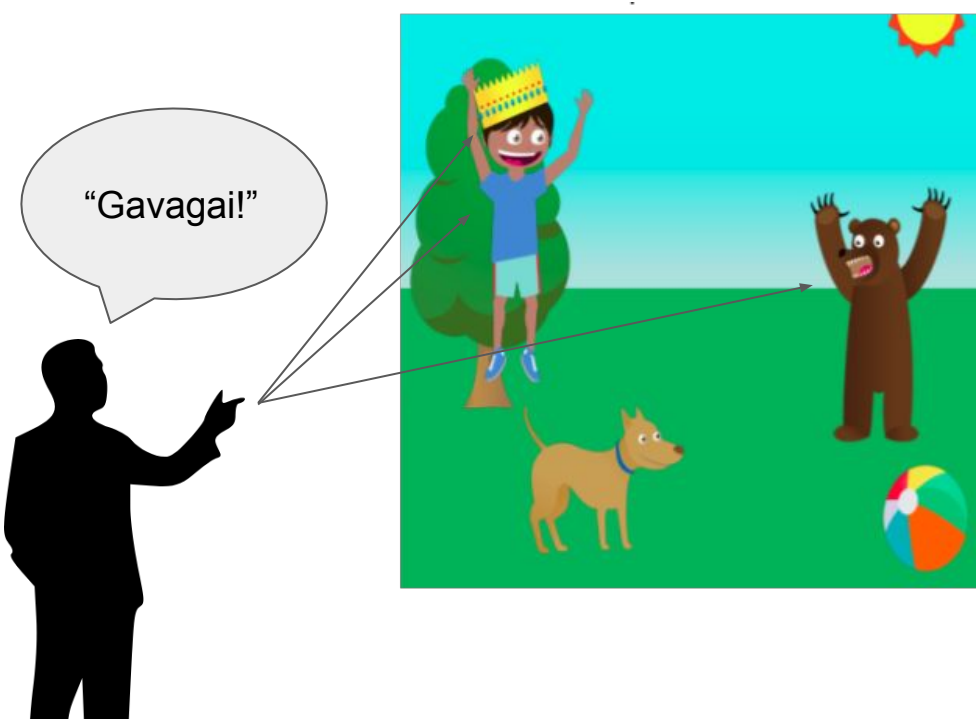
Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in ANNs

Mitja Nikolaus & Abdellah Fourtassi

Learning of Semantics

- Learning semantics in a largely unsupervised way from ambiguous input



Quine (1960)

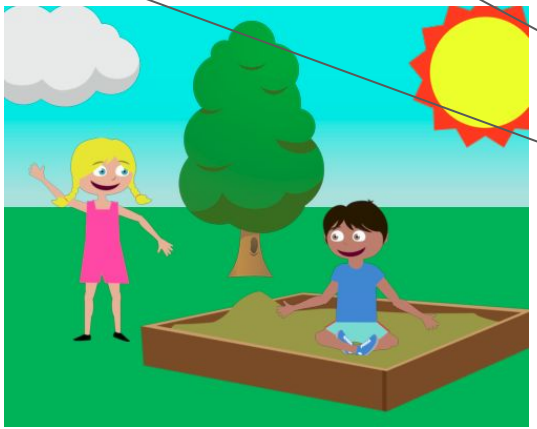
Evaluation: Two-alternative forced choice (2AFC)

“apple”

Bergelson and Swingley (2012)

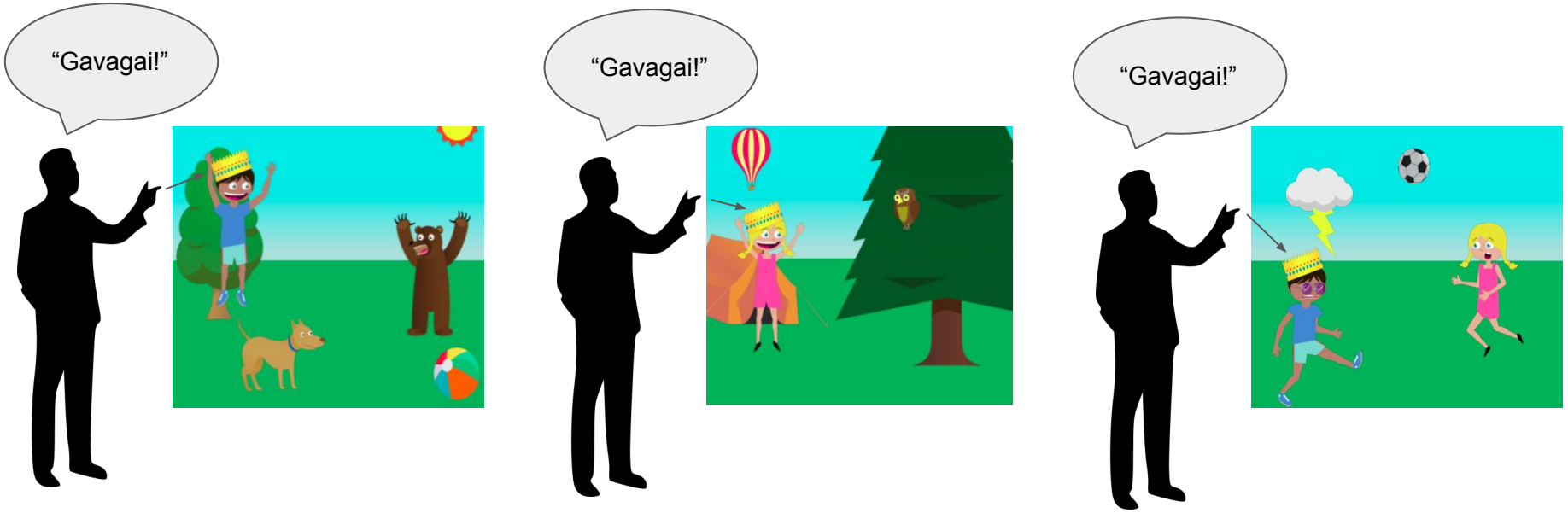
Evaluation: Two-alternative forced choice (2AFC)

"**Mike** is waving to **Jenny**"



e.g. Noble et al. (2011); Gertner and Fisher (2012)

Cross-situational learning



Tested in laboratory studies (e.g. Yu and Smith (2007))

→ Plausibility in more natural large-scale learning contexts?

Related work and contributions

- Modeling cross-situational word learning (from images and text!)
- Evaluation of word-level semantics:
  - Using a reference dictionary (Lazaridou et al., 2016)
  - Comparison to human similarity judgments (Kádár et al., 2015; Chrupała et al. 2015)
- Evaluation of sentence-level semantics:
  - Image-sentence retrieval (+ using scrambled sentences) (Chrupała et al. 2015)
  - Comparison to human similarity judgments (Merkx and Frank, 2019)

- **Here: Fine-grained testing of phenomena using 2AFC**
  - **Wider range of word-level semantics (nouns, adjectives, verbs)**
  - **Dependencies between predicates and arguments**
  - **Semantic roles**

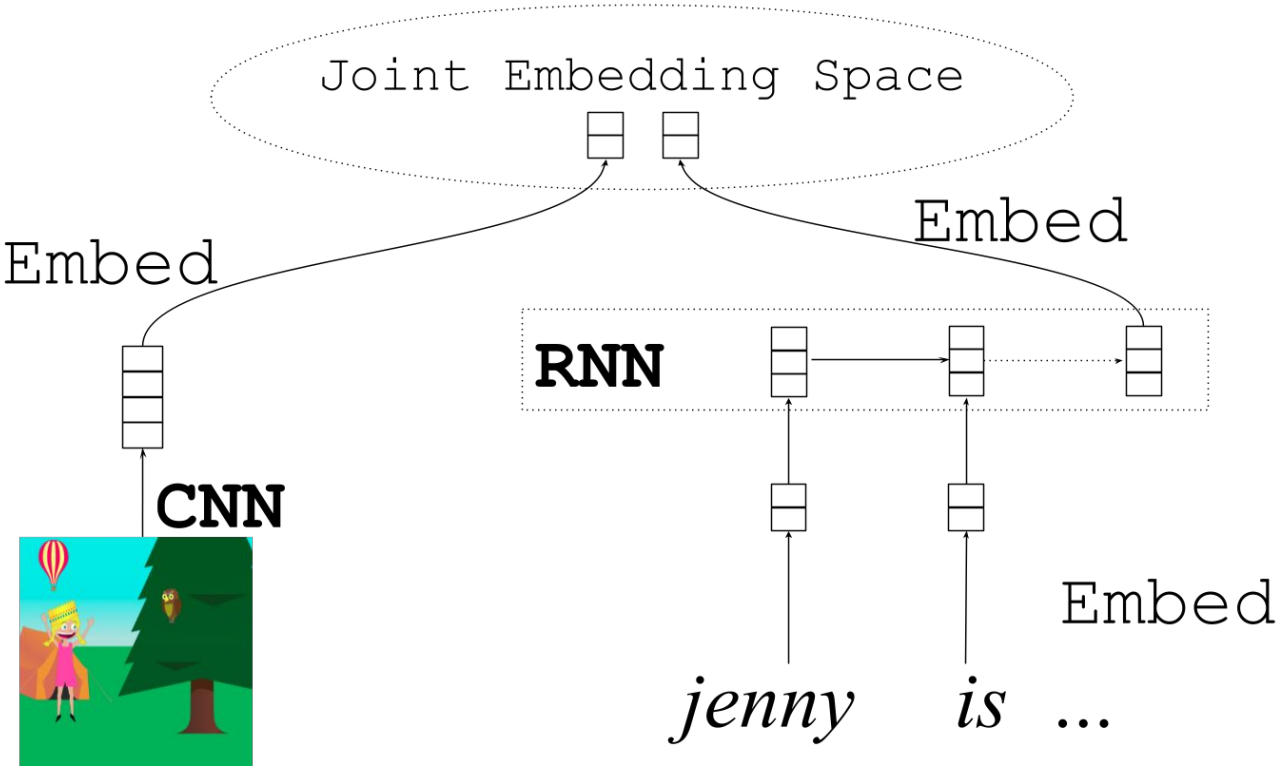
Dataset

Abstract Scenes (Zitnick and Parikh, 2013)

- 10K crowd-sourced images, 6 descriptive captions per image
- Train (80%), validation (10%), test set (10%)

Mike kicks the soccer ball to Jenny
Jenny is sad because Mike is mad
Mike is near an apple tree
The sun is partly behind a tree
Mike is angrily kicking the soccer ball
Jenny is crying as she plays with Mike

Cross-situational Learner Model: Architecture



e.g. Karpathy and Fei-Fei (2015), Faghri et al. (2018)

Cross-situational Learner Model: Training objective

$$\mathcal{L}(\theta) = \sum_a [\sum_b max(0, \gamma(i_a, s_b) - \gamma(i_a, s_a) + \alpha)$$
$$+ \sum_b max(0, \gamma(i_b, s_a) - \gamma(i_a, s_a) + \alpha)] \quad (1)$$

$\gamma(i_a, s_b)$ = cosine similarity between image i and sentence s

e.g. Karpathy and Fei-Fei (2015), Faghri et al. (2018)

Evaluation

Test trial: Image, target sentence, distractor sentence: (i, s_t, s_d)

Example

Target: **jenny** is wearing a crown
Distractor: **mike** is wearing a crown

Counter-Example

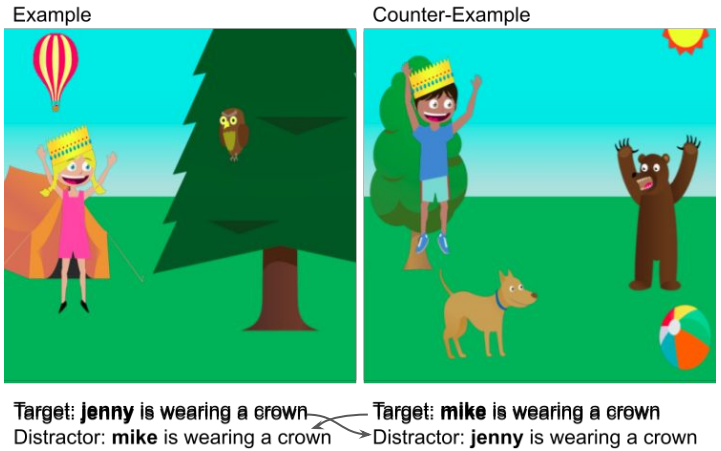
Target: **mike** is wearing a crown
Distractor: **jenny** is wearing a crown

Evaluation: Search for minimal pairs

Search the test set for image-sentence pairs $[(i_x, s_x), (i_y, s_y)]$ with **minimal differences**

Generate 2 counter-balanced test trials:

(i_x, s_x, s_y) and (i_y, s_y, s_x)



Evaluation: Nouns

Nouns: Persons

Target: **jenny** is wearing a crown
Distractor: **mike** is wearing a crown

Nouns: Animals

Target: the **cat** is looking at jenny
Distractor: the **dog** is looking at jenny

Nouns: Objects

Target: jenny has a **pizza**
Distractor: jenny has a **hat**

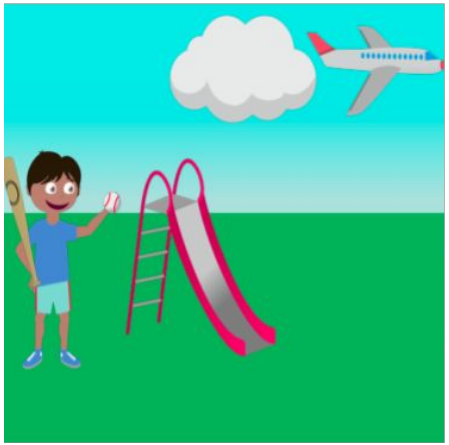
Evaluation: Adjectives & Verbs

Trimming: “mike is eating an apple” → “mike is eating”

Verbs

Target: mike is **sitting**
Distractor: mike is **standing**

Adjectives



Target: mike is **happy**
Distractor: mike is **sad**

Evaluation: Sentence-level semantics

Adjective-Noun Dependency

Target: mike is **happy**
Distractor: mike is **sad**

Verb-Noun Dependency

Target: jenny is **sitting**
Distractor: jenny is **standing**

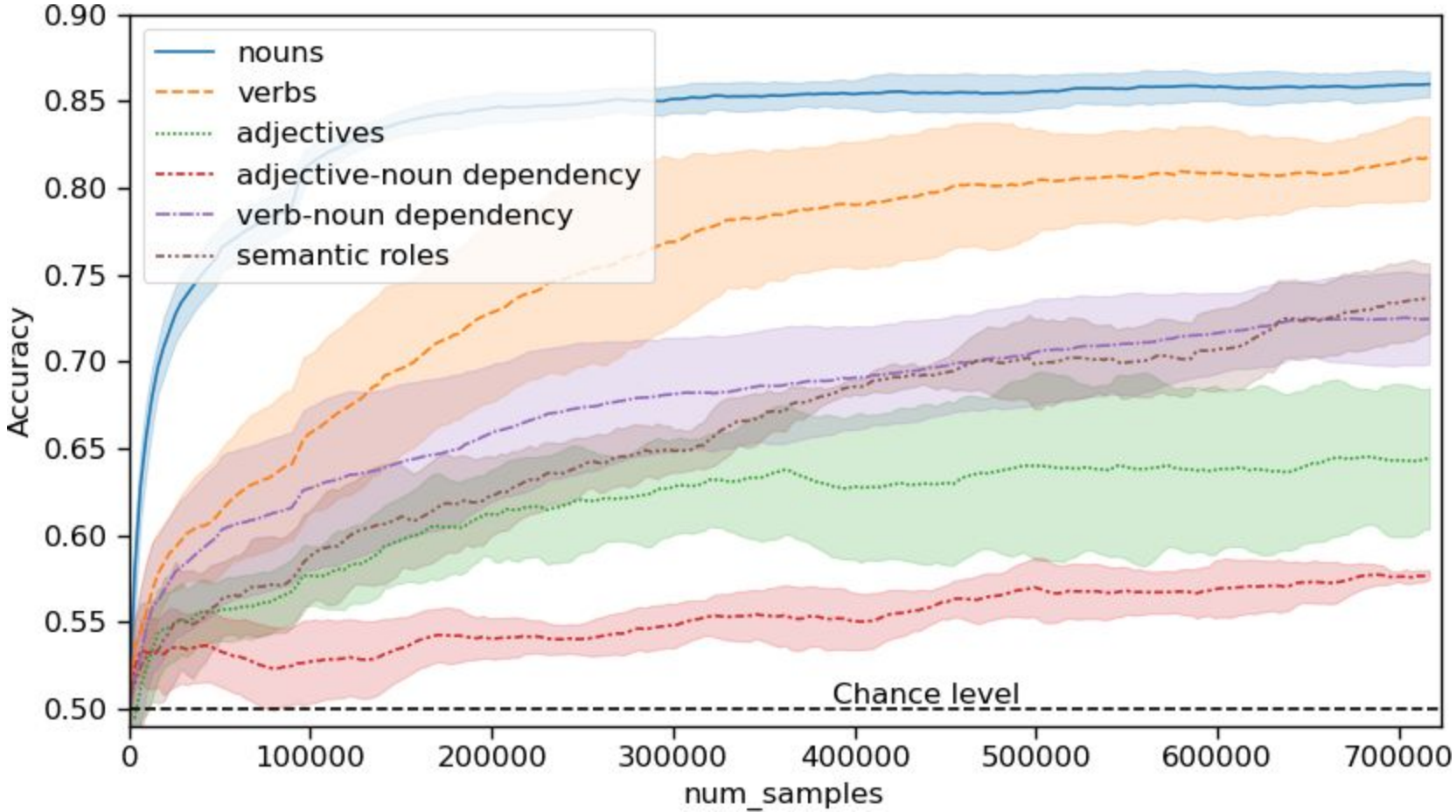
Semantic Roles

Target: **jenny** is waving to **mike**
Distractor: **mike** is waving to **jenny**

Results

Evaluation task	Accuracy	p (best)	p (worst)	Size
Nouns: Persons	0.78 ± 0.05	< 0.001	< 0.01	50
Nouns: Animals	0.93 ± 0.02	< 0.001	< 0.001	360
Nouns: Objects	0.86 ± 0.01	< 0.001	< 0.001	372
Verbs	0.83 ± 0.05	< 0.001	< 0.001	77
Adjectives	0.64 ± 0.06	< 0.01	0.25	56
Adjective-noun dependencies	0.57 ± 0.01	< 0.05	< 0.05	192
Verb-noun dependencies	0.72 ± 0.04	< 0.001	< 0.001	400
Semantic roles	0.75 ± 0.06	< 0.001	< 0.05	50

Results: Learning Trajectory



Discussion

Observation	**Findings in child language acquisition**
Model learns nouns earlier and better than predicates	Noun bias (Gentner, 1982; Bates et al., 1994; Frank et al., 2021)
Model learns semantic roles after nouns	Children become able to assign semantic roles at around 2 years and 3 months (Noble et al., 2011)

- Adjectives are learned poorly due to limited availability
  - “Happy” and “sad” are harder to detect than “sitting” and “standing”

Discussion: What has the model learned?

- Order of nouns as a cue for semantic roles:
  - “Jenny is waving to Mike” vs. “Mike is waving to Jenny”
  - Children use partial representations of sentence structure (i.e., rudimentary syntax) to guide semantic interpretation (e.g. Gertner and Fisher, 2012)

- Important to distinguish genuine learning heuristics from dataset bias!

Semantic Roles

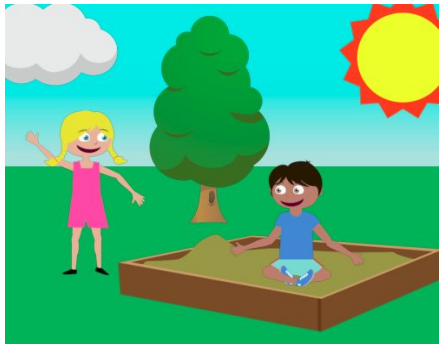
Target: **jenny** is waving to **mike**
Distractor: **mike** is waving to **jenny**

Discussion: Linguistic Bias

Training set

"jenny is waving to mike"
"jenny is kicking the ball to mike"
"jenny is laughing at mike"
"jenny is running to mike"
"mike is mad at jenny"
"jenny is cheering mike"

→ Model could learn that Jenny is usually the agent of an action
→ Model could exploit this dataset bias to achieve high performance (without actually understanding the semantics)
→ We controlled for linguistic bias by **counter-balancing** all test trials:



2AFC for ANNs

Example

Target: **jenny** is waving to **mike**

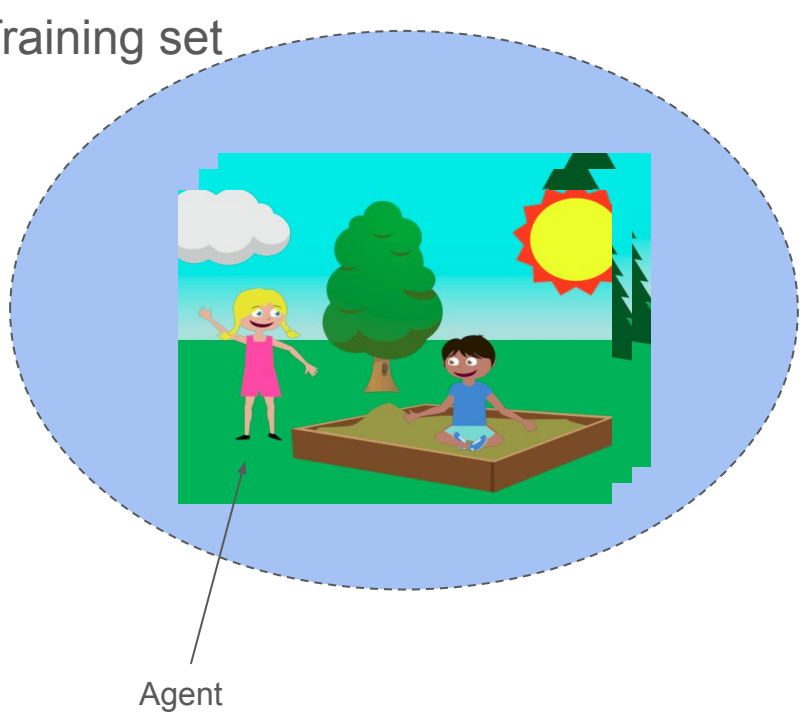
Distractor: **mike** is waving to **jenny**

Counter-example

Target: **mike** is waving to **jenny**

Distractor: **jenny** is waving to **mike**

Discussion: Visual Bias



→ Model could learn that the agent is usually on the left side of the image
→ Model does not learn agency, but position in the image
→ Agent occurs roughly equally on the right and left side (52% / 48%) of the images in the semantic roles test set

→ **Possible other biases?**

Conclusion

- Evaluation for models of cross-situational learning
  - Inspired by 2AFC paradigm in child language acquisition

- Simple cross-situational learner model learns word-level and sentence-level semantics from images and text
  - Learning trajectory mirrors patterns of learning in early childhood

Future work

- Extension to other datasets: MS COCO, Visual Genome, Conceptual Captions
- Learning from speech data (Chrupała et al. 2017, Khorrami and Räsänen 2021)
- Learning using social interaction
    - Joint **functional** and **structural** language learning (Lazaridou et al. 2020)

Nouns: Persons

Target: **jenny** is wearing a crown
Distractor: **mike** is wearing a crown

Nouns: Animals

Target: the **cat** is looking at jenny
Distractor: the **dog** is looking at jenny

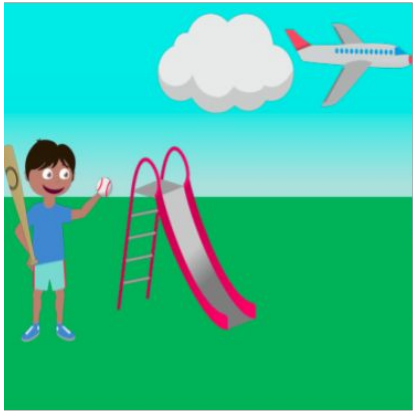
Nouns: Objects

Target: jenny has a **pizza**
Distractor: jenny has a **hat**

Semantic Roles

Target: **jenny** is waving to **mike**
Distractor: **mike** is waving to **jenny**

Adjectives



Target: mike is **happy**
Distractor: mike is **sad**

Adjective-Noun Dependency

Target: mike is **happy**
Distractor: mike is **sad**

Verbs

Target: mike is **sitting**
Distractor: mike is **standing**

Verb-Noun Dependency

Target: jenny is **sitting**
Distractor: jenny is **standing**