Catalytic Role Of Noise And Necessity Of Inductive Biases In The Emergence Of Compositional Communication

Anonymous Author(s)
Affiliation
Address
email

Abstract

Communication is compositional if complex signals can be represented as a combination of simpler subparts. In this paper, we theoretically show that inductive biases on both the training framework and the data are needed to develop a compositional communication. Moreover, we prove that compositionality spontaneously arises in the signaling games, where agents communicate over a *noisy channel*. We experimentally confirm that a range of noise levels, which depends on the model and the data, indeed promotes compositionality. Finally, we provide a comprehensive study of this dependence and report results in terms of recently studied compositionality metrics: topographical similarity, conflict count, and context independence.

1 Introduction

In emergent communication studies, one often considers agents who can share information about a set of objects described by the common features. Such a situation is common in multi-agent systems with partial observation (Foerster et al. [2016], Lazaridou et al. [2016], Jaques et al. [2018], Raczaszek-Leonardi et al. [2018]) and it is the major theme in signaling games (Fudenberg et al. [1991], Lewis [1969], Skyrms [2010], Lazaridou et al. [2018]). In a signaling game, one agent (a sender) conveys information about an object to another agent (a receiver), which then has to infer the object's features. Typically, agents are rewarded if some of the features are correctly identified. During this process, the agents develop a communication protocol. A recent line of work has studied conditions under which compositionality emerges (Batali [1998], Kottur et al. [2017], Choi et al. [2018], Korbak et al. [2019], Li and Bowling [2019], Słowik et al. [2020b,a], Guo et al. [2020]).

Compositionality is a crucial feature of natural languages and it has been investigated extensively in cognitive science (see e.g. Chomsky [1957] Fodor and Pylyshyn [1988]). It is often measured using dedicated metrics such as topographic similarity (Brighton and Kirby [2006], Lazaridou et al. [2018], Kriegeskorte [2008], Bouchacourt and Baroni [2018]), context independence Bogin et al. [2018], conflict count Kuciński et al. [2020], or positional disentanglement (Chaabouni et al. [2020]). In signaling games it bears a strong resemblance to the concept of disentangled representations, see (Higgins et al. [2017], Kim and Mnih [2018], Locatello et al. [2019]). In machine learning context, compositionality is perceived as a generalization mechanism (Lake et al. [2016]) and has been used e.g. for goal composition (Jiang et al. [2019]) or knowledge transfer (Li and Bowling [2019]).

In this paper, we theoretically show that inductive biases on both the training framework and the data are needed for compositionality to emerge. A similar observation has been made by Kottur et al. [2017]; however, our result is more fundamental and points out a common misconception that compositionality can be learned in a purely unsupervised way. Such a result can be perceived as a discrete analog of Locatello et al. [2019], applicable in the communication context.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

We then prove that adding an inductive bias in the loss function coupled with communication over a noisy channel leads to the spontaneous emergence of compositionality. This shows the catalytic role of noise in this process. Intuitively, this can be attributed to the (partial) robustness of compositional language with respect to message corruption caused by a noisy channel.

We experimentally verify that a certain range of noise levels, dependent on the model and the data, promotes compositionality. We provide a wide range of experiments that illustrate the influence of different priors. For the inductive biases in the training framework, we look into the impact of the network architecture as well as implementation and temporal variation in noise. On the data side, we study the effect of scrambling visual input or its description. We also study the generalization properties of the proposed training framework.

2 Related work

The topic of communication is actively studied in multi-agent RL, see Hernandez-Leal et al. [2020, Table 2] for a recent survey. Compositionality is often investigated in the context of signaling games (Fudenberg et al. [1991], Lewis [1969], Skyrms [2010], Lazaridou et al. [2018]). Recent research has shown that strong inductive biases or grounding of communication protocols are necessary for the protocol to be compositional (see e.g. Kottur et al. [2017], Słowik et al. [2020b]). The inductive bias can be imposed into the architecture of the agents or the training procedure. For instance, Das et al. [2017] place pressure on agents, to use symbols consistently across varying contexts, by a frequent reset of the agent's memory. A model-based approach was proposed by Choi et al. [2018] and Bogin et al. [2018], who build upon the obverter algorithm (Oliphant and Batali [1997], Batali [1998]). Słowik et al. [2020a] explore games with hierarchical inputs and shows how agents implemented as graph convolutional networks obtain good generalization. Korbak et al. [2019] implemented the idea of template transfer [Barrett and Skyrms, 2017] by pre-training the agents on simpler subtasks before the target task. Kirby [2001] studied the iterative learning paradigm, where each generation of agents learns the language spoken by the previous generation before starting to communicate. In the machine learning literature, this idea was explored by Li and Bowling [2019], Cogswell et al. [2019] and Ren et al. [2020] with the generation transfer typically implemented as reinitializing the weights of agents' neural networks. Such an approach inevitably introduces noise into the learning process. This naturally leads to a question of whether the noise itself may be a sufficient mechanism of compositionality, which we will try to address in this paper. Guo et al. [2020] have shown that the choice of a game has a large impact on the properties of a communication protocol emerging in that game, foreshadowing what we call grounding.

The noisy channel model of communication was famously introduced by Shannon [1948]. The idea of noise as a driving force in the emergence of communication was first proposed by Nowak and Krakauer [1999], who showed that word-level compositionality is the optimal solution to the problem of communication in a noisy environment under a particular fitness function. Noise is also used in deep learning, e.g. as a regularizer (see e.g. dropout [Srivastava et al., 2014]) or a mechanism allowing backpropagation through a discrete latent (see e.g. Salakhutdinov and Hinton [2009], Kaiser and Bengio [2018]). Noise in the latter context was used in Foerster et al. [2016] in order to learn to communicate. The authors observed that it is essential for successful training.

3 Noisy channel method

3.1 Language and compositionality

Consider a set of objects described by some factors, and let a set $\mathcal{F}$ contain a combination of these factors' values. For example, the factors could represent shape, say `squares` and `circles`, and color, say `red` and `green`. Then $\mathcal{F} = \{\texttt{red square}, \texttt{red circle}, \texttt{green square}, \texttt{green circle}\}$ and the factors could be identified with partitions of $\mathcal{F}$. We will assume that objects are defined in a space $\mathcal{X}$ and are generated by a two-stage process: first, the factor values are sampled from $f \in \mathcal{F}$, then an element of $\mathcal{X}$ is sampled according to a distribution conditioned on f.

A language is defined as a mapping from objects to strings over some finite alphabet $\mathcal{A}_s$, $\ell : \mathcal{X} \to \mathcal{A}_s^*$. We will say that ℓ is compositional with respect to a given factor if a change in i-th factor only impacts a corresponding j-th index of the message. Continuing the previous example, let $\mathcal{A}_s = \{\texttt{a}, \texttt{b}\}$ and consider a language mapping red squares, red circles, green squares, and green circles to $\texttt{aa}, \texttt{ba}, \texttt{ab}$,

and bb, respectively. Then ℓ is compositional with respect to color and shape factors since the change of color only impacts the first index in the message (and analogously for the shape).

Consider what happens when we permute elements of $\mathcal{F}$ before generating elements of $\mathcal{X}$. In our running example, suppose that `red circle` is swapped with `green circle` (with other elements remaining unchanged) and denoted this permutation by π. The resulting language ℓ_π would then map red squares, red circles, green squares, and green circles to aa, bb, ab, and ba, respectively. This language is not compositional with respect to color and shape factors, since if we change a shape value in $f =$ `red circle`, both symbols in the message will change (from bb to aa). However, if instead, we look at a change along with one factor *after* applying π, then ℓ_π is compositional.

3.2 Inductive biases and compositionality

In this section, we investigate, whether compositionality can spontaneously emerge in an unsupervised fashion. Theorem 1, which can be viewed as a discrete version of Locatello et al. [2019, Theorem 1].

Theorem 1. *For a uniform distribution μ on $\mathcal{F}$ and a permutation $\pi\colon \mathcal{F} \to \mathcal{F}$, the distribution $\mu \circ \pi^{-1}$ is also uniform.*

While this result is elementary to prove (see Appendix F) it has deep implications for the emergence of compositionality in the learning process. If $\mu \circ \pi^{-1}$ is the same as μ, then the distribution of the data remains unchanged whether we permute factors or not. Consequently, it is not possible to identify from data whether factors were permuted or not. This means that any emergent compositional language ℓ with respect to some factors is not compositional with respect to other (permuted) factors. We arrive at the following conclusion:

Any learning process which hopes to achieve compositionality must involve some priors related to the data and learning framework.

There are multiple ways of grounding the learning process, including imposing inductive biases on the agents and designing loss functions to disentangle features. We study this, together with a new mechanism: injecting noise into the communication channel. In the next section, we show that this mechanism provably achieves compositionality.

3.3 Compositionality and communication over a noisy channel

Our major conceptual finding is that compositional communication spontaneously emerges when introducing a relatively simple mechanism – a noisy channel. This is proved in Theorem 2, provided that the loss function penalizes agents' mistakes, but also rewards for (partially) correct guesses.

We now define the noisy channel. It is located between the sender and the receiver and may scramble messages. A message $\mathbf{s}$ is transformed into a corrupted message $\mathbf{s}'$, by replacing each symbol with a different symbol, independently and with probability $\varepsilon \in (0, 1)$.

Recall that the definition of compositionality connects the change in factor values with the change in message symbols. It is thus reasonable to look for loss functions that are somehow factorized in terms of individual symbols (and their corruption probabilities). The corrupted message corresponding to $\mathbf{f}$ is $\ell(\mathbf{f})'$ and the features corresponding to this corrupted message are defined as $\mathbf{f}' = \ell^{-1}(\ell(\mathbf{f})')$. We will consider two loss types. The first one is of the form

$$J_1(\ell, \mathbf{f}) = F(e_1, \dots, e_K), \quad e_j := \mathbb{P}(\mathbf{f}'_j \neq \mathbf{f}_j), \tag{1}$$

where F is a non-negative function such that $F(\varepsilon, \dots, \varepsilon) = 0$. Example of such a function could be

$$F(x_1, \dots, x_K) = -\frac{1}{K} \sum_{i=1}^{K} \log \big(x_i / \sum_{j=1}^{K} x_j \big) + \log(1/K),$$

i.e. the Kullback-Leibler divergence between a uniform distribution and a probability distribution $p_i \propto x_i$. We use this type of loss in our experiments. It measures the probability of a flip of each factor separately (e_j, and aggregates these probabilities over all features (with function F).

The second type of a loss function is of the following form

$$J_2(\ell, \mathbf{f}) = \mathbb{E}[H(\rho(\mathbf{f}', \mathbf{f}))], \tag{2}$$

where H is a non-negative, increasing function and ρ is the Hamming distance[1]. Intuitively, ρ measures the number of flipped factors in the corrupted message and H controls the degree in which we penalize this number. Notice that J_2 takes into account the joint distribution of flipped factors.

Theorem 2. *Assume $\mathcal{F} = \mathcal{X}$ and the noise corruption probability ϵ is small enough. Then a compositional language minimizes J_1 and J_2 over all one-to-one languages. Conversely, if ℓ is optimal for J_2 then ℓ is compositional.*

Informally, Theorem 2 states that optimization of loss functions J_1 and J_2 promotes compositionality (assumption on the one-to-one property is technical). What makes this possible is the factorized nature of the losses and the introduction of a noisy channel. The proof of Theorem 2 can be found in Appendix F.

3.4 Training pipeline

In this section, we sketch the training pipeline and postpone the details to Appendix C. The dataset of images observed by the sender is denoted by $\mathcal{D}$. Each element of $\mathcal{D}$ has K independent features $f_1, \ldots, f_K$ (here we consider $K = 2$). Both the sender and the receiver are modeled as neural networks (for details see Appendix B). The sender network takes an image from $\mathcal{D}$ as input and returns a distribution over the space of messages of length L (here we assume $L = K = 2$). We assume that that conditionally on the image, the symbols in the message are independent and take values in a finite alphabet $\mathcal{A}_s = \{1, \ldots, d_s\}$. This distribution is then distorted by the noisy channel, which is a function that maps probability vectors into d_s-dimensional logits. Unless otherwise stated, we assume that noise is be defined as a dense layer with a specific choice of (not learnable) weights matrix and a $\log$ activation function:

$$\texttt{noise}(x) = \log(Wx). \tag{3}$$

Here we assume that $W \in \mathbb{R}^{d_s \times d_s}$ is a fixed matrix, which takes a probability vector to a probability vector with positive entries. An example of such W is a stochastic matrix. In this paper we use

$$W_{ij} = \begin{cases} 1 - \varepsilon, & i = j, \\ \frac{\varepsilon}{d_s - 1}, & i \neq j. \end{cases} \tag{4}$$

Alternative implementations of noise architecture are possible, see Appendix C. The noisy logits are then used to sample a message and pass it to the receiver. To make this operation differentiable we use Gumbel-Softmax (Jang et al. [2016]) with Straight-Through mode (Kaiser and Bengio [2018]).[2] Upon receiving the message, the receiver outputs a distribution over possible values of the features, using an alphabet $\mathcal{A}_r = \{1, \ldots, d_r\}$. Finally, the neural networks are trained using a linear combination of cross-entropy loss for the receiver, cross-entropy loss for the sender, and L_2-regularization. The loss term for the sender incentives the language to be a one-to-one mapping.

4 Experimental setup

4.1 Datasets

We use two datasets: shapes3d (used in Burgess and Kim [2018] and included in the TensorFlow datasets package) and a dataset used by Choi et al. [2018][3] and Korbak et al. [2019], which we will refer to as the obverter dataset, see Figure 1. The datasets are similar but offer different forms of visual variability. Shapes3d dataset includes images of 3D shapes. Each element is a $(64, 64, 3)$ RGB image, and is characterized by multiple factors, such as shape or object hue. We choose images with values for both factors ranging in $\{0, 1, 2, 3\}$. The obverter dataset contains images of four shapes (box, cylinder, ellipsoid, sphere) in four colors (blue, cyan, gray, green). Each image has dimensions $(128, 128, 3)$, and we use 1000 images for each shape–color pair. For details see Appendix A.

4.2 Training details

[1]The Hamming distance between two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^K$ is defined as $\rho(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^{K} \mathbf{1}(v_i \neq w_i)$.

[2]We believe that Gumbel-Softmax approach is now an established choice in emergent communication research, see e.g. Lee et al. [2017], Mordatch and Abbeel [2017]. Chaabouni et al. [2020] report that Gumbel-Softmax converges to similar solutions as REINFORCE but faster and is more stable.

[3]The dataset is available at `https://github.com/benbogin/obverter`. We used the code provided in the repository to generate 1000 images for each color-shape pair.

The sender and the receiver are implemented as feed-forward neural networks. To ensure the diversity of random initializations we run each experiment with 100 random seeds. For each seed, there were 100 evaluation runs, once every 2000 network updates. We report metrics that are then averaged over the last 20 evaluation runs. For details concerning architecture and the choice of hyperparameters, see Appendix B.

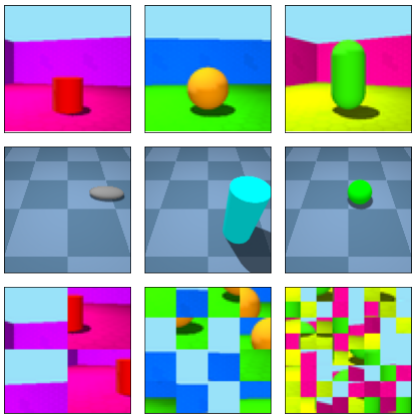


Figure 1: Top: shapes3d. Middle: obverter. Bottom: Scrambled shapes3d (32, 16, 8).

4.3 Compositionality measures

We measure compositionality in terms of four metrics used in emergent communication literature: topographic similarity, conflict count, context independence, and positional disentanglement. For all metrics, the higher values the better, except for conflict count, for which the reverse is true. The results presented in Section 5 indicate that the metrics agree in the assessment of our results. Parallel to compositionality metrics, we also report accuracy, which we often refer to as *acc*.

Topographic similarity Topographic similarity [Brighton and Kirby, 2006, Lazaridou et al., 2018], or *topo* for short, is a popular measure of structural similarity between messages and features. Let $L_f : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ be a distance over features and $L_m : A_s^* \times A_s^* \to \mathbb{R}_+$ be a distance over messages. The topographical similarity is the Spearman ρ correlation of L_f and L_m measured over a joint uniform distribution over features and symbols. We choose L_m to be the Levenshtein [1966] distance and treat features $f \in \mathcal{F}$ as ordered pairs or factors so we can choose L_f to be the Hamming distance. We use topo as the main metric.

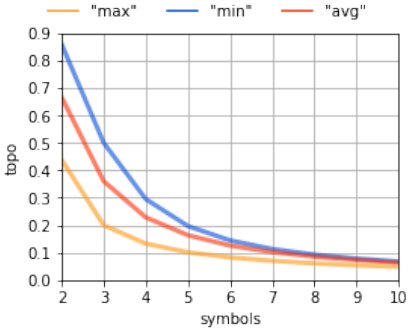


Figure 2: Expected value of topographic similarity for a random bijective language with message length 2, as a function of the alphabet size. "min", "max", and "avg" stand for different ways of computing ranks.

Figure 2 shows the expected value of topographic similarity for a random bijective language with a message length equal to 2. For 5-symbol alphabet, this value does not exceed 0.2 which sets a point of reference for compositionality results. For derivation see Appendix E.

Conflict count Let the principal meaning of a symbol s at position j assuming ϕ be $\mathtt{m}(s, j; \phi) := \arg\max_f \mathtt{count}(s, j, f; \phi)$, where $\mathtt{count}(s, j, f; \phi)$ is defined as $\sum_{\mathtt{img} \in \mathcal{D}} \mathbf{1}(s_{\mathtt{img},j} = s, f_{\mathtt{img},\phi(j)} = f)$, and ties in $\arg\max$ are broken arbitrarily. Then, conflict count is metric is defined as $\mathtt{conf} = \min_\phi \sum_{s,j} \mathtt{score}(s, j; \phi)$, where $\mathtt{score}(s, j; \phi) = \sum_{f \neq \mathtt{m}(s,j;\phi)} \mathtt{count}(s, j, f; \phi)$. Intuitively, `score` measures how many times the feature assigned to a symbol s at a position j diverts from its principal meaning $\mathtt{m}(s, j; \phi)$. `conf` sums these errors and takes $\min$ over possible orderings ϕ. This metric was introduced by Kuciński et al. [2020], and we will often refer to it as *conf*.

Context independence Context independence (Bogin et al. [2018]), abbreviated here as *cont*, measures the alignment between symbols forming a message and features $f_1, \ldots, f_K$. By $p(s|f)$, we mean the probability that the sender maps a feature f to a message containing symbol $s \in \Sigma$. We define the inverse probability $p(f|s)$ similarly. Finally, we define $s^k := \arg\max_s p(f|s)$; s^f is the symbol most often sent in presence of a feature f. Then, context independence is $\mathbb{E}(p(v^k|k) \cdot p(k|v^k))$; the expectation is taken with respect to the joint uniform distribution over features and symbols. Intuitively, context independence measures the consistency of associating symbols with shapes irrespective of colour and vice versa.

Positional disentanglement Let s_j denote the j-th symbol of a message $f(d)$, and c_1^j the feature with the highest mutual information with s_j, and c_2^j with the second highest mutual information: $c_1^j = \arg\max_c \mathcal{I}(s_j; c)$, $c_2^j = \arg\max_{c \neq c_1^j} \mathcal{I}(s_j; c)$ where $\mathcal{I}(\cdot;\cdot)$ is mutual information and c is a feature value. Then, positional disentanglement Chaabouni et al. [2020] is defined as $\frac{1}{L} \sum_{j=1}^{L} \left(\mathcal{I}(s_j; c_1^j) - \mathcal{I}(s_j; c_2^j) \right) / \mathcal{H}(s_j)$, where L is the maximum message length and $\mathcal{H}(s_j)$ is

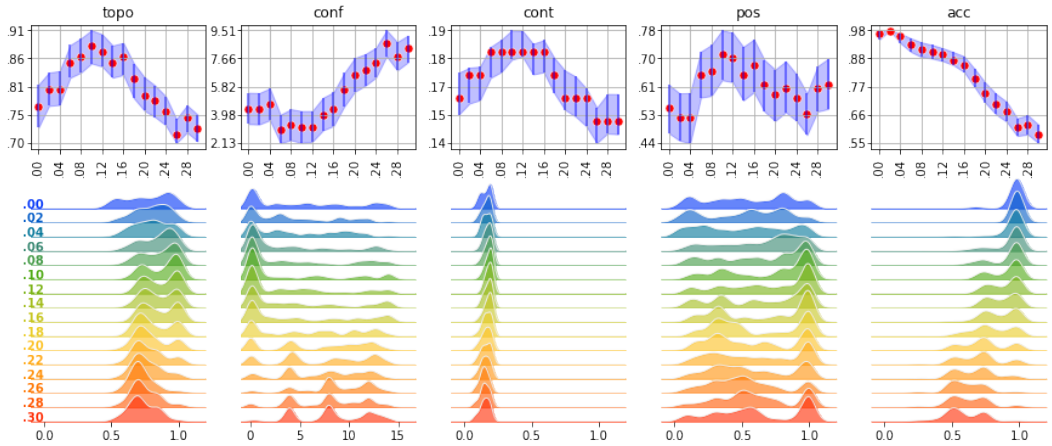


Figure 3: Results of the main experiment on shapes3d dataset. Top panel: average value of metrics for various noise levels. The shaded area corresponds to bootstrapped 95%-confidence intervals. Bottom panel: kernel density estimators for metrics and noise levels across seeds. Here *topo* stands for topographic similarity, *conf* for conclict count, *cont* for context intependence, *pos* for positional disentanglement and *acc* for accuracy.

entropy over the distribution of symbols at j-th place in messages for each feature. We ignore positions with zero entropy.[4] We will call this measure *pos*, for short.

5 Experiment results

In this section, we study how different inductive biases for the model and the data influence compositionality. The main experiment is presented in Section 5.1 and the experiments in Section 5.2-5.4 are its variation. In the main experiment we assume that the message length is $K = 2$, $|\mathcal{A}_s| = 5$, and $|\mathcal{A}_r| = 8$. Notice that both $\mathcal{A}_s$ and $\mathcal{A}_r$ differ from one another and from the number of possible feature values (which equals 4). It turns out that the qualitative results are similar for both the shapes3d and obverter datasets, and in the interest of brevity we only report the results for both in the master experiment. Supplementary material for this section can be found in Appendix D.

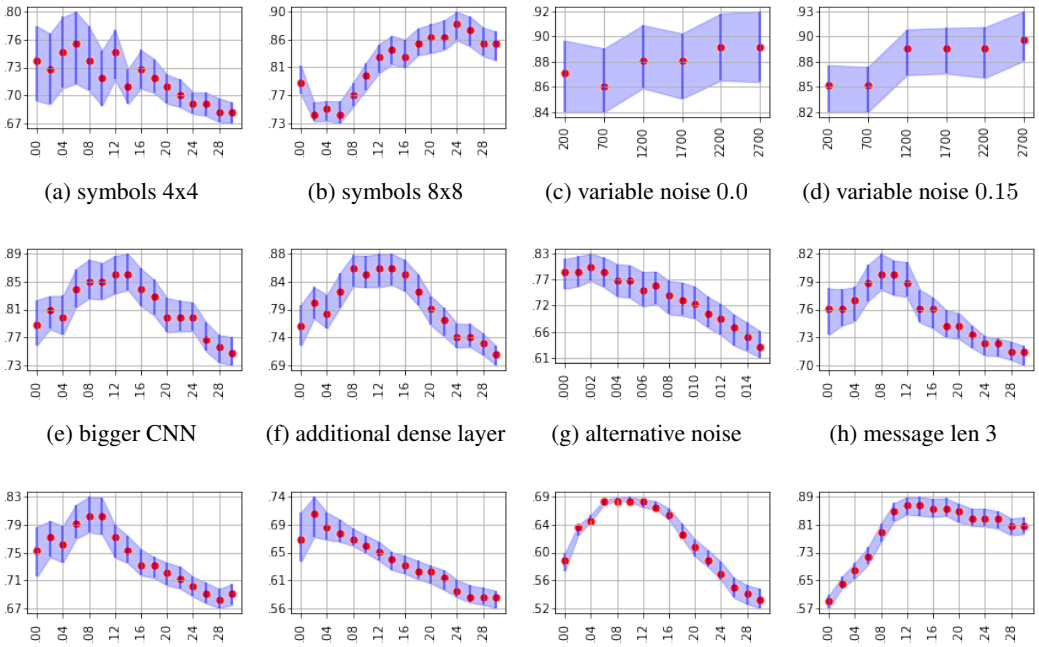
5.1 Main experiment: emergence of compositionaliy

The results, presented in Figure 3, illustrate how noise catalyzes the emergence of compositionality.The top panel of Figure 3 shows the important patterns for metrics: they improve until an extremal point is reached, and decline afterward. Topographic similarity achieves extremum for noise level 0.1, reaching value 0.87, which is a significant improvement upon 0.79 for the lack of noise. The accuracy drops down with an increase of the noise level, as expected, however the speed of the decline increases. This shows that there is an interesting compositionality-accuracy trade-off. The bottom panel of Figure 3 complements the overall picture with a visualization of metrics' distribution. We see interesting dynamics in topographic similarity distribution with respect to change in the noise levels. Namely, it starts by accumulating mass at the higher spectrum of its values, reaching a peak for noise 0.1, after which it transitions to a bimodal distribution, finally shifting its mass more towards the mediocre end of the spectrum. An interesting observation is that the results for experiments with achieved high accuracy, are not only better but also the effect of noise is more pronounced (for instance there are 47 seeds with accuracy exceeding 0.9, for which noise 0.16 yields 0.96 topo). For the sake of brevity, we defer the discussion of this phenomenon to Appendix D.1. The accuracy undergoes a similar transformation as a topographic similarity. The detailed numerical analysis, as well as corresponding results for the obverter dataset, can be found in Appendix D.1.

5.2 Influence of model inductive bias

Different number of symbols Here we study the impact of a different number of symbols on compositionality. In our experiments we used the communication channel with $|\mathcal{A}_s| = 5$, giving

[4]Positional disentanglement is related to residual entropy proposed by Resnick et al. [2020]. Chaabouni et al. [2020] also proposed bag-of-words disentanglement, which assumes order-invariance of messages. Due to our architecture choice, this assumption is not met, hence we decided not to report this metric.



(a) symbols 4x4 (b) symbols 8x8 (c) variable noise 0.0 (d) variable noise 0.15

(e) bigger CNN (f) additional dense layer (g) alternative noise (h) message len 3

(i) scramble 32 (j) scramble 16 (k) scramble 8 (l) scrambled labels

Figure 4: Topographic similarity for experiments in Sections 5.2 and Section 5.3. The shaded areas correspond to bootstrapped 95%-confidence intervals for average topo.

a total of $25 = (5 \times 5)$ possible messages. This is slightly redundant since only $16 = (4 \times 4)$ is required, so this is the first case that we study here. It turns out, that allowing for only 16 messages makes the training less stable. For topographic similarity, see Figure 4(a), small to medium values of noise exhibit wide confidence intervals and it is statistically hard to distinguish between the metric values (this might be attributed to a bimodal distribution of topo in this noise range, see Figure 9 in Appendix D.2). For larger values of noise (greater than 0.16), topo starts to visibly decline. As the second experiment, we considered the total of $64 = (8 \times 8)$ messages. Interestingly, the topographic similarity values for the small noise regime (up to 0.08) do not improve over the baseline value (0.79), see Figure 4(b). This behavior changes for medium to large values of noise, where we can observe a visible increase in topo, peaking at 0.88 with a noise level of 0.24. For details see Appendix D.2.

Variable noise Understating what happens under varying noise is an interesting and subtle problem. It would most probably arise in more complex situations when a communication channel is a part of a bigger system. Additionally, it might be similar to a phenomenon observed in supervised learning, indicating that training can benefit from learning rate warmup (see e.g. Goyal et al. [2017], Frankle et al. [2020]). While we lay theoretical ground justifying the benefits of noise in the emergence of compositionality, it is less clear how it will work in practice. The gap stems from implementation details, the use of neural networks, which may converge to suboptimal solutions. In the previous sections, we have seen an interesting dependence on the noise level: too high noise may impede learning, while it cannot be too small, as it vanishes within other signals and sources of noise.

In this paragraph, we discuss a simple experiment with increasing noise and leave a more nuanced study for further work. More precisely, in the initial stage of training the noise is kept at the level $\epsilon_0 \in \{0, 0.15\}$, and after $T \in \{200, 700, 1200, 1700, 2200, 2700\}$ network updates it is switched to 0.1. The results for topographic similarity are presented in Figure 4(c) and (d). In $\epsilon_0 = 0.15$ case, we see that topo increases from 0.85 for $T = 200$, to 0.9 for $T = 2700$. The effect for $\epsilon_0 = 0$ is weaker and the variance in the results is quite high. The rest of the details can be seen in Appendix D.3.

Sensitivity with respect to small architecture changes This set of experiments aims to check the impact of changing the parameters of CNN and the dense layers in the agents' network. In the former experiment, we change the number of filters in the sender's CNN architecture from two layers with 8 filters, to two layers with 16 filters. This results in the slight change of topographic similarity profile, see Figure 4(e), with the highest average value of 0.86 for noise range 0.14 significantly

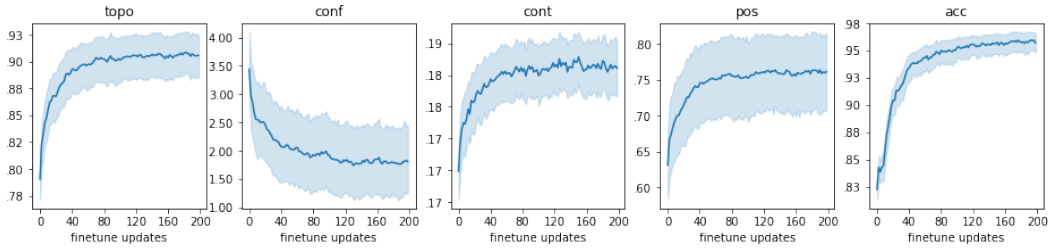


Figure 5: Fine-tuning for a baseline agent trained with noise 0.1 and with several features removed from the training set (the diagonal from the shape-by-color matrix). On the x-axis is the number of finetune updates.

outperforming the zero-noise case (0.79). For the second experiment, we added a dense layer (with 64 neurons) to the receiver's architecture. This again shows improvement of compositionality due to noise, although the noise in the range $[0.08, 0.16]$ performs roughly the same (see also Appendix D.4).

Sensitivity to noisy channel implementation Here we consider an alternative noisy channel implementation, where the noise permutes the sampled symbols, as opposed to distorting the distribution of the symbols (see Appendix C). It turns out that with this change, the scale of noise where the interesting things happen changes as well, see Figure 4(g). Namely, when compared with the main experiment, noise values are in the range smaller by the order of magnitude (we report results for noise levels in $\{0.000, 0.001, \ldots, 0.015\}$. The results suggest that the small values of noise can help, while the larger noise levels lead to a decline in compositionality, see also Appendix D.5.

Longer message In this section, we discuss the experiment with additional factor (with floor color acting as the third factor) and a message of length 3. The result for topographic similarity can be seen in Figure 4(h), see also Appendix D.6). The overall topographic similarity level is lower than in the main experiment, however, the distinctive peak is visible, here for noise level 0.08.

5.3 Data inductive biases

Visual priors This experiment was intended to check how much the CNN-backed input is relevant in the compositionality context. We could conjecture that CNN may facilitate shape recognition and therefore be the driving force in the emergence of languages compositional with respect to the canonical shape, color split. To check this we impair the prior by scrambling images, as depicted in the bottom panel Figure 1. An image is scrambled by splitting it into $(64/x)^2$ disjoint tiles of height and width equal to x, and randomly reshuffling them. This procedure significantly distorts the accuracy profile of the method (see Appendix D.7). In particular, each transition from coarser to finer tiles, the accuracy decreases significantly: for zero-noise it drops from 0.97 for no tiling, to 0.95 for $x = 32$, to 0.79 for $x = 16$ and 0.53 for $x = 8$. The overall compositionality metrics decrease as well, but the characteristic peak for some positive noise level is still present, see Figure (i)-(k). Having said that, the metrics incur a significant boost, when computed for a subset of experiments with high accuracy. We conjecture that the explanation is that the CNN prior is indeed relevant but is not the only one (the output considered in the next section is another).

Scrambled labels In this experiment, we aimed to understand how much the overall architecture output is important in the emergence of compositionality. The receiver's network has a two-headed output and the training framework uses a factorized loss function (1). In the standard setting, we compare the heads' outputs with 'color' and 'shape' respectively, therefore we reflect human priors from the data. In this experiment, we distort this setting, permuting the set of $(\text{color}, \text{shape})$ and factorizing them into new labels (see Appendix D.8). We use a random permutation, so the new labels are abstract and correspond to some joint color-shape concepts. In our experiments, we show that languages, which emerge are compositional with respect to these new concepts, see Figure 4(l). Consequently, they are *not* compositional in the standard color-shape framework. This is in line with the claims of Section 3.2 and further highlights that the output inductive bias is essential.

5.4 Generalization

Network features In this paragraph, we present visualizations of the sender network, to gain some insights into what the network is doing, and whether we can observe any obvious overfit. In the upper panel of Figure 6, a t-SNE (Van der Maaten and Hinton [2008]) visualization is shown for

both the raw shapes3d dataset and the last dense layer of a trained agent with noise 0.1. We observe that the network successfully disentangles features of the data, which in the raw dataset appears to be entangled. This could be a good sign in the context of generalization[5]. Furthermore, Figure 6 (bottom panel) shows occlusion and saliency maps for a sample image from shapes3d dataset. The network seems to pay attention to parts of the image that could be relevant for the task.

Zero-shot and fine-tuning In this experiment, we study the generalization properties of the training protocol. We trained the baseline agent with noise 0.1 with several features removed from the training set (a diagonal from the shape-by-color matrix) and used the removed set to analyze the out-of-distribution performance. In the zero-shot setup, the messages on these observations were not in line with the compositional structure acquired on the training set, which is reflected by rather unimpressive outcomes for each metric (initial values for each metric presented in Figure 5). This may suggest that for a zero-shot generalization stronger biases might be required (see e.g. Słowik et al. [2020a,b]). However, after a few network updates on the whole dataset (without the removed diagonal) we observed a quite significant increase in compositionality metrics, see Figure 5. The highest improvement can be seen roughly in the first 50 additional network updates.

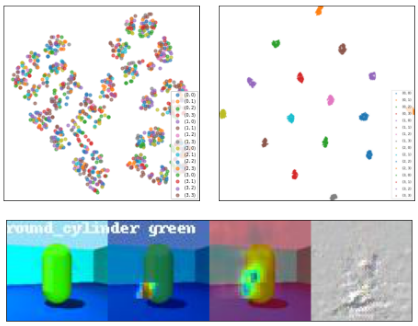


Figure 6: t-SNE visualization (with perplexity 50 and 10000 steps). Top left: shapes3d, top right: features of the last dense layer. Bottom: occlusion and saliency maps.

6 Limitations of the method

Here we summarize a few limitations of the method, that we hope to overcome in future research.

Theory Theorem 2 uses several restrictive assumptions, e.g. fixed message length or alphabet size equal to the range of feature values. In a general setting, it could be the case that noise would promote non-compositional, error-correcting, communication protocols. Stating the general conditions for the emergence of compositionality is an open problem.

Choice of datasets Confirmation on non-synthetic datasets is needed to fully underpin our method.

Communication protocol We assume a simple communication protocol with messages of length two and two features, each taking four values (except for experiments in Section 5.2).

Compositionality model We assume a rather simple compositionality model based of independent features. Going beyond assumption of independence and allowing for hierarchical concepts might be an important conceptual development (see e.g. Andreas [2019]).

Architectures Our architecture might implicitly exhibit some unknown biases that increase compositionality. On the other hand, the architecture might be too simple to achieve strong results in tasks such as zero- or few- shot generalization.

Multi-agent interactions We used supervised training setup, it would be natural to test the influence of noise in reinforcement learning scenarios.

7 Conclusions

In this paper, we theoretically show that inductive biases on both the training framework and the data are needed for the compositionality to emerge spontaneously in signaling games. We then formulate inductive biases in the loss function and prove that they are sufficient to achieve compositionality when coupled with communication over a noisy channel. Consequently, we highlight the catalytic role of noise in the emergence of compositionality. We perform a series of experiments in order to understand different aspects of the proposed framework better. We empirically validate that, indeed, a certain range of noise levels, dependent on the model and the data, promotes compositionality. Our work is foundational research and does not lead to any direct negative applications.

[5]It is known that t-SNE does not necessarily represent cluster sizes, distances, and respective positions (see Wattenberg et al. [2016]). Hence, t-SNE cannot be expected to serve as a compositionality metric.

References

Jacob Andreas. Measuring Compositionality in Representation Learning. *International Conference on Learning Representations*, April 2019. URL http://arxiv.org/abs/1902.07181. arXiv: 1902.07181.

Jeffrey A. Barrett and Brian Skyrms. Self-assembling Games. *The British Journal for the Philosophy of Science*, 68(2):329–353, June 2017. ISSN 0007-0882, 1464-3537. doi: 10.1093/bjps/axv043. URL https://academic.oup.com/bjps/article-lookup/doi/10.1093/bjps/axv043.

John Batali. Computational simulations of the emergence of grammar. *Approach to the Evolution of Language*, pages 405–426, 1998.

Ben Bogin, Mor Geva, and Jonathan Berant. Emergence of Communication in an Interactive World with Consistent Speakers. *arXiv:1809.00549 [cs]*, September 2018. URL http://arxiv.org/abs/1809.00549. arXiv: 1809.00549.

Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. *arXiv:1808.10696 [cs]*, August 2018. URL http://arxiv.org/abs/1808.10696. arXiv: 1808.10696.

Henry Brighton and Simon Kirby. Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings. *Artificial Life*, 12(2):229–242, January 2006. ISSN 1064-5462, 1530-9185. doi: 10.1162/artl.2006.12.2.229. URL http://www.mitpressjournals.org/doi/10.1162/artl.2006.12.2.229.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and Generalization In Emergent Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.407. URL https://www.aclweb.org/anthology/2020.acl-main.407.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. Compositional Obverter Communication Learning From Raw Visual Input. *ICLR 2018*, April 2018. URL http://arxiv.org/abs/1804.02341. arXiv: 1804.02341.

Noam Chomsky. *Syntactic structures*. 1957. ISBN 978-1-61427-804-7. OCLC: 934673149.

Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. Emergence of Compositional Language with Deep Generational Transmission. *ArXiv*, abs/1904.09067, 2019.

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. *2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017*, March 2017. URL http://arxiv.org/abs/1703.06585. arXiv: 1703.06585.

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28(1-2):3–71, March 1988. ISSN 0010-0277.

Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. *NIPS'16 Proceedings of the 30th International Conference on Neural Information Processing Systems*, May 2016. URL http://arxiv.org/abs/1605.06676. arXiv: 1605.06676.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.

D. Fudenberg, J. Tirole, J.A. TIROLE, and MIT Press. *Game Theory*. Mit Press. MIT Press, 1991. ISBN 9780262061414. URL https://books.google.pl/books?id=pFPHKwXro3QC.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Shangmin Guo, Yi Ren, Agnieszka Słowik, and Kory Mathewson. Inductive bias and language expressivity in emergent communication. *arXiv preprint arXiv:2012.02875*, 2020.

Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. A very condensed survey and critique of multiagent deep reinforcement learning. In Amal El Fallah Seghrouchni, Gita Sukthankar, Bo An, and Neil Yorke-Smith, editors, *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, pages 2146–2148. International Foundation for Autonomous Agents and Multiagent Systems, 2020. URL `https://dl.acm.org/doi/abs/10.5555/3398761.3399105`.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Sy2fzU9gl`.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. *arXiv:1611.01144 [cs, stat]*, November 2016. URL `http://arxiv.org/abs/1611.01144`. arXiv: 1611.01144.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. *arXiv:1810.08647 [cs, stat]*, October 2018. URL `http://arxiv.org/abs/1810.08647`. arXiv: 1810.08647.

Yiding Jiang, Shixiang Gu, Kevin Murphy, and Chelsea Finn. Language as an abstraction for hierarchical deep reinforcement learning. *arXiv preprint arXiv:1906.07343*, 2019.

Łukasz Kaiser and Samy Bengio. Discrete autoencoders for sequence models. *arXiv preprint arXiv:1801.09797*, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

S. Kirby. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110, 2001. doi: 10.1109/4235.918430.

Tomasz Korbak, Julian Zubek, Łukasz Kuciński, Piotr Miłoś, and Joanna Raczaszek-Leonardi. Developmentally motivated emergence of compositional communication via template transfer. *arXiv preprint arXiv:1910.06079*, 2019.

Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. *arXiv:1706.08502 [cs]*, June 2017. URL `http://arxiv.org/abs/1706.08502`. arXiv: 1706.08502.

Nikolaus Kriegeskorte. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2008. ISSN 16625137. doi: 10.3389/neuro.06.004.2008. URL `http://journal.frontiersin.org/article/10.3389/neuro.06.004.2008/abstract`.

Łukasz Kuciński, Paweł Kołodziej, and Piotr Miłoś. Emergence of compositional language in communication through noisy channel. 2020. URL `https://larel-ws.github.io/accepted-papers/`.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building Machines That Learn and Think Like People. *arXiv:1604.00289 [cs, stat]*, April 2016. URL `http://arxiv.org/abs/1604.00289`. arXiv: 1604.00289.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. *arXiv:1612.07182 [cs]*, December 2016. URL http://arxiv.org/abs/1612.07182. arXiv: 1612.07182.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. *arXiv:1804.03984 [cs]*, April 2018. URL http://arxiv.org/abs/1804.03984. arXiv: 1804.03984.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. *Emergent Translation in Multi-Agent Communication*. 2017. _eprint: 1710.06922.

Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

David K. Lewis. *Convention: a philosophical study*. Blackwell, Oxford, nachdr. edition, 1969. ISBN 978-0-631-23257-5 978-0-631-23256-8. OCLC: 837747718.

Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems*, pages 15825–15835, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

Igor Mordatch and Pieter Abbeel. Emergence of Grounded Compositional Language in Multi-Agent Populations. In *AAAI*, 2017.

Martin A. Nowak and David C. Krakauer. The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14):8028–8033, 1999. ISSN 0027-8424. doi: 10.1073/pnas.96.14.8028. URL https://www.pnas.org/content/96/14/8028.

Michael Oliphant and John Batali. Learning and the Emergence of Coordinated Communication. *Center for Research on Language Newsletter*, 11, 1997.

Joanna Raczaszek-Leonardi, Iris Nomikou, Katharina J. Rohlfing, and Terrence W. Deacon. Language Development From an Ecological Perspective: Ecologically Valid Ways to Abstract Symbols. *Ecological Psychology*, 30(1):39–73, January 2018. ISSN 1040-7413, 1532-6969. doi: 10.1080/10407413.2017.1410387. URL https://www.tandfonline.com/doi/full/10.1080/10407413.2017.1410387.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model, 2020.

Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew M. Dai, and Kyunghyun Cho. Capacity, Bandwidth, and Compositionality in Emergent Language Learning. *arXiv:1910.11424 [cs, stat]*, April 2020. URL http://arxiv.org/abs/1910.11424. arXiv: 1910.11424.

Ruslan Salakhutdinov and Geoffrey Hinton. Semantic hashing. *International Journal of Approximate Reasoning*, 50(7):969–978, 2009.

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948.

Brian Skyrms. *Signals: evolution, learning, & information*. Oxford University Press, Oxford ; New York, 2010. ISBN 978-0-19-958082-8 978-0-19-958294-5. OCLC: ocn477256653.

Agnieszka Słowik, Abhinav Gupta, William L Hamilton, Mateja Jamnik, and Sean B Holden. Towards graph representation learning in emergent communication. *arXiv preprint arXiv:2001.09063*, 2020a.

Agnieszka Słowik, Abhinav Gupta, William L Hamilton, Mateja Jamnik, Sean B Holden, and Christopher Pal. Exploring structural inductive biases in emergent communication. *arXiv preprint arXiv:2002.01335*, 2020b.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. doi: 10.23915/distill.00002. URL `http://distill.pub/2016/misread-tsne`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] see Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] see Section 7.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] see Section 3.3 and Appendix F.
   (b) Did you include complete proofs of all theoretical results? [Yes] see Appendix F and Appendix E.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is included in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] see Appendix A and Appendix B.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We used 100 seeds. Confidence intervals are reported in figures and tables in the paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Appendix B.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] see Section 4.1 and Appendix A.
   (b) Did you mention the license of the assets? [Yes] We use MIT license, see Appendix
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]