Grounding

CMSC 723 / LING 723 / INST 725

Hal Daumé III [he/him]

21 Nov 2019

Announcements, logistics

- Exam (grading released just now) statistics
  - Late: rob_min=35, rob_max=95, median=82, mean=77 *(does not include make-up)*
  - Early: rob_min=57, rob_max=98, median=86, mean=84
  - Remember: early worth 10% of grade, late worth 15% of grade
- Homework 4/5:
  - HW4-programming
  - HW5-written
  - Both out tomorrow
- P4 due Dec 3

Last time

- Semantic parsing from denotations
  - Assume predicates in domain are known
  - Given (sentence, validation) pairs, learn a good parser
    - That maps sentence → logical form
    - With deterministic logical form → validation
  - Key challenges:
    - Where does the lexicon come from
    - How do you learn when you only have denotations and not full parses
  - Idea:
    - Guess and check

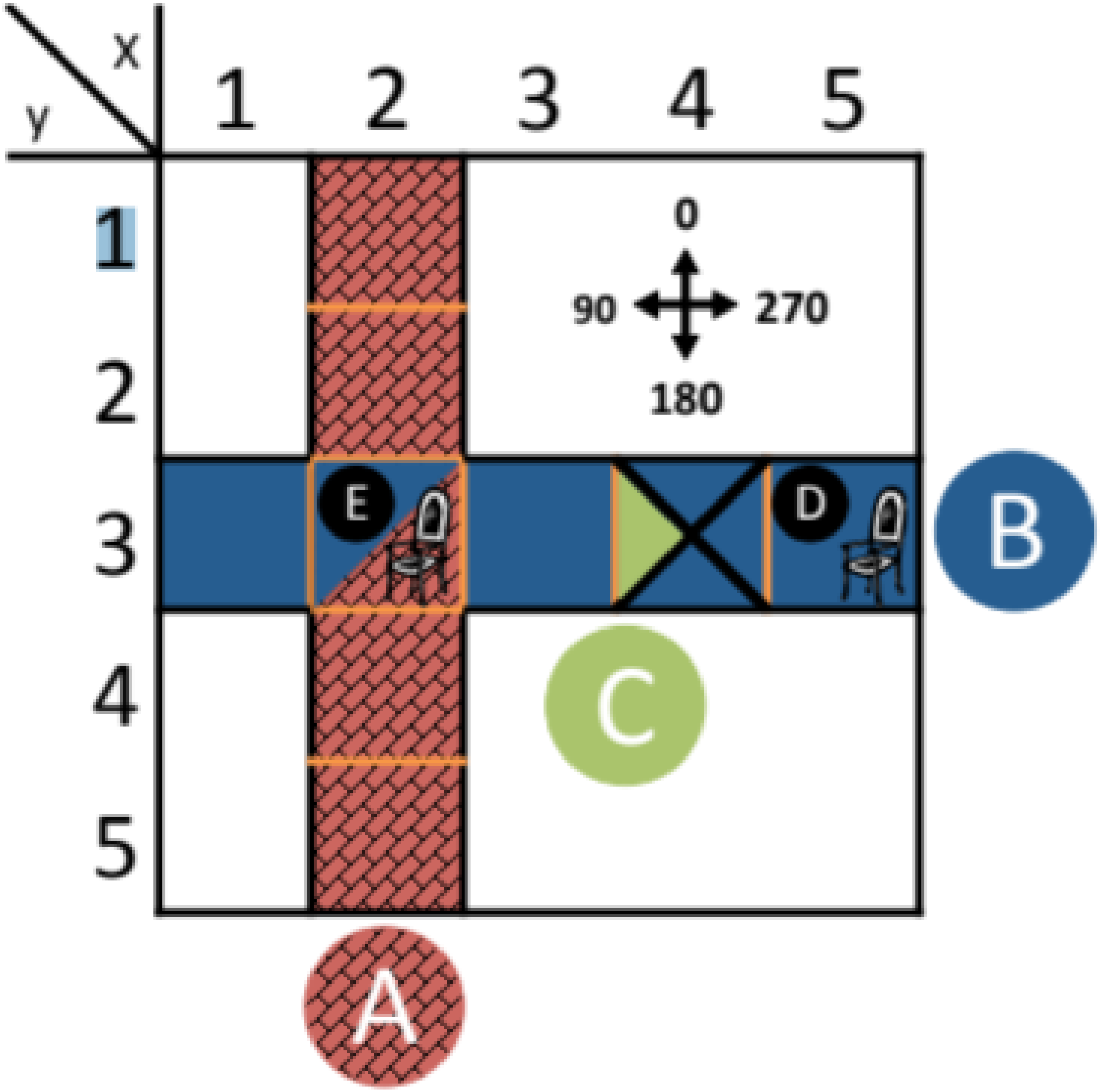
Today

- Grounding
  - Relationship between linguistic symbols and stuff in the real world
- Spatial language
  - Relationships between objects in the world
- Implicit vs explicit communication

Semantic parsing of instructions

Weakly Supervised Learning of Semantic Parsers for Mapping Instructions to Actions

Yoav Artzi and **Luke Zettlemoyer**



(a) chair
$\lambda x.chair(x)$ → {D, E}

(b) hall
$\lambda x.hall(x)$ → {A, B}

(c) the chair
$\iota x.chair(x)$ → E

(d) you
you → C

(e) blue hall
$\lambda x.hall(x) \wedge blue(x)$ → {B}

(f) chair in the intersection
$\lambda x.chair(x) \wedge intersect(\iota y.junction(y), x)$ → {E}

(g) in front of you
$\lambda x.in\_front\_of(you, x)$ → {A, B, E}

Very strong assumption: predicates

- Very common in semantics:
  - “chair” means CHAIR
  - “walk” means WALK
  - “dog” means DOG
  - “blue” means BLUE
  - …

Imp.: move from the sofa to the chair

LF: $\lambda a.move(a) \wedge to(a, \iota x.chair(x)) \wedge from(a, \iota y.sofa(y))$

- The smallcaps predicates are assumed known ahead of time

- (Some work even assumes that these are spelled similarly)

In some cases this makes sense

- Natural language interfaces to databases
  - You know the rows/columns of the db
  - You know SQL
  - Only challenge is to learn that “chair” means CHAIR (rather than, eg, BLUE)

- Execution of commands in a computational environment
  - TODO Regina paper
  - TODO web navigation

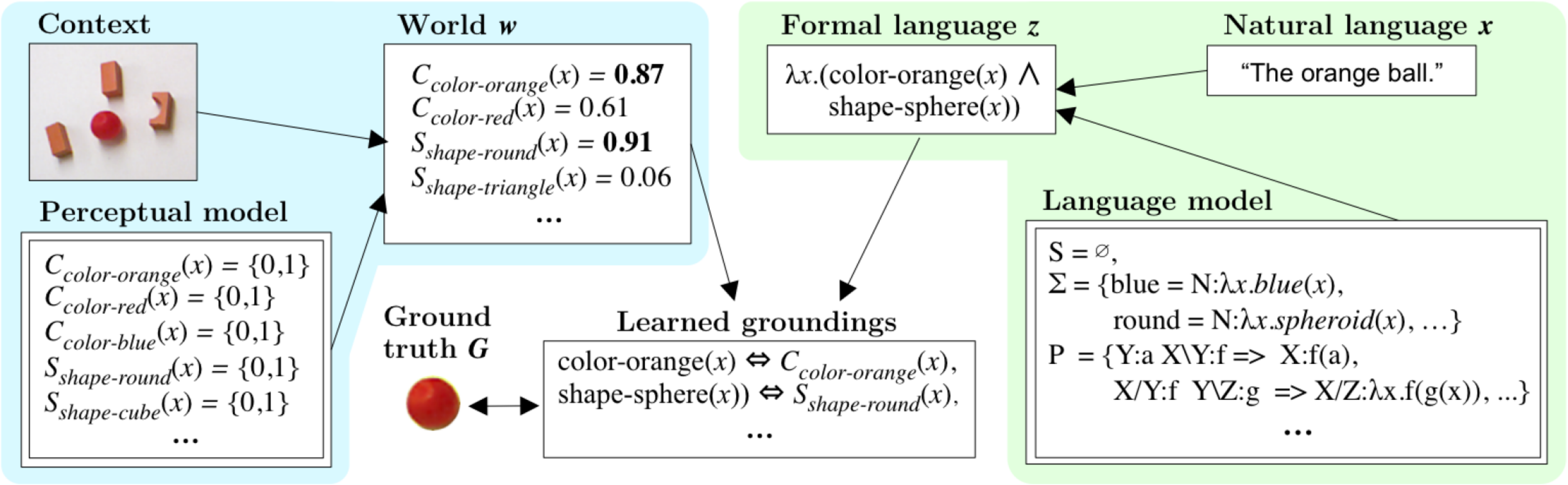
So… what is the meaning of “blue”?

- Option 1: “blue” means $\lambda x \,.\, r382328(x)$ for some fixed predicate *r382328*

- Option 2: “ground” language in more language
  - Monolingual via distributed representations:
    - “blue” means [1.815, 0.938, 0.312, -0.319, -0.019, ………., 0.414, 0.485, 1.023, -0.451, 0.443]
  - Multilingual via translations:
    - “blue” means { …. }

- Option 3: ground language in perception
  - “blue” means $\lambda x \,.\, H_{blue}(x)$ for some **classifier** H_{blue} that we can learn

Classifier-based grounding



Grounded Language Learning: Where Robotics and NLP Meet

Cynthia Matuszek
University of Maryland, Baltimore County, Baltimore, MD
cmat@umbc.edu

Data/interaction

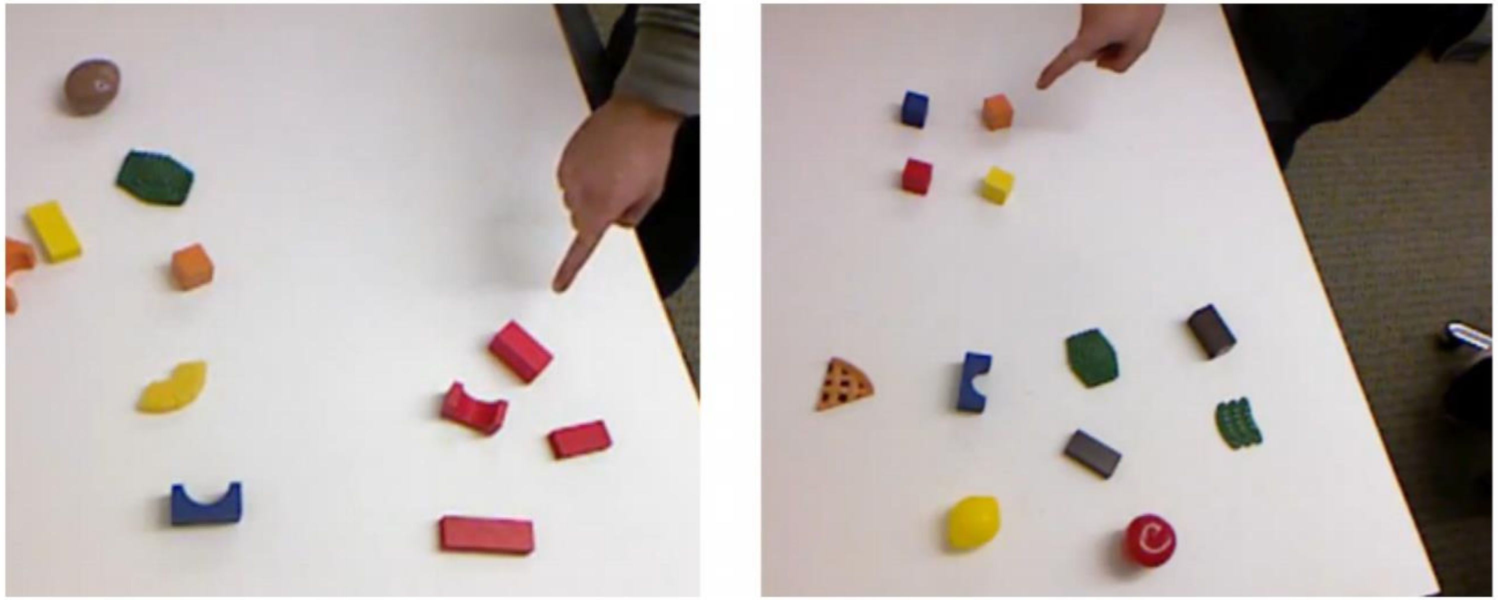


Figure 3. Example scenes presented on Mechanical Turk. Left: A scene that elicited the descriptions “here are some red things” and “these are various types of red colored objects”, both labeled as $\lambda x.color(x, red)$. Right: A scene associated with sentence/meaning pairs such as “this toy is orange cube” and $\lambda x.color(x, orange) \wedge shape(x, cube)$.

Possible worlds model

G = set of objects

O = scene

w = possible world (set of classifier outputs in {T,F})

z = logical forms

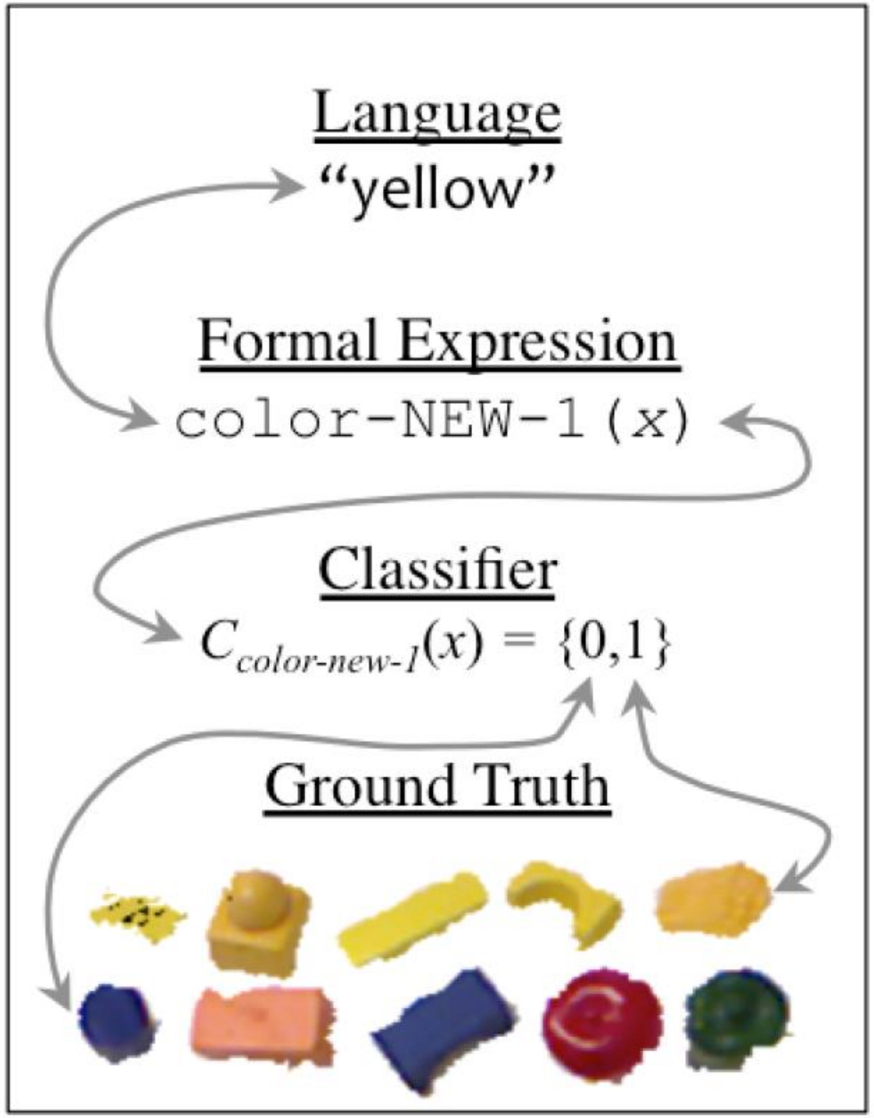
$$P(G \mid x, O) = \sum_{z} \sum_{w} P(G, z, w \mid x, O)$$

$$P(G, z, w \mid x, O) = P(z \mid x) P(w \mid O) P(G \mid z, w)$$

Challenge: learning new predicates

- Get people to label objects or scenes with language

- Hypothesize new classifiers for each new formal expression



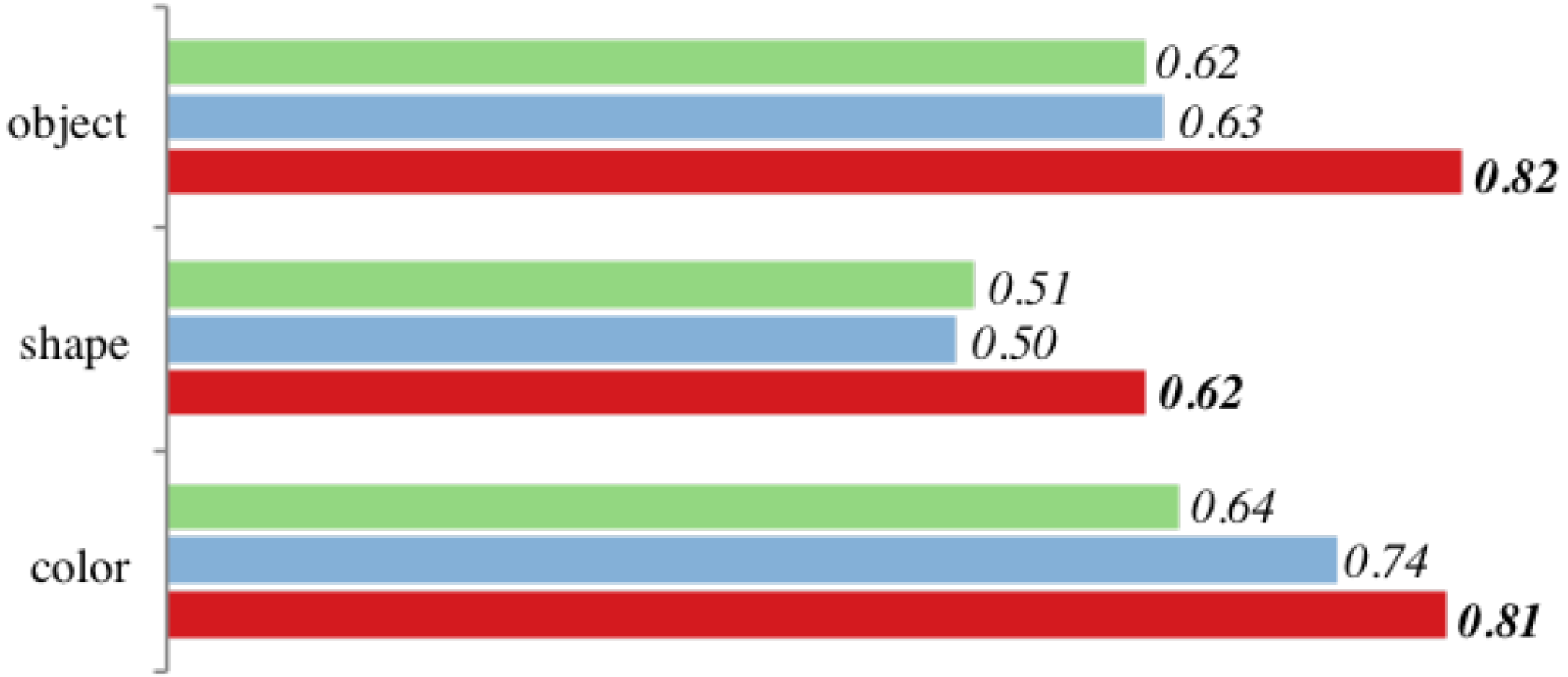
Language + Vision combo helps

	Precision	Recall	F1-Score
Vision	**0.92**	0.41	0.55
Language	0.52	0.09	0.14
Joint	0.82	**0.71**	**0.76**

- The ablations mean “don’t update that part of the model during learning”

Where do you get negative examples

- “This is a lemon” does not mean it’s not yellow
- Baselines: All/random other objects not identified
- Core idea: only negative-train on “similar” items (e.g., “apple”)



Challenges

- Some things are hard to ground out….
  - What do $H_{freedom}$(x) and $H_{justice}$(x) look like and how are they trained?

- Some things depend on broader pragmatic context
  - H_{above}(x, y)?

Some other related grounding stuff

Phenomenon	Overall c	Overall μ	Example from TOUCHDOWN
Reference to unique entity	25	10.7	*. . . You'll pass* ***three trashcans*** *on your left . . .*
Coreference	22	2.4	*. . . a brownish colored brick building with a black fence around* ***it****. . .*
Comparison	6	0.3	*. . . The bear is in the middle of the* ***closest*** *tire.*
Sequencing	22	1.9	*. . . Turn left at the* ***next*** *intersection . . .*
Count	11	0.5	*. . . there are* ***two*** *tiny green signs you can see in the distance . . .*
Allocentric spatial relation	25	2.9	*. . . There is a fire hydrant, the bear is* ***on top***
Egocentric spatial relation	25	4.0	*. . . up ahead there is some flag poles* ***on your right hand side****. . .*
Imperative	25	5.3	*. . .* ***Enter*** *the next intersection and* ***stop*** *. . .*
Direction	24	3.7	*. . . Turn* ***left****. Continue* ***forward*** *. . .*
Temporal condition	21	1.9	*. . . Follow the road* ***until you see*** *a school on your right. . .*
State verification	21	1.8	*. . .* ***You should see*** *a small bridge ahead . . .*

Turn and go with the flow of traffic. At the first traffic light turn left. Go past the next two traffic light, As you come to the third traffic light you will see a white building on your left with many American flags on it. Touchdown is sitting in the stars of the first flag.

TOUCHDOWN: Natural Language Navigation and Spatial Reasoning in Visual Street Environments

Howard Chen*
ASAPP Inc.
New York, NY
hchen@asapp.com

Alane Suhr Dipendra Misra Noah Snavely Yoav Artzi
Department of Computer Science & Cornell Tech, Cornell University
New York, NY
{suhr, dkm, snavely, yoav}@cs.cornell.edu

Grounding with dialog

Lightly Supervised Learning of Procedural Dialog Systems

Svitlana Volkova
CLSP
Johns Hopkins University
Baltimore, MD
svitlana@jhu.edu

Pallavi Choudhury, Chris Quirk, Bill Dolan
NLP Group
Microsoft Research
Redmond, WA
pallavic,chrisq,
billdol@microsoft.com

Luke Zettlemoyer
Computer Science and Engineering
University of Washington
Seattle, WA
lsz@cs.washington.edu

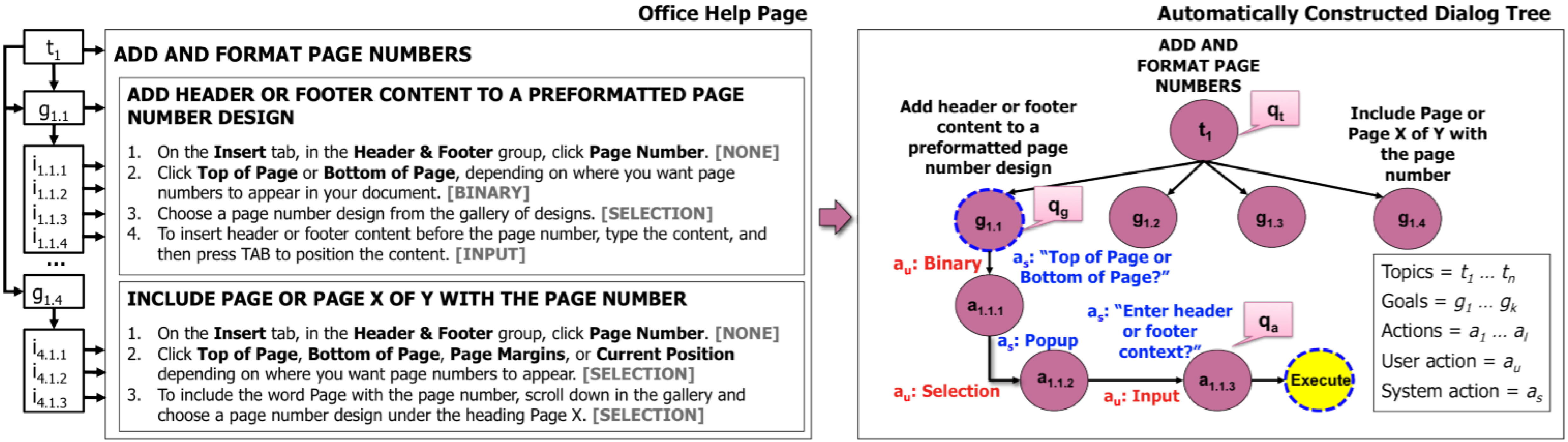
- Assumed data:
  - Instructional web pages
  - Examples of user intents

U: "I want to add page numbers and a title"
S: "Top or Bottom of the page?"
U: "Top"
S: "Please select page design from the templates" (*System shows drop down menu*)
U: *User selects from menu*
S: "Enter header or footer content"
U: "C.V."
S: "Task completed."

Figure 1: An example dialog interaction between a system (S) and user (U) that can be automatically achieved by learning from instructional web page and query click logs.

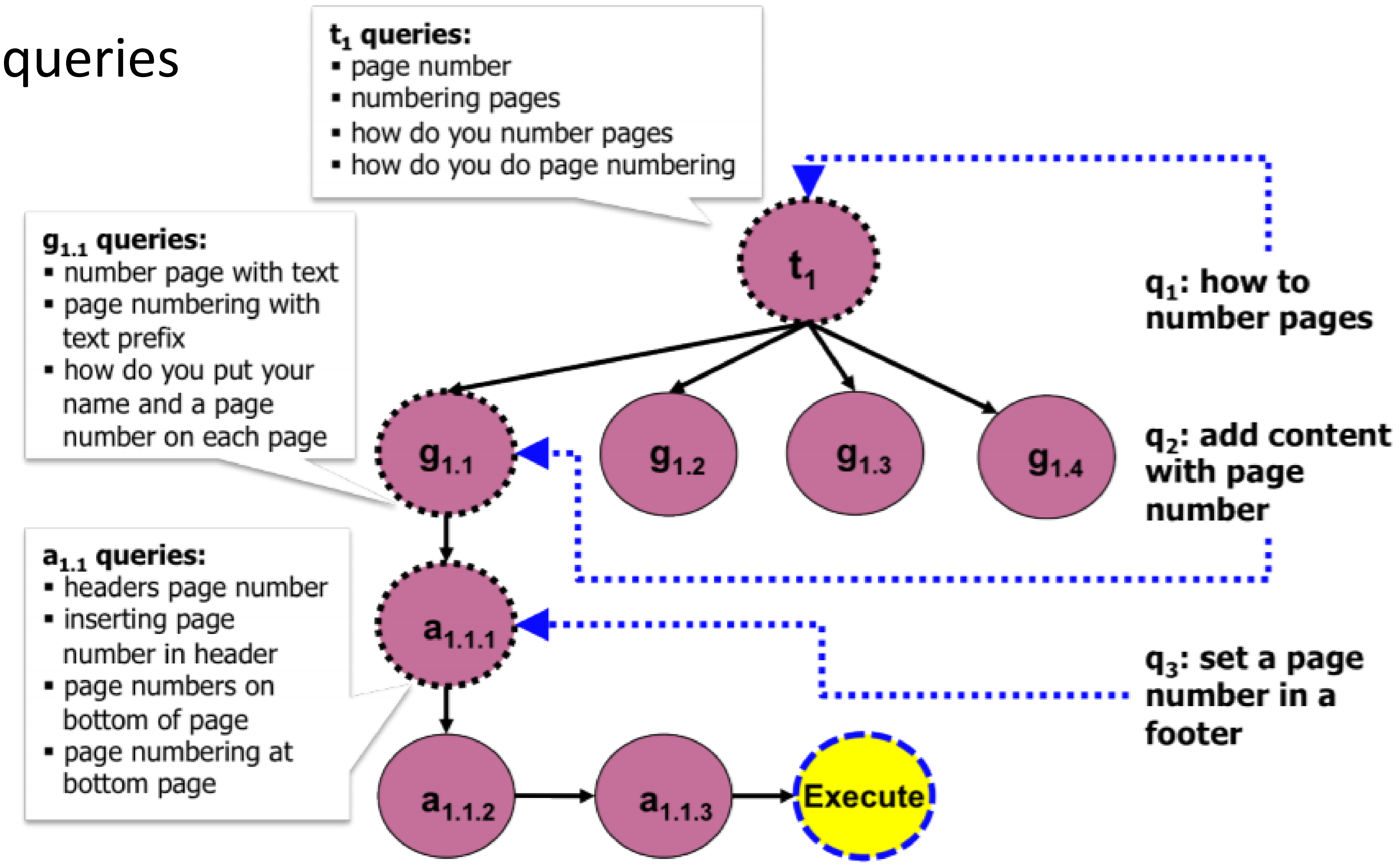
Grounding with dialog

- Build dialog trees from instructions

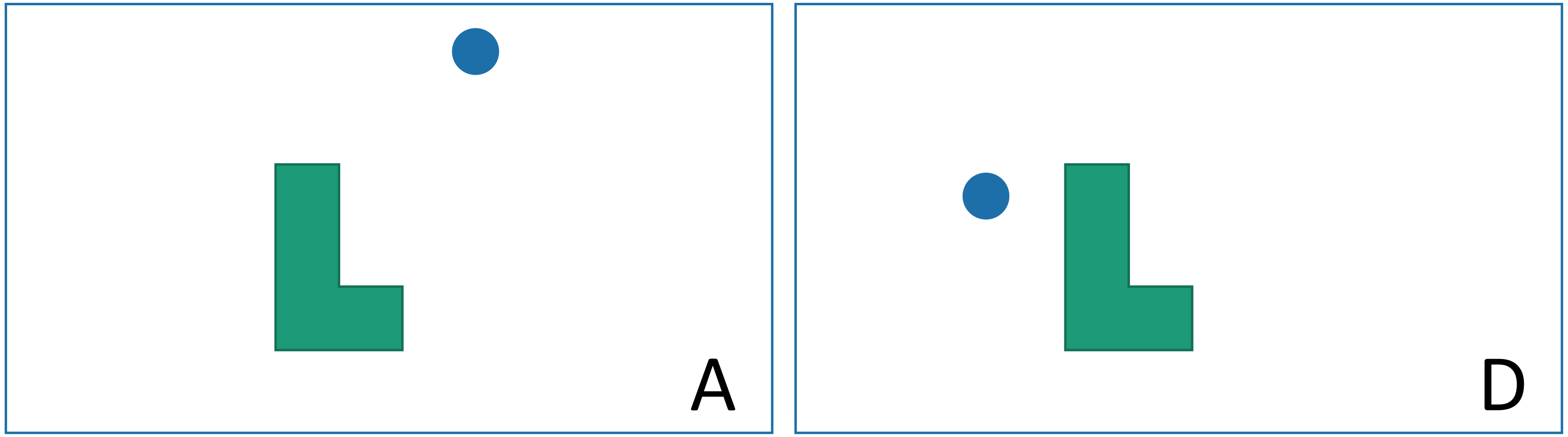


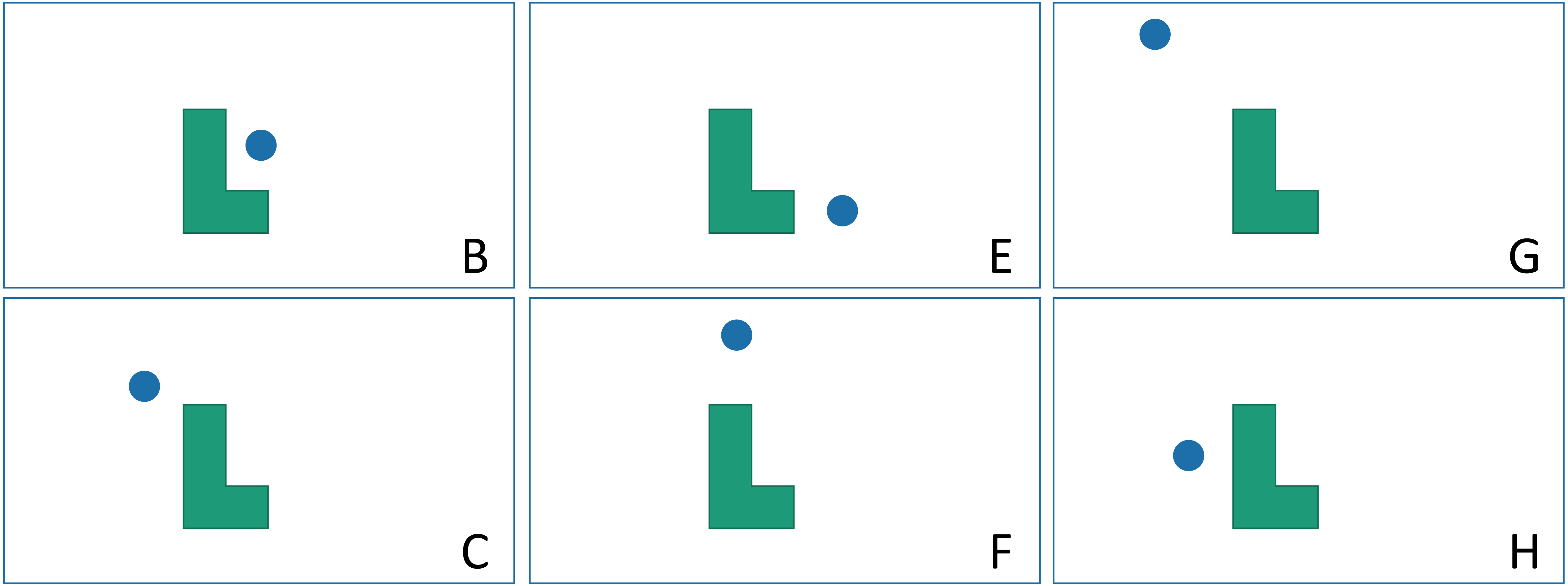
Grounding with dialog

- Understanding initial queries

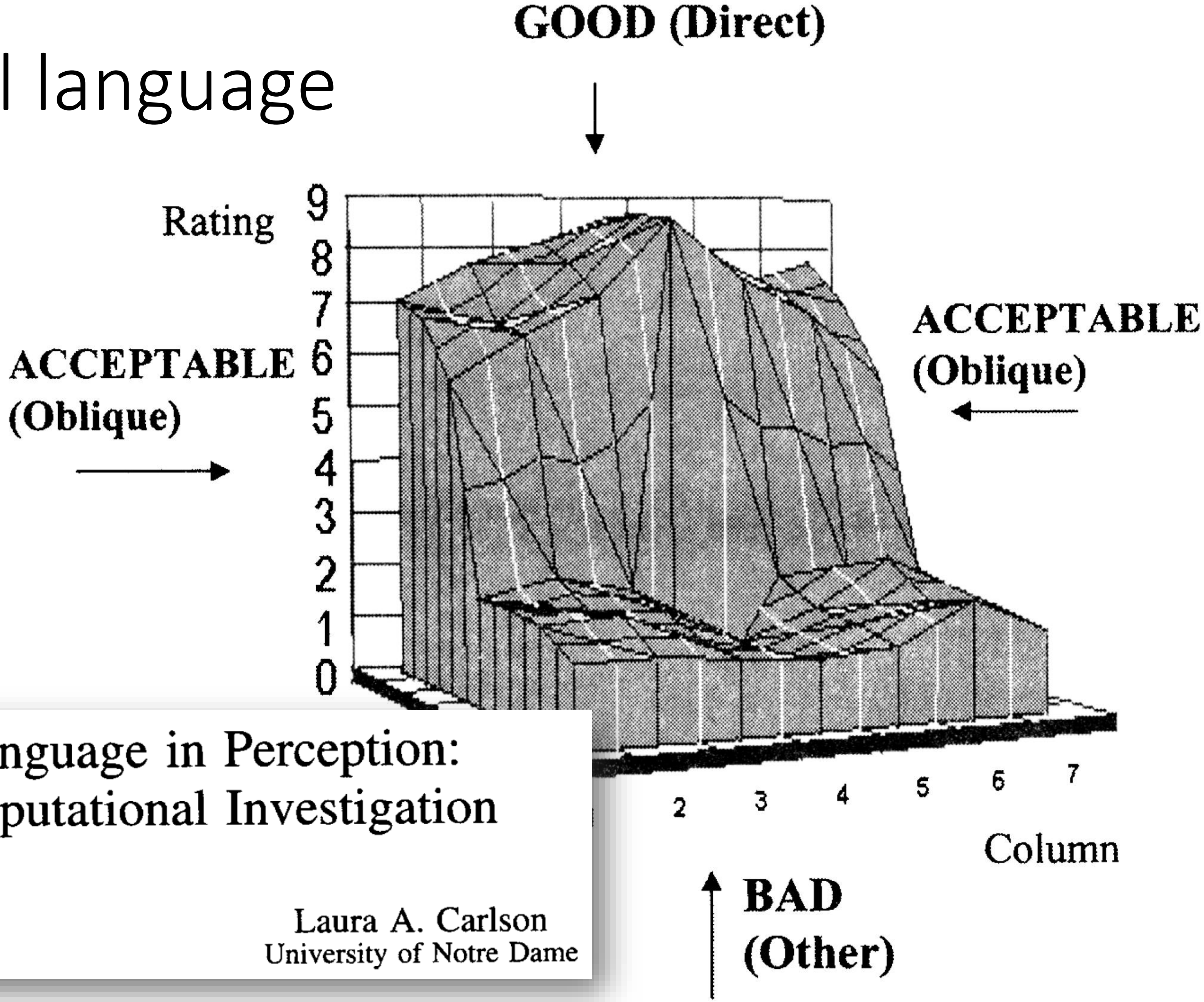


go.umd.edu/cl1above





Grounding spatial language



Grounding Spatial Language in Perception: An Empirical and Computational Investigation

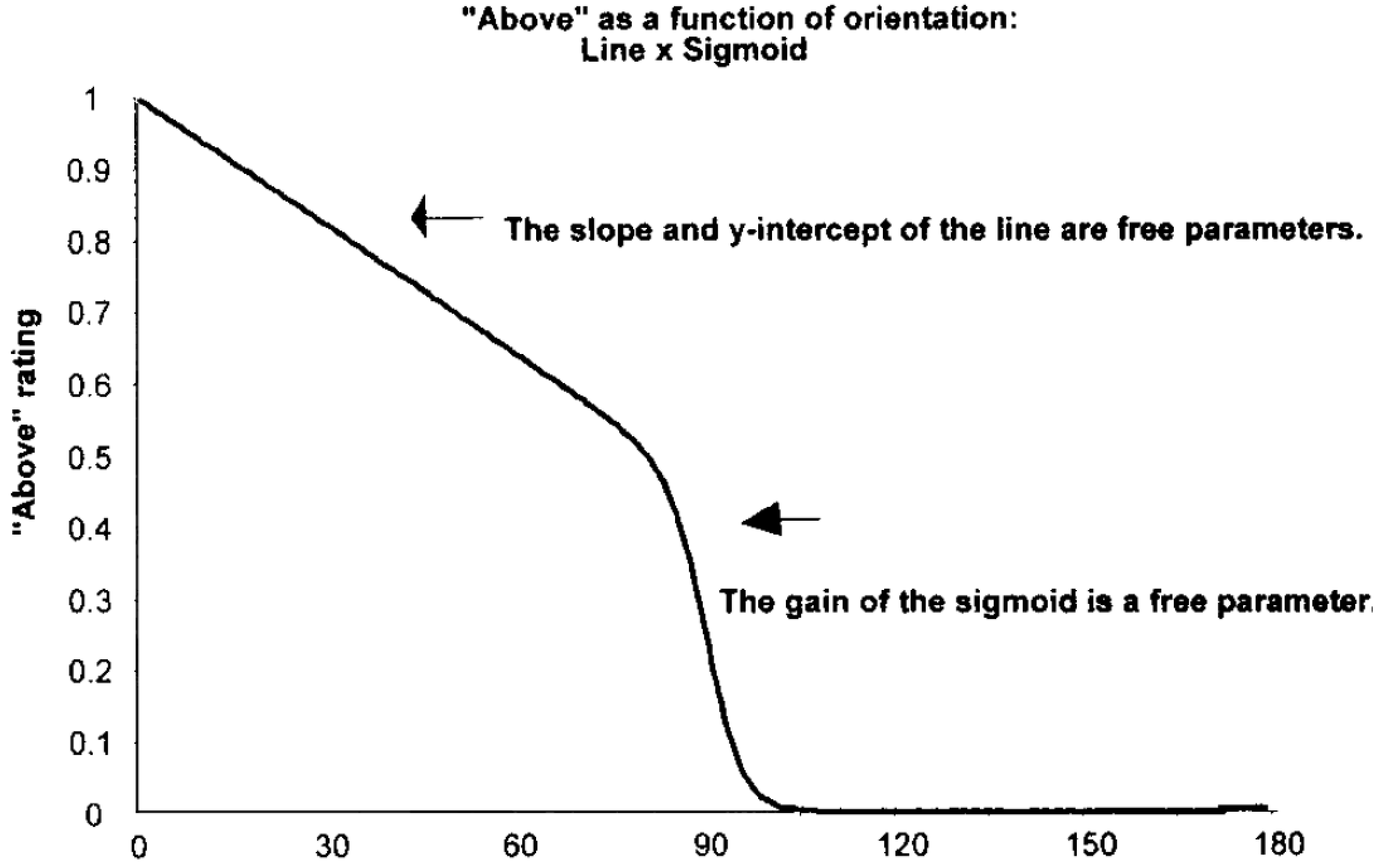
Terry Regier
University of Chicago

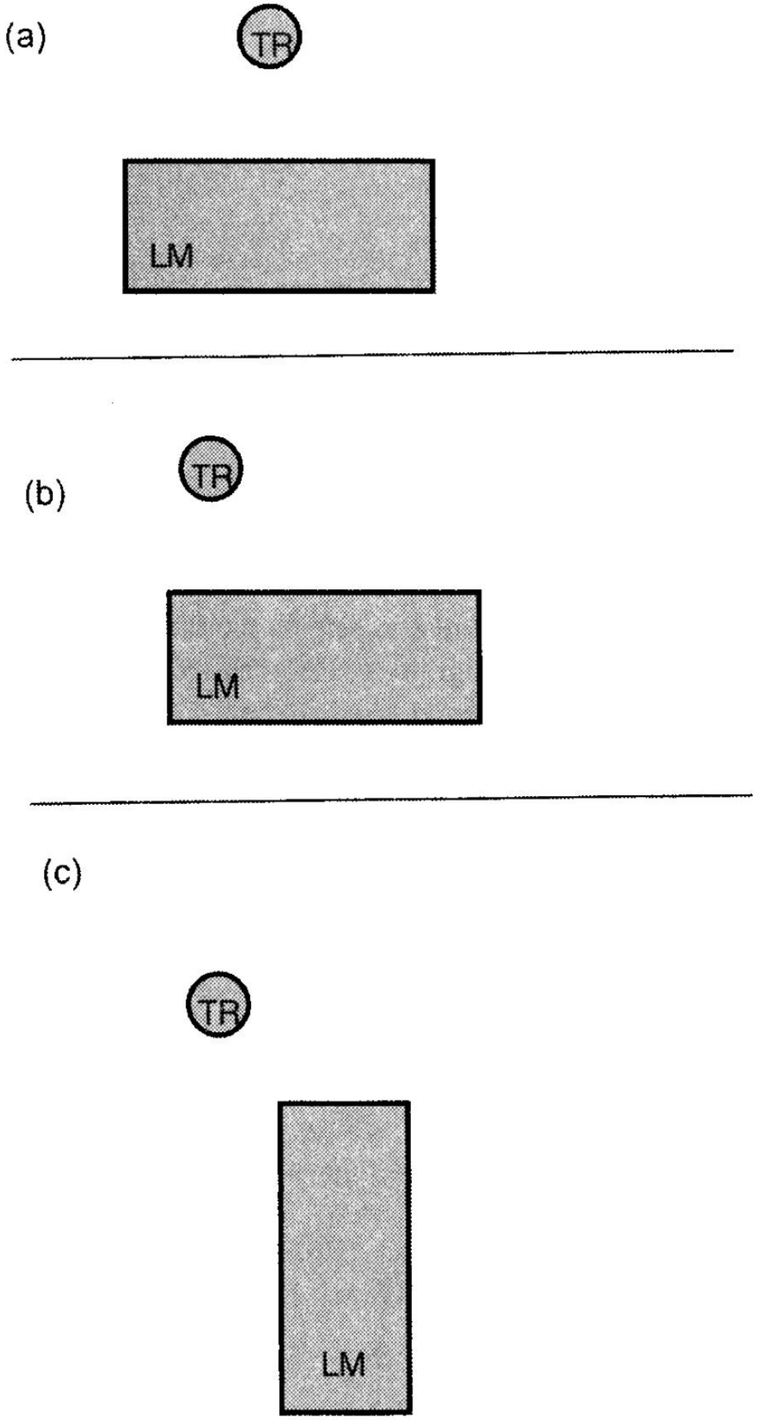
Laura A. Carlson
University of Notre Dame

Models of above(x,y)

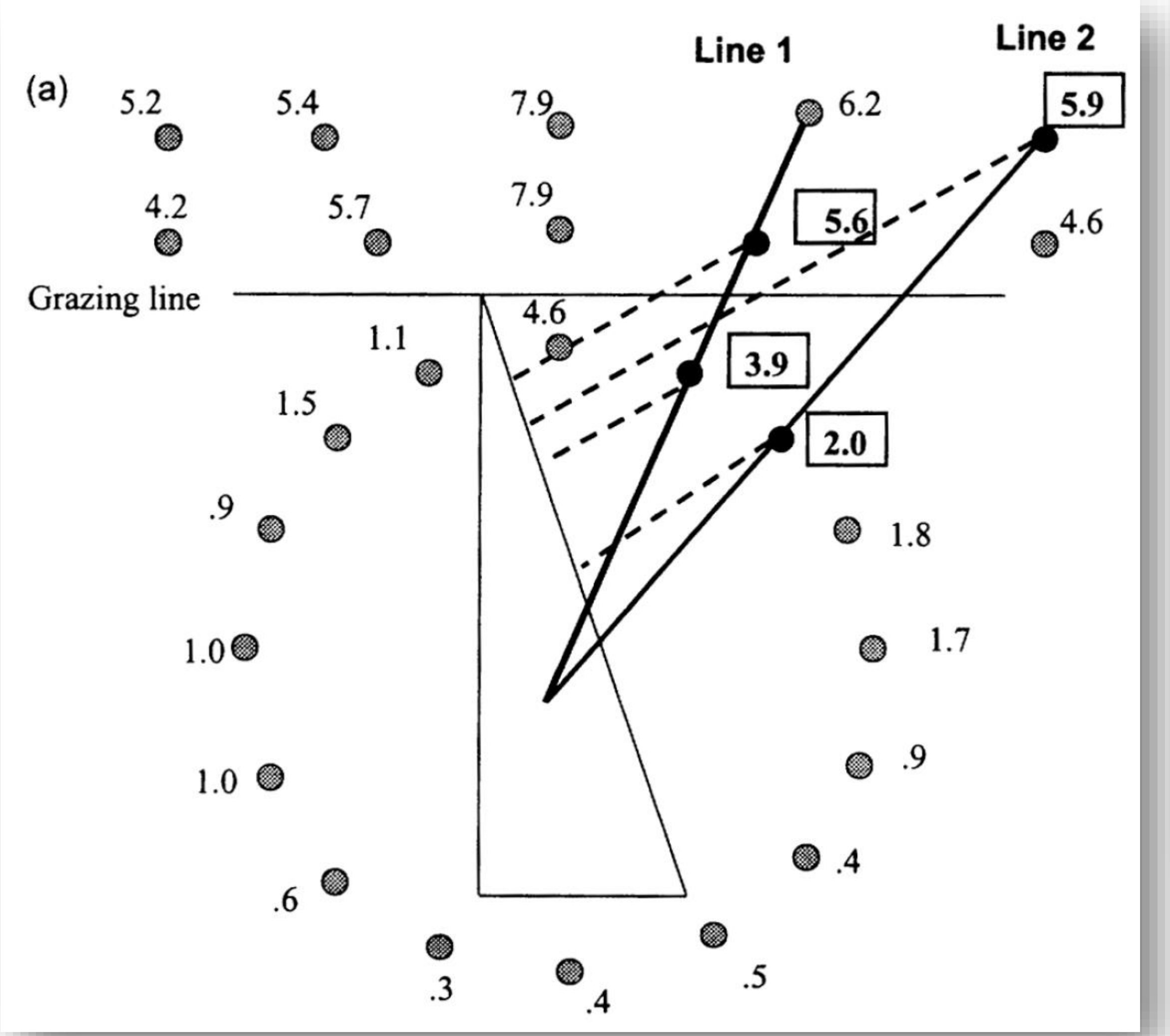
- Bounding box: above highest point, use betweenness of left-/right-most points
- Proximal/CoM: use angle between CoM of LM and TR
- Hybrid: take CoM model, then apply height as a feature
- Attention Vector-Sum model:
  - Human judgments involve *attention*: where the person focuses
  - Direction is represented as a vector sum of a set of constituent directions (neuro)

Bounding box model

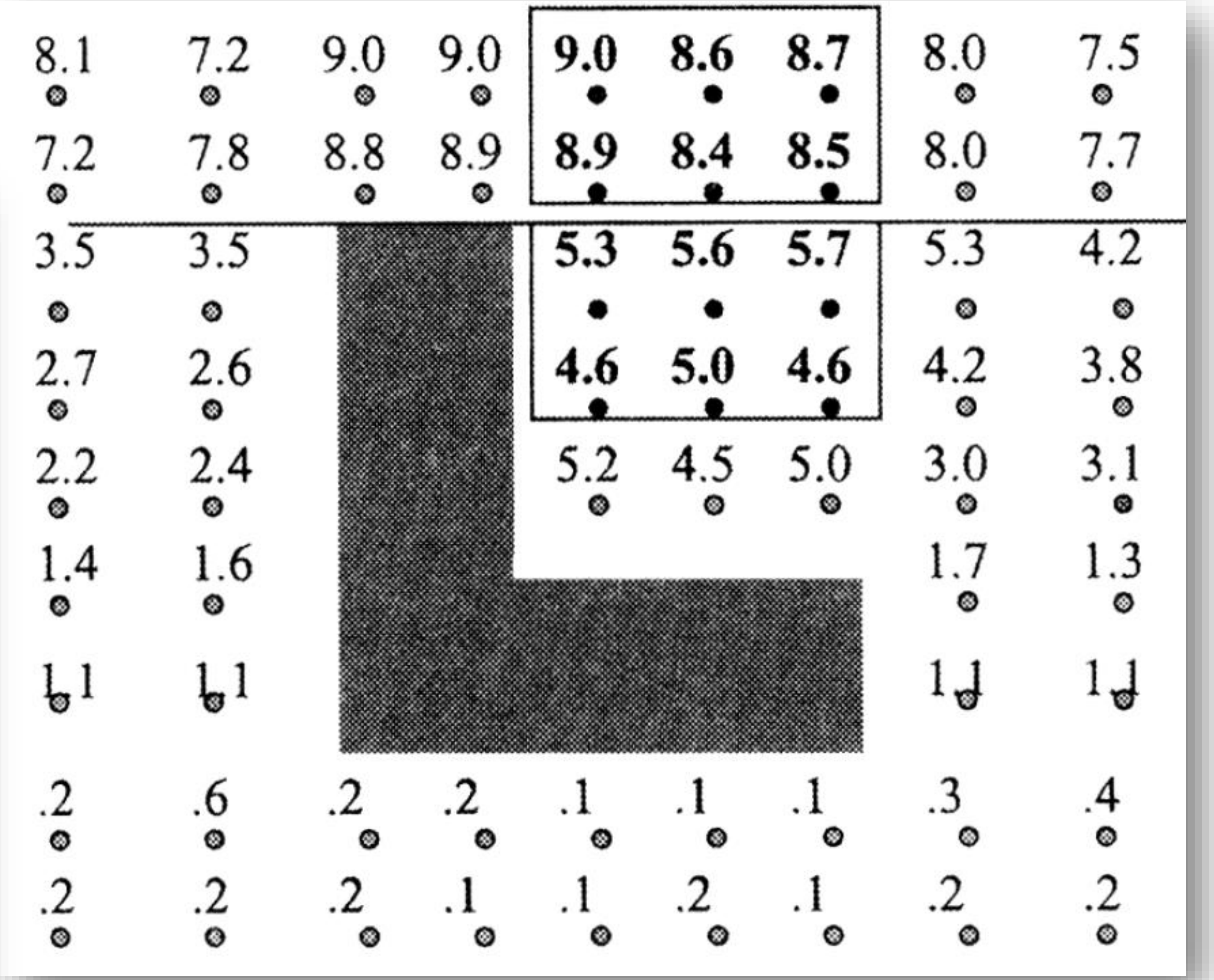




CoM Model



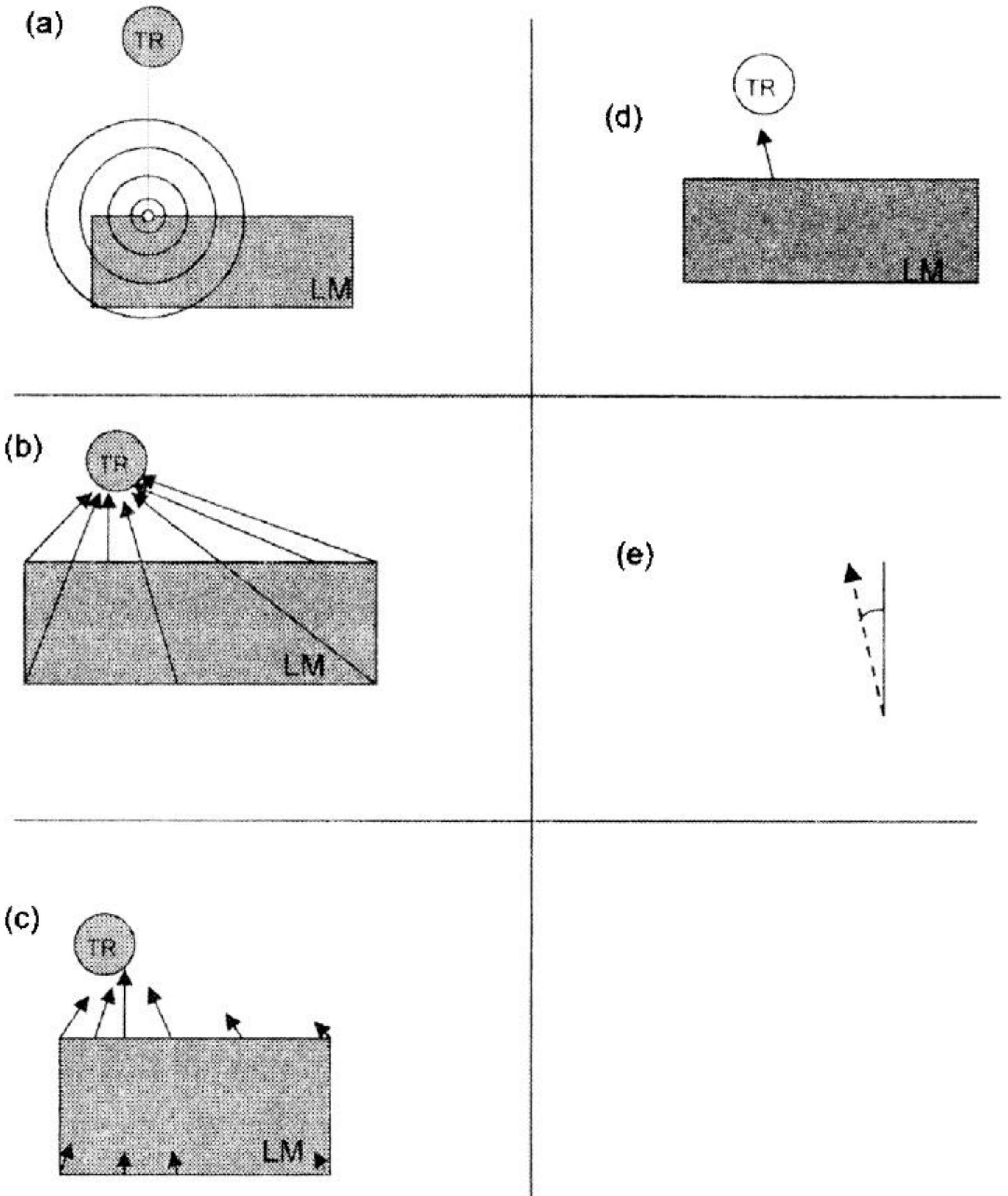
(a)
Line 1
Line 2
5.2
5.4
7.9
6.2
5.9
4.2
5.7
7.9
5.6
4.6
Grazing line
4.6
1.1
3.9
1.5
2.0
.9
1.8
1.0
1.7
1.0
.9
.4
.6
.3
.4
.5



8.1 7.2 9.0 9.0 9.0 8.6 8.7 8.0 7.5
7.2 7.8 8.8 8.9 8.9 8.4 8.5 8.0 7.7
3.5 3.5 5.3 5.6 5.7 5.3 4.2
2.7 2.6 4.6 5.0 4.6 4.2 3.8
2.2 2.4 5.2 4.5 5.0 3.0 3.1
1.4 1.6 1.7 1.3
1.1 1.1 1.1 1.1
.2 .6 .2 .2 .1 .1 .1 .3 .4
.2 .2 .2 .1 .1 .2 .1 .2 .2

Models of above(x,y)

- Bounding box: above highest point, use betweenness of left-/right-most points
- Proximal/CoM: use angle between CoM of LM and TR
- Hybrid: take CoM model, then apply height as a feature
- Attention Vector-Sum model:
  - Human judgments involve *attention*: where the person focuses
  - Direction is represented as a vector sum of a set of constituent directions (neuro)



Fit of models to data

Parameter Settings for Each Model

Model parameter	Parameter value: Logan & Sadler (1996)	Parameter value: Experiment 7
BB		
Gain on left–right sigmoids	0.109	0.065
Gain on top sigmoid	0.066	0.373
Exponent for left–right sigmoids	0.062	0.220
PC		
α, relative weight of P and C	0.500	0.174
y-intercept of alignment function	0.969	0.929
Slope of alignment function	−0.005	−0.006
Gain on sigmoid	0.112	3.265
PC-BB		
		0.115
		0.909
		−0.005
		6.114
		0.512
		1.224
		−0.007
		0.002

Model Fits to Above Data From Logan and Sadler (1996)

Model	R^2	Adj R^2	Slope	y-intercept
BB	.904	.897	0.907	0.038
PC	.959	.955	1.011	−0.024
PC-BB	.963	.959	1.030	−0.036
AVS	.963	.959	1.030	−0.036

Fits to lots more data examples

Model Fits to Data from Experiments 1–7, Broken Down by Landmark Shape

Model	R^2	Adj R^2	Slope	y-intercept
Experiment 1				
Tall rectangle (24 data points)				
BB	.982	.979	0.945	−0.387
PC	.963	.955	0.975	0.073
PC-BB	.995	.994	1.075	−0.616
AVS	.996	.995	1.088	−0.614
Wide rectangle (24 data points)				
BB	.971	.966	0.981	−0.076
PC	.954	.944	0.989	0.202
PC-BB	.992	.992	1.024	−0.431
AVS	.993	.992	1.017	−0.407
Experiment 4				
Upright triangle (4 data points)				
BB	.963		1.048	0.212
PC	.967		0.697	2.373
PC-BB	.993		1.485	−3.723
AVS	.991		1.402	−2.859
Inverted triangle (4 data points)				
BB	.999		1.102	−0.329
PC	.987		1.037	−0.161
PC-BB	.986		1.150	−0.909
AVS	.990		1.150	−0.907

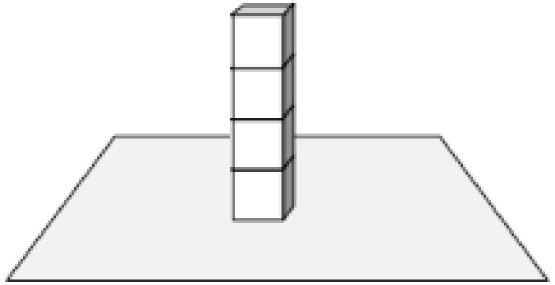
Model	R^2	Adj R^2	Slope	y-intercept
Experiment 5				
L shape (65 data points)				
BB	.943	.941	0.960	−0.141
PC	.862	.853		
PC-BB	.943	.940		
AVS	.976	.975		
Experiment 6				
Tall triangle (31 data points)				
BB	.750	.723		
critical points only (14 data points)				
BB	.784	.719	1.635	−5.103
PC	.400	.133	0.428	4.552
PC-BB	.367	.086	0.243	6.015
AVS	.888	.838	1.138	−1.287
Composite (337 data points)				
BB	.953	.952	1.007	−0.242
PC	.910	.909	0.926	0.443
PC-BB	.959	.958	1.01	−0.450
AVS	.970	.970	1.031	−0.439

Composite (337 data points)	R^2	Adj R^2	Slope	y-intercept
BB	.953	.952	1.007	−0.242
PC	.910	.909	0.926	0.443
PC-BB	.959	.958	1.01	−0.450
AVS	.970	.970	1.031	−0.439

Some more recent stuff on spatial language

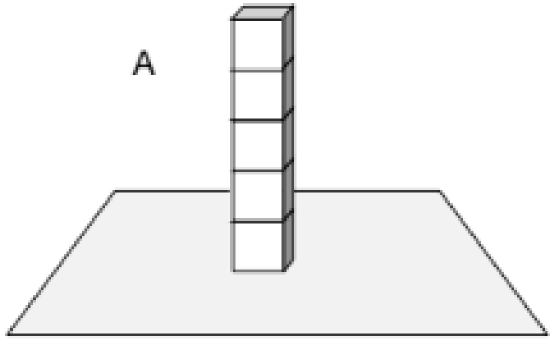
Item 1

Someone shows you this configuration and asks you to “add a block”

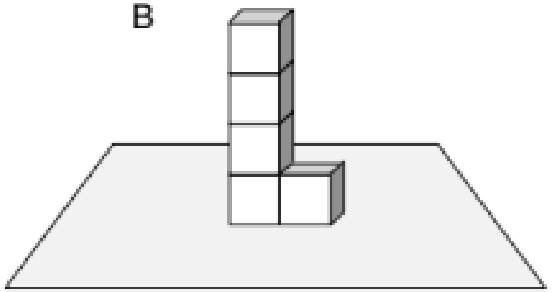


Which of these configurations do they probably have in mind?

A

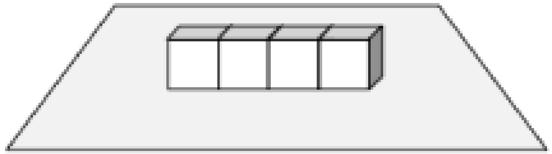


B



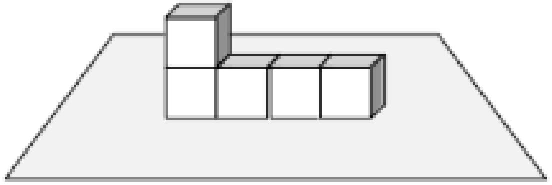
Item 2

Someone shows you this configuration and asks you to “add a block”

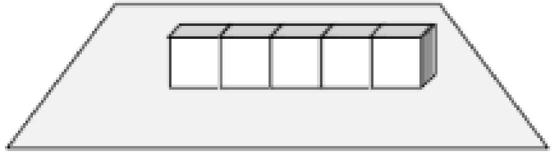


Which of these configurations do they probably have in mind?

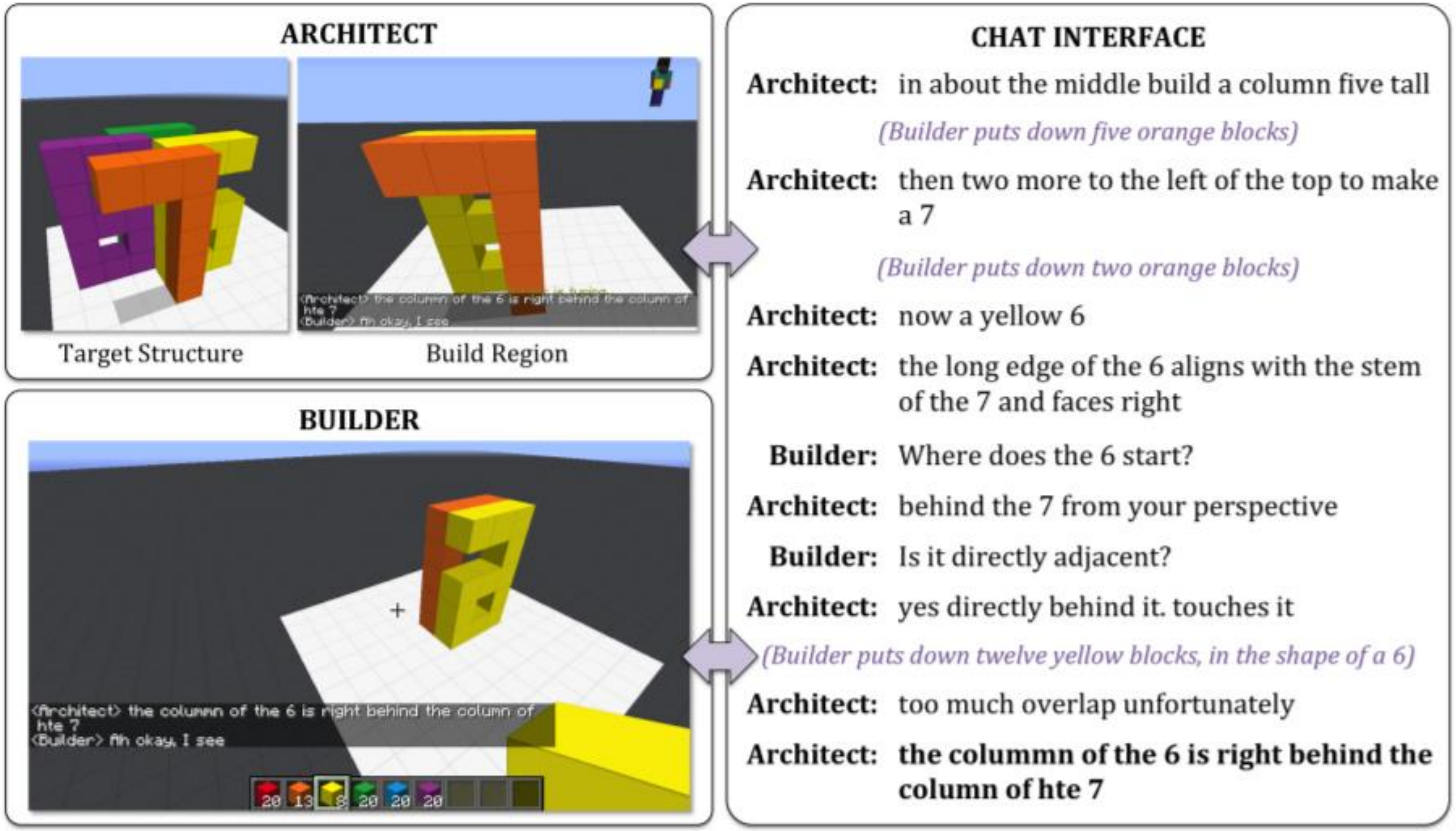
A



B



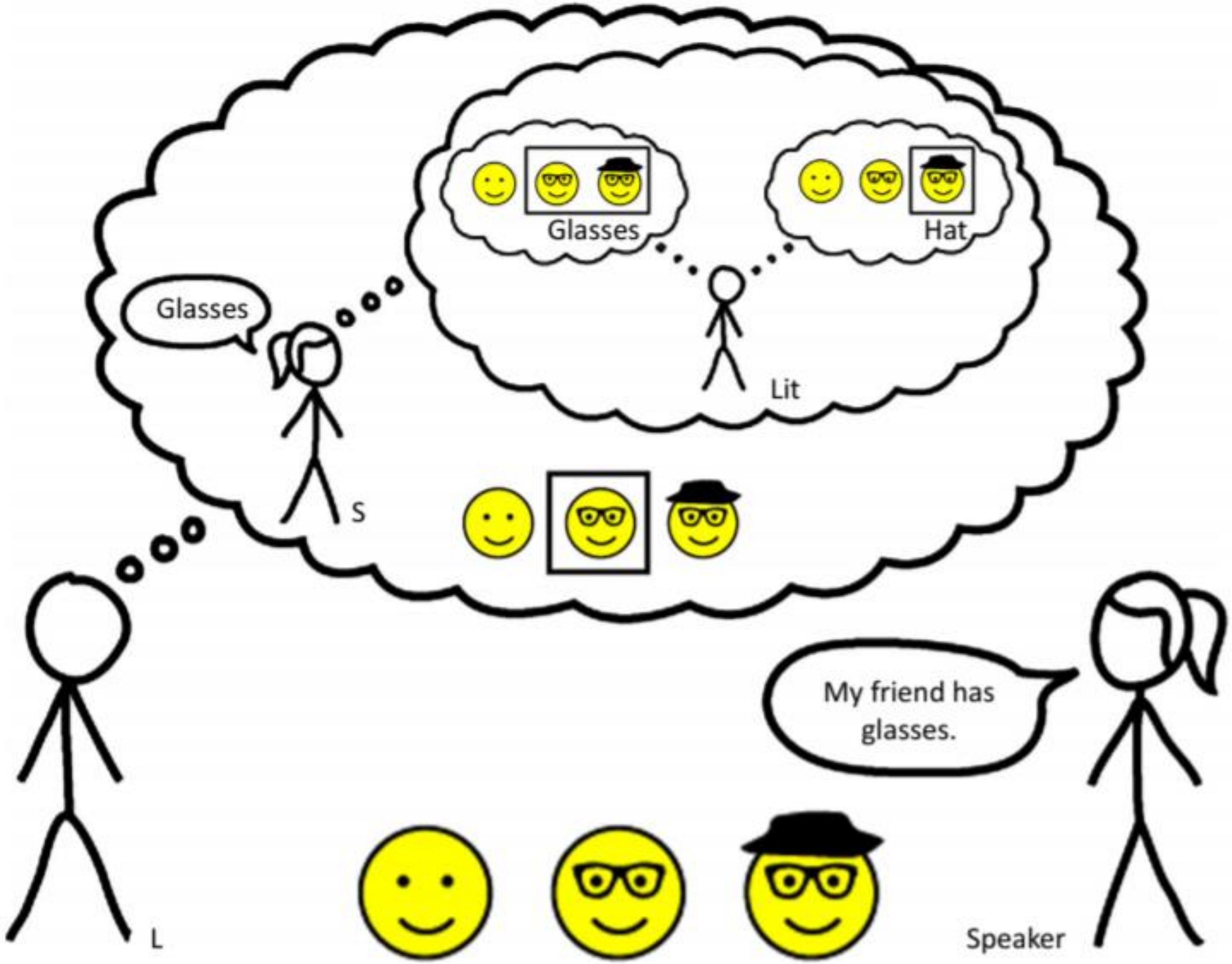
Some more recent stuff on spatial language



Collaborative Dialogue in Minecraft

Anjali Narayan-Chen* **Prashant Jayannavar*** **Julia Hockenmaier**
University of Illinois at Urbana-Champaign
{nrynchn2, paj3, juliahmr}@illinois.edu

Grounding based on pragmatics

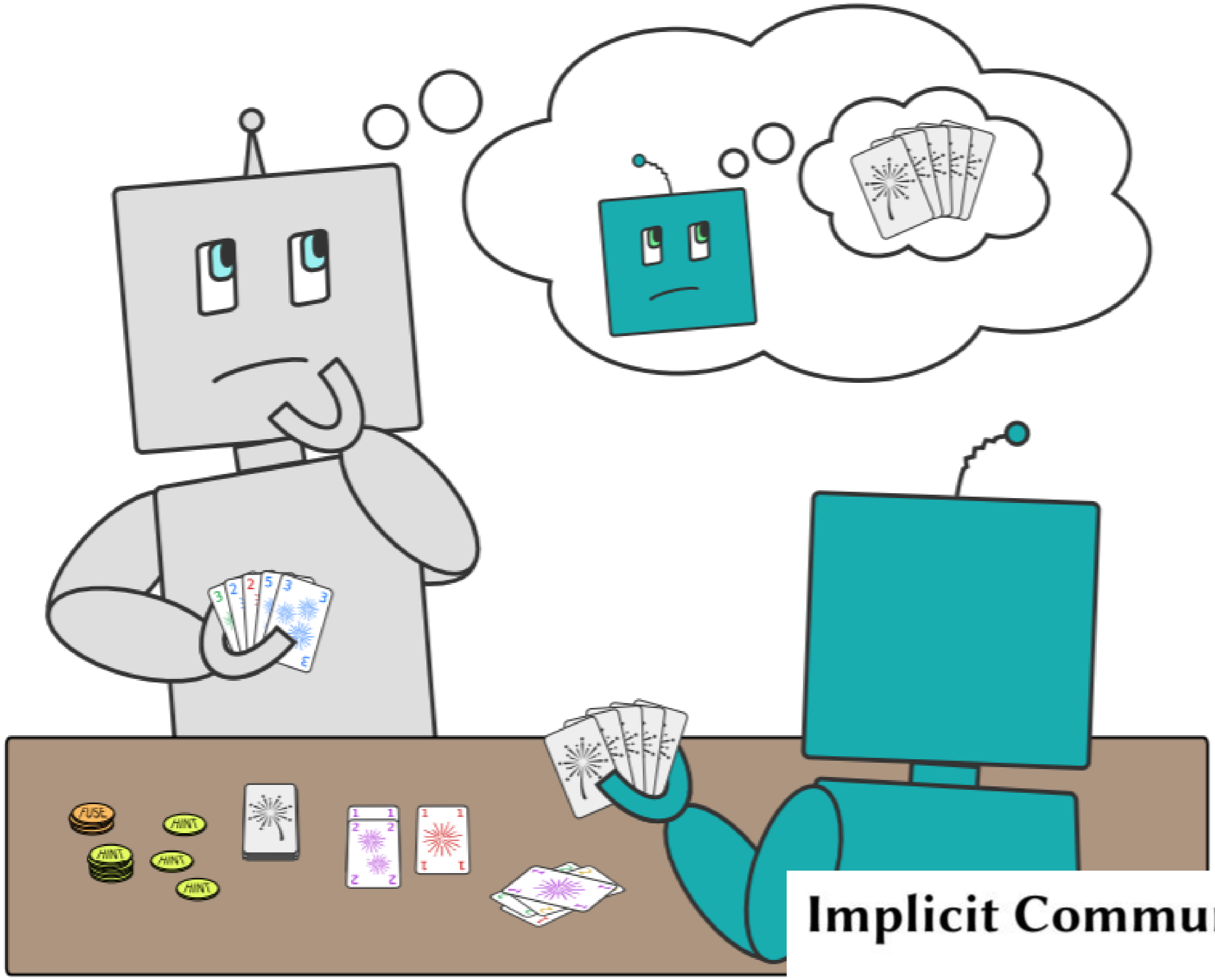


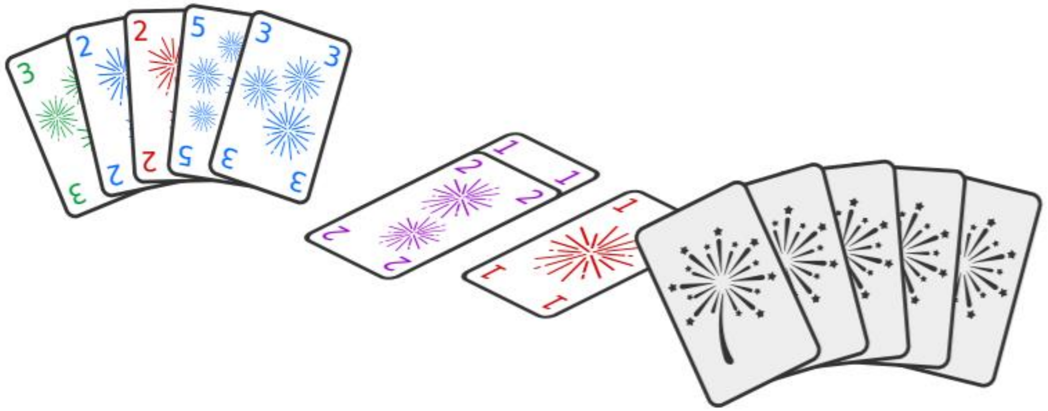
Review

Pragmatic Language Interpretation as Probabilistic Inference

Noah D. Goodman[1,*] and Michael C. Frank[1]

Grounding based on pragmatics in teams





Implicit Communication of Actionable Information in Human-AI teams

Claire Liang, Julia Proft, Erik Andersen, and Ross A. Knepper
Department of Computer Science, Cornell University
cyl48@cornell.edu,{jproft,eland,rak}@cs.cornell.edu

Today

- Grounding
  - Relationship between linguistic symbols and stuff in the real world
- Spatial language
  - Relationships between objects in the world
- Implicit vs explicit communication