Emergent Communication of Generalizations

Jesse Mu[1] and Noah Goodman[1,2]
Departments of Computer Science[1] and Psychology[2]
Stanford University
{muj,ngoodman}@stanford.edu

Abstract

Artificial agents are often trained to play Lewis-style referential games, but such games often lead to uninterpretable communication and encourage development of only a limited subset of the rich capabilities of human language. Instead, we propose games where agents must learn to communicate about *sets* of objects encoding abstract visual concepts, and show that speaking in generalizations improves systematicity of the learned languages.

1 Introduction

To build agents that can communicate and collaborate effectively with others, recent research has trained agents to communicate for Lewis (1969)-style signaling games (Figure 1a). A general consensus of this work is that without carefully managed environmental pressures (see Lazaridou et al. 2020 for review), agents tend to develop communication protocols distinctly unlike human language: non-compositional (Andreas, 2019; Chaabouni et al., 2020), anti-Zipfian (Chaabouni et al., 2019), and generally uninterpretable (Kottur et al., 2017).

We argue that the reference games typically used in these studies are ill-suited for linguistic systematicity. One reason is perceptual: agents can exploit inscrutable patterns in single inputs, which leads to communication via spurious features (Bouchacourt and Baroni, 2018). Another reason is cognitive: human language can convey abstract ideas, such as kinds and causes, not just referring expressions. Simple reference games are unlikely to drive emergence of such language. In particular, *generalization* about categories is a crucial part of language (Tessler and Goodman, 2019), helping us transfer knowledge that will be useful in unseen environments. Some have even argued that language emerged precisely due to the need to teach others through generalizations (Laland, 2017).

As an alternative to reference games, we propose training agents in extensions of Lewis-style signaling games to *sets*. In the *set reference* (setref) game, a teacher must communicate to a student not just a single object, but rather a group of objects belonging to a concept (Figure 1b). In the *concept* game, each agent sees disjoint images (Figure 1c). Inspired by human teaching (Chopra et al., 2019), our core insight is that requiring generalization to combinatorially large sets of (possibly unseen) objects encourages agents to learn and communicate rich abstractions across inputs (e.g. *seagulls*), instead of low-level features (e.g. *color #FDA448*). These tasks are more difficult than traditional reference games, and we will show that the resulting languages are more systematic and compositional.

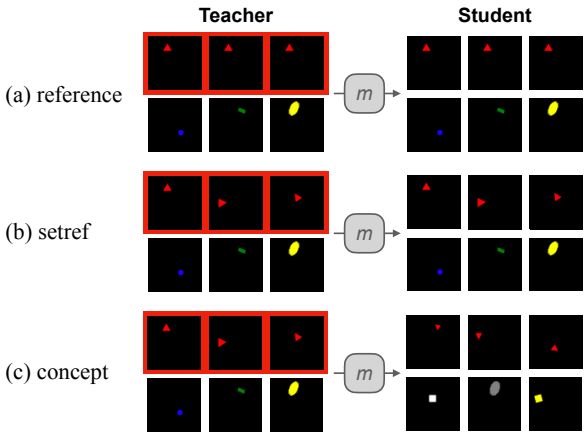

Figure 1: Communication games for the concept *red triangle*. Given a set of targets (red borders) and distractors, a teacher must send a message to help a student identify the targets. In (a) reference games, targets are identical; in (b) setref, there are multiple targets; and (c) concept, the agents see different inputs.

2 Communication Games

We will first describe a generic communication game between a teacher T and student S. Let $G = (c, X^T, Y^T, X^S, Y^S)$ be a communication game, where c is a latent concept to be communicated, $X^T = \{x_1^T, \ldots, x_n^T\}$ is a set of n inputs presented to the teacher, and $Y^T = \{y_1^T, \ldots, y_n^T\}$ is a set of

labels for the teachers' inputs. We call x_i^T a *target* if $y_i^T = 1$, which indicates that x_i^T is a member of the concept c; otherwise x_i^T is a *distractor* and $y_i^T = 0$. X^S and Y^S are defined similarly for the student. Given targets and distractors, the teacher must send a message m to a student that allows them to correctly identify their own targets, where $m = (m_1, \ldots, m_n)$ is a discrete sequence over a fixed alphabet $m_i \in \mathcal{M}$.

Reference game. Here, the teacher and student see the same examples ($X^T = X^S$, $Y^T = Y^S$) and every target input is the same, i.e., $x_i^T = x_j^T$ for all i, j where $y_i^T = y_j^T = 1$.[1]

Set reference (setref) game. The teacher and student see the same examples, but there are multiple target images (e.g. different *red triangles*).

Concept game. The teacher and student see *different examples* ($X^T \neq X^S$, $Y^T \neq Y^S$) of the same concept. When X^T and Y^T contain a single positive and negative example, this is a reference game with separate inputs for each agent, a setup which has previously been shown to encourage linguistic systematicity (Lazaridou et al., 2017) .

3 Models

We will now formalize our models of the teacher and student. In this context, a teacher is a distribution over messages given inputs $p^T(m \mid X^T, Y^T)$, and a student is a distribution over targets given an message $p^S(Y^S \mid X^S, m) = \prod_i p^S(y_i^S \mid x_i^S, m)$.

Teacher. The teacher encodes all inputs with a convolutional neural network (CNN) f_θ^T; embeddings for targets and distractors are averaged to form positive and negative class *prototypes* (Snell et al., 2017), which then conditions a recurrent neural network (RNN). Let X_+^T and X_-^T denote the sets of targets and distractors in X^T; then define a prototype embedding $c_+^T = \frac{1}{|X_+^T|} \sum_{x_i \in X_+^T} f_\theta^T(x_i)$ (analogously for c_-^T). Then $p_T(m \mid X_T, Y_T) = \text{RNN-DECODE}(m \mid \text{proj}([c_+^T; c_-^T]))$ where proj is a linear projection to the RNN hidden state.

Student. The student takes a message and makes predictions about the label $\hat{y}_i^S$ for each input x_i^S. Given the teacher message and an input image, define $p^S(y_i^S \mid x_i^S, m) = \sigma(\text{RNN-ENCODE}(m) \cdot f_\phi^S(x_i^S))$, where f_ϕ^S is a separate CNN.

We train a teacher-student pair end-to-end with the Gumbel-Softmax trick (Jang et al., 2017), to maximize the likelihood of the student picking the correct targets over a set of games in the domains described below. For full details, see Appendix A.

[1] For ease of comparison, we present an atypical formulation of a reference game with multiple identical targets and student target decisions made independently for each input, instead of the single-target forced-choice setting. Appendix C shows that results do not change with traditional games.

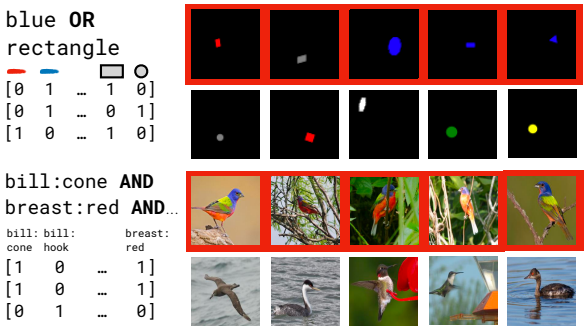

Figure 2: Example sets of targets (red border) and distractors for ShapeWorld (top) and Birds (bottom), with corresponding concepts represented as intensional logical formulas and extensional sets of input features.

4 Tasks

We explore two datasets: first, an artificial shape dataset which allows us to evaluate communication over cleanly defined logical concepts; second, a dataset of birds to test agents' ability to extract concepts from realistic visual input.

ShapeWorld. We use the ShapeWorld visual reasoning dataset (Kuhnle and Copestake, 2017, Figure 2, top). For reference games, target images are a single object, defined by a conjunction of a shape and a color; of the 30 possible shapes, we hold out 6 (20%) for testing. For setref and concept games, concepts are defined as disjunctions or conjunctions of (possibly negated) shapes and/or colors. This produces 312 concepts, 20% of which are held out. For each game, sets consist of 10 randomly sampled targets depicting shapes that satisfy the concept, and 10 distractors. We specifically sample "hard" targets and distractors to test understanding of conjunctions or disjunctions (see Appendix B for details). For this dataset, we use an agent vocabulary of 14 tokens and maximum length 5 (mimicking the true concept formulas).

Birds. We next use the Caltech-UCSD Birds dataset (Wah et al., 2011) which contains 200 classes of birds with 40–60 images (Figure 2, bottom). As before, reference games involve a single target; setref and concept game targets are mem-

bers of a specific bird class. We use 100 classes at train and 50 at test, sampling 5 targets and 5 distractors per game. The dataset contains boolean attributes (e.g. *beak*, *size*) for individual birds and classes.[2] Thus, we represent reference game concepts as the feature vector of the target bird, and setref/concept game concepts as the feature vector of the class. In our evaluation, we will measure how well agent languages capture these features.

For these experiments, we set the vocabulary size to 20 and the maximum message length to 7.

5 Evaluation

Besides measuring communication success, defined by student accuracy on held-out games from seen and unseen concepts, we evaluate the systematicity of the agents' languages with the following:

H and AMI. We compute simple information theoretic quantities between the agent messages and concepts: the conditional entropy of messages given concepts, $H(M|C)$, and the adjusted mutual information $\text{AMI}(M,C) = (I(M,C) - \mathbb{E}(I(M,C)))/(\max(H(M),H(C)) - \mathbb{E}(I(M,C)))$ (Vinh et al., 2010). A lower $H(M|C)$ indicates more consistent language for each concept; higher AMI indicates better alignment between messages and concepts.

Topographic ρ. A finer measure of compositionality often used in the literature (Lazaridou et al., 2018; Li and Bowling, 2019; Lazaridou et al., 2020) is *topographic* ρ (Brighton and Kirby, 2006). We define a distance metric between game concepts $d_C(c_1, c_2)$ and another between agent messages $d_M(m_1, m_2)$, compute pairwise distances between concepts and messages, and measure their alignment with Spearman's ρ. A high ρ indicates that similar messages are produced for similar concepts.

For messages, we use the **Edit** (Levenshtein) distance with equal insert/delete/replace costs. For game concepts, we define two distances based on intensional and extensional concept representations (Figure 2). First, **Edit** distances between string representations of logical formulas. Second, **Hausdorff** distance d_H between the sets of inputs defined by concepts. Let $Z^a = \{z_i^a\}$ be the set of feature-based inputs belonging to concept a. Then d_H is the maximum distance from a point in one set to the closest point in the other: $d_H(Z^a, Z^b) = \max(\sup_i d(z_i^a, Z^b), \sup_j d(z_j^b, Z^a))$, where $d(a, B) = \inf_{b \in B} \text{EditDistance}(a, b)$.

[2] Feature vectors for individual birds in a class vary due to the visibility of features in the image; class vectors are averaged across all individual birds, then rounded to 1 or 0.

	Acc Seen	Unseen	$H(M\|C)$	AMI
ShapeWorld				
Ref	**97** (0.5)	**96** (0.8)	6.2 (0.5)	0.07 (0.0)
Setref	84 (2.0)	83 (1.8)	4.0 (0.7)	0.30 (0.1)
Concept	80 (1.9)	78 (1.3)	**1.6** (0.2)	**0.53** (0.0)
Birds				
Ref	**93** (0.3)	**89** (0.1)	5.9 (0.2)	0.05 (0.0)
Setref	89 (0.2)	78 (0.2)	5.2 (0.1)	0.17 (0.0)
Concept	88 (0.1)	73 (0.3)	**4.1** (0.2)	**0.26** (0.0)

Table 1: Student accuracy (seen and unseen concepts), conditional entropy of messages given concepts (lower is better), and adjusted mutual information score ($\in [0,1]$; higher is better), with (SD) across 5 runs.

6 Results

Table 1 shows test accuracy on communication games over seen and unseen concepts for the best validation models. Reference game performance is high across both datasets, and agents are able to generalize well to unseen concepts at test time. Accuracy on setref and concept games are lower for both datasets, with a considerable drop of 10–15 points for Birds when generalizing to novel classes. Overall, communicating sets and concepts seems to be a much *harder* task than concrete reference, and thus an interesting avenue for further work.

The ability to communicate accurately, even on held-out games, is not necessarily indicative of more *systematic* communication; generalization without compositional language is a common finding (Kottur et al., 2017; Andreas, 2019; Chaabouni et al., 2020). Instead, the more difficult games produce more systematic language. Compared to ShapeWorld reference games, concept game entropy over messages is much lower (1.6 vs. 6.2), and AMI much higher (0.53 vs. 0.07); setref is somewhere in the middle; this difference also occurs in Birds. Additionally, Figure 3 shows topographic ρ between the languages and the edit and Hausdorff concept distances, which is consistently higher for concept and setref throughout training.

Figure 4 (more examples in Appendix E) shows messages generated by agents for the concept *red triangle*. At one extreme, reference agents use a huge variety of messages to refer a red triangle, while concept agents consistently use *deeee/edeee*, and setref is somewhere in the middle.[3]

[3] We also show that concept and setref agents are more sys-

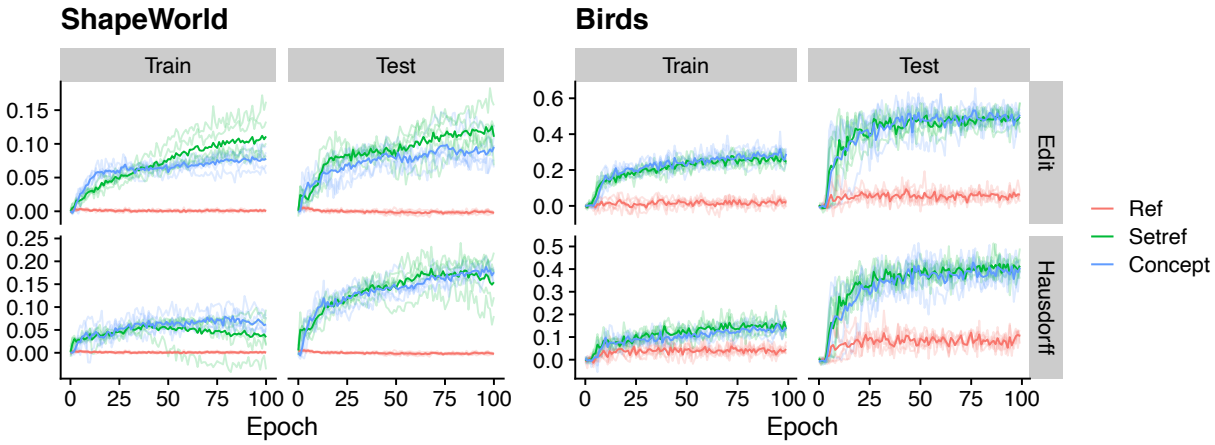

Figure 3: Topographic ρ between language and edit (top) or Hausdorff (bottom) concept distances for training and test splits across both datasets. Results from 5 runs plotted, with averages in bold.

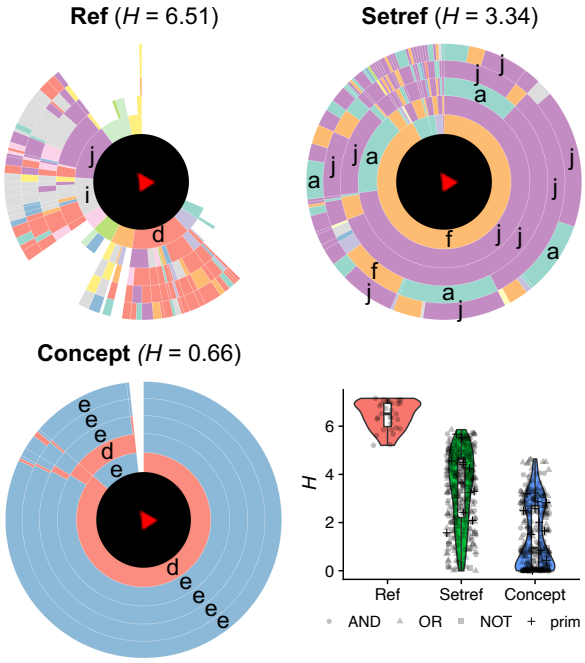

Figure 4: Distribution of 300 messages (starting from center and proceeding outwards) and entropies for the concept *red triangle* in ref, setref, and concept settnigs. Each color corresponds to a unique token. Bottom right: distribution of entropies over concepts for a single model (a more detailed view of the data in Table 1.)

Set Size. Figure 5 shows how the number of positive targets n (with equal numbers of distractors) affects language systematicity. n has a statistically significant effect on topographic ρ for ShapeWorld setref (Spearman $\rho = 0.82, p < 10^{-7}$) and concept ($\rho = 0.71, p < 10^{-4}$) and Birds setref ($\rho = 0.89, p < 10^{-5}$) and concept ($\rho = 0.54, p = 0.017$). When $n = 1$, the setref game is equivalent to a reference game with 1 target and 1 distractor, and the concept game is similar, but with agents

tematic when evaluated on reference games; see Appendix D.

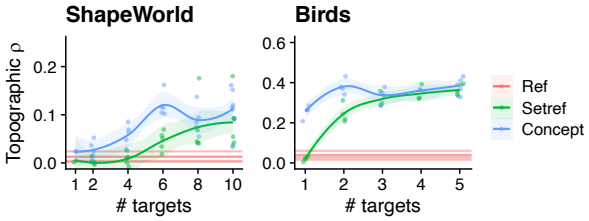

Figure 5: Topographic ρ (concept edit distance) with varying number of targets (average over seen/unseen splits). Each point and Ref line is an independent run.

given separate inputs. Our results suggest that this decoupling, as done in Lazaridou et al. (2017), promotes systematicity in some cases (Birds) but not others (ShapeWorld). We additionally show that (1) sets are an *alternative* way of encouraging systematicity without needing this separation, and (2) even with this separation, larger set sizes improve systematicity across both datasets.

7 Conclusion

We have proposed extensions of referential games to sets of objects, and found that the need to convey generalizable categories leads to the development of more systematic languages, whether targets are shared (setref) or unshared (concept) across agents. One interesting avenue for follow up work is identifying whether the structure of the more sophisticated logical concepts are reflected in the agent languages, perhaps using recently proposed tools for measuring compositionality (Andreas, 2019).

8 Reproducibility

Code and data are available at `github.com/jayelm/emergent-generalization`.

Acknowledgments

We thank Alex Tamkin and anonymous reviewers for helpful comments and discussions. This research was supported by an NSF Graduate Research Fellowship for JM, the Stanford HAI-AWS Cloud Credits for Research program, and the Office of Naval Research grant ONR MURI N00014-16-1-2007.

References

Jacob Andreas. 2019. Measuring compositionality in representation learning. In *International Conference on Learning Representations (ICLR)*.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985.

Henry Brighton and Simon Kirby. 2006. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial Life*, 12(2):229–242.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4427–4442.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

Sahil Chopra, Michael Henry Tessler, and Noah D Goodman. 2019. The first crank of the cultural ratchet: Learning and transmitting concepts through language. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci)*, pages 226–232.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

E. Jang, S. Gu, and B. Poole. 2017. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2962–2967.

A. Kuhnle and A. Copestake. 2017. ShapeWorld - a new test methodology for multimodal language understanding. *arXiv preprint*.

Kevin N Laland. 2017. The origins of language in teaching. *Psychonomic Bulletin & Review*, 24(1):225–231.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations (ICLR)*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations (ICLR)*.

Angeliki Lazaridou, Anna Potapenko, and Oliver Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

D. Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4077–4087.

Michael Henry Tessler and Noah D Goodman. 2019. The language of generalization. *Psychological Review*, 126(3):395.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research (JMLR)*, 11:2837–2854.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The Caltech-UCSD Birds-200-2011 dataset.