Generalization and Translatability in Emergent Communication via Informational Constraints

Mycal Tucker **Julie Shah** **Roger Levy** **Noga Zaslavsky**
Massachusetts Institute of Technology

Abstract

Traditional emergent communication (EC) methods often fail to generalize to novel settings or align with representations of natural language. Here, we show how controlling the Information Bottleneck (IB) tradeoff between complexity and informativeness (a principle thought to guide human languages) helps to address both of these problems in EC. Using VQ-VIB, a recent method for training EC agents while controlling the IB tradeoff, we find that: (1) increasing pressure for informativeness, which encourages agents to develop a shared understanding beyond task-specific needs, leads to better generalization to more challenging tasks and novel inputs; (2) VQ-VIB agents develop an EC space that encodes some semantic similarities and facilitates open-domain communication, similar to word embeddings in natural language; and (3) when translating between English and EC, greater complexity leads to improved performance of teams of simulated English speakers and trained VQ-VIB listeners, but only up to a threshold corresponding to the English complexity. These results indicate the importance of informational constraints for improving self-play performance and human-agent interaction.

1 Introduction

We wish to develop artificial agents that communicate in grounded settings, via communication that enables high task utility, generalizability to novel settings, and good human-agent cooperation. Emergent communication (EC) methods, wherein agents learn to communicate with each other in an unsupervised manner by maximizing a reward function, take a step towards this vision by producing agents that use grounded communication to solve a particular task [1–3], but they still fall short of the vision of generalizable and human-interpretable communication [4, 5]. In this work, we take steps towards addressing these limitations by building on the information-theoretic EC approach of Tucker et al. [6]. This approach connects EC with the Information-Bottleneck [IB, 7] framework for semantic systems [8, 9], via the vector-quantized variational Information Bottleneck (VQ-VIB) neural architecture [6]. VQ-VIB agents are trained to optimize a tradeoff between maximizing utility (how well they perform a task), maximizing informativeness (how well a listener can infer a speaker's meaning, independently of any downstream task), and minimizing communicative complexity (roughly the number of bits allocated for communication). Given broad evidence suggesting that human languages are guided by the IB informativeness-complexity tradeoff [8, 10–13], we hypothesize that taking into account informativeness could improve EC generalizability to novel settings while adjusting complexity could improve the translatability between EC and human languages. Results from our experiments support this hypothesis. First, we show that encouraging informativeness allows EC agents to generalize beyond their training distribution to handle more challenging tasks and out-of-distribution objects, with VQ-VIB achieving the best performance compared to alternative EC methods. Second, we propose a simple method for translating natural language word embeddings [e.g., GloVe, 14] into EC signals and use that to simulate human-agent communication in a cooperative object-discrimination task. We find that team performance for

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

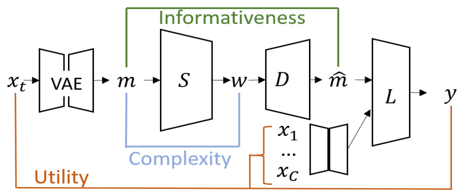


Figure 1: Agents communicate to identify a target input (e.g., an image), x_t, among a set of candidates $\mathcal{C} = \{x_1, \ldots, x_C\}$ (see main text).

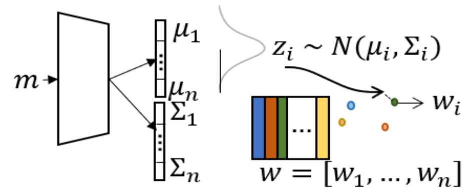


Figure 2: Our modified VQ-VIB architecture combines n quantized vectors into a single message, enabling k^n unique messages.

English speakers and trained VQ-VIB agents improves with the communicative complexity of the EC system, but only up to a certain threshold, which corresponds to the complexity of the English object naming system. Together, our findings suggest that training EC agents while controlling the informativeness-complexity tradeoff, in addition to maximizing utility, may simultaneously support improved self-play performance as well as human-agent interaction.

2 Related Work

Several researchers test EC generalization by evaluating agents on novel combinations of symbolic inputs held out during training [15–18]. Chaabouni et al. [4] find that using harder tasks at training time is important for improving test-time performance and that using population-based voting improves cross-domain transfer. Complementing research of EC generalization in self-play, many researchers explore connections between EC and natural language. Some top-down methods combine pretrained language models with finetuning in grounded environments, but such methods can suffer from "drift" wherein agents learn to ascribe new meanings to words [19–21]. Other researchers train agents in self-play and then seek to connect learned EC to human-interpretable concepts or natural language [5, 22, 23]. Based on information-theoretic analysis of human naming systems, we investigate whether producing EC that matches humans' complexity and informativeness enables better generalization and translation.

3 Background: Information-theoretic emergent communication

Our work builds on the information-theoretic framework of [8] for semantic systems, and especially on its extension to scalable emergent communication in artificial neural agents, proposed in [6].

Efficient compression and semantic systems Zaslavsky et al. [8] argued that languages are shaped by the need of speakers and listeners to efficiently compress meanings into words. They formulated this objective using the Information Bottleneck [IB, 7] principle, which can be interpreted as a tradeoff between the informativeness and complexity of the lexicon, and follows from rate-distortion theory [24, 25]. In this framework, a speaker is characterized as a probabilistic encoder $S(w|m)$ that, given a meaning $m \sim p(m)$, generates a communication signal, w. A listener is characterized as a probabilistic decoder $D(\hat{m}|w)$, that reconstructs the speaker's meaning from w. Complexity is computed as the mutual information between the speaker's meanings and signals, $I(m;w)$. Informativeness measures how well the listener's interpretation, $\hat{m}$, matches the speaker's intended meaning, such that maximizing informativeness amounts to minimizing the expected Kullback-Leibler (KL) divergence between the agents' belief states over meanings, $\mathbb{E}[D[m\|\hat{m}]]$. Optimal speakers and listeners balance the tradeoff between complexity and informativeness by minimizing $I(m;w) - \beta\mathbb{E}[D[m\|\hat{m}]]$, where $\beta \geq 0$ controls the tradeoff. This optimization problem is equivalent to the IB principle. This theoretical framework has gained broad empirical support across human languages in multiple domains [8, 10–12], as well as artificial agent communication [26, 27].

Vector-quantized variational Information Bottleneck (VQ-VIB) While solving the IB optimization problem in high-dimensional settings is challenging, Tucker et al. [6] proposed vector-quantized variational Information Bottleneck (VQ-VIB), a deep learning method that combines notions from vector quantization variational autoencoders [VQ-VAE, 28] and variational Information Bottleneck (VIB), and applied it to EC. This framework can be applied to Lewis reference games [29], a standard

EC setting, depicted in Figure 1. Here, m (the speaker's meaning representation) is produced by passing a target input, x_t, through a pretrained VAE to model perceptual noise. A decoder, D, seeks to reconstruct m given the speaker's communication, w. Using this reconstruction, a listener agent, L, attempts to identify x_t from a set $\mathcal{C} = \{x_1, ..., x_C\}$ of (noisily observed) candidate inputs by selecting $y \in \mathcal{C}$. In VQ-VIB, a probabilistic speaker, S, maps from m to parameters of a Gaussian distribution, μ and σ, in a continuous latent space, $\mathbb{R}^Z$. A latent vector, z, is sampled from $\mathcal{N}(\mu(m), \sigma(m))$ and discretized by looking up the nearest element of a (trainable) codebook of k quantized vectors $\zeta \in \mathbb{R}^Z$. The final communication vector output by the speaker, w, is this nearest quantized vector. That is, the VQ-VIB encoder is defined by $S(w|m) = \mathbb{P}(w = \text{argmin}_\zeta[||z - \zeta||^2]|m)$.

In the information-theoretic emergent communication (ITEC) framework, agents maximize a combination of utility and informativeness while minimizing complexity (although variational bounds are often used in practice) as shown in Equation 1, with λ_U, λ_I, and λ_C controlling the relative weight of each term. Utility, $U(x_t, y)$, is a task-specific performance measure, e.g., the listener's accuracy in identifying the target input. Informativeness, approximated as $-(m - \hat{m})^2$, is a task-agnostic measure of the listener's ability to reconstruct the speaker's intention, m. Lastly, for VQ-VIB agents, we upper bound complexity as the KL divergence between $\mathcal{N}(\mu(m), \sigma(m))$ and a unit Gaussian [30].

$$\text{maximize} \quad \lambda_U \mathbb{E}[U(x_t, y)] - \lambda_I \mathbb{E}[(m - \hat{m})^2] - \lambda_C \mathbb{E}[D_{\text{KL}}[\mathcal{N}(\mu(m), \sigma(m))|\mathcal{N}(0, 1)]] \qquad (1)$$

4 Technical Approach

We modify the VQ-VIB architecture to enable larger codebooks via combinatoriality, and introduce a translation framework for analyzing alignment between natural language and EC.

Combinatorial Codebook for Larger Vocabulary While VQ-VIB agents represent an important step towards information-bounded EC, the architecture in [6] limits agents to only communication via one of the quantized vectors in their codebook. Here, we modify the VQ-VIB architecture to allow agents to select $n \geq 1$ vectors from the codebook before concatenating them into a single message (Figure 2). Given a meaning, m, a speaker computes n means and variances (μ_i, σ_i) from which n latent, continuous vectors z_i are sampled, each in $\mathbb{R}^{Z/n}$. These continuous vectors are discretized via standard vector-quantization by returning the nearest entry and then concatenated into a single message $w = [w_1, ..., w_n] \in \mathbb{R}^Z$. For a codebook of size k, this change increases the number of messages the speaker may emit from k to k^n, without increasing the network size.

Human-agent translation Because VQ-VIB agents communicate via discrete signals that are embedded in a continuous space [see also 23], much like grounded word embeddings, we propose a simple translation mechanism from natural language to EC. Using a dataset of images with natural-language labels, one can construct a *translation dataset*, associating EC vectors and natural language embeddings, by (i) mapping each image's label to its word embedding, and (ii) passing the input image through an EC speaker to map it to an EC vector. Using this dataset, we train a least-squares linear "translator" model to map from natural language to EC. In evaluation, we test the translatability of an EC system by combining a human speaker with a pre-trained EC listener in a Lewis reference game, mediated by a trained translator. Evaluating the performance of this hybrid team would reflect the degree to which our EC agents are suited for communication and cooperation with humans.

5 Experiments

Generalizing to out-of-distribution inputs We tested agents' ability to generalize to out-of-distribution (OOD) using the ManyNames dataset [31], consisting of 25,000 images of natural objects with approximately 36 human-generated labels per image (see Figure 3). Using the most common response (dubbed the `topname`), we split the dataset: 20% of the `topnames` were selected for training data, while the test set was constructed from all images for which no label matched a training `topname`. Agents were trained with two candidates ($C = 2$) in the Lewis reference game, with candidate images drawn from disjoint `topname` sets. In evaluation, we measured team performance using the test set, for various C, and without restrictions on the candidates' `topnames`. Thus, in our hardest evaluation settings, agents needed to simultaneously generalize to a larger number of candidates, drawn from non-distinct categories, which were not seen during training.

(a) **bird** (20), duck (14) (b) **couch** (21), sofa (9)

Figure 3: Examples from the ManyNames dataset, with naming responses (and counts).

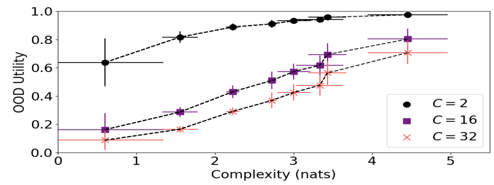


Figure 4: OOD accuracy improved with complexity. Results generated by varying λ_I.

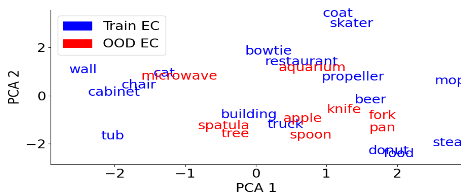


Figure 5: 2D PCA of EC vectors for training (blue) and OOD (red) images at complexity that roughly matches English (2.1 nats, $\lambda_I = 0.5$).

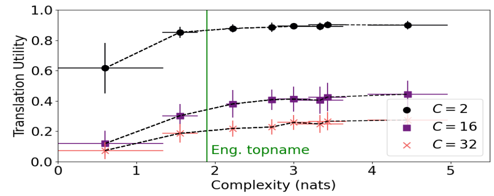


Figure 6: Performance of human-agent teams using GloVe to EC translation. Performance flattened after passing English complexity.

Figure 4 shows the utility of VQ-VIB agents ($n = 4$) on OOD images, evaluated with varying numbers of candidates. By increasing $\lambda_I \in \{0, 0.1, 0.5, 1, 1.5, 2, 3, 10\}$, we induced more informative and complex communication, which in turn supported greater utility. Further results with different n or other architectures (Appendices C, D) show similar trends: greater complexity supported greater OOD utility, with VQ-VIB outperforming alternative EC methods. Lastly, we recorded the mean EC vector associated with images for each `topname`, as shown in Figure 5 (using 2D PCA [32]), for training (blue) or OOD (red) images. The agents learned a semantically structured space, much like word embeddings, by clustering words with similar meanings. Remarkably, this holds for OOD images, revealing how VQ-VIB agents may generalize to novel inputs.

English-EC Translation We measured team performance for a simulated English speaker and our trained VQ-VIB listeners by sampling in-distribution images, looking up the GloVe embedding for each image's `topname`, translating the embedding to EC via the linear translator (see Section 4), and then passing on the translated communication to the listener. The teams's utility (listener's accuracy) is plotted in Figure 6. Increasing complexity supported greater utility, but only up to roughly 2 nats, after which the team's performance stopped improving. This cutoff is noteworthy because it nearly equals the complexity of the English naming system for the ManyNames domain, calculated to be 1.9 nats using the MINE estimator [33]. Thus, the team's utility appears to be bottlenecked by the complexity of the English speaker, and further increasing complexity in self-play affords no benefits.

6 Contributions

In this work, we explored the influence of informational constraints on two desired properties of emergent communication (EC): generalization to novel inputs and translation between EC and natural language. We found that (1) encouraging informativeness allows agents to better generalize to out-of-distribution inputs; (2) the structure of the emergent VQ-VIB communication vectors encodes some semantic similarities and facilitates open-domain communication, similar to word embeddings in natural language; and (3) performance for teams of simulated English speakers and trained VQ-VIB listeners improves with the complexity of the EC system, but only up to the complexity level of the English speaker. These results suggest that taking into account the IB informativeness-complexity tradeoff in EC, in addition to maximizing utility, may support both improved self-play performance and human-agent interaction.

Acknowledgements M.T. is supported by an Amazon Science Hub Fellowship. N.Z. is supported by a K. Lisa Yang Integrative Computational Neuroscience (ICoN) Postdoctoral Fellowship.

References

[1] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Neural Information Processing Systems (NIPS)*, 2017.

[2] Ryan Lowe, Abhinav Gupta, Jakob N. Foerster, Douwe Kiela, and Joelle Pineau. On the interaction between supervision and self-play in emergent communication. In *ICLR*. OpenReview.net, 2020. URL `http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#LoweOFKP20`.

[3] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*, 2020.

[4] Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *International Conference on Learning Representations*, 2021.

[5] Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1321. URL `https://aclanthology.org/D17-1321`.

[6] Mycal Tucker, Roger Levy, Julie Shah, and Noga Zaslavsky. Trading off utility, informativeness, and complexity in emergent communication. *To Appear in: Neural Information Processing Systems (NeurIPS)*, 2022.

[7] Naftali Tishby, Fernando C. Pereira, and William Bialek. The Information Bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.

[8] Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, 115(31): 7937–7942, 2018. doi: 10.1073/pnas.1800521115. URL `https://www.pnas.org/doi/abs/10.1073/pnas.1800521115`.

[9] Noga Zaslavsky. *Information-Theoretic Principles in the Evolution of Semantic Systems*. Ph.D. Thesis, The Hebrew University of Jerusalem, 2020.

[10] Noga Zaslavsky, Terry Regier, Naftali Tishby, and Charles Kemp. Semantic categories of artifacts and animals reflect efficient coding. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 2019.

[11] Noga Zaslavsky, Mora Maldonado, and Jennifer Culbertson. Let's talk (efficiently) about us: Person systems achieve near-optimal compression. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, 2021.

[12] Noga Zaslavsky, Karee Garvin, Charles Kemp, Naftali Tishby, and Terry Regier. The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, 2022.

[13] Francis Mollica, Geoff Bacon, Noga Zaslavsky, Yang Xu, Terry Regier, and Charles Kemp. The forms and meanings of grammatical markers support efficient communication. *Proceedings of the National Academy of Sciences*, 118(49), 2021.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

[15] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=HJGv1Z-AW`.

[16] Jacob Andreas. Measuring compositionality in representation learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz05o0qK7.

[17] Łukasz Kuciński, Tomasz Korbak, Paweł Kołodziej, and Piotr Miłoś. Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. *Advances in Neural Information Processing Systems*, 34:23075–23088, 2021.

[18] Sam Spilsbury and Alexander Ilin. Compositional generalization in grounded language learning via induced model sparsity. *arXiv preprint arXiv:2207.02518*, 2022.

[19] Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *ACL*, 2020.

[20] Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1259. URL https://aclanthology.org/D17-1259.

[21] Athul Paul Jacob, Mike Lewis, and Jacob Andreas. Multitasking inhibits semantic drift. *arXiv preprint arXiv:2104.07219*, 2021.

[22] Jacob Andreas, Anca Dragan, and Dan Klein. Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 232–242, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1022. URL https://aclanthology.org/P17-1022.

[23] Mycal Tucker, Huao Li, Siddharth Agrawal, Dana Hughes, Katia P. Sycara, Michael Lewis, and Julie Shah. Emergent discrete communication in semantic spaces. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=EuvW3-Lmoxm.

[24] Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

[25] Peter Harremoës and Naftali Tishby. The Information Bottleneck revisited or how to choose a good distortion measure. In *IEEE International Symposium on Information Theory*, pages 566–571, June 2007.

[26] Tiwalayo Eisape, Roger Levy, Joshua B. Tenenbaum, and Noga Zaslavsky. Toward human-like object naming in artificial neural systems. In *Bridging AI and Cognitive Science workshop at the International Conference on Representation Learning (ICLR)*, 2020.

[27] Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, 118(12):e2016569118, 2021.

[28] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[29] David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 2008.

[30] Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. URL https://arxiv.org/abs/1612.00410.

[31] Carina Silberer, Sina Zarieß, and Gemma Boleda. Object Naming in Language and Vision: A Survey and a New Dataset. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, 2020. URL https://aclanthology.org/2020.lrec-1.710/.

[32] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[33] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/belghazi18a.html`.

[34] Tom Eccles, Yoram Bachrach, Guy Lever, Angeliki Lazaridou, and Thore Graepel. Biases for emergent communication in multi-agent reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

A Training Hyperparameters

Here, we report the hyperparameters used in our experiments. In general, hyperparameters were chosen to maximize self-play performance on the training task, without considering performance on more challenging validation setups (e.g., with more distractors or out-of-distribution images). Thus, we were able to fairly test how successfully learning one task enabled agents to generalize to other settings.

We used a Resnet18, available through TorchVision that had been pretrained on ImageNet, for extracting 512-dimensional features from images. We trained a VAE with a symmetric encoder and decoder with fully connected ReLU layers of dimensions 128, 64, and 32 to reconstruct such features. Thus, in reference games, the speaker and listener actually observed the result of taking an image, passing it through the Resnet18 feature extractor, and then passing it through the VAE reconstruction to get a slightly noisy version of the true features.

All speakers were parametrized with a common backbone comprising three, 64-dimensional fully-connected ReLU layers, feeding into the communication "head." Details of the particular head architectures are included below. The communication decoder comprised 2 fully-connected ReLU layers with hidden dimension 64, mapping from communication vectors to reconstructed images. The listener agent comprised two separate linear layers that mapped from communication or image features to 16 dimensions; the listener's output was computed via the softmax of the cosine similarity between the communication embedding and each image's embedding.

The decoding loss was calculated as the mean squared error (MSE) between the decoder's output and the speaker's observation. The classification loss was the categorical crossentropy of the listener's output and the onehot label for which candidate image was the target. Agents were trained with batch size 64 for 100,000 batches.

A.1 VQ-VIB Speaker

The VQ-VIB Speaker generated communication vectors by first sampling in a 64-dimensional latent space using the reparametrization trick for sampling from a Gaussian and then quantizing to one of the vectors in the codebook. We parametrized VQ-VIB agents with 1024 codebook entries.

Traditional VQ architectures sometimes suffer from "codebook collapse" in which only a small fraction of quantized vectors are used. We found that simply using a small positive $\lambda_C = 0.01$ overcame such issues, likely because regulating complexity increased the stochasticity of encodings, preventing premature settling to local minima.

As in standard VQ architectures, we set the hyperparameter β to tradeoff between the committment and embedding losses. In experiments for $\lambda_I > 0$, we set $\beta = 0.25$, the default value. In experiments for $\lambda_I = 0$, we set $\beta = 0.01$; this lower value allowed prototypes to move more, which appeared necessary given the weaker training signal when the informativeness loss was not present.

In all experiments, we used an Adam optimizer with default parameters

A.2 Onehot Speaker

Although the results in our main paper focused on VQ-VIB agents, we conducted some baseline experiments with onehot-based communication (results reported in Appendix C). The onehot speaker produced onehot vectors by passing through a "hard" Gumbel-softmax layer. We used a 1024-dimensional Gumbel-softmax layer to allow for up to 1024 unique messages.

Training agents with default settings for this layer often failed to achieve better-than-random chance; the speaker often devolved to outputting the same message for all inputs. We explored several methods for addressing such failures, including losses for the message entropy (as discussed in Eccles et al. [34], among others) and increasing the temperature parameter of the Gumbel softmax layer (as explored in reference games by Chaabouni et al. [27], among others). Sweeping over entropy weights from 0.0001 to 0.1 (at powers of 10), and temperatures in the set `[0.1, 1.0, 10.0, 20.0, 50.0]` we achieved the best self-play results by penalizing the message entropy with weight 0.001 and setting the temperature to 20 throughout training, while setting $\lambda_C = 0$. We note that onehot-based communication appeared much more sensitive to hyperparameter tuning than VQ-VIB agents.

Agents were trained using an Adam optimizer with default hyperparameters, except for the learning rate, which we set to 0.0001. This rate led to higher self-play scores compared to 0.001 (the default, which led to unstable training that caused communication collapse) and to 0.00001, which never achieved greater-than-random scores.

A.3 Prototype Speaker

Similarly to using onehot agents as a baseline, we conducted additional experiments with prototype-based agents, proposed by Tucker et al. [23]. Such agents used an internal Gumbel-softmax layer with 1024 units to select a trainable prototype in 64-dimensional communicative space (to match VQ-VIB). We used default parameters for all settings, except the Adam optimizer learning rate which, as for onehot agents, we set to 0.0001. We also set $\lambda_C = 0$ to encourage highly informative (and complex) communication.

B Appendix: Visualization of Learned Communication

In this section, we further explored visualization of learned communication via the visualization method used for Figure 5, wherein we recorded the mean EC vector for images associated with different topnames. In Figure 7, we plotted the 2D PCA of EC for in-distribution images (blue), EC for OOD images (in red), and the result of translating English words into the EC space (in green).

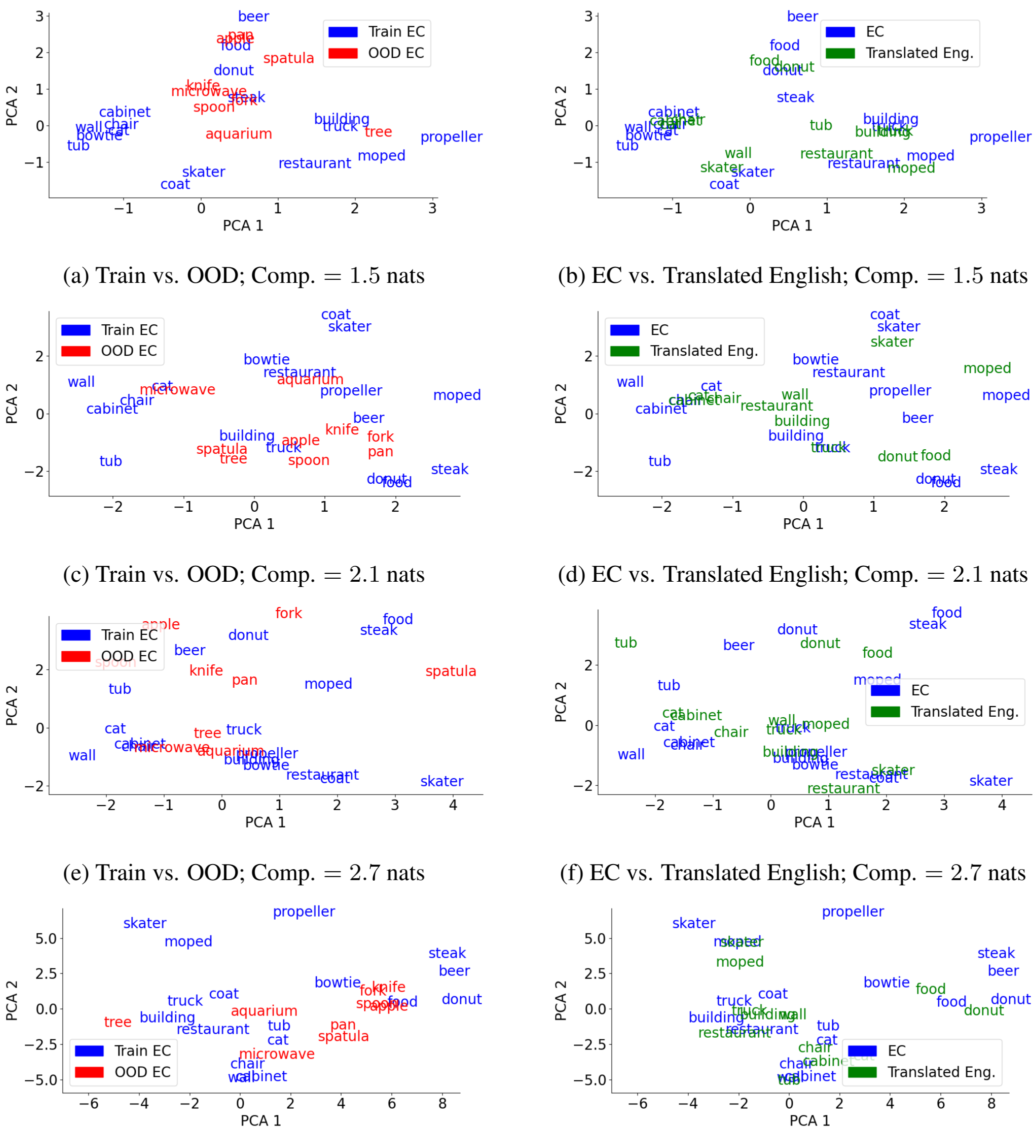
Across models of varying complexity, we consistently found that agents, which had been trained on images of walls, dressers, steaks, etc. (but not cribs, spoons, forks, etc.) learned a semantically-meaningful communication space. Words for household items like “dresser" and “‘cabinet” are near each other, while foods form another distinct cluster. Furthermore, this semantic space generalized to held-out classes as well. An image of an apple, for example, induced similar communication to images of steaks, donuts, and food in the training set. This explains how agents were able to generalize to held-out classes so well: the form of communication vectors encoded information.

Inspection of the translated GloVe embeddings gives a sense of what EC listeners observed in translation experiments. While admittedly imperfect, the linear translator model was largely able to locate translated communication in the right semantic area. For example, in Figure 7 d, the translated EC for “donut” and “food” are not perfectly aligned with EC for donuts or food, but the two representations are still near each other.

C Appendix: Other communication architectures

Although the main focus of our work is on the relationship between information-theoretic properties of EC and generalization or translation (and therefore independent of the exact form of communication), we performed some experiments with onehot- and prototype-based (the architecture introduced by [23]) communication. Results with such agents reinforced the main findings of our paper establishing a link between complexity and generalization, regardless of the specific speaker architecture. Further, our results indicated that VQ-VIB agents enabled far greater control over communication complexity compared to onehot or prototype agents.

Recreating the OOD experiments for onehot- and prototype-based agents, team utility, plotted against complexity, is shown in Figure 8, with VQ-VIB results copied from the main paper for reference. Generally, all agent architectures achieved roughly similar utility for the same complexity, but non-VQ-VIB agents failed to learn as complex communication. (Interestingly, prototype agents seemed to learn more complex communication than onehot, but not as complex as VQ-VIB.) Note that simply



(a) Train vs. OOD; Comp. $= 1.5$ nats

(b) EC vs. Translated English; Comp. $= 1.5$ nats

(c) Train vs. OOD; Comp. $= 2.1$ nats

(d) EC vs. Translated English; Comp. $= 2.1$ nats

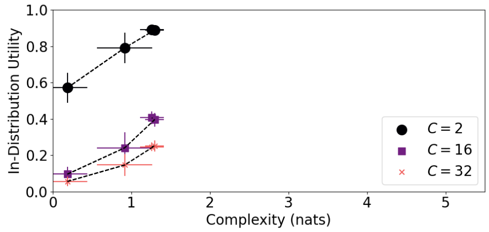
(e) Train vs. OOD; Comp. $= 2.7$ nats

(f) EC vs. Translated English; Comp. $= 2.7$ nats

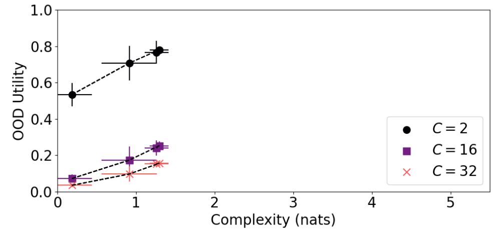
(g) Train vs. OOD; Comp. $= 4.6$ nats

(h) EC vs. Translated English; Comp. $= 4.6$ nats

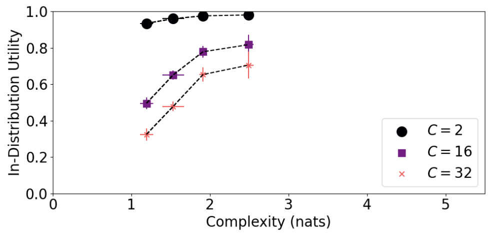
Figure 7: 2D PCA of communication for OOD inputs (left) or translated GloVe embeddings (right) at various complexity levels (different rows).



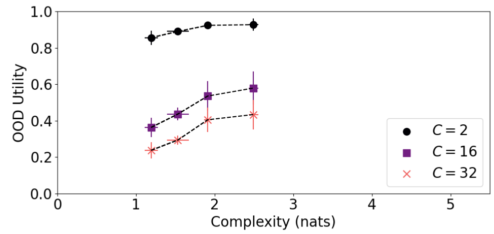
(a) In-distribution evaluation; onehot



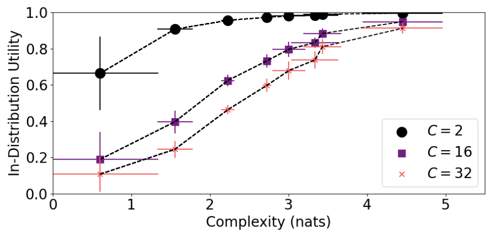
(b) Out-of-distribution evaluation; onehot



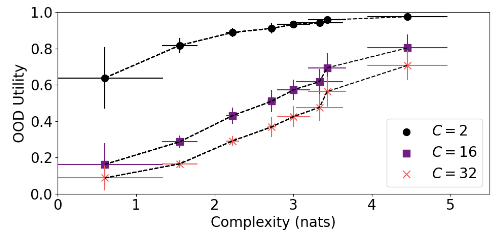
(c) In-distribution evaluation; prototype



(d) Out-of-distribution evaluation; prototype



(e) In-distribution evaluation; VQ-VIB $n = 4$



(f) Out-of-distribution evaluation; VQ-VIB $n = 4$

Figure 8: In- and out-of-distribution utility for onehot (top), proto (middle), or VQ-VIB agents (bottom). All agents exhibited similar utility for the same complexity, but VQ-VIB agents learned far more complex communication, which allowed them to perform better in harder tasks.

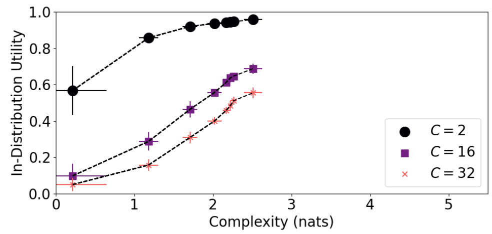
increasing λ_I failed to produce the desired effect for onehot agents, which converged to similar behavior for $\lambda_I = 10$ and $\lambda_I = 100$.

Lastly, translation performance with onehot-based agents was predictably poor, with utility for $C = 2$ never exceeding 75%, even for $\lambda_I = 100$. Such failure can be attributed to two limitations of onehot agents. First, as already shown in earlier experiments, onehot agents learned lower-complexity communication than VQ-VIB agents, which limited the maximum utility. Second, specifically in the context of translation, communicating via onehot vectors inherently limits the semantic relationships between messages. VQ-VIB agents, with discrete representations in a continuous space, could leverage the alignment dataset to learn a transformation of the whole space. Conversely, because every onehot vector is by definition orthogonal to every other onehot vector, translation necessarily failed to capture relationships between messages.

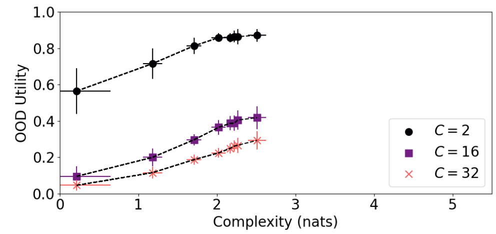
Ultimately, based on these results, we omitted discussion of onehot- and prototype-based communication from the main paper. Our primary focus was the effect of different complexity and informativeness on generalization; given the consistent results across architectures but more limited range of complexities for non-VQ-VIB agents, VQ-VIB agents simply presented more interesting results.

D Appendix: Combining Quantized Vectors

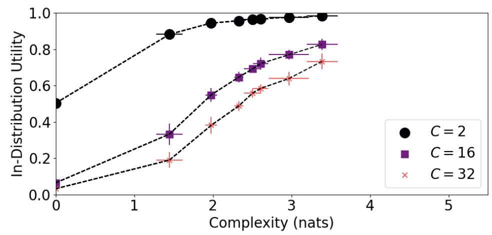
In the main paper, we presented results generated for VQ-VIB agents with $n = 4$. In Figure 9, we include results from our `topname` experiments, generated for $n = 1, 2,$ and 4. In general, we



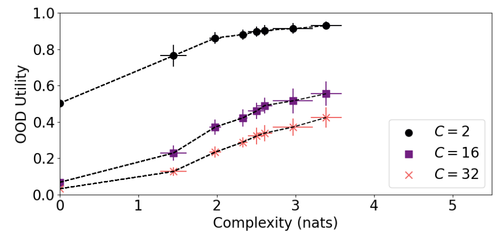
(a) In-distribution evaluation; $n = 1$



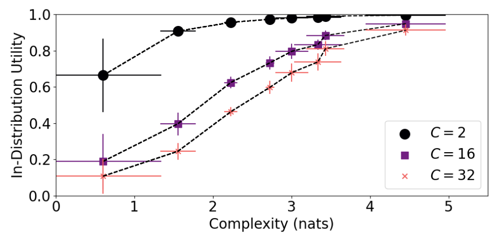
(b) Out of distribution evaluation; $n = 1$



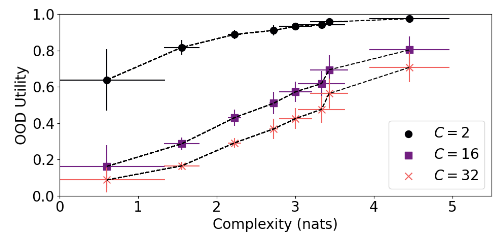
(c) In-distribution evaluation; $n = 2$



(d) Out of distribution evaluation; $n = 2$



(e) In-distribution evaluation; $n = 4$



(f) Out of distribution evaluation; $n = 4$

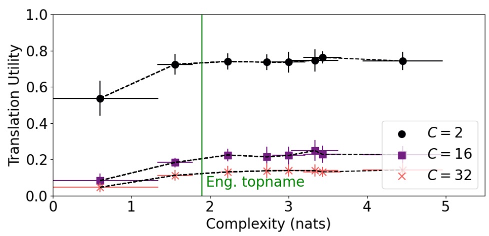
Figure 9: Team performance on in-distribution (a) and out-of-distribution (b) images, for various n (different rows). The general complexity-utility tradeoff remained the same, but using greater n allowed greater complexity.

found that using greater n increased communication informativeness and complexity, which in turn supported greater utility. Most importantly, across n, we found a similar relationship between utility and complexity. Thus, we view this architectural change of using larger n as allowing greater control over the complexity and informativeness of communication (which, as we discuss in the main paper, is important for downstream tasks).

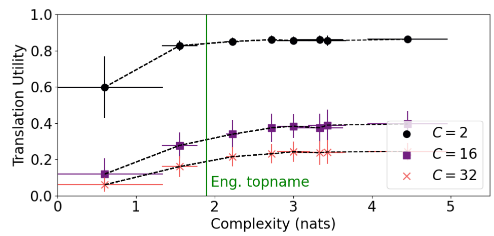
E Appendix: Varying Translation Alignment Data

In Section 5, we presented results for a simulated English speaker, with communication translated for an EC listener. The translator was a linear transformation, fitted with N randomly-drawn examples from the training dataset. In the main paper, we presented results for $N = 100$; here, in Figure 10, we present results for various N.

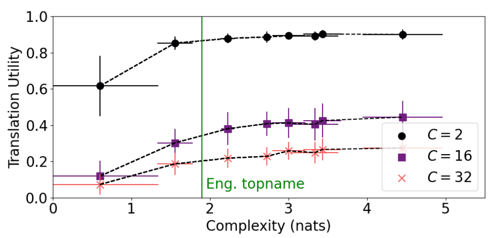
Ultimately, while increasing N helped translation somewhat, the general trends from the main paper held true. That is, as EC complexity increased from 0 to just over 2 nats, team utility increased, but at greater complexity values, team performance was bottlenecked by the English speaker. Furthermore, beyond a certain value, increasing N further did not appear to improve performance, as demonstrated by the nearly identical performance for $N = 100$ and $N = 1000$.



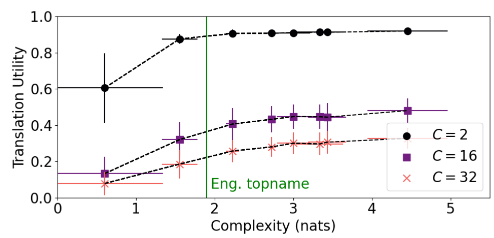
(a) $N = 10$



(b) $N = 50$



(c) $N = 100$



(d) $N = 1000$

Figure 10: Utility of human-agent teams using English GloVe to EC translation, for various translation dataset sizes. Increasing the dataset size helped up to a point, but performance plateaued around 2 nats in all cases.