CMPT 983

Grounded Natural Language Understanding

March 4, 2021

Speaker listener models

Today

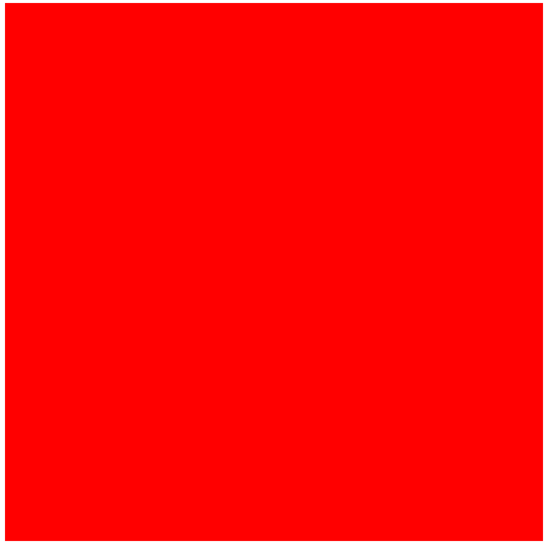
- Bayesian models for color

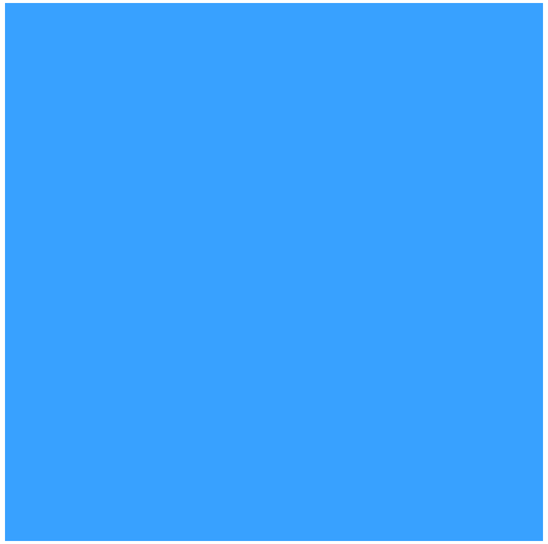
- Rational Speech Acts (RSA)

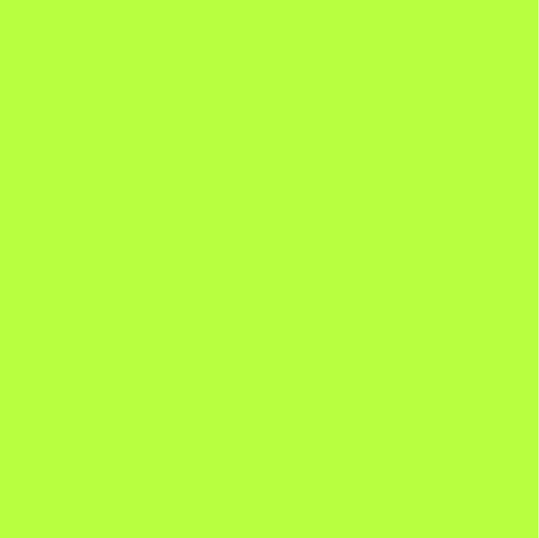
Colors

Color test

- What color is this?

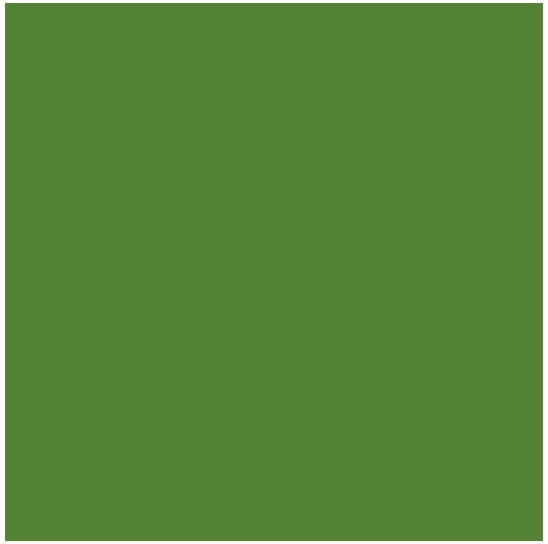


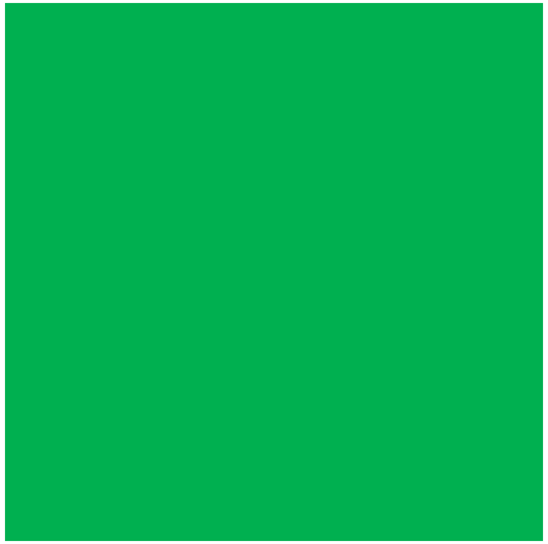


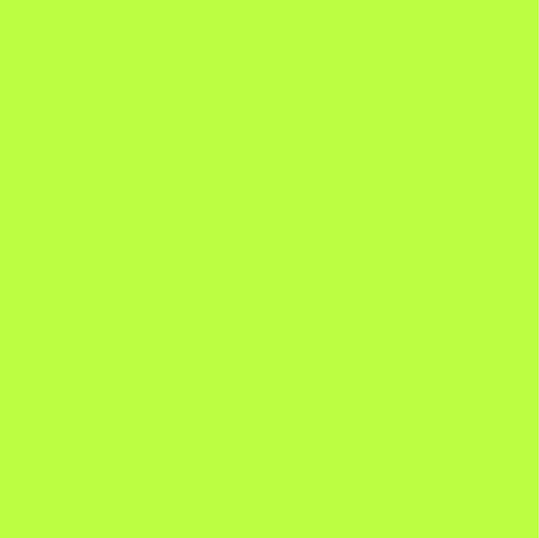


Color test

- What color is this?

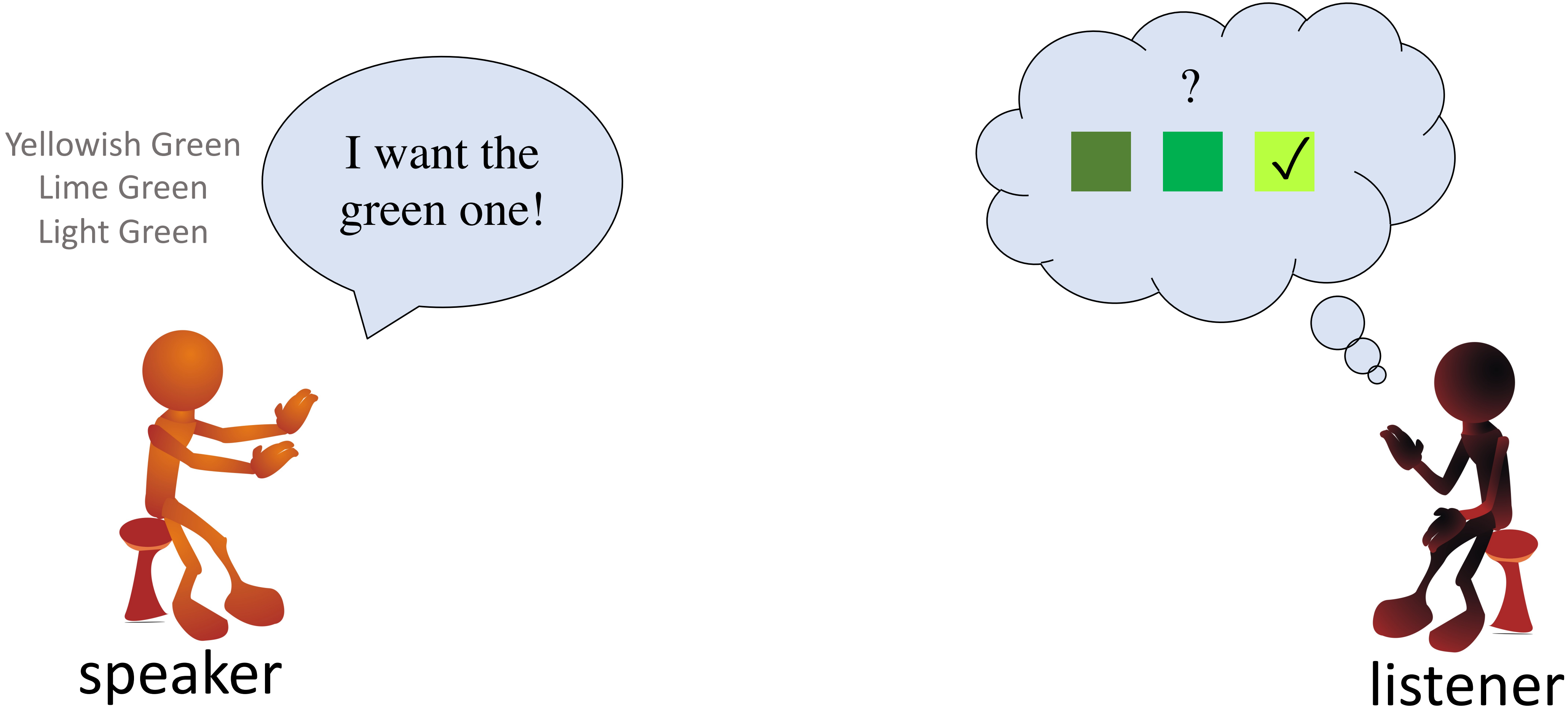






Effective communications

- What you say depend on context and what the listener knows.
- Want to select words that are informative, clear and unambiguous.



Gricean Maxims

Guidelines for cooperative, effective communication

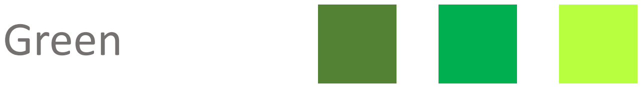
- Maxim of quantity: Give as much information as need, and no more
- Maxim of quality: Provide truthful information, supported by evidence
- Maxim of relation: Be relevant, say things pertinent to discussion
- Maxim of manner: Be clear, brief and orderly, avoid obscurity and ambiguity

To communicate clearly, we must have a shared convention of mapping of symbols to meanings.

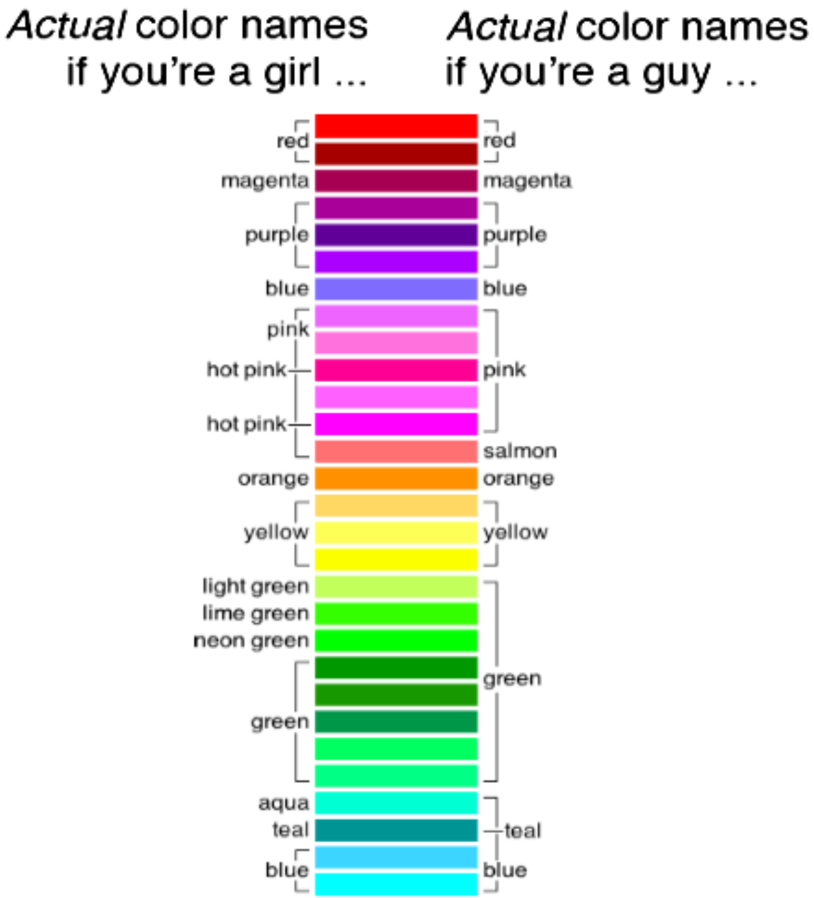
Grounding color

Is there a true mapping of words to a single meaning?

- Given the same word, will two listeners have the same interpretation?



- Given the same stimuli, will two speakers choose to use the same word?



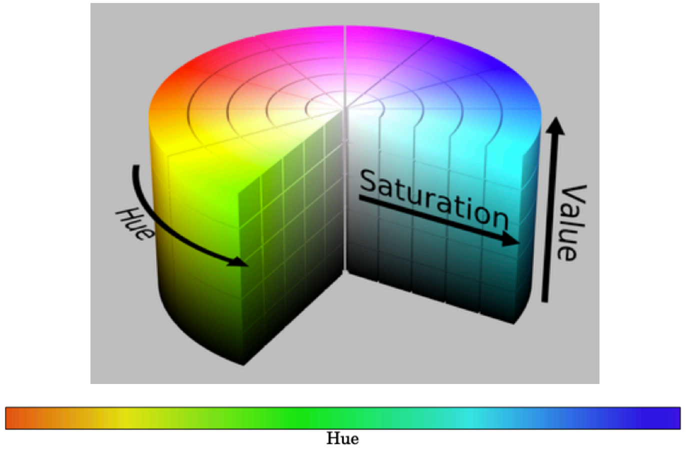
(XKCD color survey, Randall Munroe, https://blog.xkcd.com/2010/05/03/color-survey-results/)

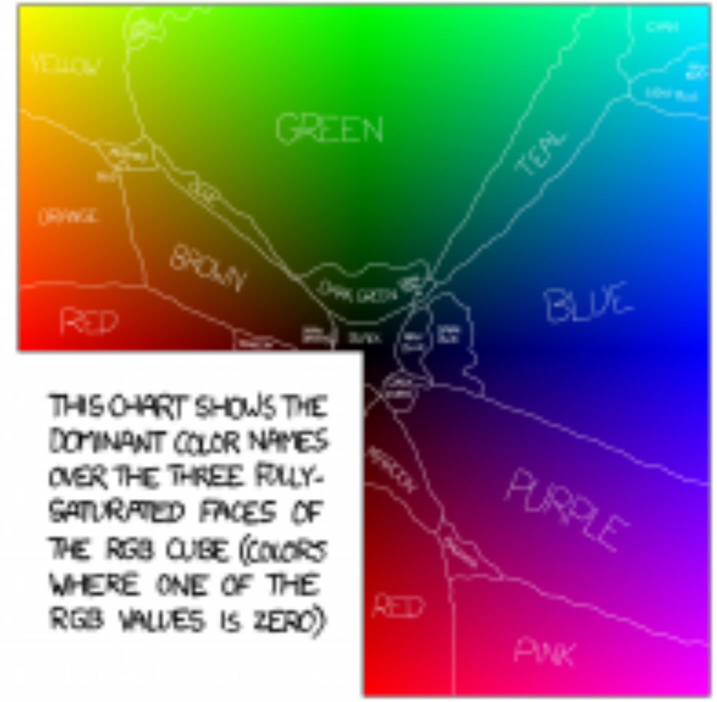
Grounding color

XKCD color survey

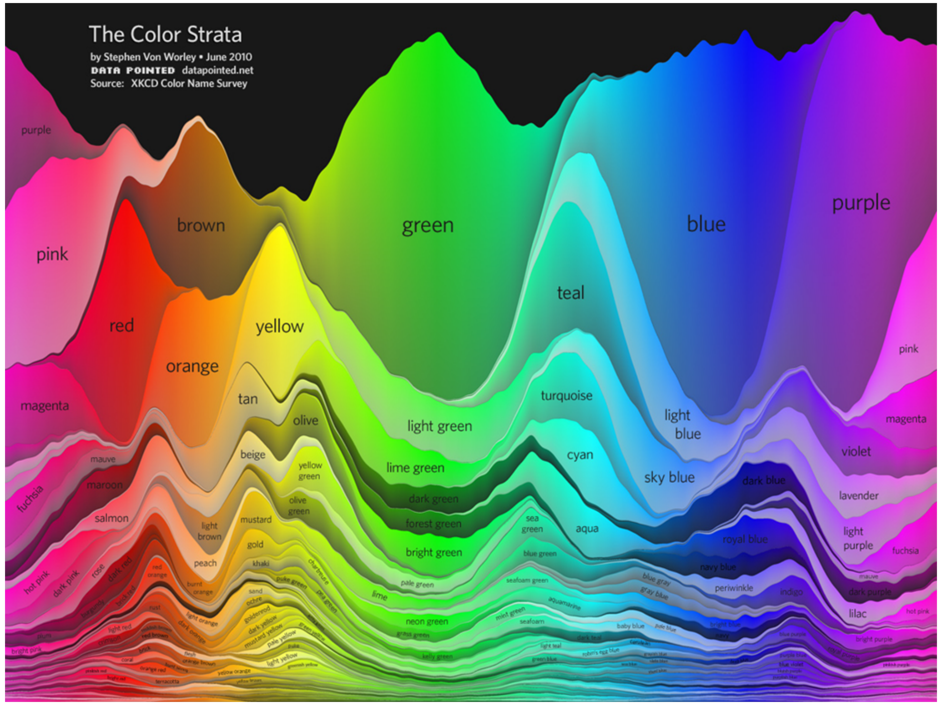
- Solicited names >5M random hues
- Got ~2.1M data points from >200K participants, with 829 distinct color names

HSV color space





THIS CHART SHOWS THE DOMINANT COLOR NAMES OVER THE THREE FULLY-SATURATED FACES OF THE RGB CUBE (COLORS WHERE ONE OF THE RGB VALUES IS ZERO)



The Color Strata

by Stephen Von Worley • June 2010

DATA POINTED datapointed.net

Source: XKCD Color Name Survey

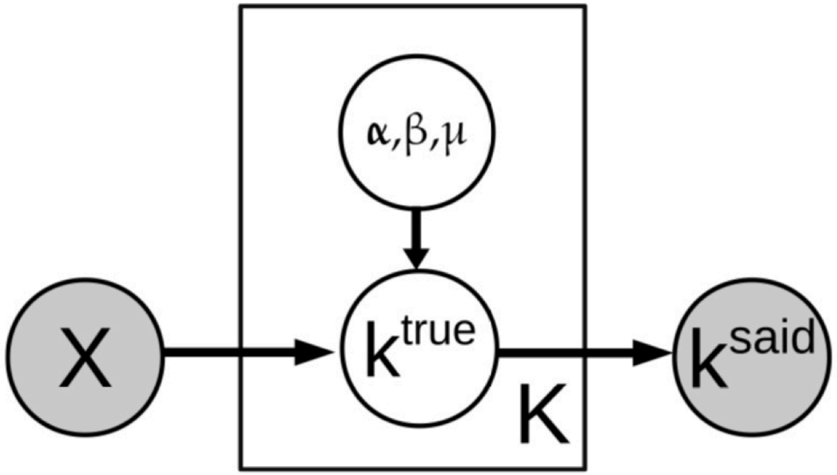
Let's use a probabilistic model!

(XKCD color survey, Randall Munroe, https://blog.xkcd.com/2010/05/03/color-survey-results/)

Grounding color

Bayesian model for grounded color semantics

- Model variation in meaning of words
- Given observed HSV color (X) and labels (k^{said}), how to learn a model of how to name colors?
- Speaker model: P(k^{said} | X)

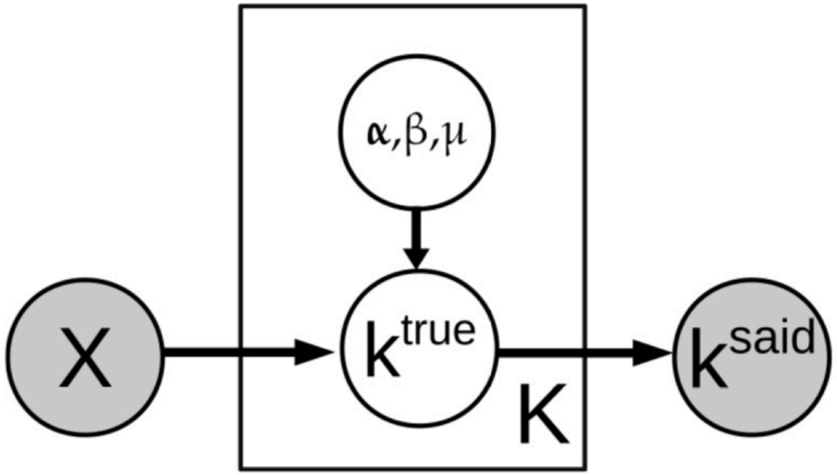


(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

Grounding color

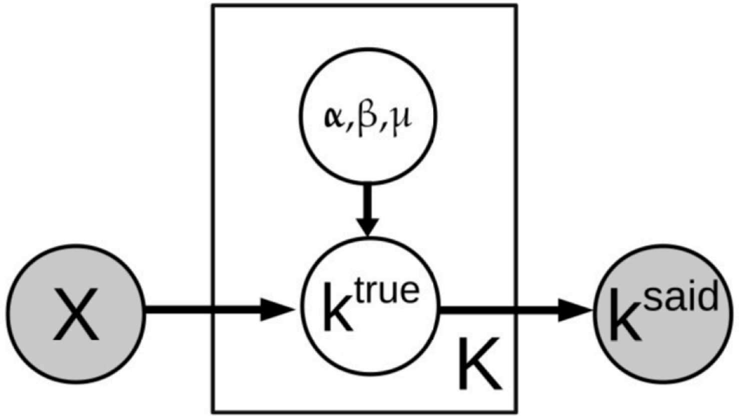
Bayesian model for grounded color semantics

- Model variation in meaning of words
- Given observed HSV color (X) and labels (k^{said}), how to learn a model of how to name colors?
- Speaker model: $P(k^{said} \mid X)$



(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

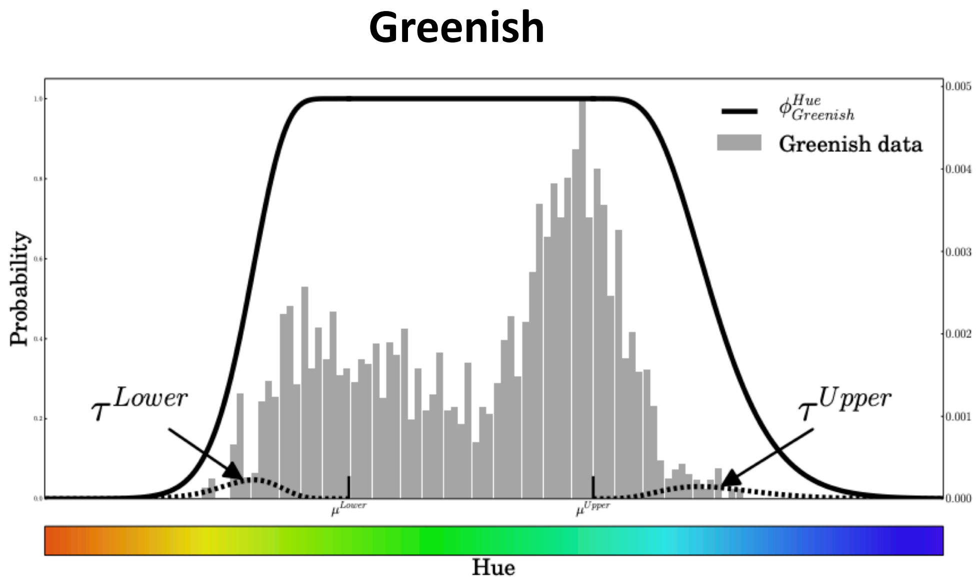
Grounding color



Bayesian model for grounded color semantics

- Model variation in meaning of words
- Model probability distribution of color being called a given name

Model color channel (HSV) referred to by a color name *k* as a noisy box with a lower and upper threshold



$$
\begin{aligned}
\tau_k^{Lower,d} &\sim \mu_k^{Lower,d} - \Gamma(\alpha_k^{Lower,d}, \beta_k^{Lower,d}) \\
\tau_k^{Upper,d} &\sim \mu_k^{Upper,d} + \Gamma(\alpha_k^{Upper,d}, \beta_k^{Upper,d})
\end{aligned}
$$

Thresholds follow a gamma distribution from the mean for each dimension $d \in \{H, S, V\}$

Parameters estimated to maximize the log-likelihood of the Munroe color data

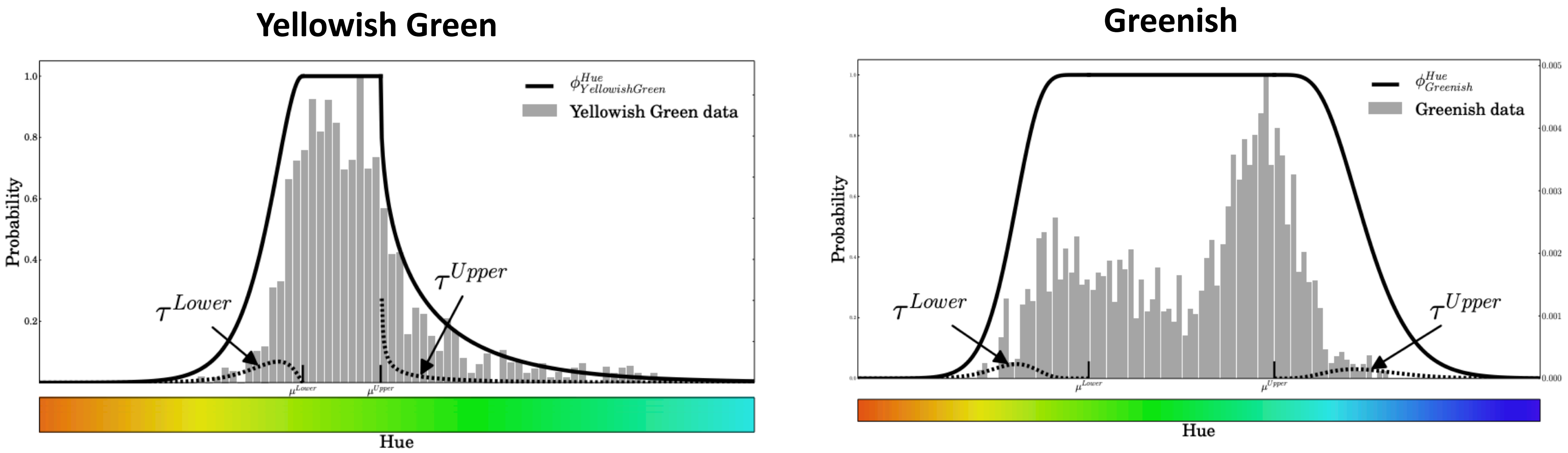
(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

Grounding color

Lexicon of Uncertain Color Standards (LUX)
semantic representations of 827 English color labels

Bayesian model for grounded color semantics

- Model variation in meaning of words
- Probability distribution of denotation for each word



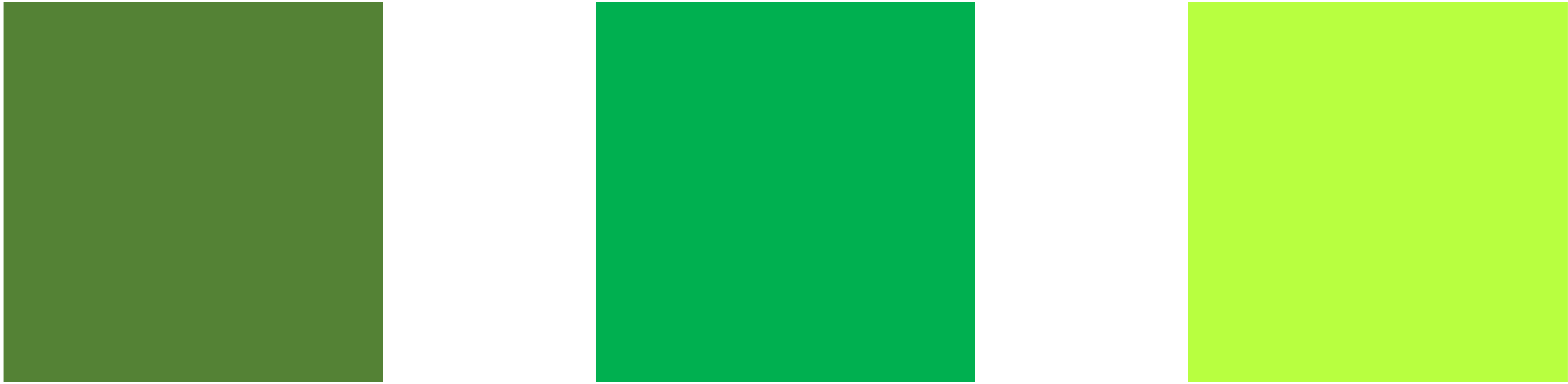
Yellowish Green

Greenish

(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

Color test

- What color is this?

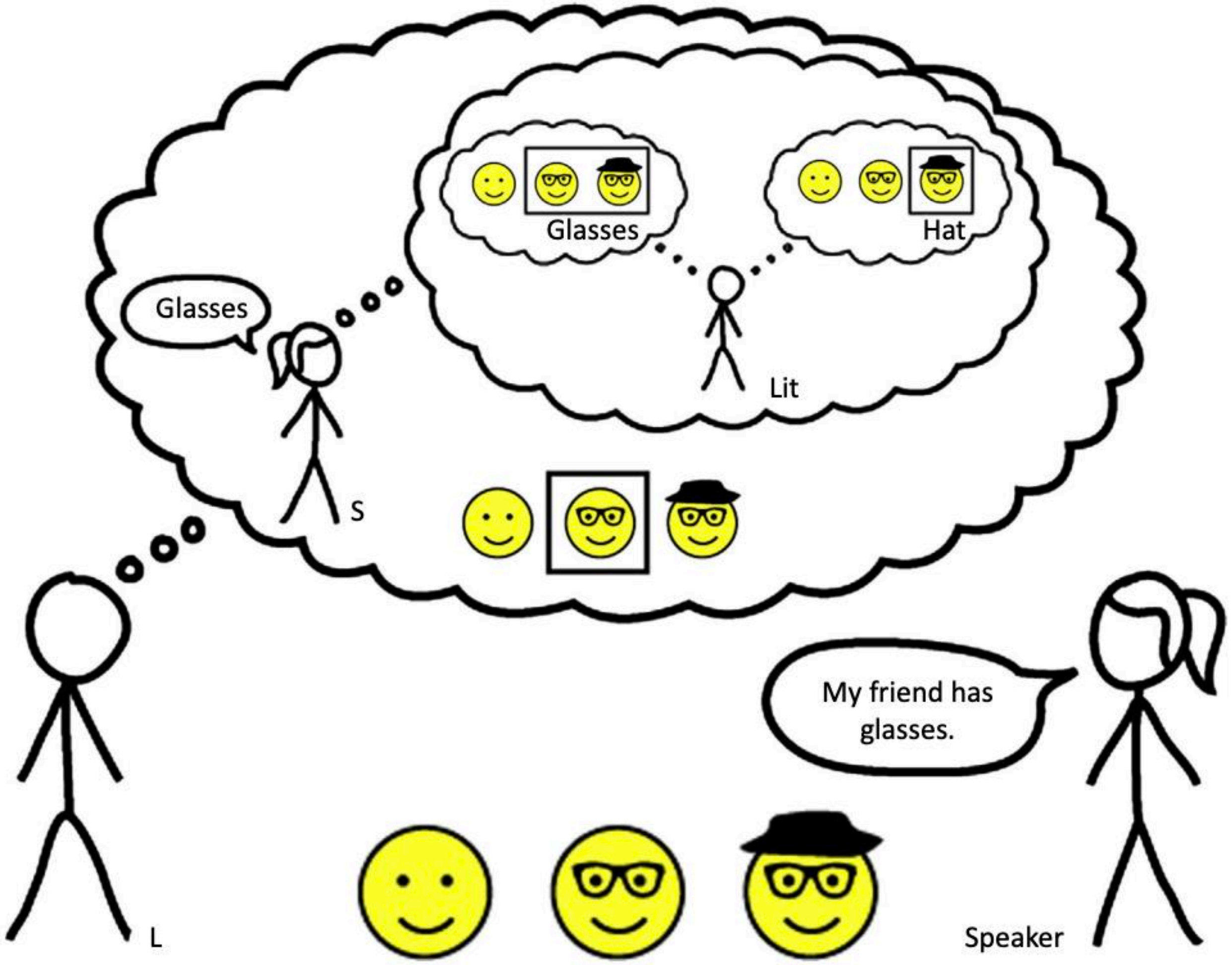


What words would a speaker select to

- indicate each of these colors?
- so that the listener can pick out the correct color given the triplet?

Rational Speech Acts Framework

Probablistic Bayesian view



[Pragmatic Language Interpretation as Probabilistic Inference, Goodman and Frank 2016, http://langcog.stanford.edu/papers_new/goodman-2016-tics.pdf]

Reflex Literal speaker and listeners

- Don't think about the other party
- Straightforward interpretation
- A bit of notation
  - u: utterance, t: world state,
  - M(u,t): meaning function connecting utterance u to world state t

    M(u,t) = 1 if u can be used to describe t, 0 otherwise

$M(u,t)$

u \ t	□	■
blue	1	1
cyan	1	0

Assume uniform priors

$$S_0(u|t,M) \propto M(u,t)P(u)$$

$$L_0(t|u,M) \propto M(u,t)P(t)$$

speaker

u \ t	□	■
blue	1/2	1
cyan	1/2	0

$S_0(u|t,M)$

u \ t	□	■
blue	1/2	1/2
cyan	1	0

$L_0(t|u,M)$

listener

Example from *Understanding the Rational Speech Act model [Monroe et al, CogSci 2018]*

Pragmatic listener and speaker

- Pragmatics: how context contributes to meaning
  - any non-local meaning phenomena

    “Can you pass the salt?”

    “Is he 21?” “Yes, he’s 25.”
- Model mental state of the other party

Literal version: “Can you pass the container with the salt in it?”

Literal version: “Is he older than 21?”

Conversational implicatures

speaker

listener

Rational Pragmatic listener and speaker

$S_2(u|t, M)$

u \ t	🩵	🟦
blue	1/4	1
cyan	3/4	0

$$S_2(u|t, M) \propto L_1(t|u, M)$$

$$L_1(t|u, M) \propto S_0(u|t, M)$$

u \ t	🩵	🟦
blue	1/3	2/3
cyan	1	0

$L_1(t|u, M)$

$$S_0(u|t, M) \propto M(u, t)P(u)$$

speaker

u \ t	🩵	🟦
blue	1/2	1
cyan	1/2	0

$S_0(u|t, M)$

Example from *Understanding the Rational Speech Act model*

listener

Pragmatic speaker and listener

$L_2(t|u, M)$

u \ t	🟦	🟦
blue	1/4	3/4
cyan	1	0

$$L_2(t|u, M) \propto S_1(u|t, M)$$

$$S_1(u|t, M) \propto L_0(t|u, M)$$

u \ t	🟦	🟦
blue	1/3	1
cyan	2/3	0

$S_1(u|t, M)$

$$L_0(t|u, M) \propto M(u, t)P(t)$$

u \ t	🟦	🟦
blue	1/2	1/2
cyan	1	0

$L_0(t|u, M)$

speaker

listener

Example from *Understanding the Rational Speech Act model*

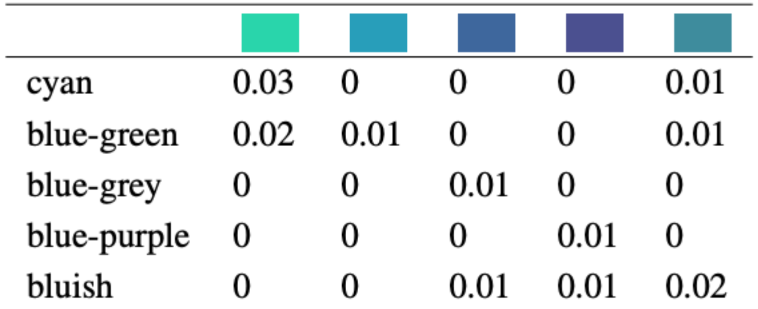
Converged speaker-listener model

After many iterations

u \ t	🟦	🟦
blue	0	1
cyan	1	0

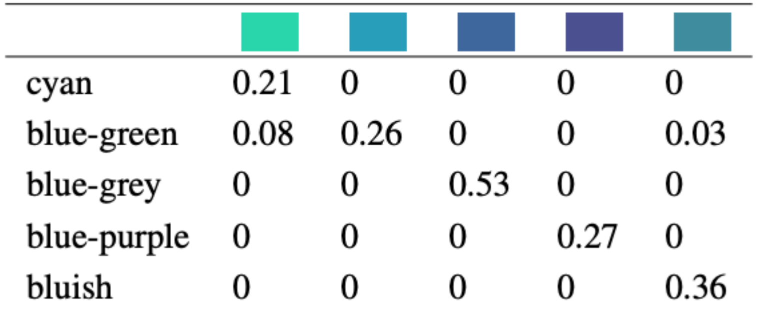
A more complex example

S_0



	🟩	🟦	🟦	🟪	🟦
cyan	0.03	0	0	0	0.01
blue-green	0.02	0.01	0	0	0.01
blue-grey	0	0	0.01	0	0
blue-purple	0	0	0	0.01	0
bluish	0	0	0.01	0.01	0.02

S_n

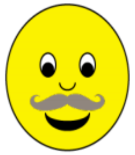
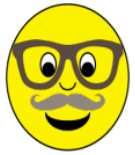
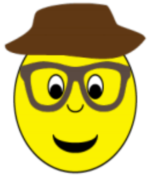


	🟩	🟦	🟦	🟪	🟦
cyan	0.21	0	0	0	0
blue-green	0.08	0.26	0	0	0.03
blue-grey	0	0	0.53	0	0
blue-purple	0	0	0	0.27	0
bluish	0	0	0	0	0.36

Example from *Understanding the Rational Speech Act model*

Moustache, Glasses, Hat example

$M(u,t)$

u \ t			
moustache	1	1	0
glasses	0	1	1
hat	0	0	1

$L_0(t|u,M)$

u \ t			
moustache	1/2	1/2	0
glasses	0	1/2	1/2
hat	0	0	1

$S_0(u|t,M)$

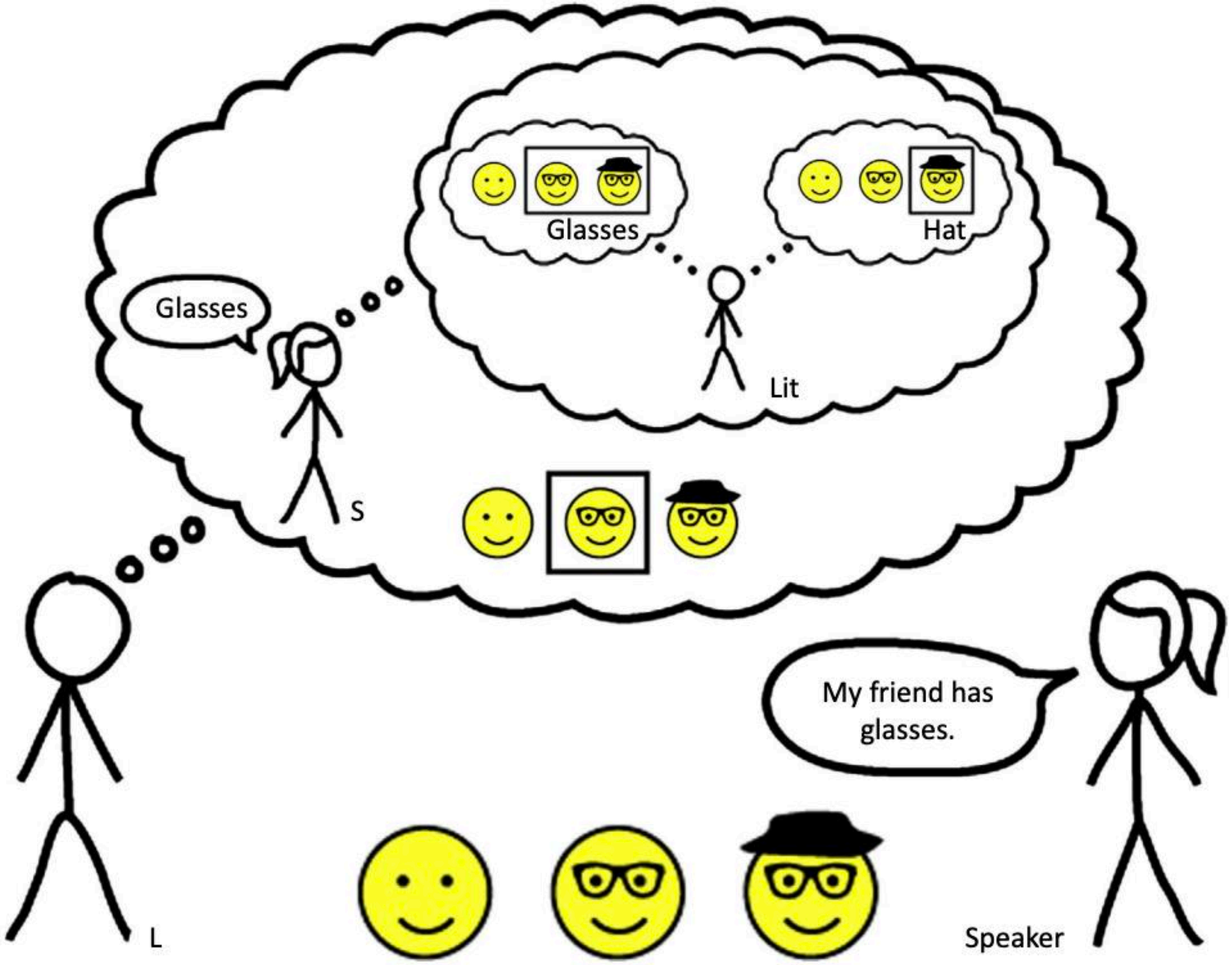
u \ t			
moustache	1	1/2	0
glasses	0	1/2	1/2
hat	0	0	1/2

Converged model

u \ t			
moustache	1	0	0
glasses	0	1	0
hat	0	0	1

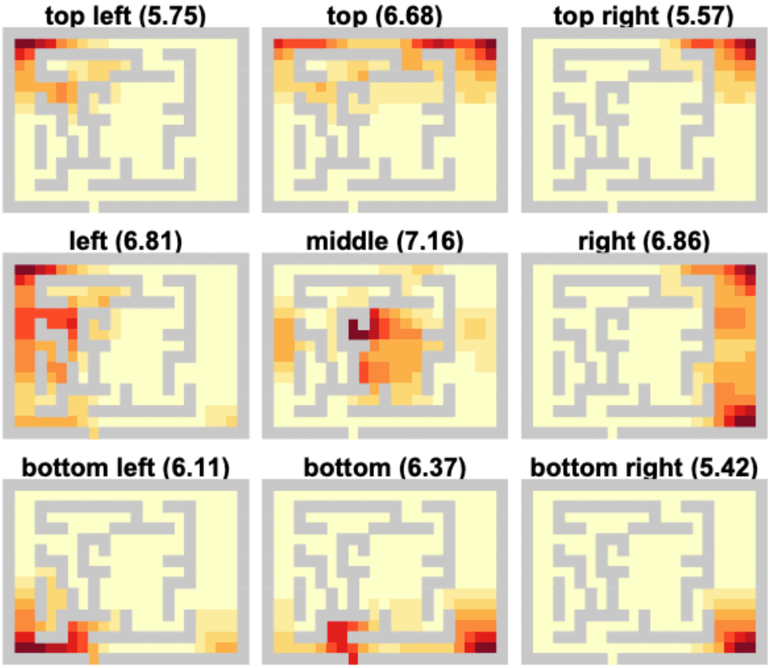
Example from *Implicatures and Nested Beliefs in Approximate Decentralized-POMDPs*, Vogel et al, ACL 2013

Do we need to keep recursing?



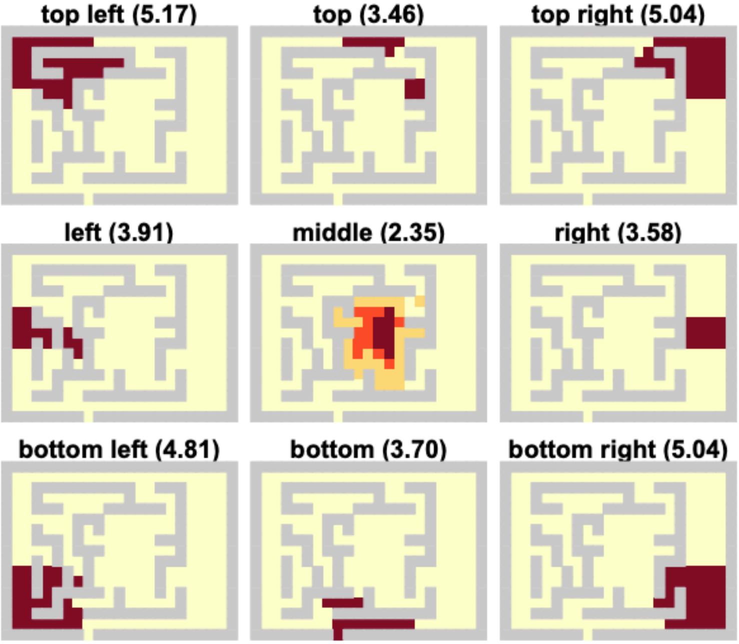
- Can be computationally expensive
- Let's consider basic level 1 speaker and listener models

Spatial references

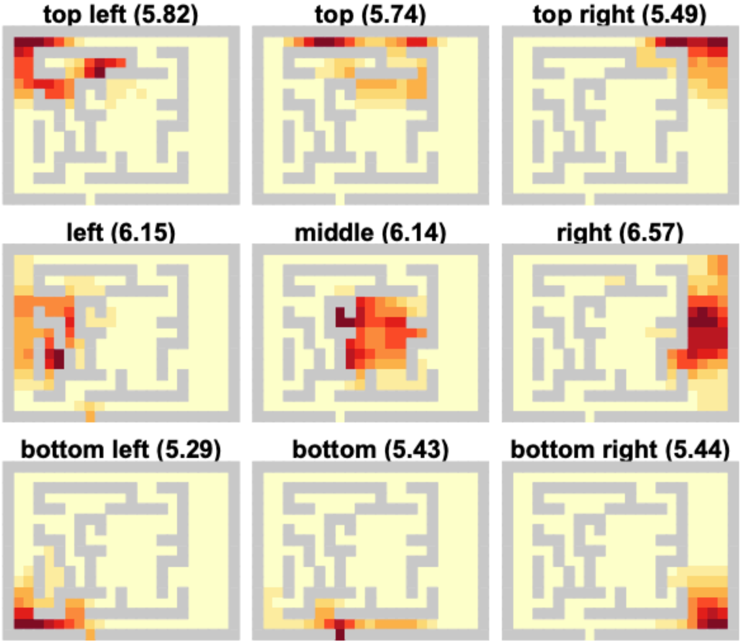


Literal listener (level 0)

Pragmatic listener (level 1)



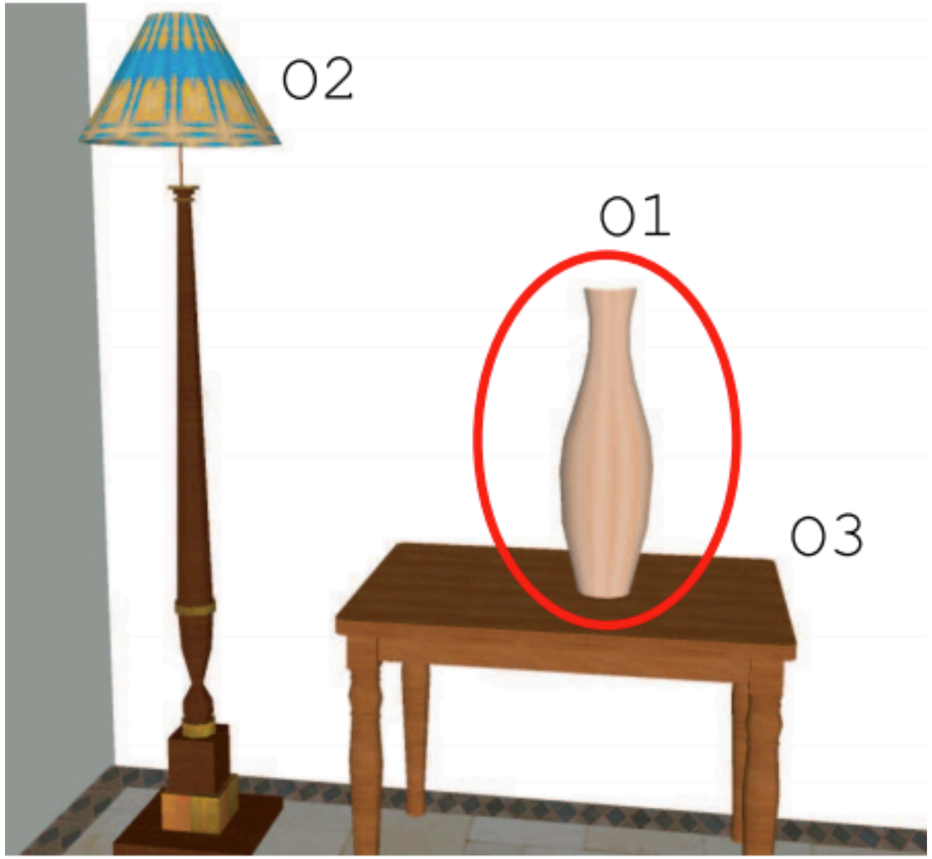
Human speakers



Example from *Implicatures and Nested Beliefs in Approximate Decentralized-POMDPs*, Vogel et al, ACL 2013

Speaker listener in applications (research papers)

Spatial relations



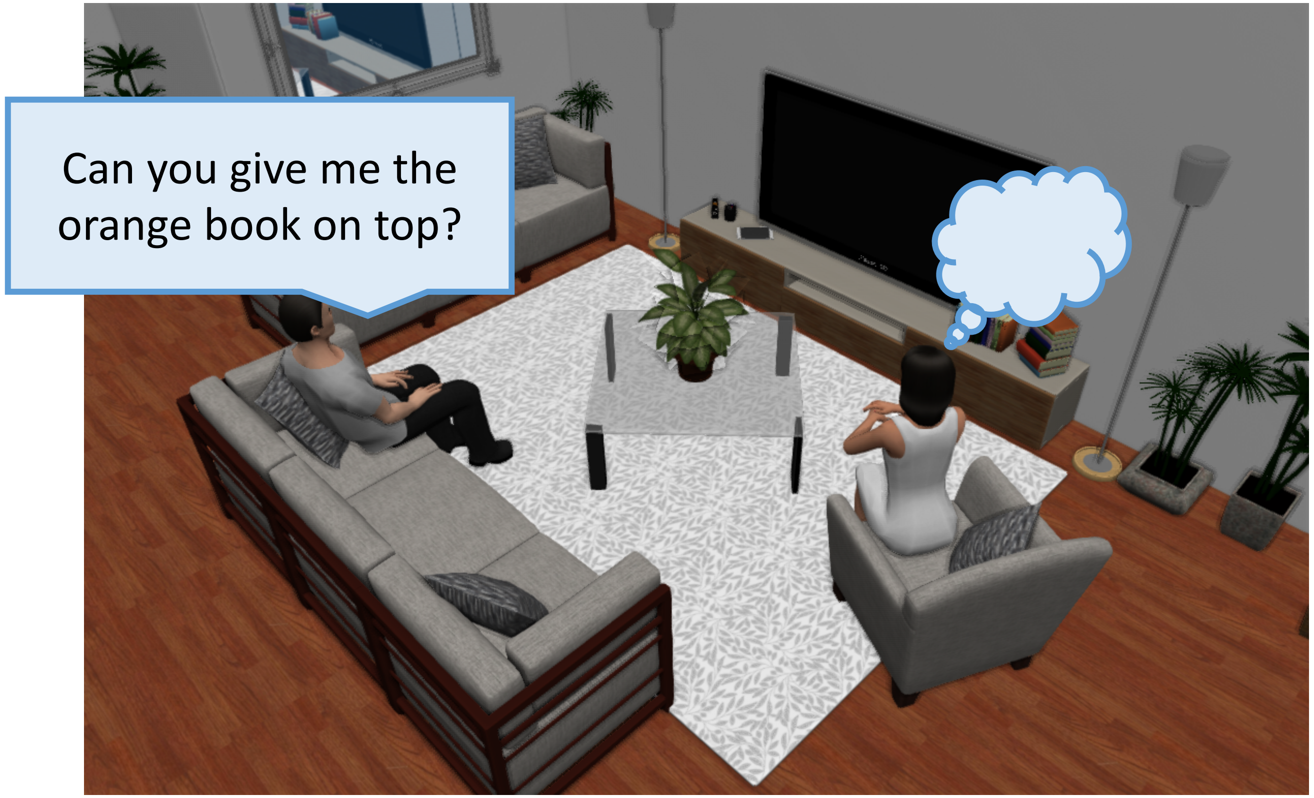
Consider only use spatial relations wrt to other objects to indicate (pick out) an object

- (i.e. do not say it is a vase or mention its color or other inherent properties)

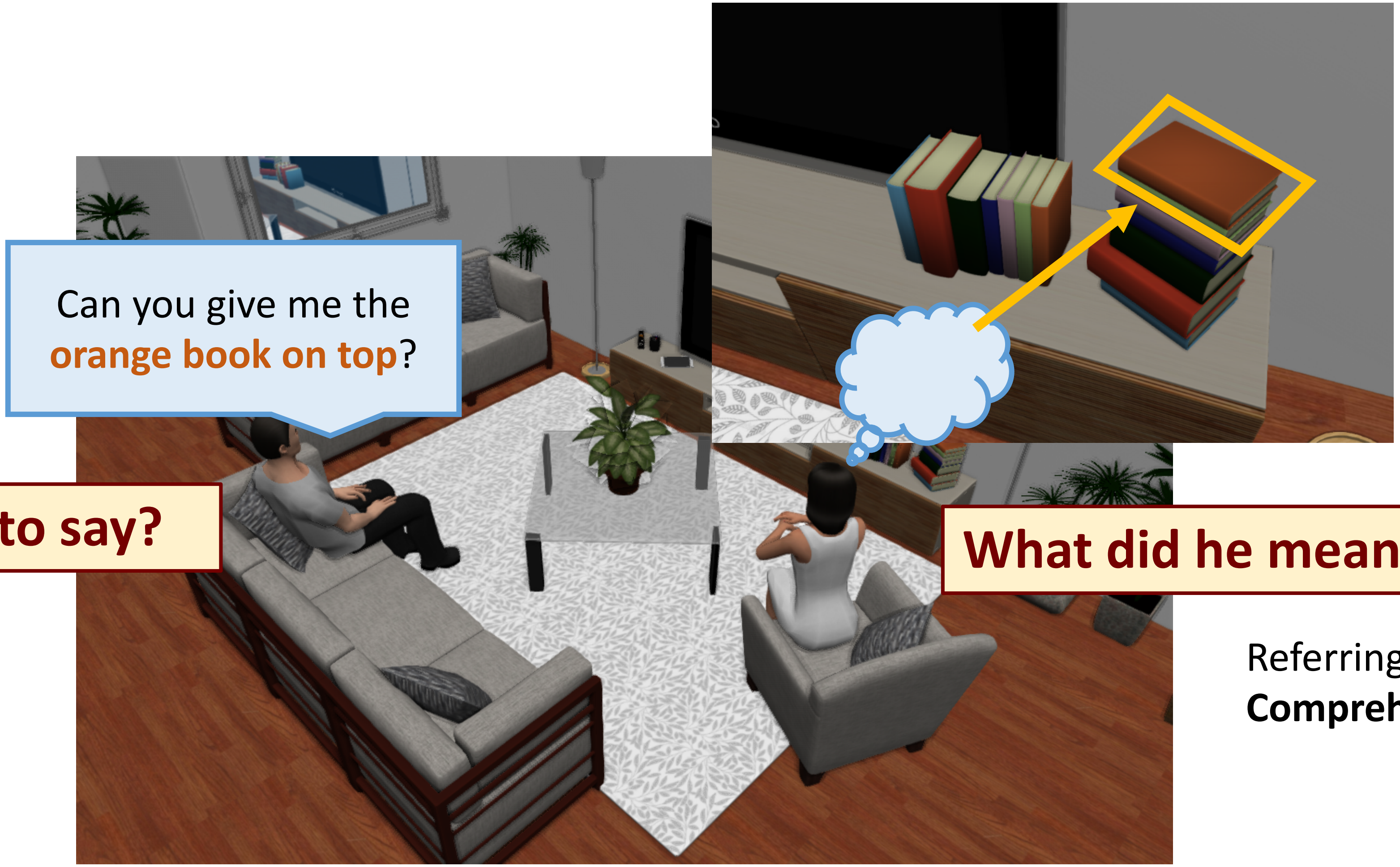
How to indicate O1?

- Requires modeling listener
- "right of O2" is not sufficient to disambiguate the object

A Game-Theoretic Approach to Generating Spatial Descriptions, Golland et al, EMNLP 2010



Can you give me the orange book on top?



What to say?

the book?

the orange book?

Referring Expression
Generation

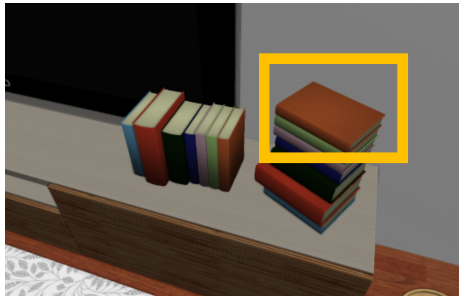
What did he mean?

Referring Expression
Comprehension

Need mental model of the other person

Referring expression generation

- Input: Image I with region R
- Output: Description S^*



orange book on top

$$S^* = \arg\max_S P(S|R, I)$$

L0 Speaker

Similar to standard image captioning task except input is a region in additional to the full image

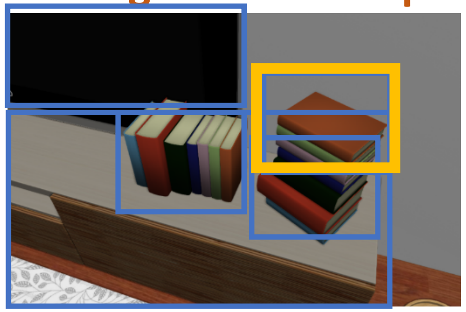
- The full image / surrounding objects are used as context

Referring expression comprehension

- Input: Image I with description S

  Generate candidate regions C
- Output: Region R^*

orange book on top



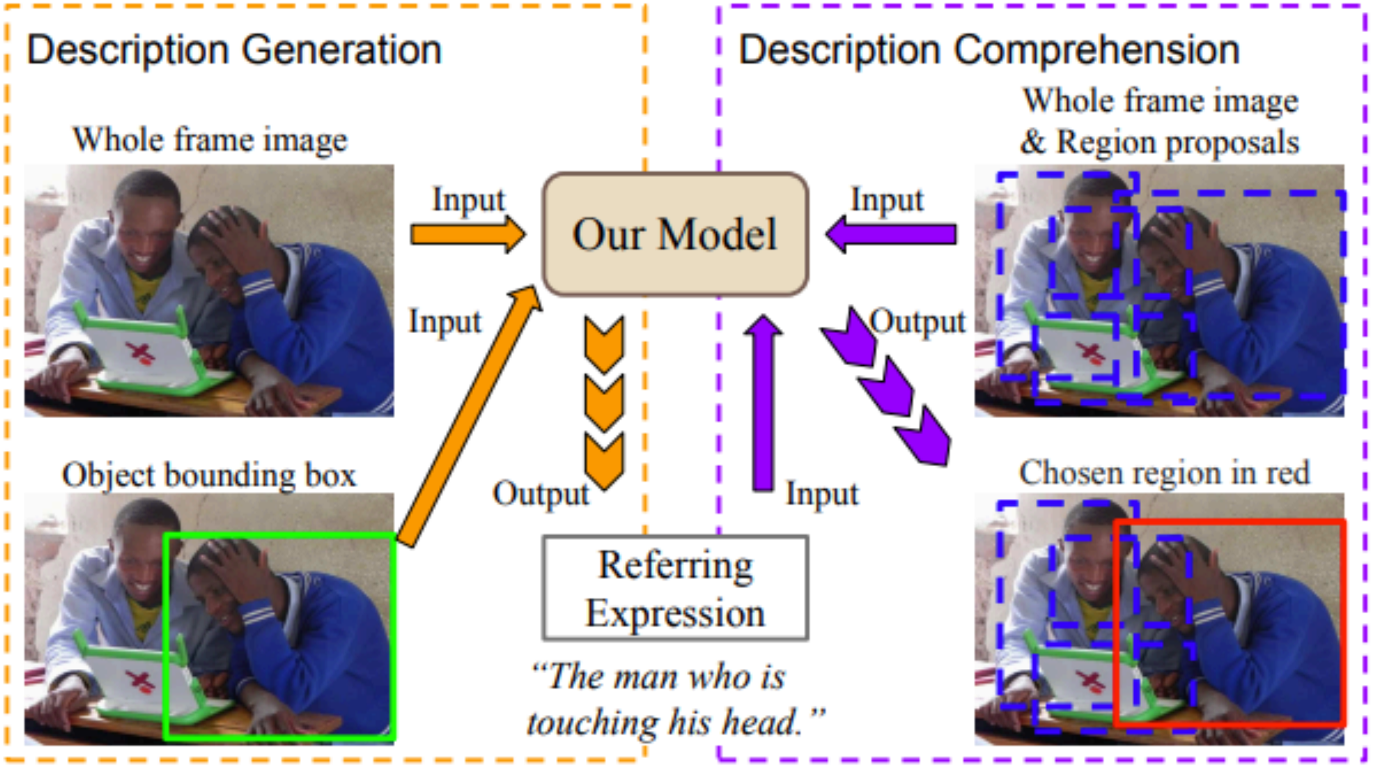
$$R^* = \arg\max_{R \in C} P(R|S,I)$$

Bayes Rule

$$P(R|S,I) = \frac{P(S|R,I)P(R|I)}{\sum_{R' \in C} P(S|R',I)P(R'|I)}$$

L1 Listener

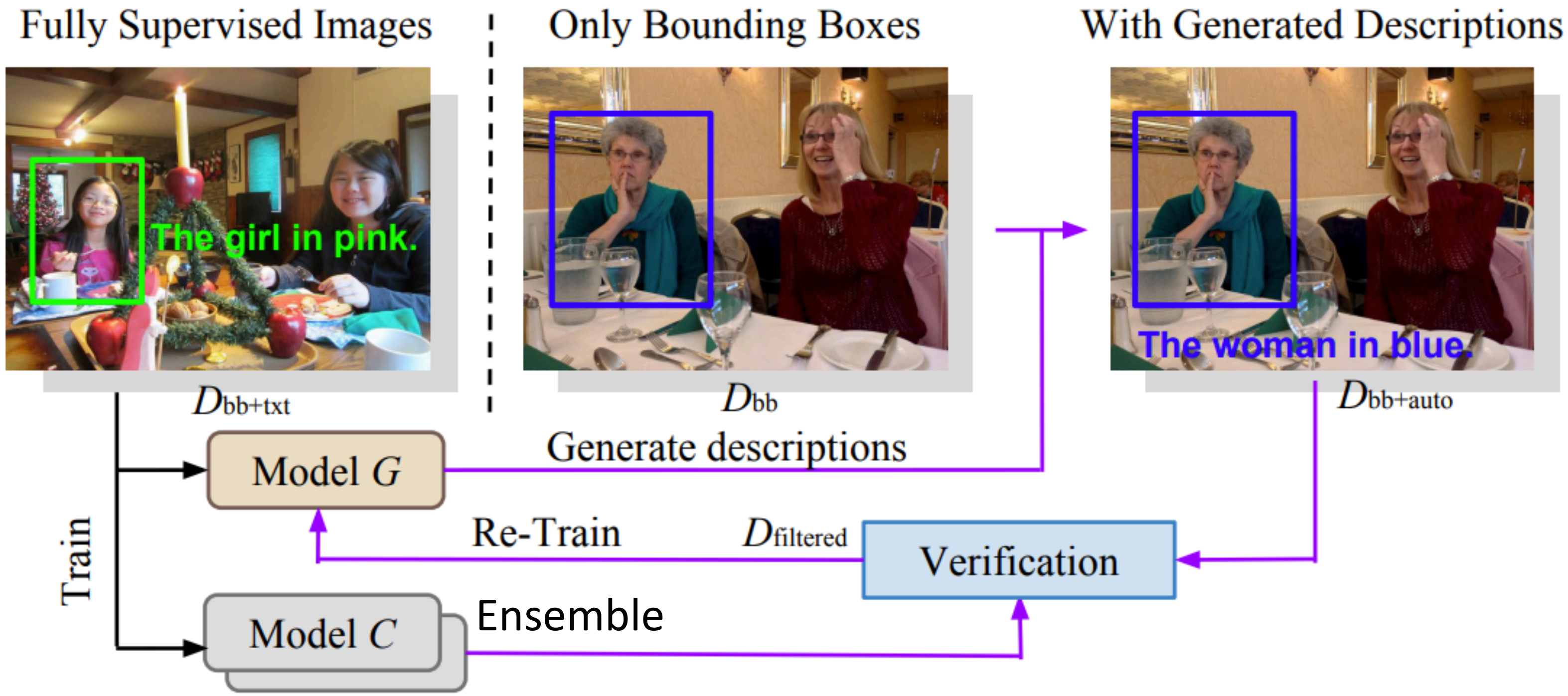
Jointly modeling speakers and listeners for referring expressions



- Will training jointly result in more discriminative descriptions?

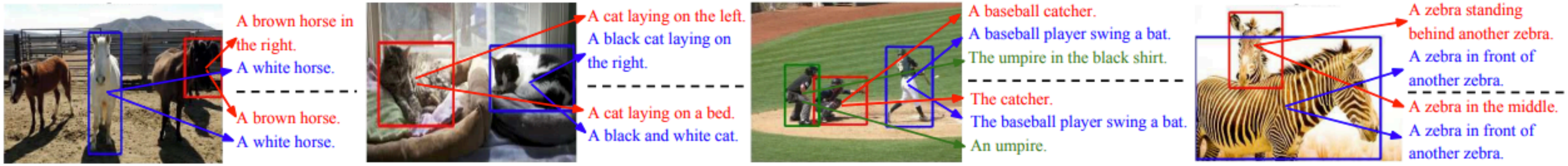
Generation and comprehension of unambiguous object descriptions, Mao et al, CVPR 2016

Jointly modeling speakers and listeners for referring expressions



Output

Full model


A brown horse in the right.
A white horse.
A brown horse.
A white horse.
A cat laying on the left.
A black cat laying on the right.
A cat laying on a bed.
A black and white cat.
A baseball catcher.
A baseball player swing a bat.
The umpire in the black shirt.
The catcher.
The baseball player swing a bat.
An umpire.
A zebra standing behind another zebra.
A zebra in front of another zebra.
A zebra in the middle.
A zebra in front of another zebra.

Baseline

Generation and comprehension of unambiguous object descriptions, Mao et al, CVPR 2016

ShapeGlot

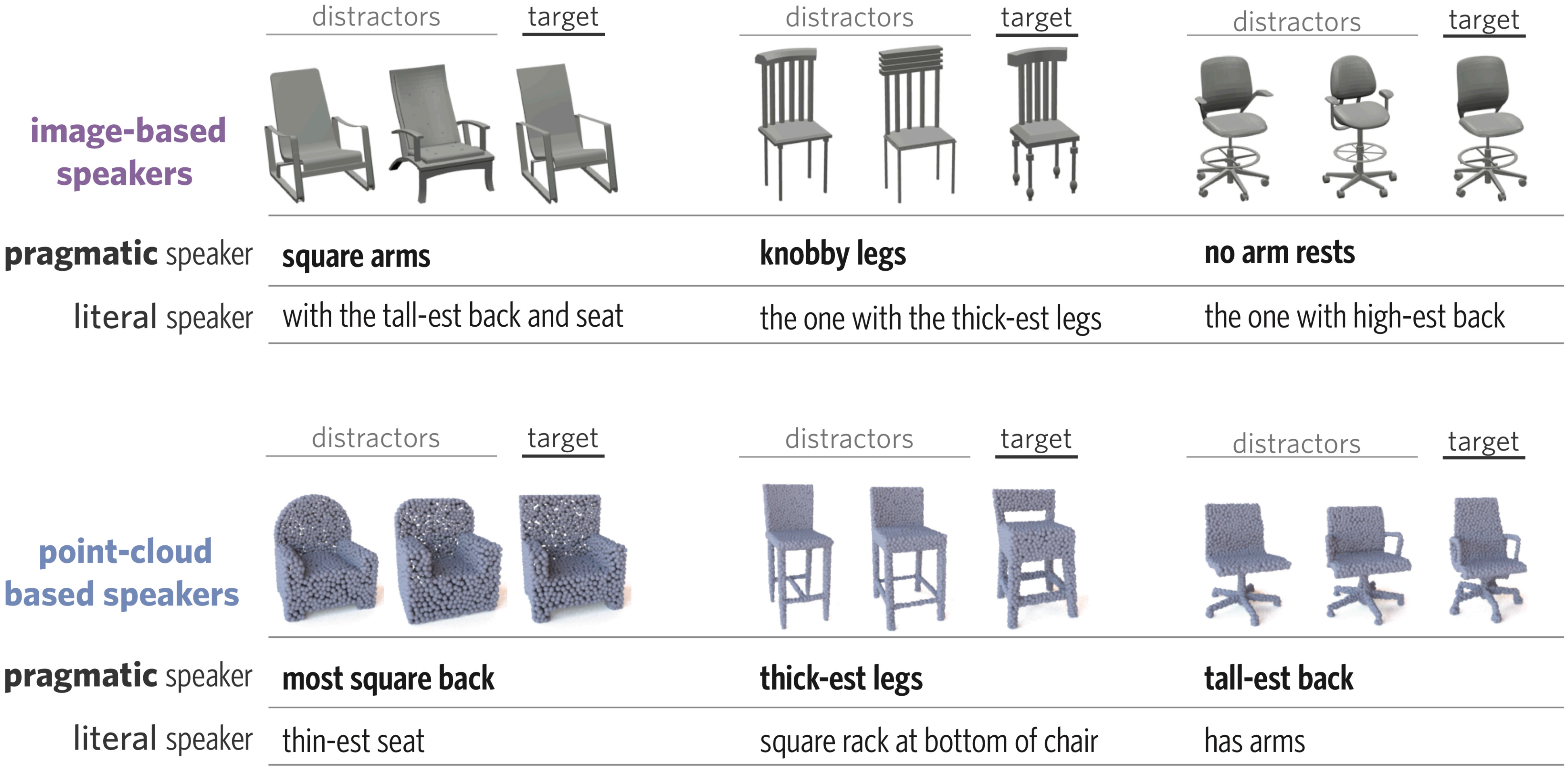


image-based speakers

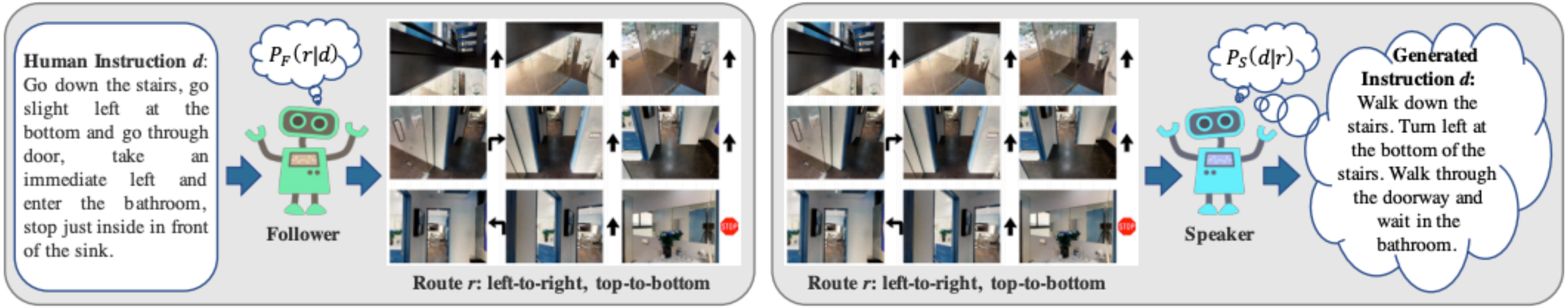
pragmatic speaker	**square arms**	**knobby legs**	**no arm rests**
literal speaker	with the tall-est back and seat	the one with the thick-est legs	the one with high-est back

point-cloud based speakers

pragmatic speaker	**most square back**	**thick-est legs**	**tall-est back**
literal speaker	thin-est seat	square rack at bottom of chair	has arms

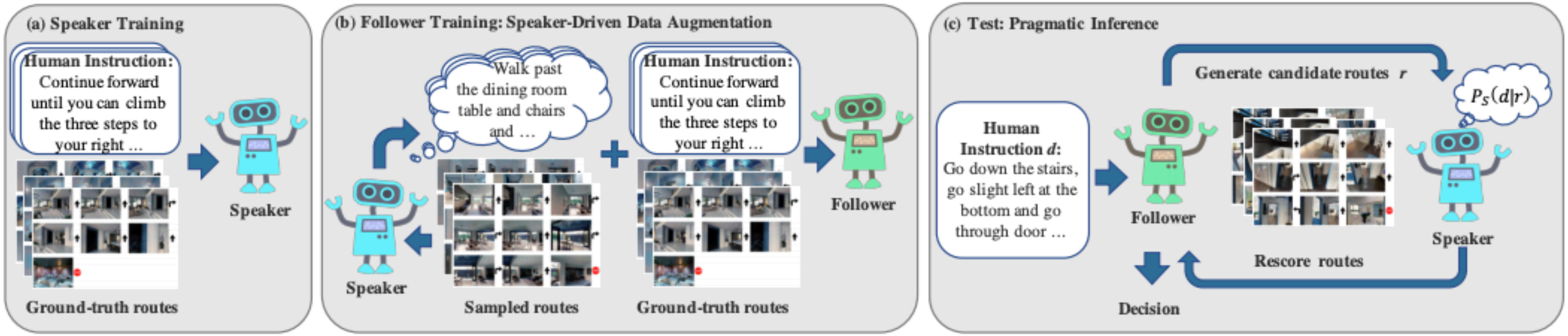
ShapeGlot: Learning Language for Shape Differentiation, Achlioptas et al, ICCV 2019

Vision-language navigation


$P_F(r|d)$
Follower
Route r: left-to-right, top-to-bottom
$P_S(d|r)$
Speaker
Route r: left-to-right, top-to-bottom

Human Instruction *d*: Go down the stairs, go slight left at the bottom and go through door, take an immediate left and enter the bathroom, stop just inside in front of the sink.

Generated Instruction *d*: Walk down the stairs. Turn left at the bottom of the stairs. Walk through the doorway and wait in the bathroom.



Human Instruction: Continue forward until you can climb the three steps to your right ...

Walk past the dining room table and chairs and ...

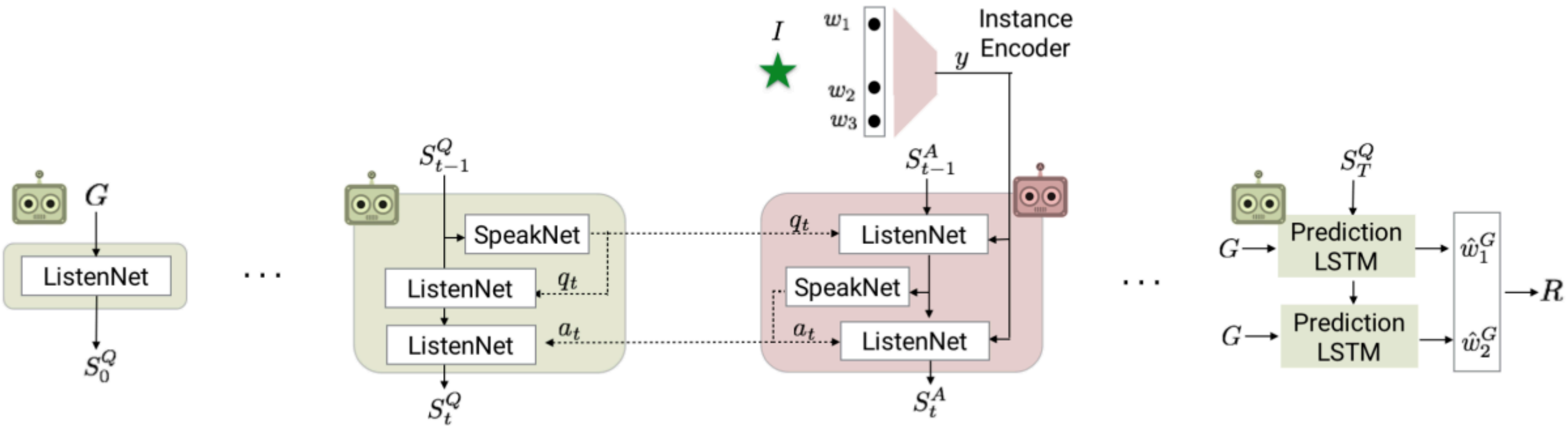
Human Instruction: Continue forward until you can climb the three steps to your right ...

Human Instruction *d*: Go down the stairs, go slight left at the bottom and go through door ...

Speaker-Follower Models for Vision-and-Language Navigation, Fried et al, NIPS 2018

Multi-agent communication

- Simulate speakers and listeners and see what happens
- Emergent communications!



Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog, Kottur et al, EMNLP 2017

Summary

- Speaker-listener
- RSA: Mental model of the other agent
- Full model computationally expensive and may not be necessary
- Simulate speakers and listener → emergent communications

Next time

- Paper presentations (3/8)
  - ShapeGlot: Learning Language for Shape Differentiation (Qirui)
  - Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog (Sonia)

- Thursday (3/11): Instruction following – review of deep RL