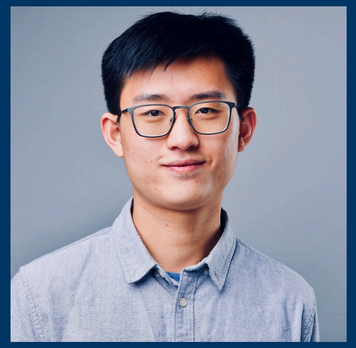
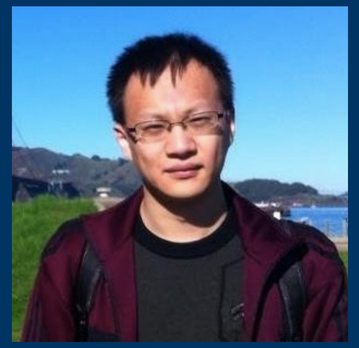
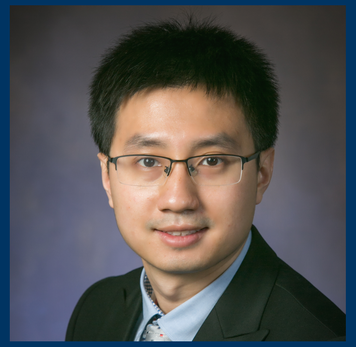
Linking Emergent and Natural Languages via Corpus Transfer

Shunyu Yao, Mo Yu, Yang Zhang,
Karthik Narasimhan, Josh Tenenbaum, Chuang Gan

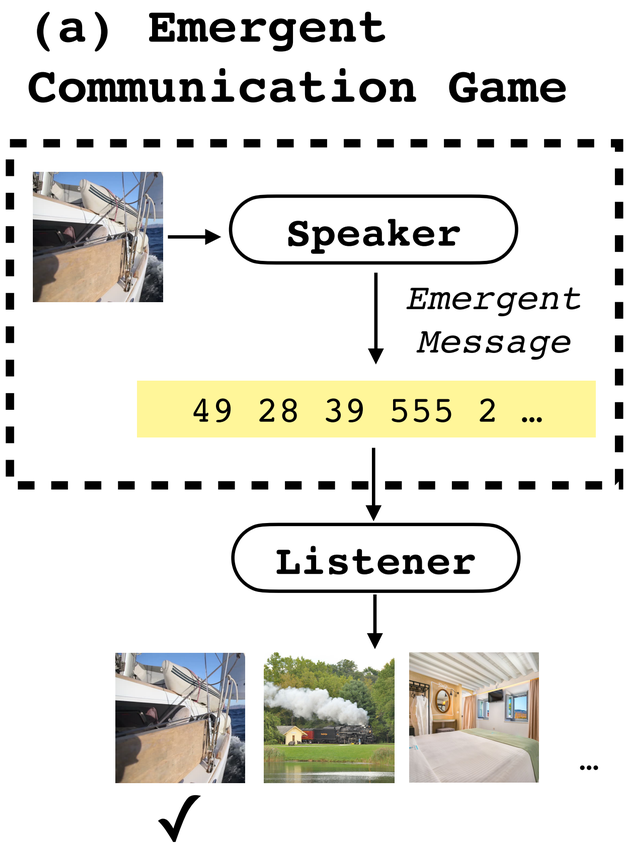

ICLR 2022 (Spotlight)

Emergent Communication (EC)

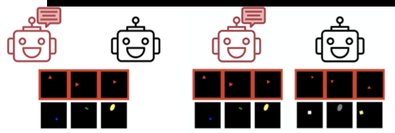
Emergent Language Generalization and Acquisition Speed are not tied to Compositionality

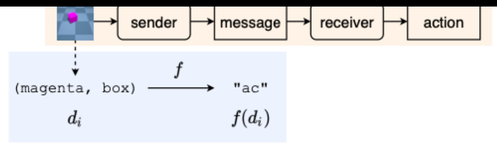
Measuring non-trivial compositionality in emergent communication

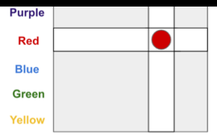
Structural Inductive Biases in Emergent Communication

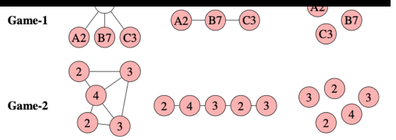


- Many papers with various setups/metrics/claims…
- Scientific: How to establish metrics/claims with respect to natural language, instead of over-approximations?
- Practical: How can emergent language be useful for real-world tasks?



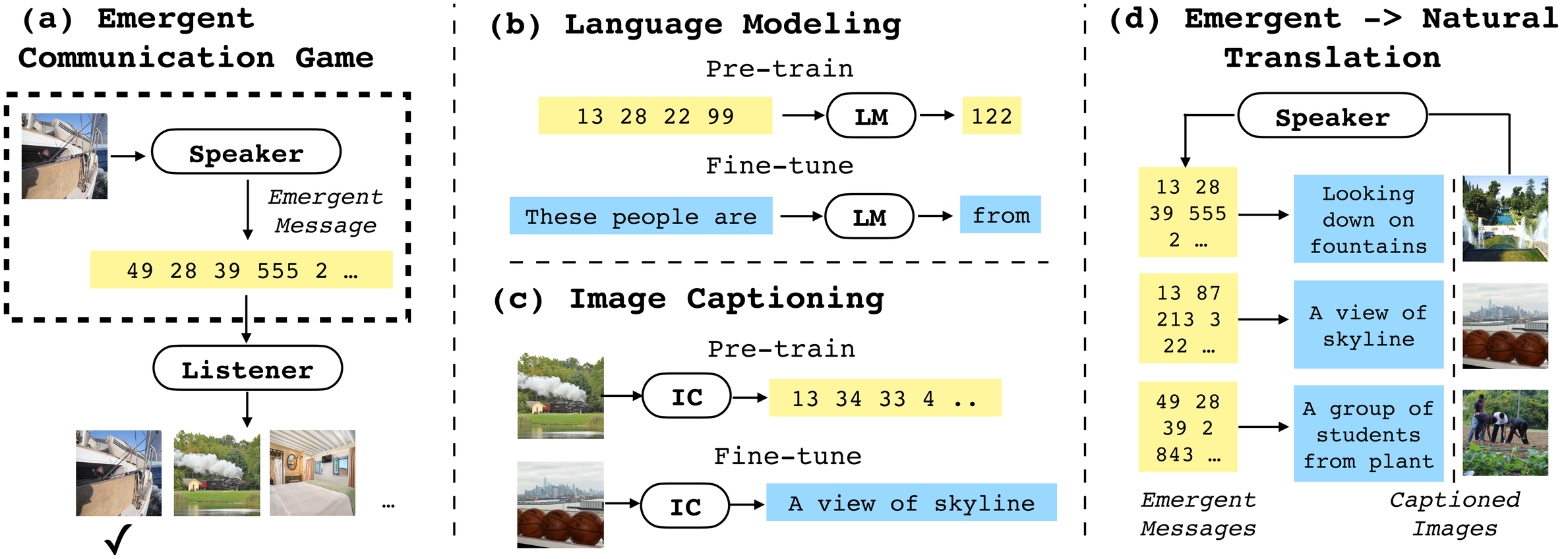






Linking Emergent and Natural Languages

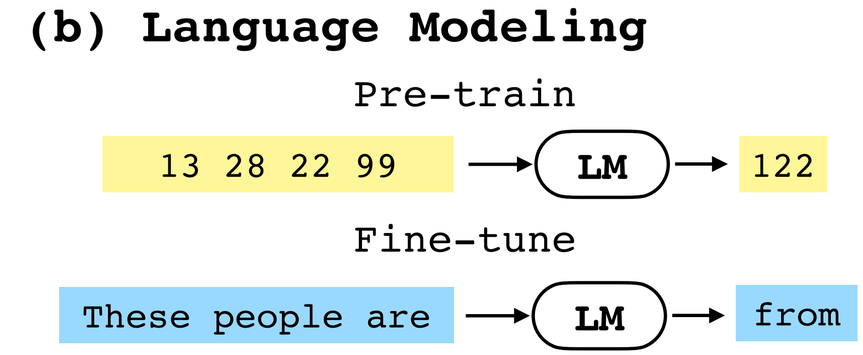
Transfer and Translate

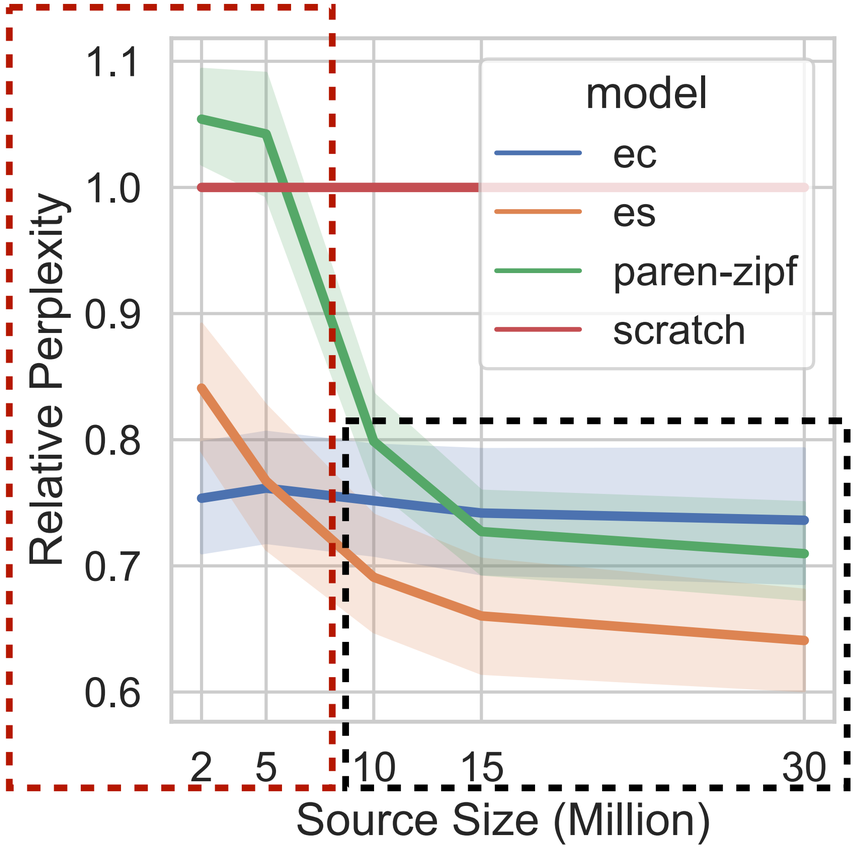


EL -> NL Transfer

Language Modeling

- Fine-tune: Wikipedia corpora of 10 languages (2M tokens, “low-resource”)
- Pre-train: corpora size 2/5/10/15/30M tokens
  - Expect ***scratch*** (no pre-train) < {***ec***, ***paren-zipf*** (nested parentheses)} < ***es*** (Spanish wikipedia)
  - Pre-train size ≤ 5M: ***ec ≥ es!!***
  - Pre-train size >15M: *ec ~ paren-zip < es*
  - More discussion & ablations in the paper



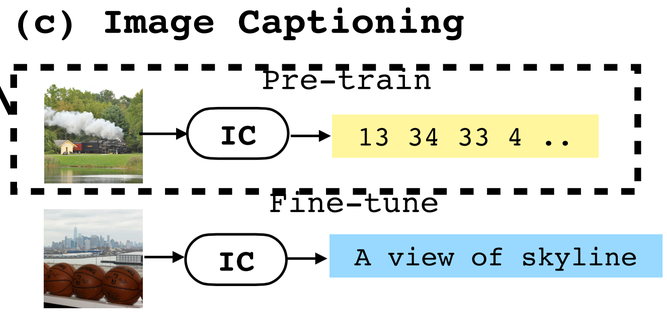


EL -> NL Transfer

Image Captioning

- Can use abundant unlabeled images!
- Cannot use ungrounded languages (dyck/music/code…)

- LM transfers **structural** properties (e.g. *dyck* has no explicit meanings), what about **semantics**?
- **Pre-train:** Conceptual Captions (~3M pairs)
  - Expect ***base*** *(no pre-train)* <***+ec*** < ***+nl***
- **Fine-tune**: MS-COCO (~500k pairs)
  - Fine-tune size 5k/50k: match expectation
  - Fine-tune size 500k: even *+nl* cannot help…
  - Transfer helps if task language data is limited



Metrics	coco (*5k*) base	+ec	+nl
BLEU4	14.1	15.3	18.5

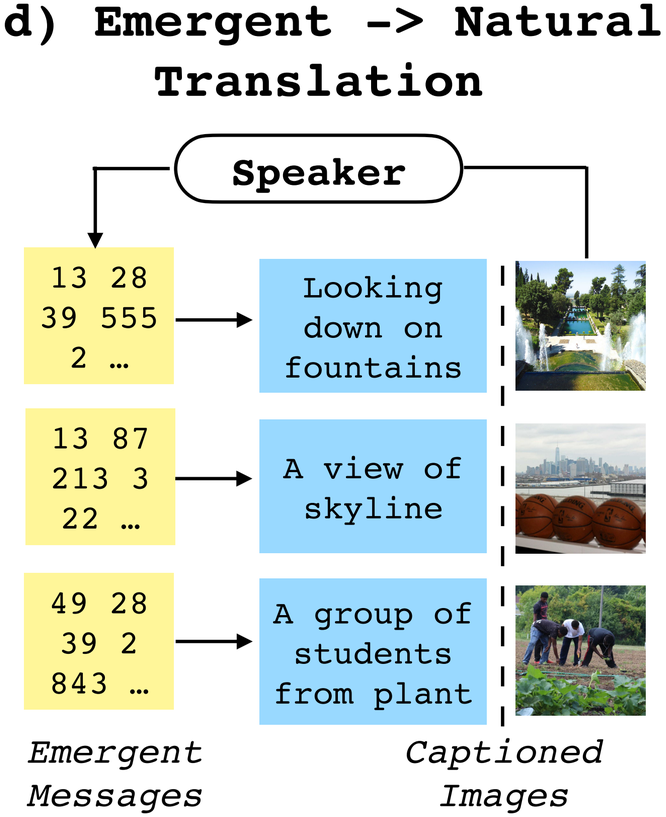
Metrics	coco (*50k*) base	+ec	+nl
BLEU4	26.3	27.2	28.2

Metrics	coco (*full*) base	+ec	+nl
BLEU4	35.8	36.2	36.0

EL -> NL Translation

Correlation across checkpoints

- Transfer tasks are “extrinsic” and varied
- **“Intrinsically”, closer languages translate easier**
- Image-caption pairs -> EL-NL pairs
- 5 setups (vocab, seqlen) x 4 trials/setup x 10 ckpts/trial = 200 ckpts of different EL
- **Translation metric** (ROUGE_L) better correlates with downstream LM performance (Romanian, Hebrew) than **validation accuracy** and **topographic similarity**
- More fine-grained analysis & discussion in paper



Metrics	LM (ro)	LM (he)
Accuracy	0.672	0.737
Topographic	0.030	0.003
Translation	**0.757**	**0.829**

Summary

- **Practical question:** how can emergent language be useful for real-world tasks?
  - Use emergent language corpus to pre-train for natural language tasks
  - Useful when natural language data (for pre-training/fine-tuning) is limited
  - **Future**: find tasks better suited for emergent pre-training
- **Scientific question:** how to measure properties resembling natural language?
  - Transfer to natural language tasks with different properties
  - Translation to natural language predicts transfer better than previous metrics
  - **Future**: disentangle language properties (syntactic/semantic/functional…)