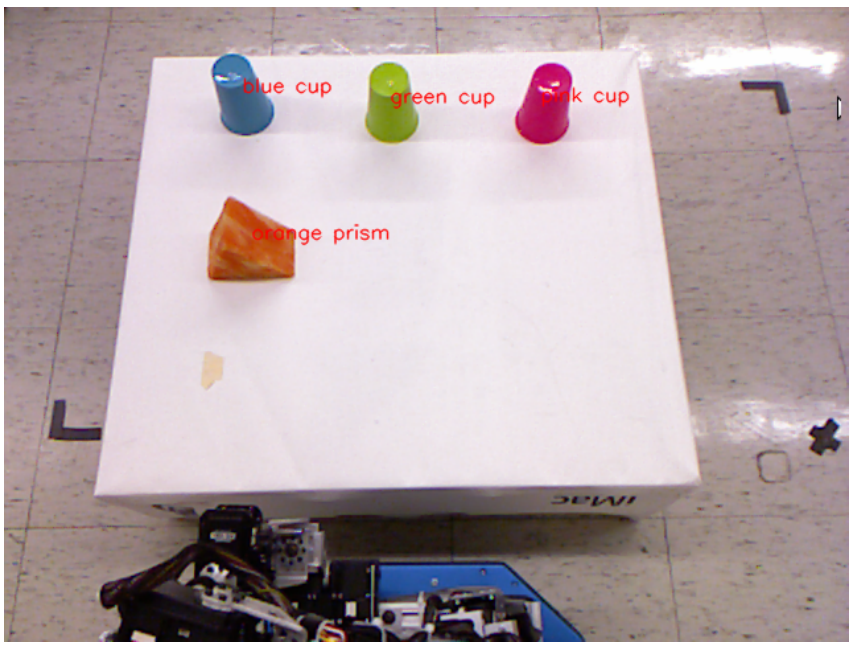
The Targets of Understanding



- What's involved in understanding utterances like (S_1) : "the blue cup in the corner of the table"

  in the context of a physical environment, π?

The Targets of Understanding



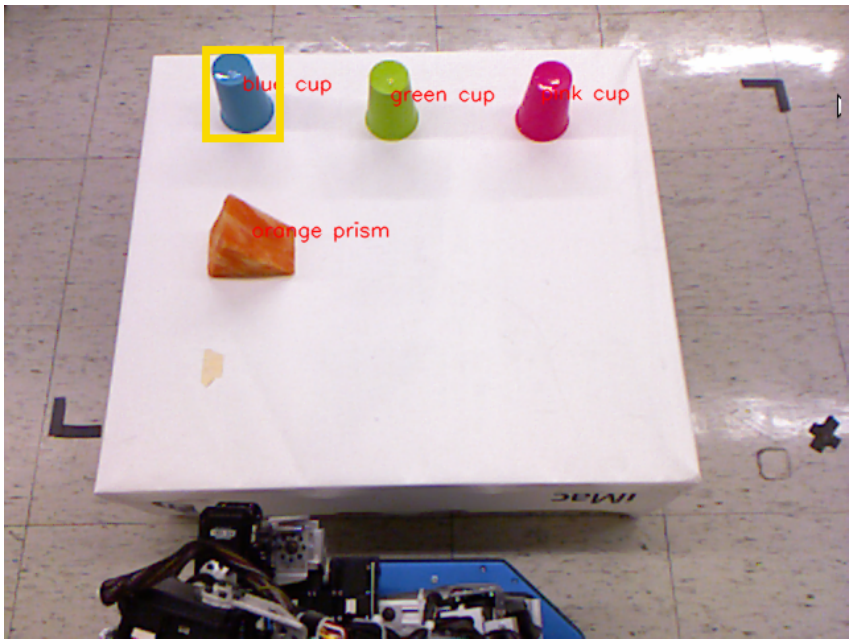
- What's involved in understanding utterances like

  (S_1) : "the blue cup in the corner of the table"

  (S_2) : "the one at the end of the line of cups"

  in the context of a physical environment, π?

The Targets of Understanding



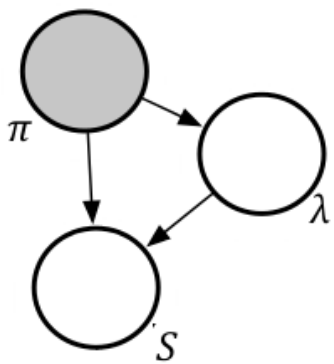
- Goal: identify a **referential intent**, λ, which is grounded in the scene
- e.g., a particular blue cup centered at coordinates (x, y)

Prediction as Understanding

- π : the scene itself, prior to constructing representations;
- λ : referential intent ; objects and spaces ; the orange wedge, an edge;
- S : sentences ; "The nearest one", "the orange one in the foreground";

shallow discriminative model: $\textsc{Predict}(S, \pi) \rightarrow \lambda$

Evaluating Meanings



Key

π	abstract perceptual representation
λ	referential intent (object/location)
S	word sequence

- We take the approach that the best referent, λ^*, is the one that does the best job of *explaining* the use of the utterance, S, in the physical context, π (so far treated as given).
- That is, we want a referential intent, λ (e.g., a particular blue cup) that satisfies

$$\lambda^* = \arg\max_{\lambda} P(\lambda|S,\pi)$$
$$= \arg\max_{\lambda} P(S|\lambda,\pi)P(\lambda|\pi)$$

Advantages of This Approach

- Bayesian statistics allows uncertainty to be handled in a principled way.

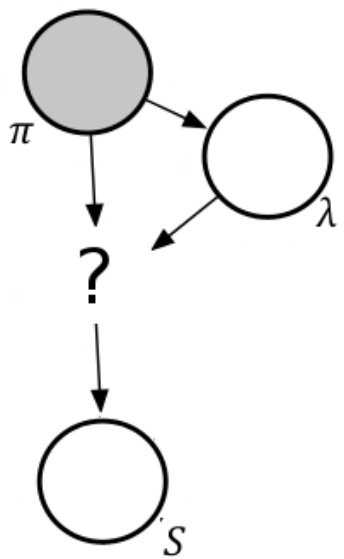
Advantages of This Approach

- Bayesian statistics allows uncertainty to be handled in a principled way.
- Counterfactual reasoning: "If the speaker meant x, she probably would have said s, but since she didn't, that's evidence against x".

Advantages of This Approach

- Bayesian statistics allows uncertainty to be handled in a principled way.
- Counterfactual reasoning: "If the speaker meant x, she probably would have said s, but since she didn't, that's evidence against x".
- Bidirectional model can generate utterances and identify locations

Beyond shallow meaning

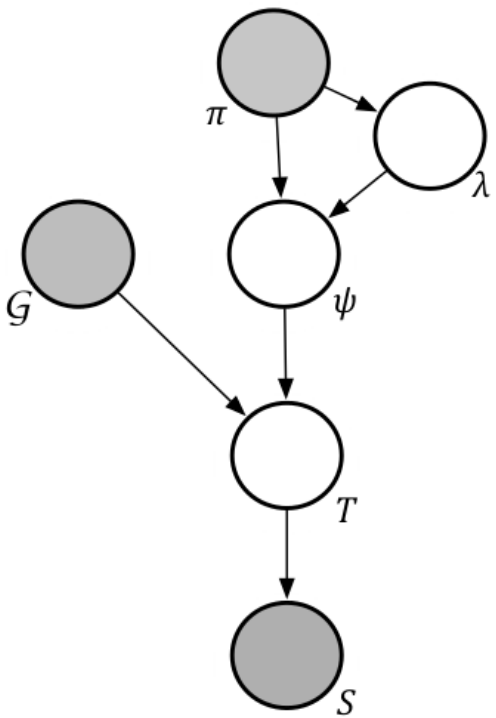


Key

π	abstract perceptual representation
λ	referential intent (object/location)
S	word sequence

- No clear way to transfer this "grounded meaning" to another domain.
- Introspection about our language understanding
- Rather than "sensors-to-sentences" we might try "sensors-to-schemas," where schemas have internal structure and role bindings.

Adding Semantics, ψ



- π : the scene itself, prior to constructing representations;
- λ : referential intent ; objects and spaces ; the orange wedge, an edge;
- ψ : meanings ; how referential intent is expressed; $\textsc{CloseTo}(\textsc{x},\textsc{me})$;
- S : sentences ; "The nearest one", "the orange one in the foreground"

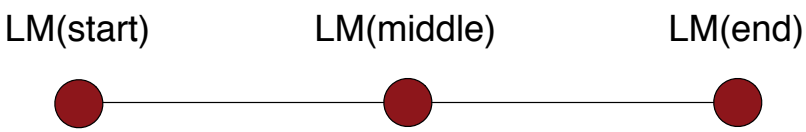
Adding Semantics, ψ

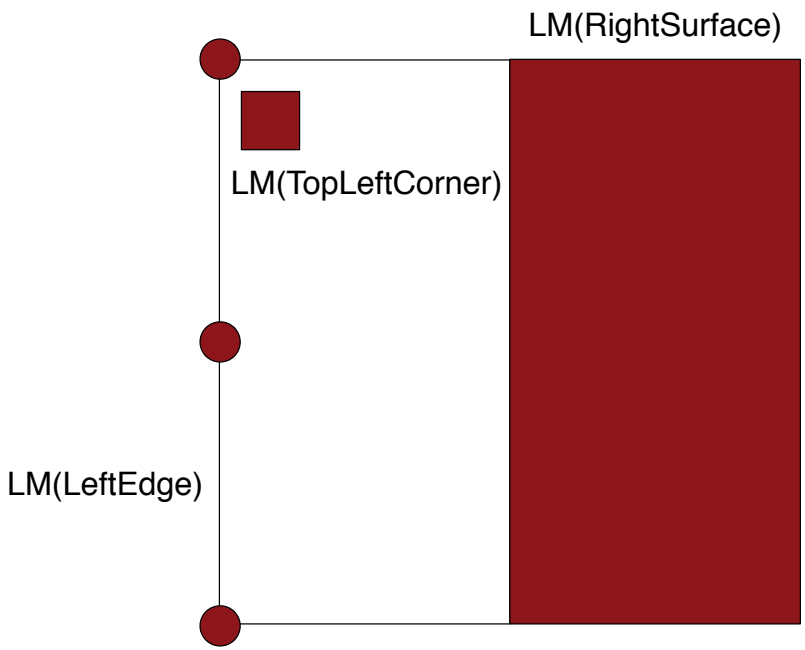
Physical configurations can be represented or **cast** as geometric schemas:

- `AbstractSchema`
  - `PointSchema`
  - `LineSchema`
    - `GroupLineSchema`
  - `RectangleSchema`
    - `SurfaceSchema`
    - `GroupRectangleSchema`
  - `CircleSchema`
  - `PolygonSchema`

Primitive Schemas and Landmarks

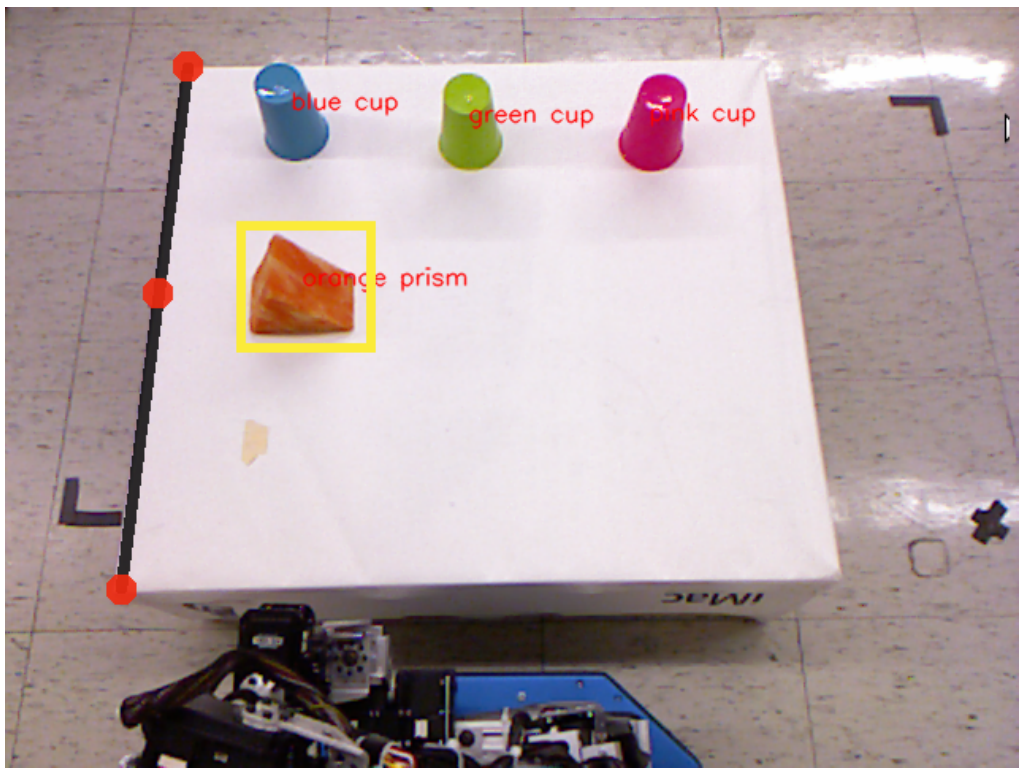
A schema has landmarks inherent to its shape, regardless of the real world object it represents. They can contain other landmarks, and can be coerced into different types.





Landmarks and Relations

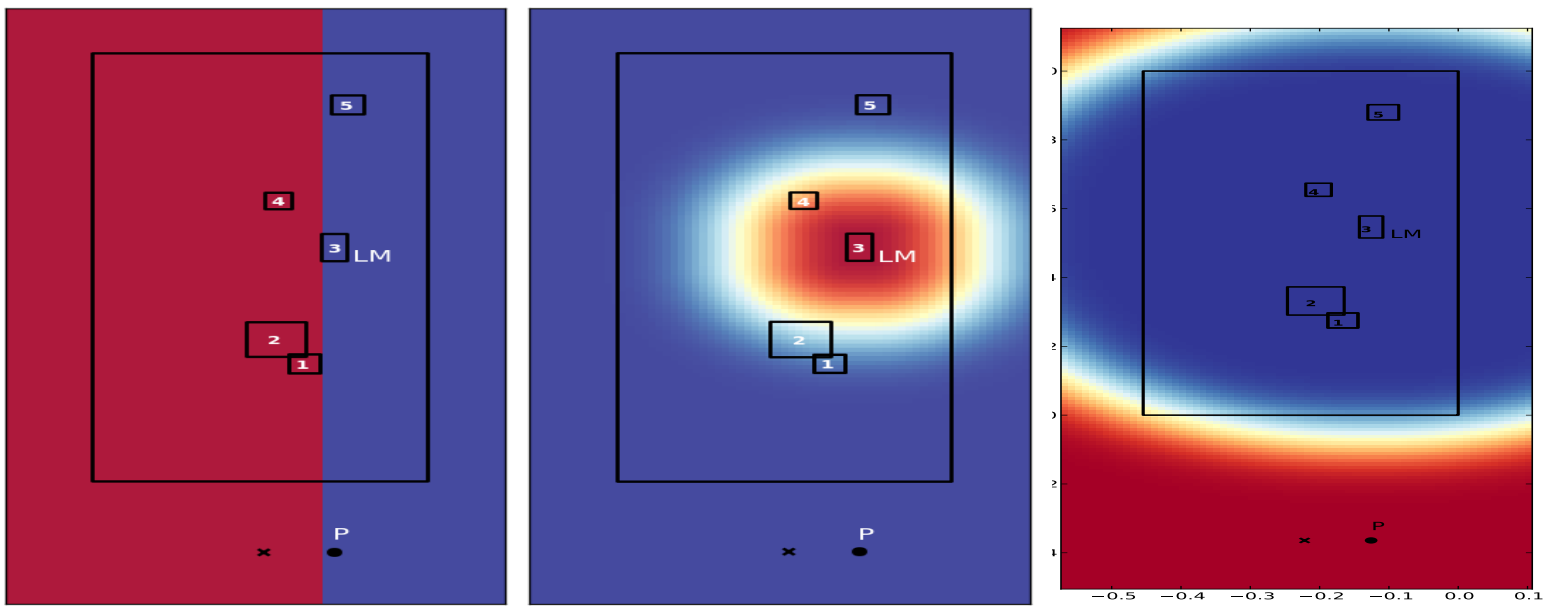
- ψ consists of a **landmark**, ψ_{lmk}, and a relation, ψ_{lmk}. We get a predicate, $\psi_{\text{rel}}(\cdot, \psi_{\text{lmk}})$ that can be applied to referents, λ.



ψ: NEAR-TO(λ, `Midpoint3`)

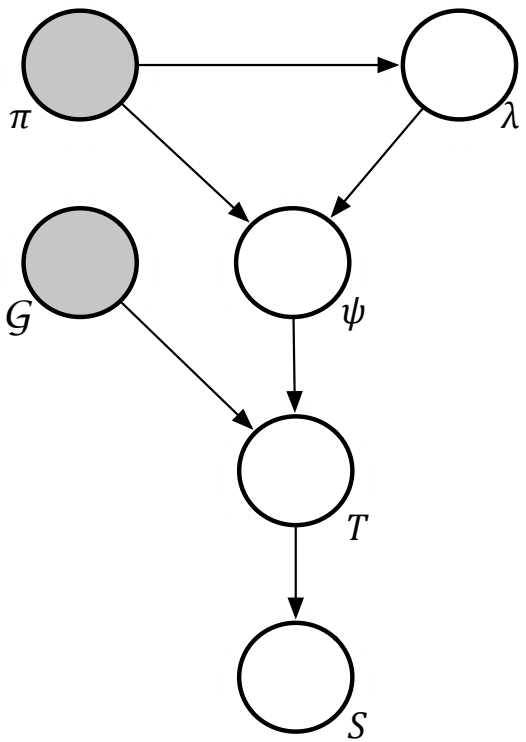
Applicability Heatmaps

Predicates are grounded using **applicability functions**, $A_{\text{rel}}(\lambda, \text{lmk})$, which take a referent, λ, and return a fuzzy truth-value for $\psi_{\text{rel}}(\lambda, \psi_{\text{lmk}})$.



(a) $A_{\text{Left-Of}}(\lambda, \texttt{Obj3})$ (b) $A_{\text{Near}}(\lambda, \texttt{Obj3})$ (c) $A_{\text{Far}}(\lambda, \texttt{Obj3})$

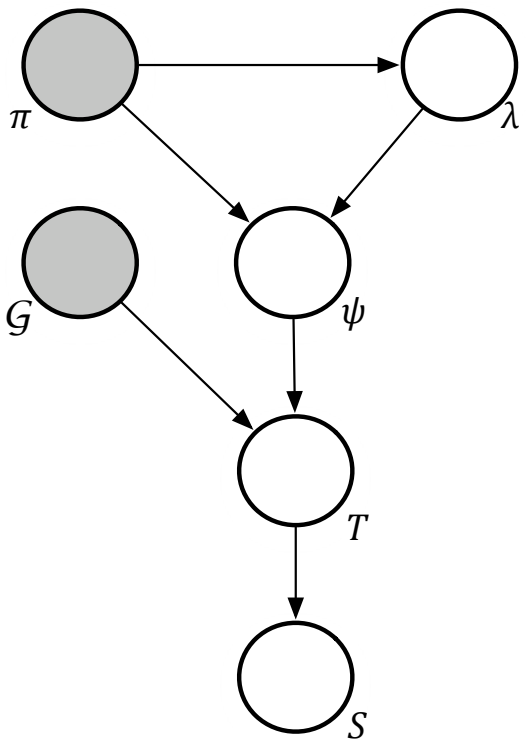
Distance (Near, Far), Orientation (Left-Of, Right-Of,Front-Of, Behind) and Containment (In)relations



Key

π	abstract perceptual representation
λ	referential intent (object/location)
$\mathcal{G}$	grammar/lexicon
ψ	semantic tree (relations, landmarks)
T	syntactic tree (phrases, PoS, words)
S	word sequence

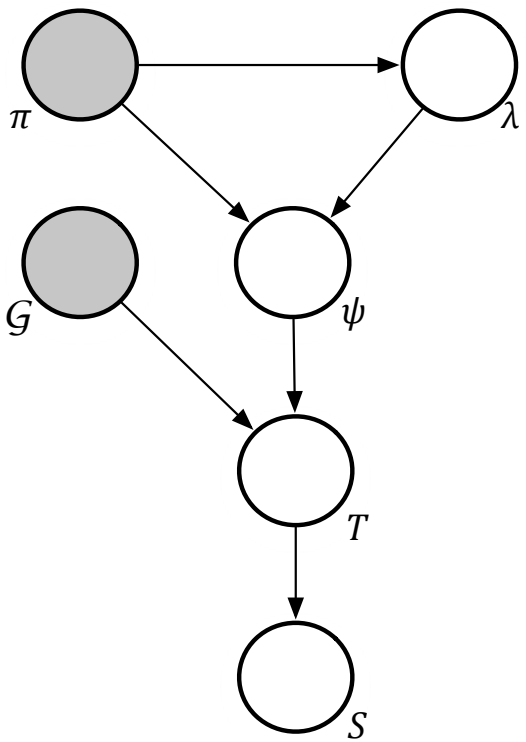
- Generation Process:
  1. Assuming a known scene, π, identify referential intent, λ (e.g., `Cup1`).



Key

π	abstract perceptual representation
λ	referential intent (object/location)
$\mathcal{G}$	grammar/lexicon
ψ	semantic tree (relations, landmarks)
T	syntactic tree (phrases, PoS, words)
S	word sequence

- Generation Process:
  1. Assuming a known scene, π, identify referential intent, λ (e.g., `Cup1`).
  2. Construct semantic representation, ψ, in terms of nearby landmarks, spatial relations to those landmarks, etc. (e.g., $\textsc{Near-To}(\lambda, \texttt{Corner3})$)



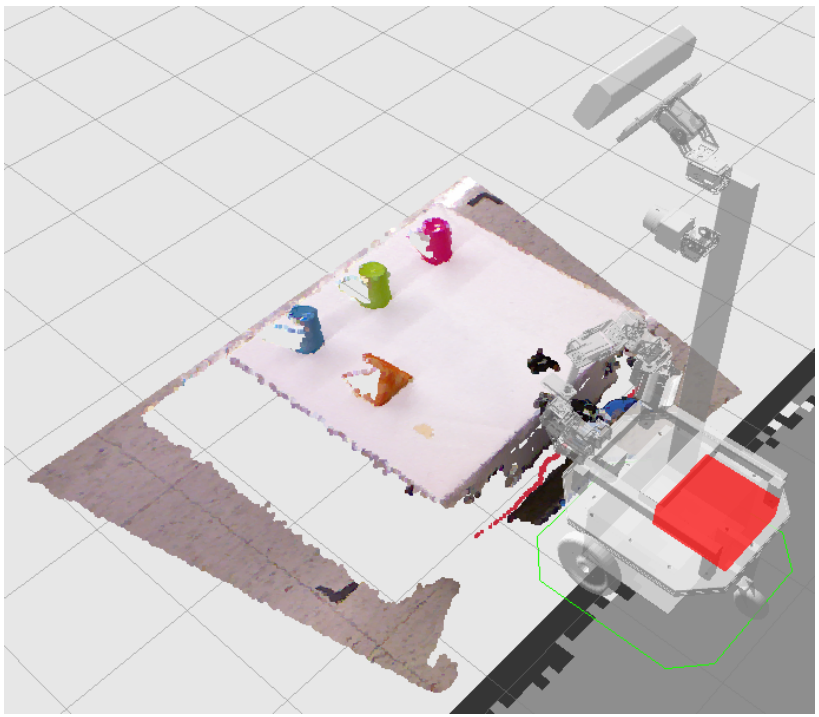
Key

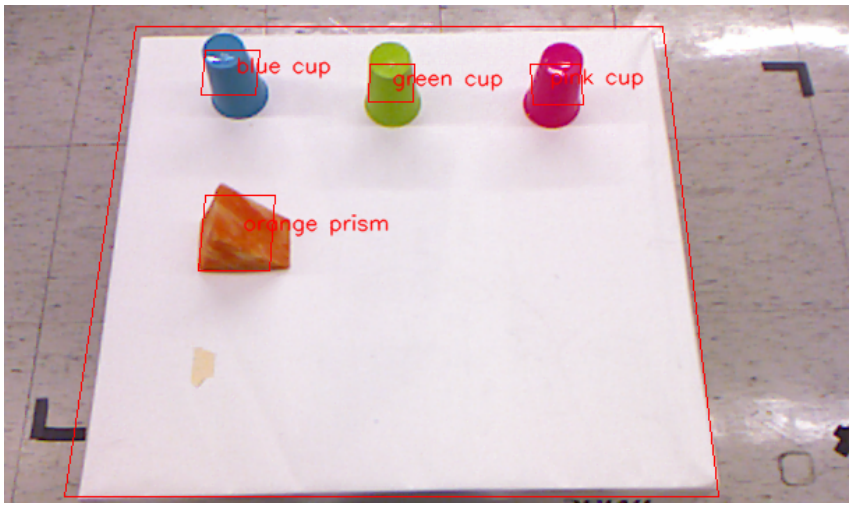
π	abstract perceptual representation
λ	referential intent (object/location)
$\mathcal{G}$	grammar/lexicon
ψ	semantic tree (relations, landmarks)
T	syntactic tree (phrases, PoS, words)
S	word sequence

- Generation Process:
  1. Assuming a known scene, π, identify referential intent, λ (e.g., `Cup1`).
  2. Construct semantic representation, ψ, in terms of nearby landmarks, spatial relations to those landmarks, etc. (e.g., $\textsc{Near-To}(\lambda, \texttt{Corner3})$)
  3. Using a probabilistic grammar $\mathcal{G}$, build a syntactic/lexical representation, T, and read off words, S.

"The blue cup in the far left corner of the table"

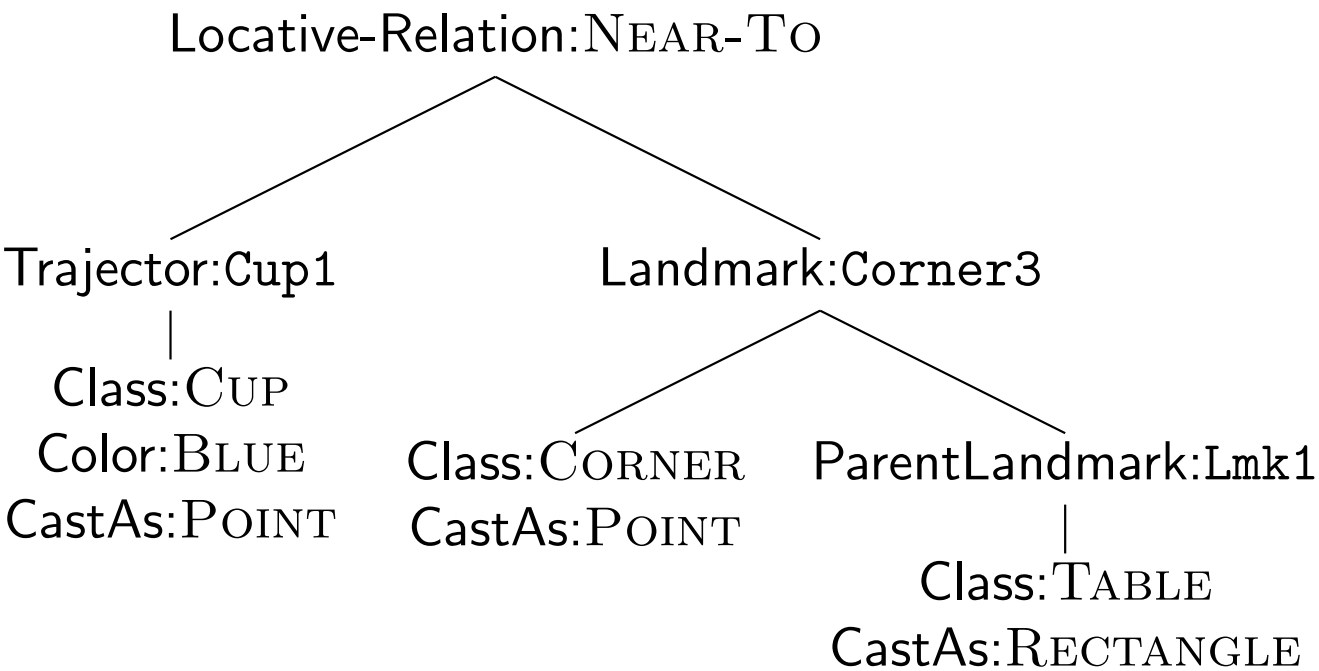
1. Construct an abstract representation of the scene (so far, this is a discriminative process for a small number of trained objects).





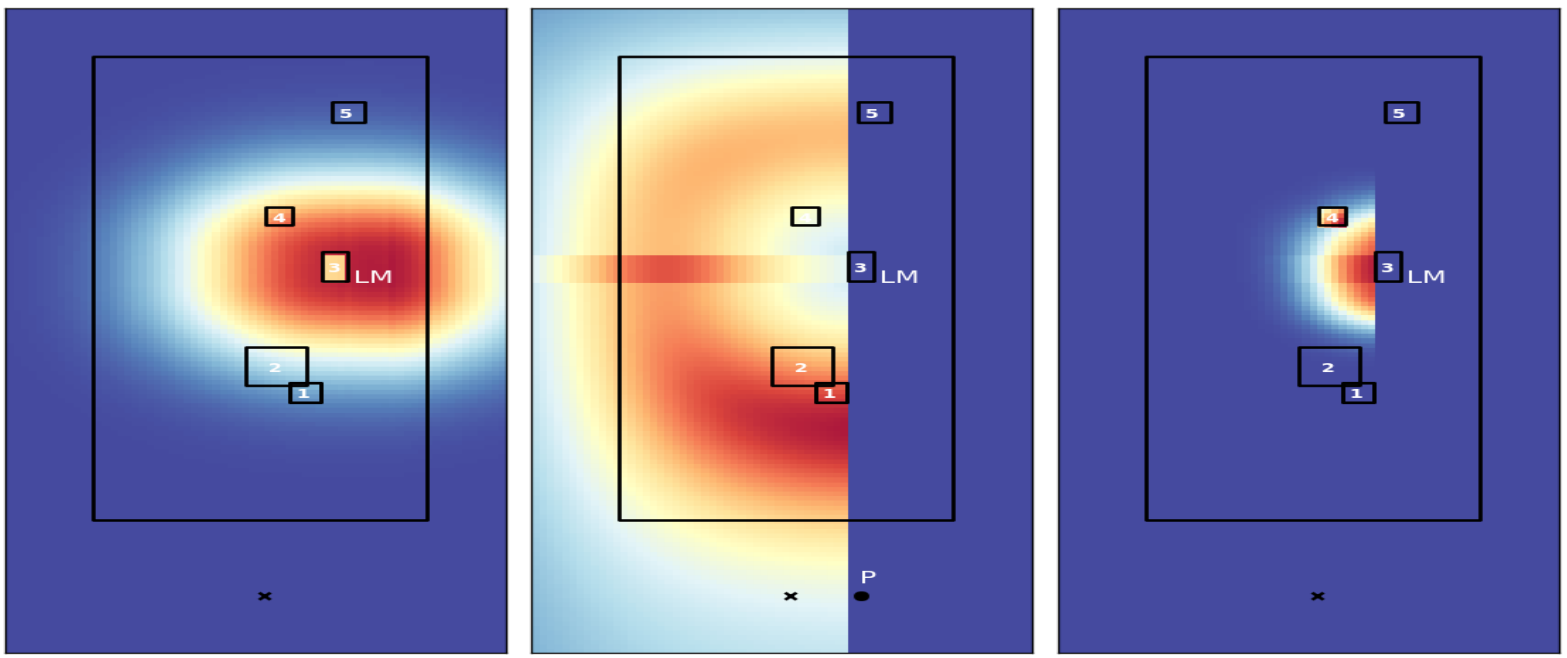
"The blue cup in the corner of the table"

2. Build a semantic "tree": in this case, sample a landmark and relation according to $P(\psi|\lambda, \pi)$.



Semantic Likelihood Calculation

- $P(\psi|\lambda,\pi) = P(\psi_{\text{lmk}}|\lambda,\pi)P(\psi_{\text{rel}}|\psi_{\text{lmk}},\lambda,\pi)$.
- $P(\psi_{\text{rel}}|\psi_{\text{lmk}},\lambda,\pi)$ obtained by normalizing each $A_{\text{rel}}(\lambda,\text{lmk})$

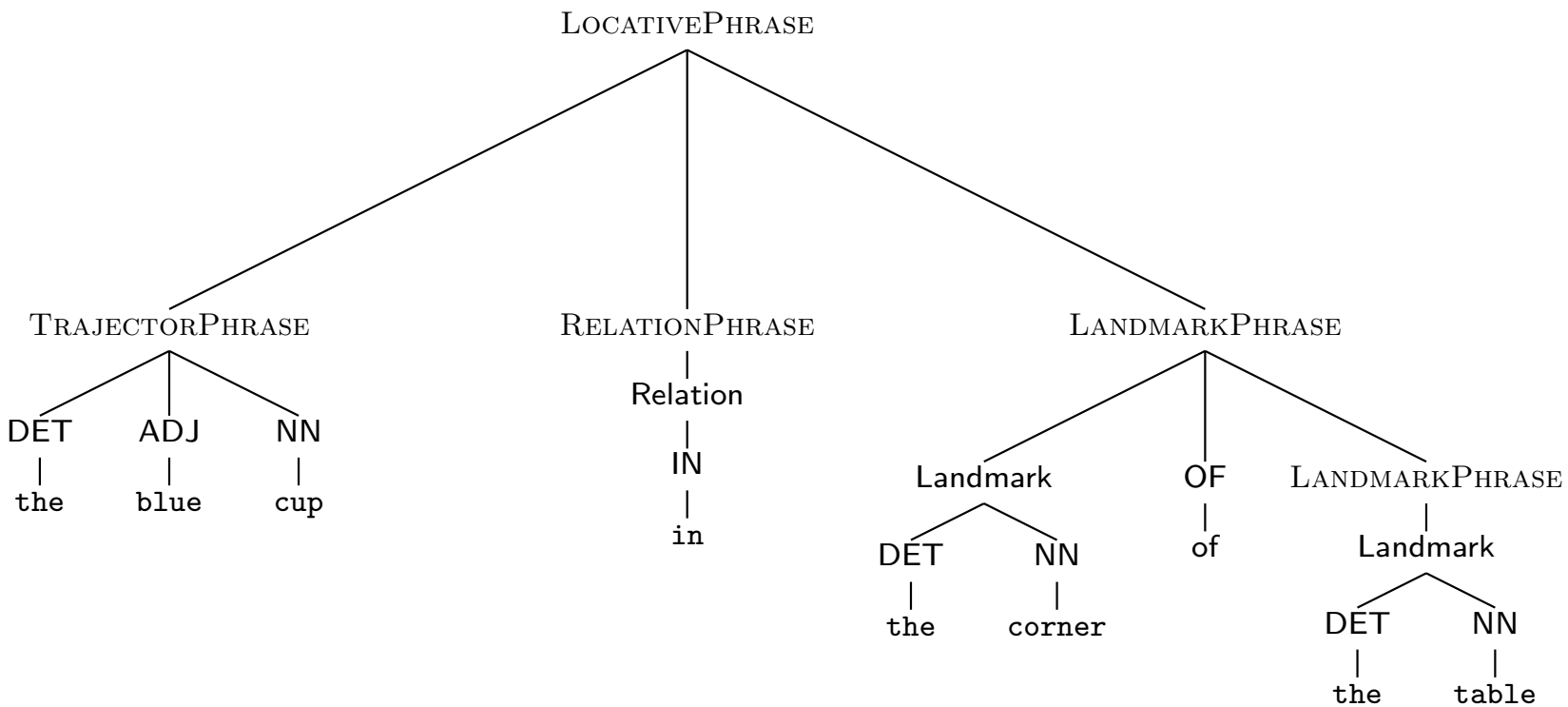


$P(\psi_{\text{lmk}}|\lambda,\pi)$ $P(\psi_{\text{rel}}|\lambda,\psi_{\text{lmk}},\pi)$ $P(\psi|\lambda,\pi)$

Likelihood calculation for $\psi_{\text{rel}} = \textsc{Left-Of}$, $\psi_{\text{lmk}} = \textsc{Object-3}$

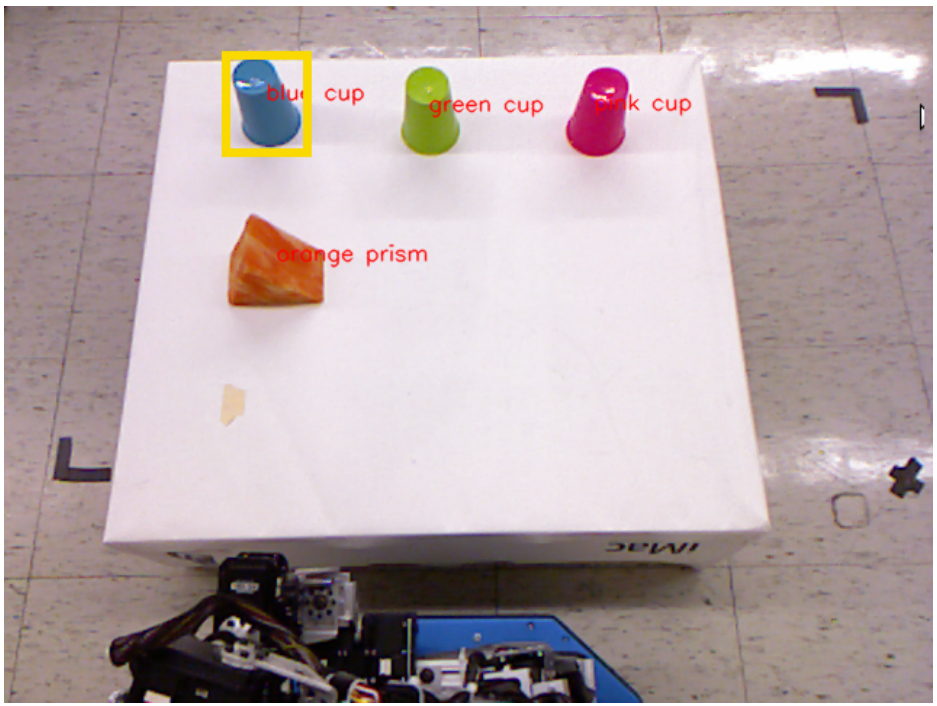
"The blue cup in the corner of the table."

3. Conditioned on the elements in the semantic tree, construct a linguistic representation according to $P(T|\psi, \mathcal{G})$.



4. Read off the tree to create the sentence, S

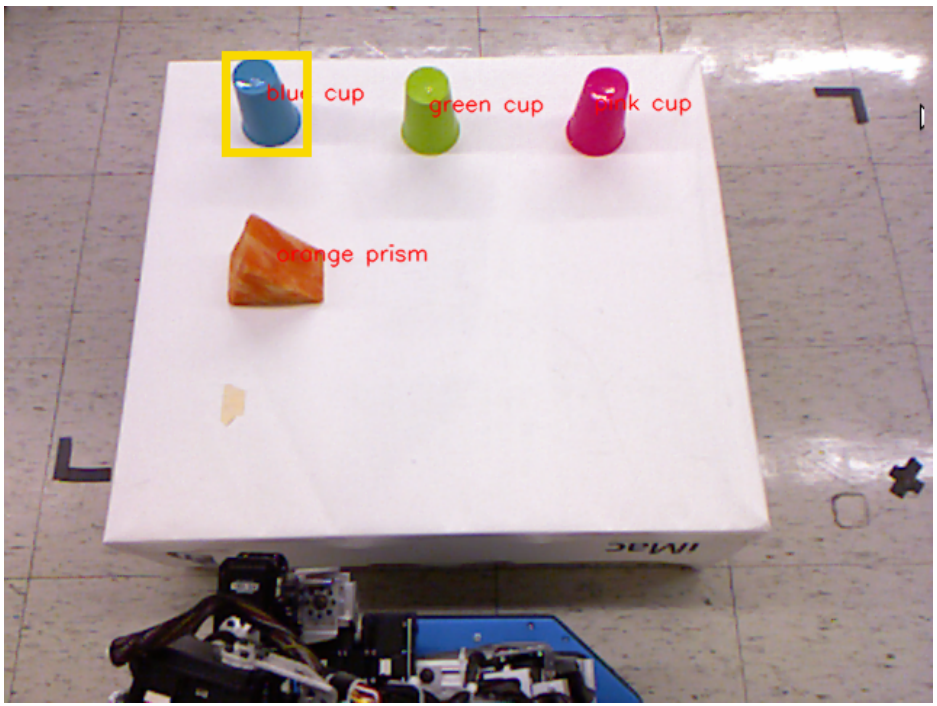
Training the Linguistic Model



- To learn about $P(T|\psi)$, need some instances of "grounded" sentences.

(λ is) ‘‘`in the corner of the table`

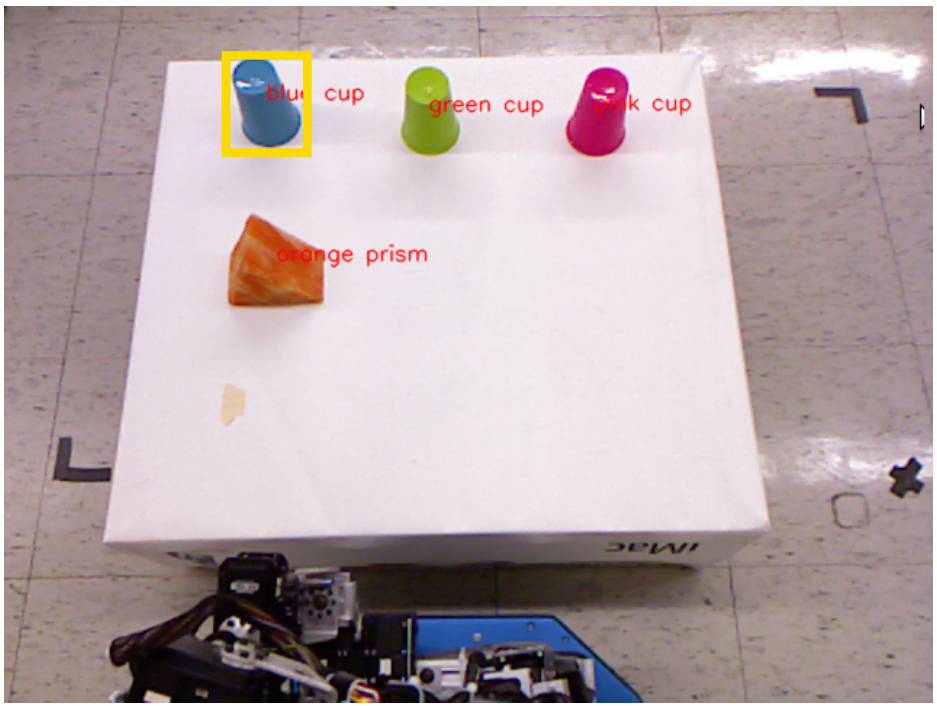
Training the Linguistic Model



(λ is) ‘‘in the corner of the table

- To learn about $P(T|\psi)$, need some instances of “grounded” sentences.
- Imagine a teacher pointing to an object or location in a scene, and producing an utterance. Result: (S_i, λ_i, π_i).

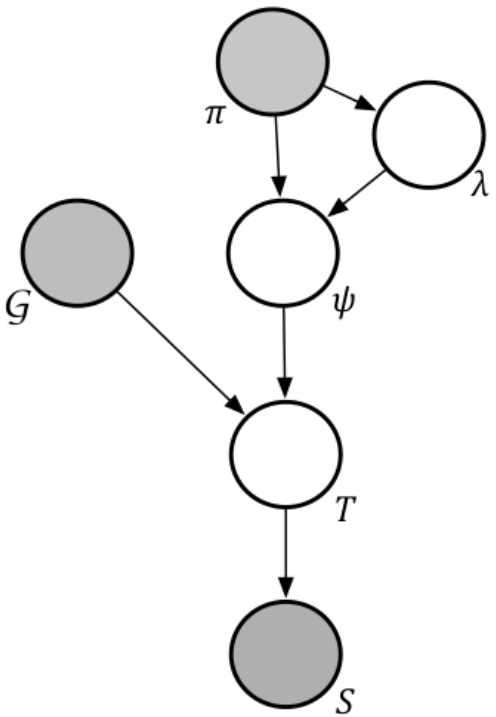
Training the Linguistic Model



(λ is) ‘‘`in the corner of the table`

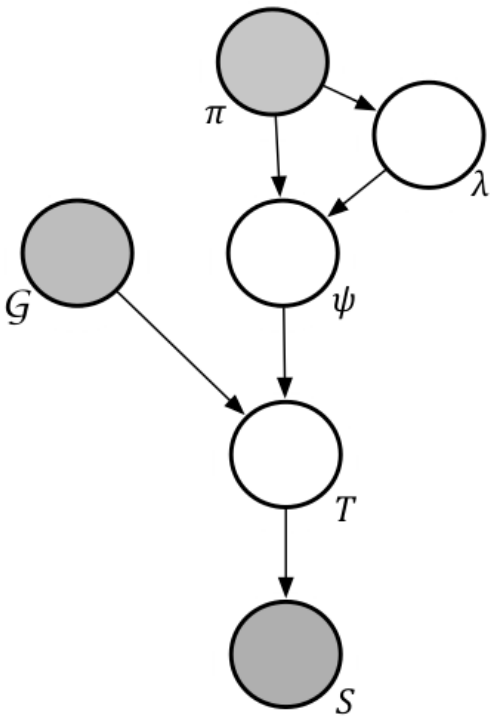
- To learn about $P(T|\psi)$, need some instances of “grounded” sentences.
- Imagine a teacher pointing to an object or location in a scene, and producing an utterance. Result: (S_i, λ_i, π_i).
- Parse the sentence, producing trees T, and model the probability $P(T|\psi)$ using syntactic and lexical features (PCFG and bigrams)

Training the Linguistic Model



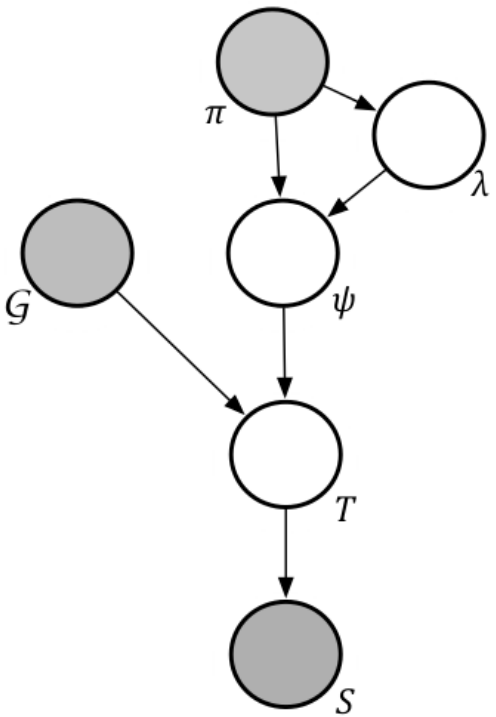
- How can we use $\{(S_i, \lambda_i, \pi_i)\}$ to learn the parameters of $\mathcal{G}$?

Training the Linguistic Model



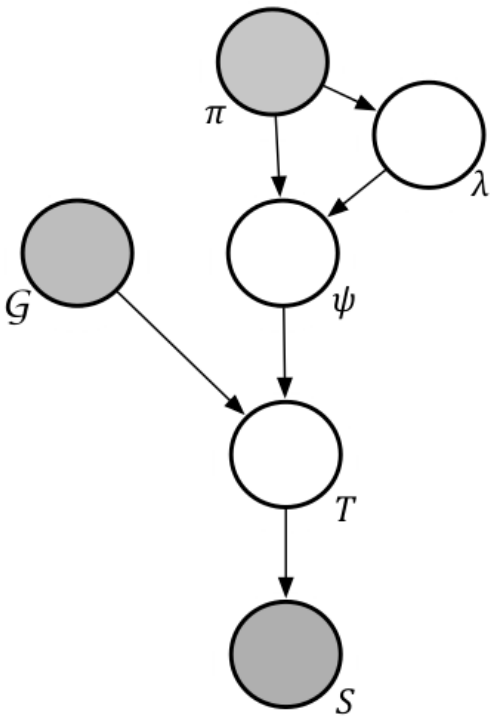
- How can we use $\{(S_i, \lambda_i, \pi_i)\}$ to learn the parameters of $\mathcal{G}$?
- If ψ were known: count cooccurrences between syntactic/lexical features and semantic features ψ and estimate conditional probabilities from the counts.

Training the Linguistic Model



- How can we use $\{(S_i, \lambda_i, \pi_i)\}$ to learn the parameters of $\mathcal{G}$?
- If ψ were known: count cooccurrences between syntactic/lexical features and semantic features ψ and estimate conditional probabilities from the counts.
- Sadly, ψ is hidden; but given a grounding model, we can construct a distribution, $P(\psi|\lambda, \pi)$.

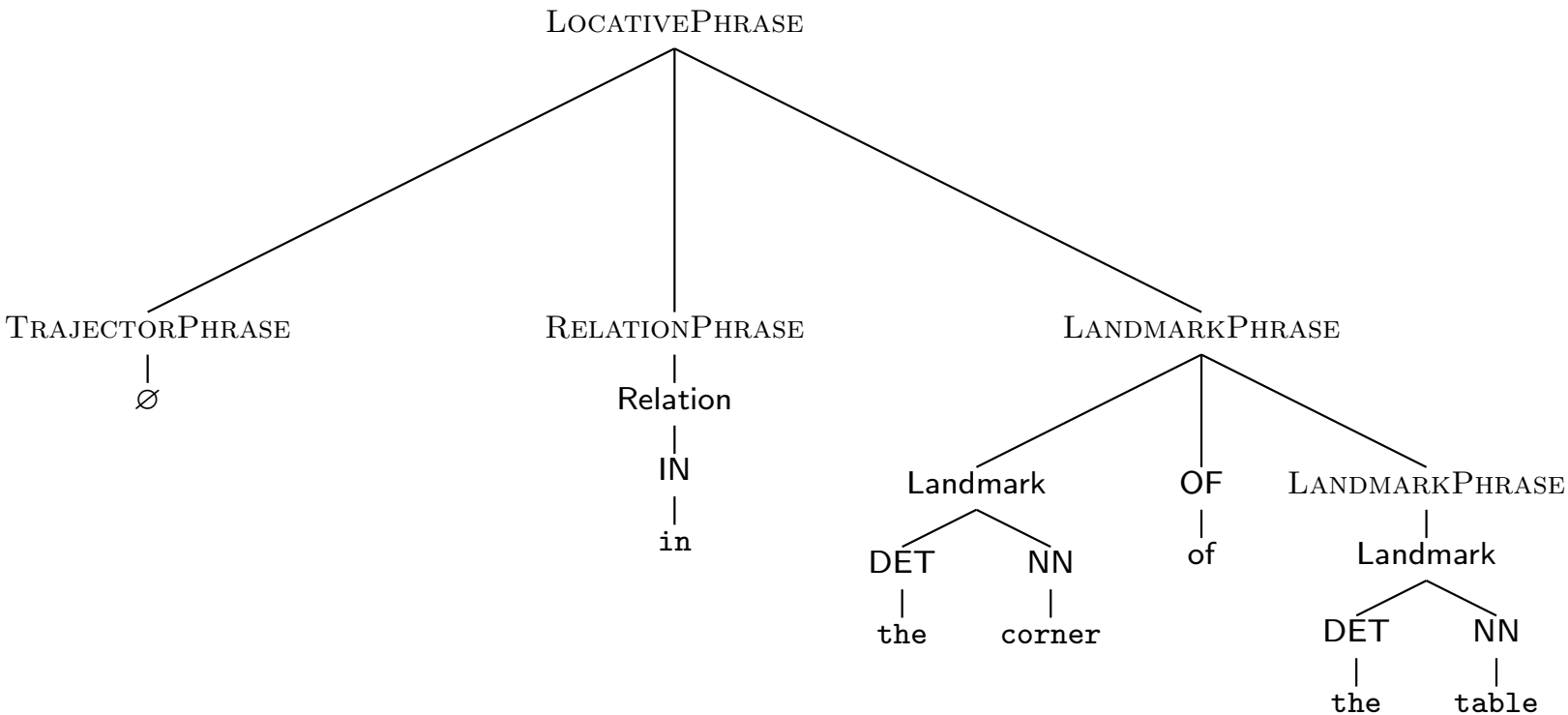
Training the Linguistic Model



- How can we use $\{(S_i, \lambda_i, \pi_i)\}$ to learn the parameters of $\mathcal{G}$?
- If ψ were known: count cooccurrences between syntactic/lexical features and semantic features ψ and estimate conditional probabilities from the counts.
- Sadly, ψ is hidden; but given a grounding model, we can construct a distribution, $P(\psi|\lambda, \pi)$.
- Then just construct a contingency table with counts weighted by $P(\psi|\lambda, \pi)$, smoothed

“in the corner of the table.”

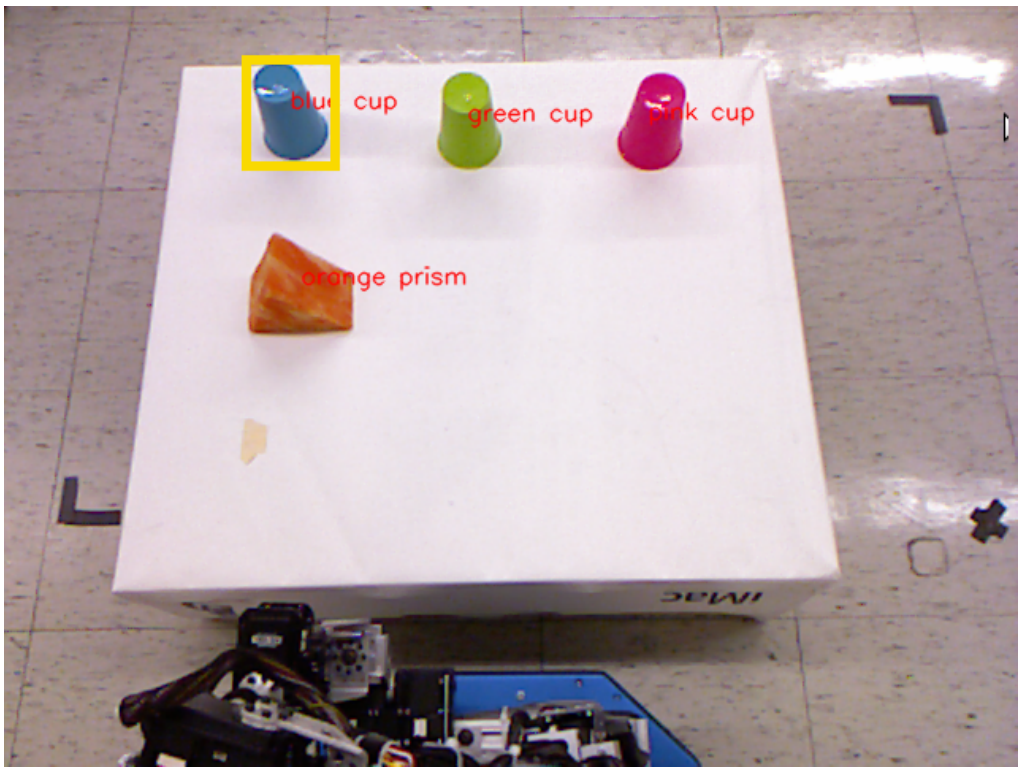
1. Parse the phrase.



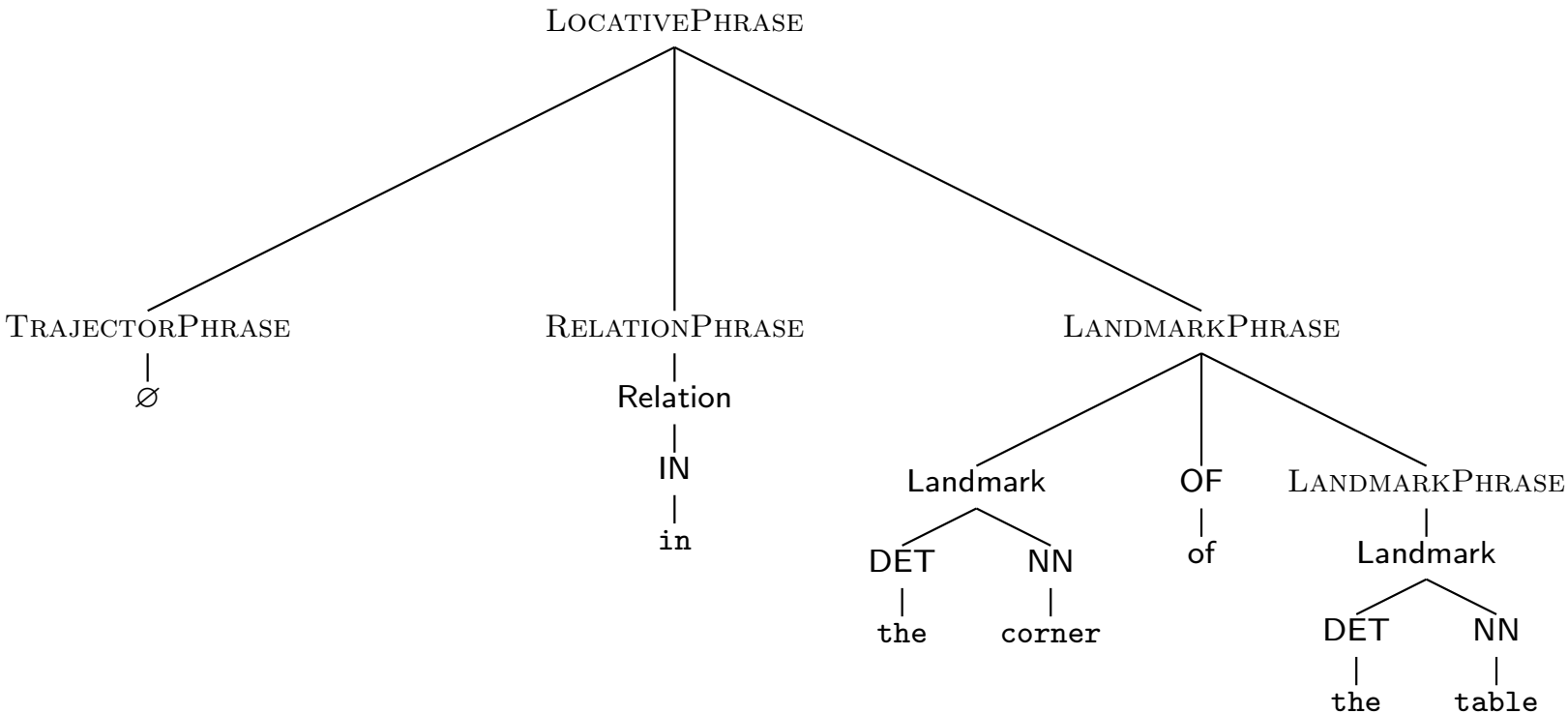
- CFG productions such as Landmark → DET NN and
- Words together with PoS tag and preceeding word

"in the corner of the table"

2. Given the referent, λ (in this case, just an (x, y) pair), sample possible landmarks and relations, e.g.,
$\psi_1 = \textsc{Near-To}(\lambda, \texttt{Corner3})$, $\psi_2 = \textsc{Right-Of}(\lambda, \texttt{Cup1})$



"in the corner of the table."



3. Add observations to a contingency table over ψ and syntactic features (CFG productions and bigrams).
4. Smooth this table to estimate $P(T|\psi, \lambda)$.

Sentence Interpretation

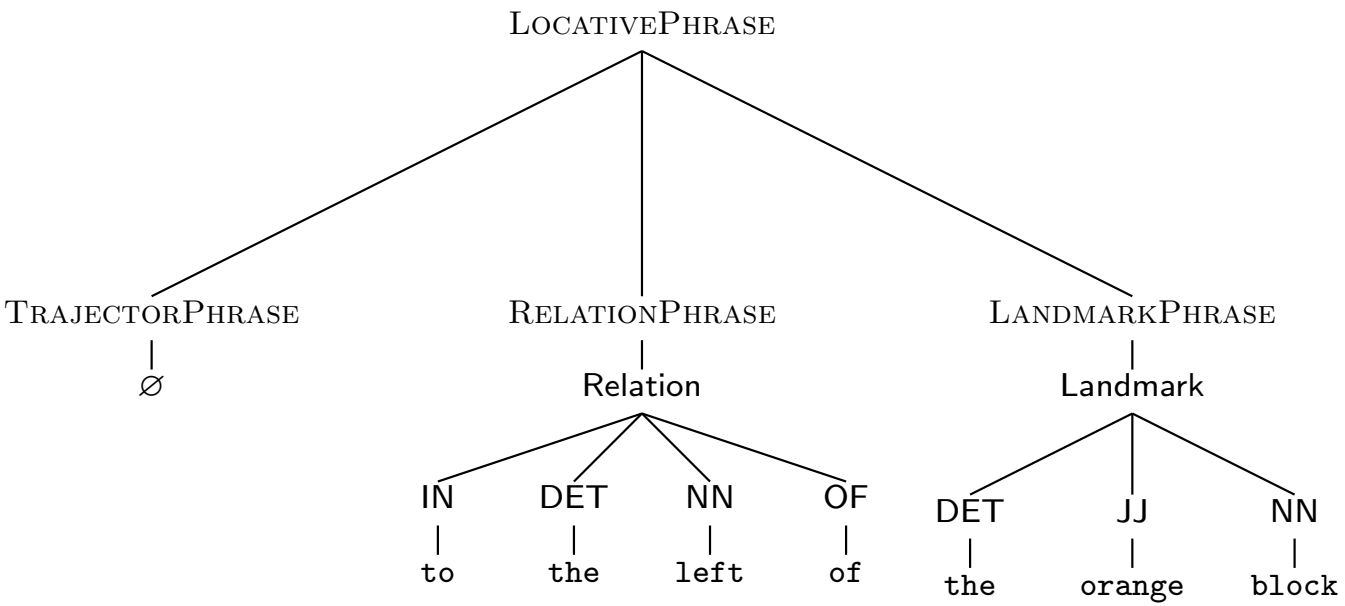
- Given a sentence about an (unknown) object/location in a (known) scene, data is S and π.

Sentence Interpretation

- Given a sentence about an (unknown) object/location in a (known) scene, data is S and π.
- Goal: infer λ (integrating out ψ and T)

"to the left of the orange block"

1. Parse the phrase.



2. Given a parse, T, evaluate ψs (e.g., $\textsc{Near-To}(\cdot, \texttt{Cup1})$) based on how well they explain T (i.e., compute $P(T|\psi)$ for each ψ consistent with the scene π).

"to the left of the orange block"

3. For each ψ_r, construct a heatmap over locations, λ, using $P(\psi_r|\lambda, \pi)$ over λ.

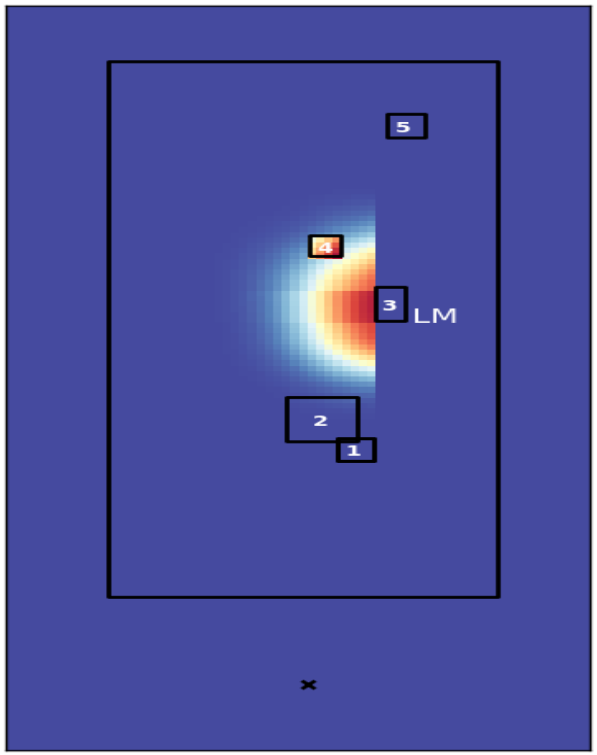
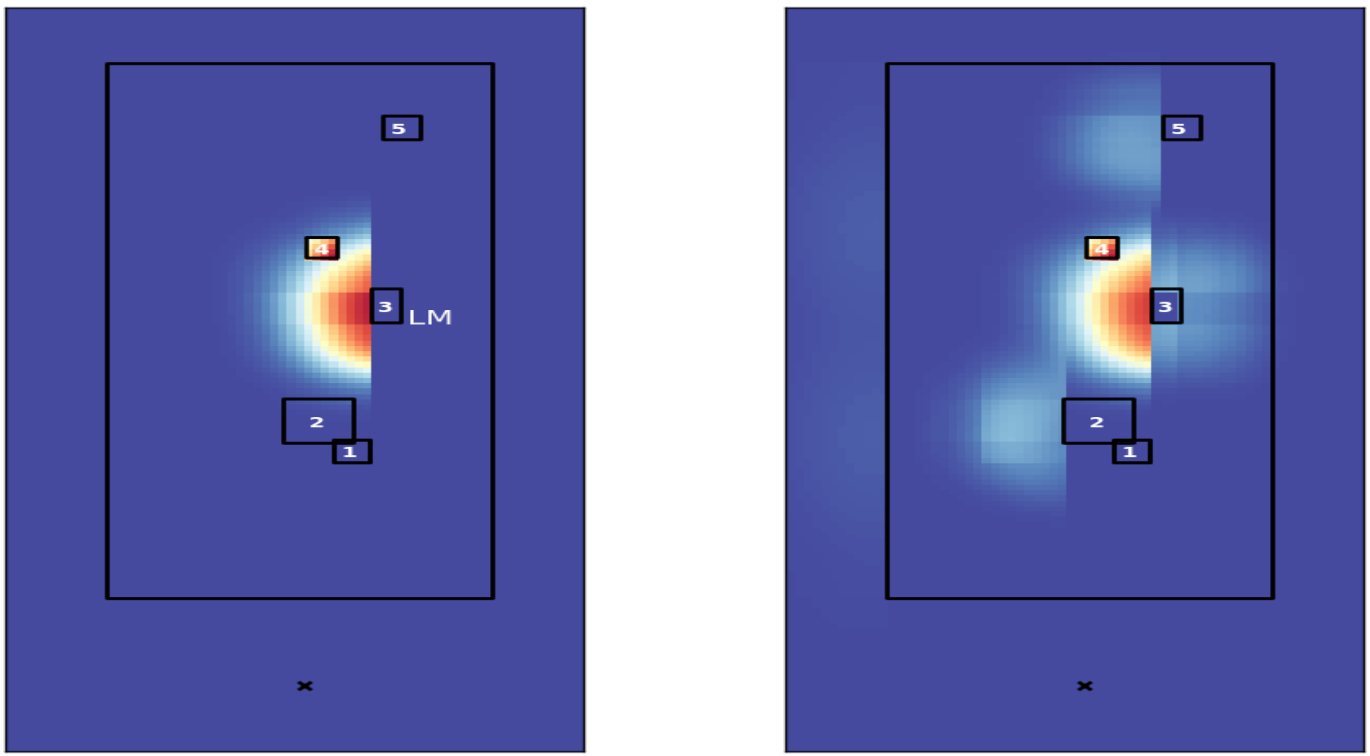


Figure: $P(\psi|\lambda, \pi)$ for $\psi = \textsc{Left-Of}(\lambda, \texttt{Obj3})$

"to the left of the orange block"

4. Integrate the heatmaps to get $P(T|\lambda, \pi)$.



(a) $P(\psi|\lambda, \pi)$ (b) $P(T|\lambda, \pi)$

Figure: Posterior for λ given the best parse of the phrase "To the left of the orange block". Left uses the "correct" ψ.

Experiment

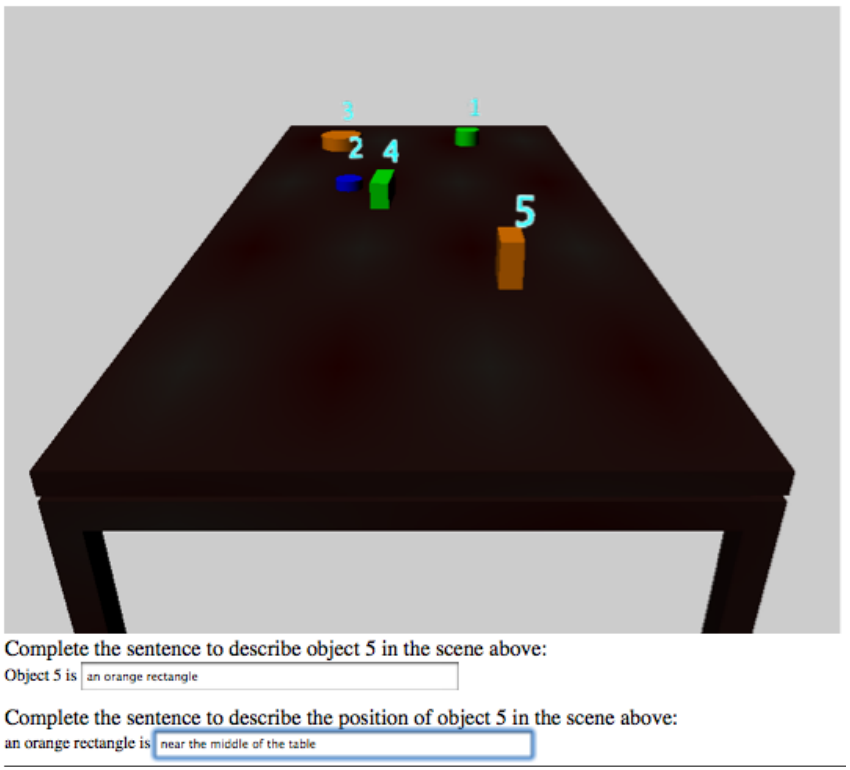


Figure: Display used to elicit sentences from human speakers

- Task: Train model on a corpus of phrases describing location of an object, elicited from human speakers (via Mechanical Turk).
- Perform a "forced choice" object selection task with held out test sentences.
- Obtain "ceiling" performance by giving the same forced choice task to human raters.

Turker Descriptions – Highly Variable

- `is middle of the tan and green`
- `on the right side of the table beyond the center of the table and between a rectangular and cylindrical block`
- `towards the back left side of the table close to the green box and orange cylinder`
- `at middle right table`
- `next to a cube`
- `is last`
- `in the middle of other objects`
- `the mideast side of the table`
- `at the center end of the table`
- `at the corner to your left`

Only about 50% (2113 of 4280 sentences) parse into a usable syntactic form

Results

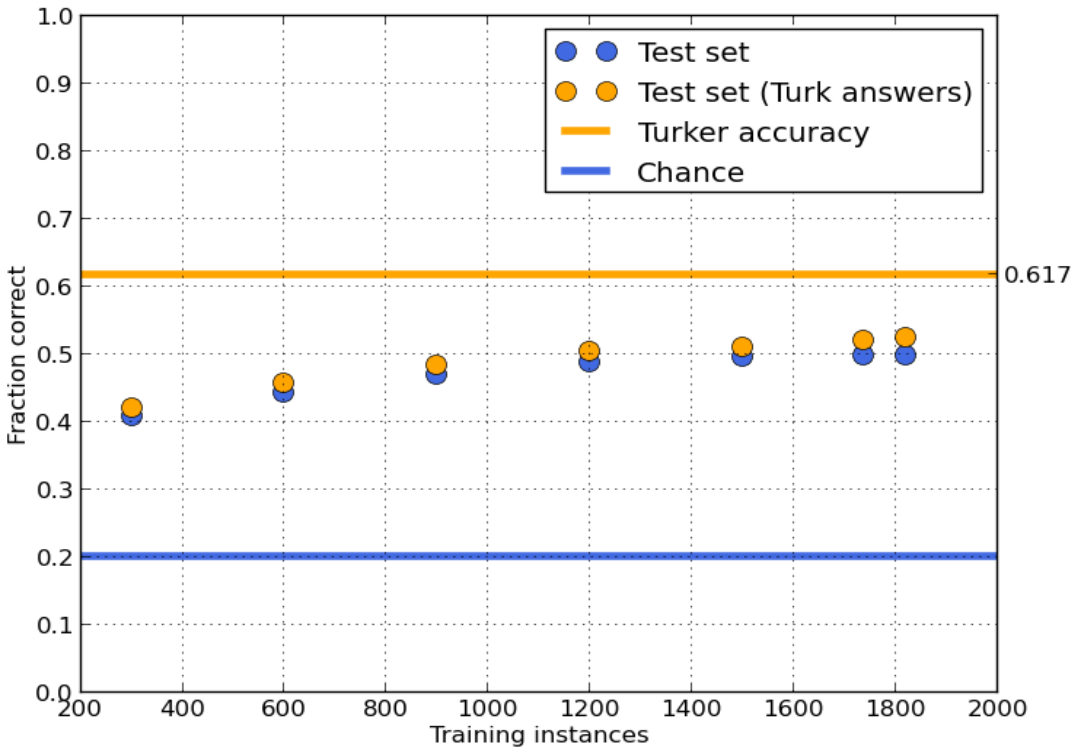


Figure: Results of forced choice object identification. Blue dots score answers as correct when the object that produced the sentence was identified. Gold dots count an answer as "correct" if the object was chosen by a plurality of humans.

Future Work

- Use a generative parser to better combine semantic and syntactic information
- More flexible grounded semantics (e.g., learn scale-dependent parameters for "near")
- Learning semantic grounding functions $P(\psi|\lambda, \pi)$

Thanks!