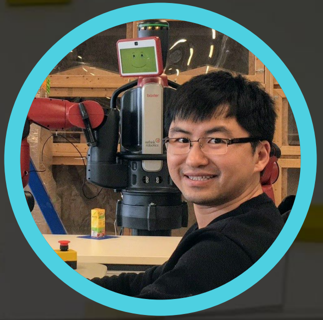
INLG 2022 – 15th International Natural Language Generation Conference

Evaluating Referring Form Selection Models in Partially-Known Environments



Zhao Han

Polina Rygina

Tom Williams

COLORADOSCHOOLOFMINES
EARTH • ENERGY • ENVIRONMENT

July 19, 2022

MIRRORLab
Mines Interactive Robotics Research

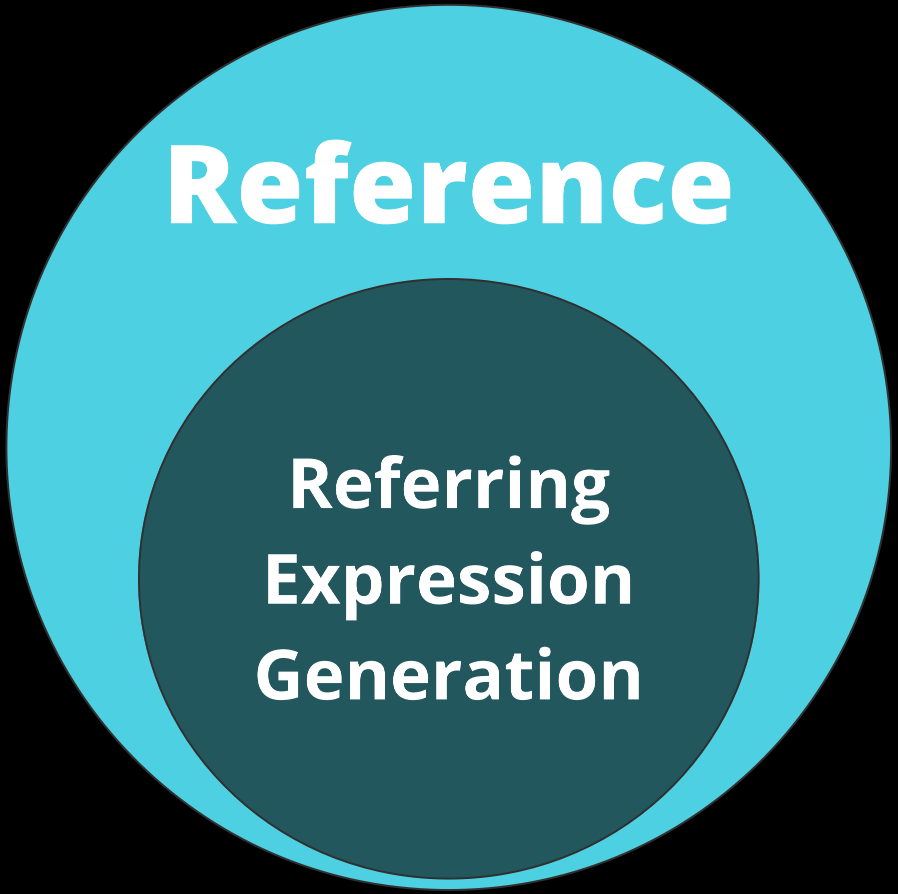
Background: Reference & Generation

Reference

- One of the most studied dimensions of natural language pragmatics
- To pick out things of interest & how to interpret/resolve references

Referring Expression Generation

- The focus of reference research
- To select the properties to be used in a generated expression
  - e.g., choosing to highlight the redness, or the boxiness, of a red box, among other possible properties



Background: Referring Form Selection

Referring Form Selection

- Important initial step during language generation
- Speaker must select a more general referring form, such as “it”, “that”, or “the ⟨N′⟩”

Little is done on its **computational modeling**

How do we model referring form selection?

Models of Referring Form Selection

Rational models:

- Egocentric, whether to use pronouns (e.g., ease of production)
- Prediction are thus mostly reduced forms, rather than used in practice

Pragmatic models:

- Allocentric, explain why certain pronouns are chosen

One pragmatic model is the **Givenness Hierarchy (GH)** theory

- A hierarchically nested set of **Cognitive Statuses**
    - {in focus $\subseteq$ activated $\subseteq$ familiar $\subseteq$ uniquely identifiable $\subseteq$ referential $\subseteq$ type identifiable}

Rational models

Pragmatic models

Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

Problem of Linguistic Models

To make matters work, both models:

- Focus on specific referential phenomena
- **Less on comprehensive model** of entire process of reference production

And, computationally, they provide **little input into algorithms that govern this process (and precisely predict)**

Critically to those **studying situated interaction**

- Previous work was assessed on **corpora without any situated features, e.g., physical distance**

Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

Problem of Computational Models

"Multifactorial process modeling"

- **Do not attempt to predict at fine-grain level** of referring forms
- Assessed in **pure text domains**, avoiding challenges in ambiguous open worlds

Recent efforts

- Achieved over 80% accuracy in predicting the referring forms
- Using data from situated interactants in human-human & human-robot interactions

Was the task domain suited?

The Task Domain Is Ill-Suited

1. Few candidate referents
(thus constantly activated)

→ ***"this"***

2. Uniquely identifiable

→ ***proper nouns***

3. Visible all the time

→ **exacerbate & discourage use of indefinite nouns**

Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

Solution

1. New task context
2. New corpus
3. Re-assess existing models

Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

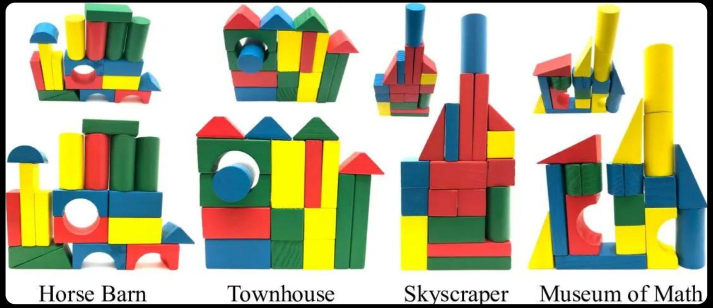
1. New Task Environment and Design

Tower construction in four quadrants

Pairs of participants: instructor teach learner to construct buildings (highly interactive)

A building (total 4):

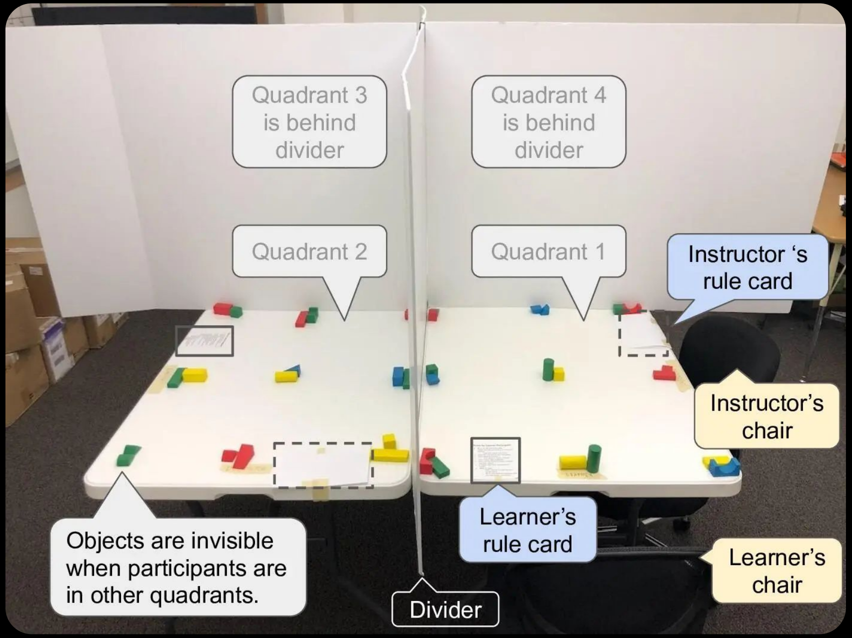
- 18 repeated blocks – for more "this", "that", "it"



Half objects in current quadrant, **the other half hidden in other quadrants**

- Visibility changes after switching quadrants

Within a quadrant: blocks are at vertices of 3 × 3 grid – more referring forms varied by distance (this, that)



Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

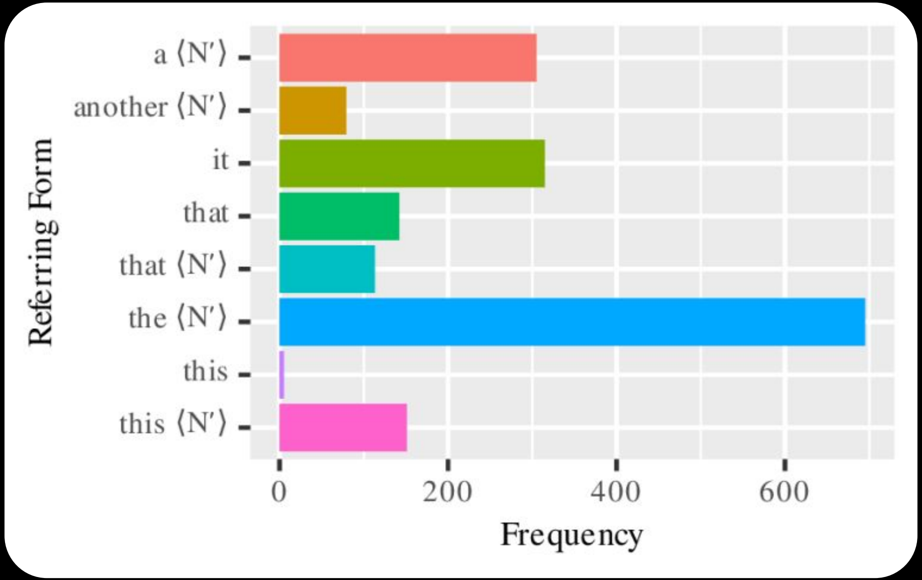
2. New Situated Corpus

Dyad corpus

- Eleven collections of four monologues
- Collection: 27:32 minutes on average
- Monologue: 6:53 on average

- 1867 referring expressions
- Each participant: 169.7 referring forms
  - Significantly more than 18 forms in previous work

Wider range of referring forms



20.5% indefinite nouns

- 16.3% "a *⟨N′⟩*", 4.2% "another *⟨N′⟩*"

3. Re-assess Existing Models

Object features

01	Cognitive status	• Most informative feature • Predicted by Pal et al.'s model
02	Number of distractors	• Objects with the same cognitive status or higher
03	Physical distance	• e.g., {near (N), middle (M), far (F)} • e.g., {left (L), middle (M), right (R)}
04	Temporal distance	• 0: if not mentioned yet • 1/n: the number of objects referred since the object was mentioned

Model

Decision tree algorithm

Types:

Model	Removed Feature
M1	N/A (full model)
M2	Cognitive status
M3	Number of distractors
M4	Physical distance
M5	Temporal distance

Five-fold cross validation

3. Re-assess Existing Models (Results)

Model with our new corpus (~60% accuracy)

	Six GH informed referring forms					*With two indefinite forms*				
	M1	M2	M3	M4	M5	M1’	M2’	M3’	M4’	M5’
Accuracy	**65.73**	64.11	**65.80**	62.98	61.72	**59.83**	58.95	**59.83**	51.30	57.29
RMSE	0.343	0.359	0.342	0.370	0.383	0.402	0.411	0.402	0.487	0.427
Precision	0.552	0.543	0.552	0.509	0.521	0.493	0.487	0.493	0.435	0.476
Recall	0.657	0.641	0.658	0.630	0.617	0.598	0.589	0.598	0.513	0.573
F1 score	0.589	0.576	0.589	0.542	0.556	0.536	0.528	0.536	0.445	0.514

Model with previous corpus
(~80% accuracy)

	M1	M2	M3	M4	M5
Accuracy	84.74	79.6	71.97	83.58	**86.07**
RMSE	0.197	0.230	0.244	0.208	**0.195**
Precision	0.858	0.820	0.710	0.840	**0.882**
Recall	0.847	0.796	0.720	0.836	**0.861**
F1 score	0.843	0.811	0.716	0.838	**0.858**

~20% accuracy drop

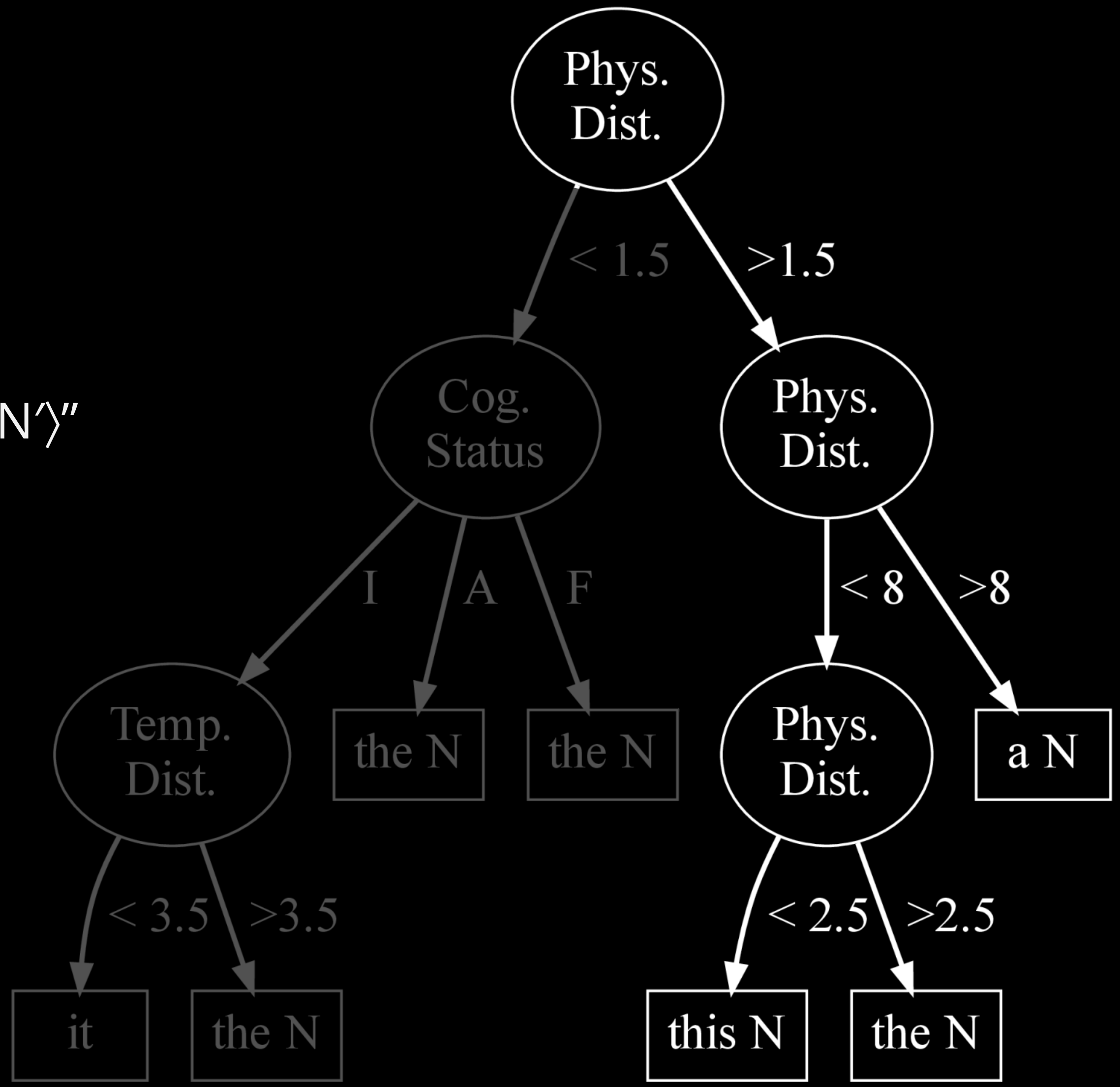
Zhao Han, Polina Rygina, Tom Williams, “**Evaluating Referring Form Selection Models in Partially-Known Environments**”, *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

Model Interpretation

Physical distance

- Rightmost branch: phys. dist. → “a ⟨*N′*⟩”
- To the left: phys. dist. → “this ⟨N′⟩” & “the ⟨N′⟩”



Zhao Han, Polina Rygina, Tom Williams, “**Evaluating Referring Form Selection Models in Partially-Known Environments**”, *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

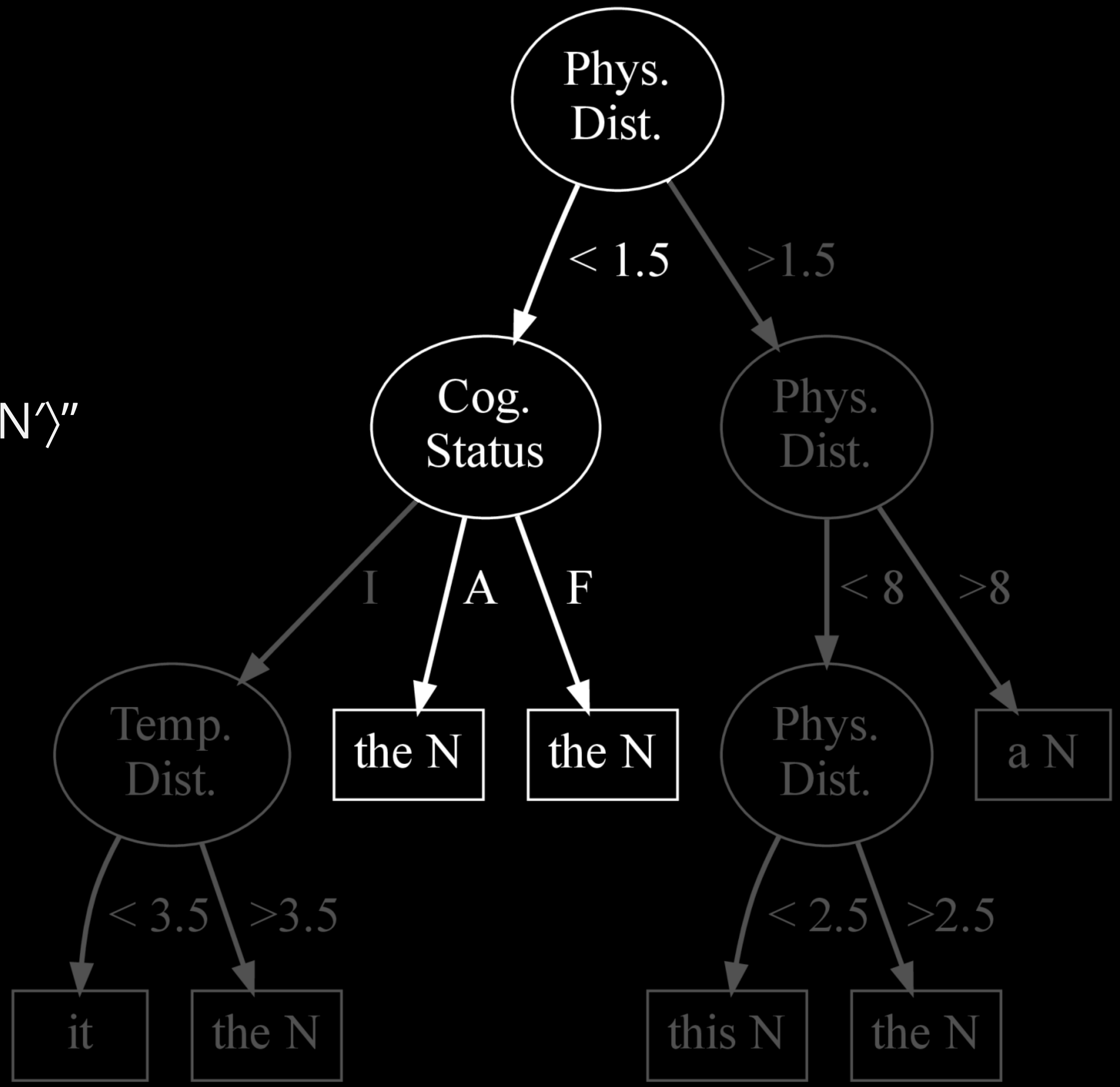
Model Interpretation

Physical distance

- Rightmost branch: phys. dist. → "a ⟨*N′*⟩"
- To the left: phys. dist. → "this ⟨N′⟩" & "the ⟨N′⟩"

Cognitive status

- Middle two branches: A & F → "the ⟨N′⟩"



Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. bit.ly/inlg2022

@hanzhao

Model Interpretation

Physical distance

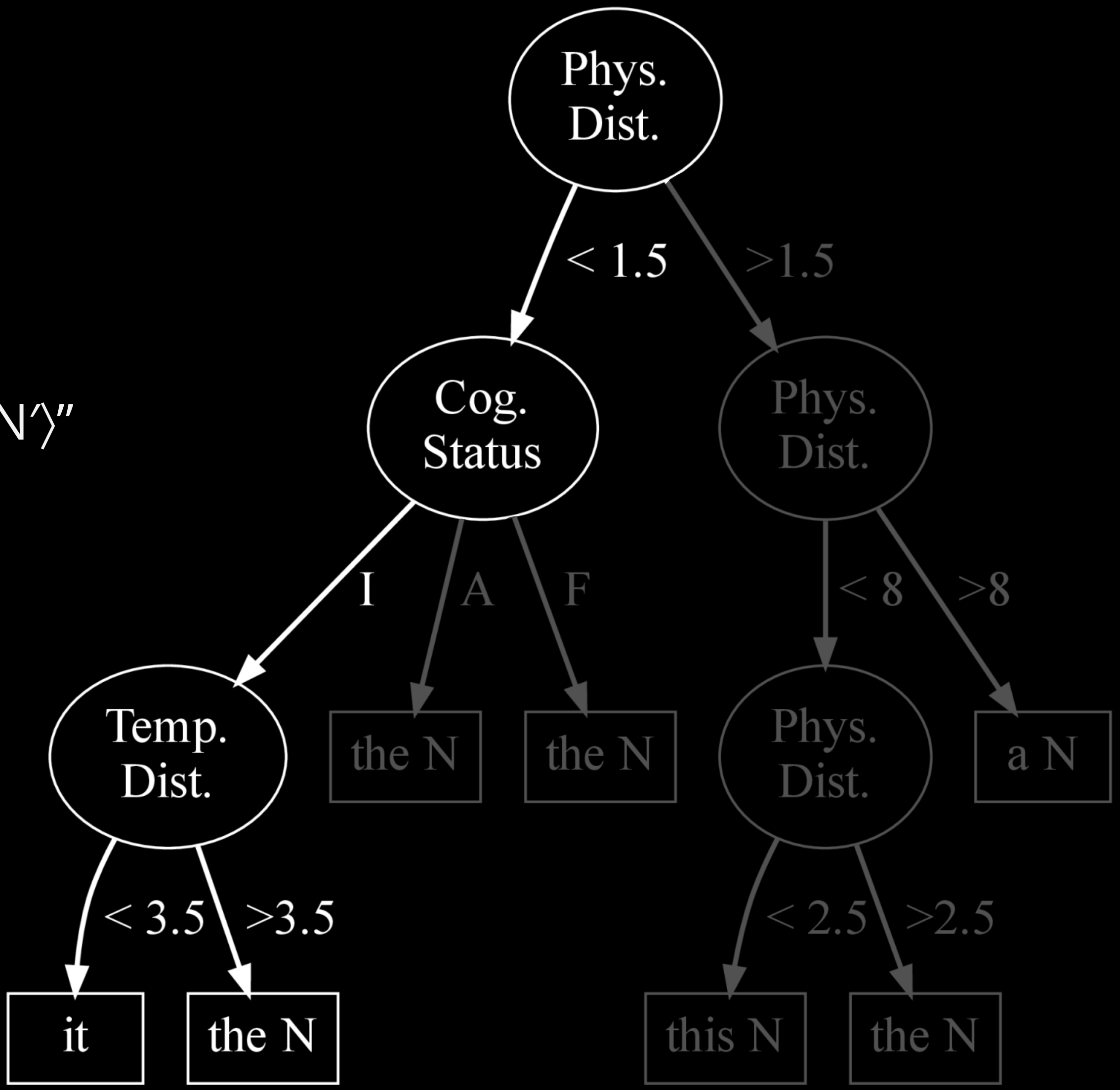
- Rightmost branch: phys. dist. → "a ⟨*N′*⟩"
- To the left: phys. dist. → "this ⟨N′⟩" & "the ⟨N′⟩"

Cognitive status

- Middle two branches: A & F → "the ⟨N′⟩"

Temporal distance

- <3.5 → "it"



Zhao Han, Polina Rygina, Tom Williams, "**Evaluating Referring Form Selection Models in Partially-Known Environments**", *2022 ACL International Conference on Natural Language Generation*. **bit.ly/inlg2022**

@hanzhao

Evaluating Referring Form Selection Models in Partially-Known Environments

Zhao Han
zhaohan@mines.edu
zhaohanphd.com
@hanzhao

MIRRORLab
Mines Interactive Robotics Research
mirrorlab.mines.edu
@mirrorlab

This work was funded in part by:

Takeaways

1. We proposed a **novel, situated task**
   a. more and invisible objects
   b. comprehensive referring form data
2. We re-assessed performance of existing model and saw **20% drop with our new corpus**
3. Performance drop showed **more, non-uniquely identifiable, repeated, invisible objects are useful** to evaluate referring form selection models