Emergence of compositional language in communication through noisy channel

Łukasz Kuciński [1] **Paweł Kołodziej** [1] **Piotr Miłoś** [1 2]

Abstract

In this paper, we investigate how communication through a noisy channel can lead to the emergence of compositional language. Our approach is end-to-end, allows for different inductive biases on the agents' architecture, and trains without periodical resets of the networks' weights. This relaxes some of the assumptions in recently developed methods. The impact on the structure of the resulting language is shown in the context of signaling games. We also develop a new metric for measuring degree of compositionality.

1. Introduction

Communication can emerge in situations that require coordination and information sharing to achieve a joint objective. This is common in multi-agent systems with partial observation Foerster et al. (2016); Lazaridou et al. (2016); Jaques et al. (2018); Rączaszek-Leonardi et al. (2018). Communication is compositional if complex signals can be represented as a combination of their constituents. It is a feature of human languages, and it facilitates generalization, productivity, and knowledge sharing. As such, it is considered essential for general intelligence (Lake et al. (2016)).

Noise is ubiquitous in non-digital communication. For example, in biology, it has profound importance being the driving force of evolution. It is also present in human communication (Rothwell (1999)) and is perhaps equally significant, if, indeed, it promotes compositionality. Our main contribution is showing experimentally that this is the case in a simple communication model: signaling games. We argue that this effect might follow from the fact that a compositional language is (partially) robust to the mistakes caused by a noisy channel.

Signaling games (Lewis (1969); Skyrms (2010); Lazaridou et al. (2016); Fudenberg et al. (1991)) are popular model of communication. In such a game, there is a sender and a receiver. The former observes a state of the world and sends a message to the latter. Upon receiving a message, the receiver performs an action, which results in a reward or a loss. Typically, both the agents are rewarded if the receiver can identify the state of the world communicated by the sender. Compositionality rarely emerges in standard versions of this model.

While signaling games are inherently a (multi-agent) reinforcement learning setup, in the simplest case, it can be represented as a supervised learning problem. More precisely, it can be viewed as a classification task, with the underlying architecture resembling a discrete auto-encoder: the encoder acts as a sender, the bottleneck corresponds to a communication channel, and a decoder plays the role of a receiver. Without additional constraints, good accuracy can be achieved with non-compositional communication protocol - it is enough that the sender distinguishes the features of the observed state in his messages. It is likely that such a mapping will show poor generalization to novel combinations of features (Kottur et al., 2017).

Our experiments were conducted using a set of images with two features: shape and color (in a similar vein to Choi et al. (2018), Bogin et al. (2018), and Korbak et al. (2019)). We obtained some promising results, which demonstrate that introducing a noisy channel to the system, indeed creates enough tension for compositionality to emerge. Additionally, we show that the noisy channel can be modeled as a simple dense layer, which takes probabilities as input, has a fixed weight matrix of a certain form, and uses a logarithmic activation function. This constitutes a small change to the supervised learning pipeline and is interesting in its own right (it can be viewed as a yet another regularizing layer). In particular, this does not require template transfer (Korbak et al. (2019)), periodical resets of the agents' weights or memory (Li & Bowling (2019), Kottur et al. (2017), Das et al. (2017)). The approach is also agnostic to the choice of neural network architectures for a receiver and sender. This makes less assumptions about the cognitive abilities of the agents, when compared to other methods (e.g. in the obverter algorithm it is assumed that an agent can use its own responses to messages to predict other agent's responses, see Batali (1998), Choi et al. (2018)).

Measuring compositionality is quite challenging. We introduce a new, *conflict counting*, measure and use it alongside

[1]Institute of Mathematics of the Polish Academy of Sciences, Warsaw, Poland [2]deepsense.ai, Warsaw, Poland. Correspondence to: Łukasz Kuciński <lkucinski@impan.pl>.

Accepted to the 1st Workshop on Language in Reinforcement Learning, ICML 2020, Copyright 2020 by the author(s).

topographical similarity, a metric commonly used in the literature (see e.g. Brighton & Kirby (2006); Lazaridou et al. (2018) and Kriegeskorte (2008); Bouchacourt & Baroni (2018), where it is called a *representational similarity*). We provide a discussion concerning these measures and use them to analyze the relationship between noise level and compositionality of emergent language.

The rest of the paper is organized as follows. In the next section, we discuss related works. In Section 3, we present details of our method. Section 4 is devoted to experimental results. We discuss future work and conclude the paper in Section 5. The neural network architecture and the choice of hyperparameters can be found in Appendix A. For comparison of the aforementioned compositionality measures, see Appendix B. Finally, theoretical results concerning optimality of compositional communication over a noisy channel, were placed in Appendix C.

2. Related work

Signaling games (Lewis (1969); Skyrms (2010); Lazaridou et al. (2016); Fudenberg et al. (1991)) are popular models of communication and have traditionally been solved using either simple reinforcement learning (Skyrms, 2010) or evolutionary optimization (Cangelosi, 2001; Grouchy et al., 2016). Early work on the subject utilizing neural networks is (Rumelhart et al., 1986), while the recent results taking advantage of deep learning progress Lazaridou et al. (2018), Bouchacourt & Baroni (2018). Kottur et al. (2017) argue that a strong inductive bias is necessary for the emergence of compositionality between communicating agents. Das et al. (2017) places pressure on agents, to use symbols consistently across varying contexts, by a frequent reset of the agent's memory. The topic of communication is inherently interesting in the context of a multi-agent RL system, see Hernandez-Leal et al. (2020, Table 2) for a recent survey.

A psychologically driven approach was proposed by Choi et al. (2018) and Bogin et al. (2018), who build upon the obverter algorithm (Oliphant & Batali (1997), Batali (1998)). This algorithm assumes that an agent can use its own policy as a model to predict the response of the other agent. This approach could be connected with *theory of mind* (Premack & Woodruff (1978)) and its variant, *simulation theory* (Gordon (1986), Heal (1986)). The obverter algorithm has several limitations (the agents must share an identical architecture and the task must be symmetric) and Korbak et al. (2019) used an alternative approach, based on the idea of template transfer (Barrett & Skyrms (2017)).

A different take on compositionality creates a bias towards protocols that are easy to teach to new agents (Kirby (2001); Kirby et al. (2008), Brighton (2002)). In the machine learning literature, this idea was explored by Li & Bowling (2019) and Cogswell et al. (2019). In a similar spirit, De Beule & Bergen (2006) gradually increased task complexity that incentivized the reuse of existing patterns of communication.

The notion of compositional language is also related to the idea of learning disentangled representations from high-dimensional inputs, see e.g. Higgins et al. (2017), Locatello et al. (2018), Kim & Mnih (2018).

The use of noise as a regularizer is a powerful concept in deep learning, see e.g. dropout (Srivastava et al. (2014)) or semantic hashing (Salakhutdinov & Hinton (2009)). The latter was used in discrete autoencoders (Kaiser & Bengio (2018)) as a mechanism allowing backpropagation through a discrete latent. A similar idea was also applied in the context of learning to communicate (Foerster et al. (2016)). The authors used noise as a mechanism to backpropagate through a discrete communication channel, and observed that it is essential for successful training. They also investigated the channel's capacity. In the context of compositionality, Li & Bowling (2019) used noise to promote the emergence of this phenomenon, by resetting the weights of the agents' neural networks. In this work, we use the noisy channel exclusively, as a mechanism to encourage compositionality. The backpropagation through a discrete communication channel is done differently, by the use of a Gumbel softmax.

3. Method

Training pipeline Consider a sender s_θ, modeled as a neural network, which observes an image $\texttt{img}$ from some dataset $\mathcal{D}$. We assume that each element of $\mathcal{D}$ has K independent features $f_1, \ldots, f_K$ (here we consider $K = 2$). The sender sends a message comprising of L symbols (here we assume $L = K$). Basing on these symbols, the receiver has to guess the values for each of K features. Both the features and symbols are discrete and enumerated with $A_f = \{1, \ldots, d_f\}$ and $A_s = \{1, \ldots, d_s\}$, respectively.

Formally, $s_\theta(\texttt{img}) = (s^i_\theta(\texttt{img}))_{i=1}^K$, where $s^i_\theta(\texttt{img}) = (s^i_{j,\theta}(\texttt{img}))_{j=1}^{d_s} \in \mathcal{P}(A_s)$[1] represents the probability distribution corresponding to the ith symbol. Define a function $\texttt{noise}\colon \mathcal{P}(A_s) \to \mathbb{R}^{d_s}$ as follows:

$$\texttt{noise}(x) = \log(Wx), \tag{1}$$

where $W \in \mathbb{R}^{d_s \times d_s}$ is a fixed matrix, such that $Wx > 0$ and $Wx \in \mathcal{P}(A_s)$, for any $x \in \mathcal{P}(A_s)$[2]. The second condition on W is satisfied, for instance, by a family of stochastic matrices; several examples are also given at the end of this section. Let $\widehat{s}^i_\theta(\texttt{img})$ denotes the distribution of ith symbol which passes through the noisy channel:

$$\widehat{s}^i_\theta(\texttt{img}) = \texttt{noise}(s^i_\theta(\texttt{img})).$$

Suppose further that $g^i = (g^i_1, \ldots, g^i_{d_s})$ is a vector of i.i.d.

[1] $\mathcal{P}(A) = \{p \in \mathbb{R}^{|A|} : p_i \geq 0, \sum_{i \in A} p_i = 1\}$.

[2] We could also define $\texttt{noise}$ for all $x \in \mathbb{R}^m$, for some m, by first applying $\texttt{softmax}$ to x, and then using equation (1).

Gumbel$(0, 1)$ random variables and define the following functions:

$$\texttt{gumbel_sample}(x; g) = \arg\max_i (\log(x_i) + g_i),$$

$$\texttt{gumbel_softmax}(x; \tau, g)_i = \frac{\exp((\log(x_i) + g_i)/\tau)}{\sum_{j=1}^{k} \exp((\log(x_j) + g_j)/\tau)}.$$

Define

$$\widehat{\mathfrak{m}}_i = \texttt{gumbel_softmax}(\widehat{s}^i_\theta(\texttt{img}); \tau, g^i) \in \mathbb{R}^{d_s}.$$

The receiver neural network is defined as $r_\psi(\mathfrak{m}) = (r^i_\psi(\mathfrak{m}))_{i=1}^K$, where $r^i_\psi(\mathfrak{m}) = (r^i_{j,\psi}(\mathfrak{m}))_{j=1}^d \in \mathcal{P}(A_f)$ represents the probability distribution on A_f, corresponding to the ith feature.

In the Straight-Through mode (see Jang et al. (2016)), r_ψ takes $\widehat{\mathfrak{m}}$ as input half of the time, and the remaining half of the time, it takes $\widetilde{\mathfrak{m}}$. Here

$$\widetilde{\mathfrak{m}} = \texttt{stop_gradient}(\widehat{\mathfrak{m}} - \overline{\mathfrak{m}}) + \overline{\mathfrak{m}},$$
$$\overline{\mathfrak{m}}_i = \texttt{one_hot}(\widehat{m}_i) \in \mathbb{R}^d,$$
$$\widehat{m}_i = \texttt{gumbel_sample}(s^i_\theta(\texttt{img}); g^i) \in A,$$

i.e. $\widehat{m} = (\widehat{m}_i)_{i=1}^K$ is a sampled noisy message. The neural networks are trained using

$$\mathcal{L} = \mathcal{L}_{xent} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{l_2}\mathcal{L}_{l_2}.$$

The cross-entropy loss is defined as

$$\mathcal{L}_{xent} = \mathbb{E}_{(\texttt{img}, f_1, \ldots, f_K) \sim \mathcal{D}} \left[\sum_{i=1}^{K} \log r_{f_i,\psi}(\widetilde{\mathfrak{m}}(\texttt{img})) \right].$$

Furthermore, $\mathcal{L}_{KL} = \mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{i=1}^K \mathrm{KL}(U(A_s) || s^i_\theta(x)) \right]$ and $\mathcal{L}_{l_2} = ||\theta||_2 + ||\psi||_2$.

Compositionality measures In this paper we, use two compositionality measures: topographical similarity (see e.g. Brighton & Kirby (2006), Lazaridou et al. (2018), Kriegeskorte (2008), Bouchacourt & Baroni (2018)), and present a new, conflict counting, measure described below and defined in (2). For more discussion on the topic, see also Section 4 and Appendix B.

We assumed $L = K$, the number of symbols generated by the sender is equal to the number of features. In such a setting, we expect that compositional language will use one symbol for each feature. To simplify we also assume that the permutation of the position of symbols to the position of features, $\phi : \{1, \ldots, L\} \mapsto \{1, \ldots, K\}$, is fixed.

We say that the principal meaning of a symbol s at position j assuming ϕ is $\mathtt{m}(s, j; \phi) := \arg\max_f \texttt{count}(s, j, f; \phi)$, where

$$\texttt{count}(s, j, f; \phi) = \sum_{\texttt{img} \in \mathcal{D}} \mathbf{1}(s_{\texttt{img},j} = s, f_{\texttt{img},\phi(j)} = f),$$

and ties in $\arg\max$ are broken arbitrarily. Our metric is defined as:

$$\texttt{conflicts} = \min_\phi \sum_{s,j} \texttt{score}(s, j; \phi), \tag{2}$$

where $\texttt{score}(s, j; \phi) = \sum_{f \neq \mathtt{m}(s,j;\phi)} \texttt{count}(s, j, f; \phi)$. Intuitively, `score` measures how many times the feature assigned to a symbol s at a position j diverts from its principal meaning $\mathtt{m}(s, j; \phi)$. `conflicts` totals these errors and takes $\min$ over possible orderings ϕ.

Noise In this paragraph, we show how to define different noise channels using the `noise` layer, as defined in (1). Consider a probability distribution $p \in \mathcal{P}(d_s)$. Then Wp can represent a diverse set of distributions[3].

A uniform noise applied to p manifests itself as a mixture distribution of p and $U\{1, \ldots, d_s\}$, with a density equal to $(1 - \varepsilon)p + \frac{\varepsilon}{d_s}$, where ε is the probability of randomly scrambling the symbol. The logits of such a distribution can be represented as $\texttt{noise}(p)$, where W, defined in (1), can take the following form: $W_{ii} = 1 - \varepsilon(1 - 1/d_s)$ and $W_{ij} = \varepsilon/d_s$ for $i \neq j$. Here, the probability of changing a symbol equals $\varepsilon(1 - 1/d_s)$. To make it equal to ε, we define W as

$$W_{ij} = \begin{cases} 1 - \varepsilon, & i = j, \\ \frac{\varepsilon}{d_s - 1}, & i \neq j. \end{cases} \tag{3}$$

The formula given in (3) is the one we use in this paper. Similarly, a formula for a bit-flipping noise can be obtained[4].

4. Experiments

Experiments setup For our experiments, we chose a subset of a dataset used by Choi et al. (2018)[5] and Korbak et al. (2019), comprised of four shapes (box, cylinder, ellipsoid, sphere) in four colors (blue, cyan, gray, green), see Figure 1. Each image has dimensions of $128 \times 128 \times 3$ and there are 100 images for each shape–color pair. The sender and the receiver are implemented as feed-forward neural networks. For details concerning architecture and the choice of hyperparameters, see Appendix A.

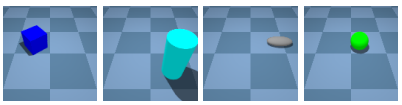

Figure 1. Samples from the dataset used in this paper.

The compositionality of emerging languages was measured on the training set as our dataset was not big enough to

[3]For instance if $p_i > 0$ and $g_i > 0$ for all i, then $Wp = g$, where $W_{ii} = g_i/p_i$ and $W_{ij} = 0$, for $i \neq j$.

[4]$W_{ij} = Y_{i \oplus j}$, where $\oplus$ is a `xor` operation, $Y_1 = 1 - \varepsilon$ and $Y_i = \varepsilon \mathbf{1}(i - 1 \text{ is a power of } 2)/\lfloor \log_2(d) \rfloor$.

[5]The dataset is available at https://github.com/benbogin/obverter.

ensure generalization (the reconstruction accuracy was low). We conjecture that this is orthogonal to compositionality and we plan to investigate this in issue in the future work.

Tight alphabet In this experiment, we trained a language with the alphabet as tight as possible. Clearly, four symbols are enough to describe four shapes in four colors. However, we found out that it is quite challenging to learn in this setup. Consequently, we loosened the constraints on the alphabet size and allowed the sender to use five symbols. This allowed the model to learn, which could be attributed to higher model capacity or better ability to escape local minima during training. An example of a language learned by the sender is presented in Table 1. Although the alphabet contains five symbols, only four of them are used to describe shape and a different four to describe color. We did not use pre-training in this experiment to demonstrate the ability of our method to work in the end-to-end setup.

Table 1. Language learned by the sender in end-to-end training. $|A_s| = 5$, $\varepsilon = 0.1$. Here $5/10$ experiments have `conflicts` $= 0$ (one of them is presented). Four symbols were used to denote color and four for shape. This language is fully compositional.

SHAPE COLOR	BOX	CYLINDER	ELLIPSOID	SPHERE
BLUE	0,1	3,1	1,1	4,1
CYAN	0,3	3,3	1,3	4,3
GRAY	0,2	3,2	1,2	4,2
GREEN	0,4	3,4	1,4	4,4

Abundant alphabet In this experiment, we trained a model using alphabet significantly bigger than necessary: $|A_s| = 7$. In half of the experiments, the sender learned a language that is compositional according to our metrics (with zero or one conflicts). The language that has the highest number of conflicts is presented in Table 2. Despite poor values of the metrics (9 conflicts and topographical similarity of 0.68) it can be argued that the language is also compositional. It may be seen from Table 2, that in the first two columns, the messages start with a color, followed by a shape. In the remaining two columns, this order is reversed. It is possible to achieve high accuracy (0.95), because symbols used in the first two columns are almost never used in the same position in the remaining columns (with the exception of a symbol 2).

Table 2. Language learned by the sender (pre-trained convnets). $|A_s| = 7$, $\varepsilon = 0.1$.

SHAPE COLOR	BOX	CYLINDER	ELLIPSOID	SPHERE
BLUE	2,3	2,5	3,1	2,1
CYAN	5,4	5,5	3,6	1,6
GRAY	6,4	6,5	3,2	1,2
GREEN	4,4	4,5	3,0	1,0

Noise vs metrics We examined the relationship between noise level (in the noisy channel as defined in (3)) and compositionality of emergent language. For each selected noise level we run a batch of experiments with 10 different seeds, and we measured compositionality using topographical similarity and conflicts count. The experiments were performed with the same pre-trained convnets. The results are presented in Figure 2. The main takeaway is that compositionality does not emerge without, or with excessive levels, of noise, see also Appendix C for theoretical justification. As a result, there is a sweet spot for noise, located in quite a narrow window $\varepsilon \in [0.08, 0.15]$, which results in compositional languages.

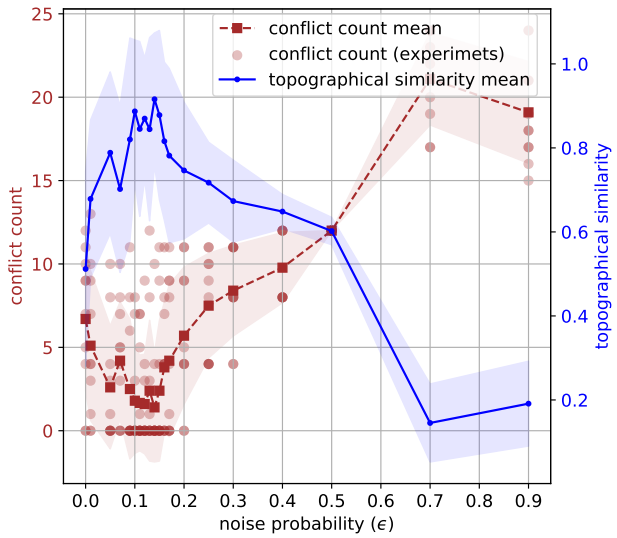

Figure 2. Compositionality metrics as functions of the channel's noise. Ten experiments were run for each noise probability. The shaded area represents standard deviation.

5. Conclusions and future work

We presented a promising method to achieve emergence of compositional language in communication between two agents. The key insight was to leverage randomness by introducing a noisy communication channel, which can be implemented as a regularization layer. We tested the approach on a version of a signaling game, where a sender observes an image, communicates with the receiver, and the receiver has to guess the image's features.

There remain many directions for future work. Two important ones, concern improving the results on the held-out dataset, and incorporating a reinforcement learning objective. A natural further extension would be to apply the method on more complex datasets, with greater number of features and their possible range of values (e.g. by using CLEVR dataset). It is also interesting to study how introducing a noisy layer can improve performance of neural networks in other contexts.

Acknowledgments

Special thanks to Joanna Rączaszek-Leonardi, Julian Zubek and Tomasz Korbak. This research was supported by the PL-Grid Infrastructure. We extensively used the Prometheus supercomputer, located in the Academic Computer Center Cyfronet in the AGH University of Science and Technology in Kraków, Poland. The work of Piotr Miłoś was supported by the Polish National Science Center grants UMO-2017/26/E/ST6/00622. We managed our experiments using https://neptune.ai. We would like to thank the Neptune team for providing us access to the team version and technical support.

References

Barrett, J. A. and Skyrms, B. 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.

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

Bogin, B., Geva, M., and Berant, J. 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.

Bouchacourt, D. and Baroni, M. 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.

Brighton, H. Compositional Syntax From Cultural Transmission. *Artificial Life*, 8(1):25–54, January 2002. ISSN 1064-5462, 1530-9185. doi: 10.1162/106454602753694756. URL http://www.mitpressjournals.org/doi/10.1162/106454602753694756.

Brighton, H. and Kirby, S. 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.

Cangelosi, A. Evolution of communication and language using signals, symbols, and words. *IEEE Transactions on Evolutionary Computation*, 5(2):93–101, April 2001. doi: 10.1109/4235.918429.

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

Cogswell, M., Lu, J., Lee, S., Parikh, D., and Batra, D. Emergence of Compositional Language with Deep Generational Transmission. *ArXiv*, abs/1904.09067, 2019.

Das, A., Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. 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.

De Beule, J. and Bergen, B. K. On the emergence of compositionality. In *The Evolution of Language*, pp. 35–42, Rome, Italy, March 2006. World scientific. ISBN 978-981-256-656-0 978-981-277-426-2. doi: 10.1142/9789812774262_0005. URL http://www.worldscientific.com/doi/abs/10.1142/9789812774262_0005.

Foerster, J. N., Assael, Y. M., de Freitas, N., and Whiteson, S. 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.

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

Gordon, R. M. Folk Psychology as Simulation. *Mind & Language*, 1(2):158–171, 1986. ISSN 1468-0017. doi: 10.1111/j.1468-0017.1986.tb00324.x.

Grouchy, P., D'Eleuterio, G. M. T., Christiansen, M. H., and Lipson, H. On The Evolutionary Origin of Symbolic Communication. *Scientific Reports*, 6:34615, October 2016. URL https://doi.org/10.1038/srep34615.

Heal, J. Replication and Functionalism. In Butterfield, J. (ed.), *Language, Mind, and Logic*, pp. 135–150. Cambridge University Press, 1986.

Hernandez-Leal, P., Kartal, B., and Taylor, M. E. A very condensed survey and critique of multiagent deep reinforcement learning. In Seghrouchni, A. E. F., Sukthankar, G., An, B., and Yorke-Smith, N. (eds.), *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems, AAMAS '20, Auckland, New Zealand, May 9-13, 2020*, pp. 2146–2148. International Foundation for Autonomous Agents and Multiagent Systems,

2020. URL https://dl.acm.org/doi/abs/10.5555/3398761.3399105.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. 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.

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

Jaques, N., Lazaridou, A., Hughes, E., Gulcehre, C., Ortega, P. A., Strouse, D. J., Leibo, J. Z., and de Freitas, N. 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.

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

Kim, H. and Mnih, A. Disentangling by factorising, 2018.

Kirby, S. 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, April 2001. doi: 10.1109/4235.918430.

Kirby, S., Cornish, H., and Smith, K. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686, August 2008. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0707835105. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0707835105.

Korbak, T., Zubek, J., Kuciński, Ł., Miłoś, P., and Rączaszek-Leonardi, J. Developmentally motivated emergence of compositional communication via template transfer. *arXiv preprint arXiv:1910.06079*, 2019.

Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. 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.

Kriegeskorte, N. 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.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. 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.

Lazaridou, A., Peysakhovich, A., and Baroni, M. 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.

Lazaridou, A., Hermann, K. M., Tuyls, K., and Clark, S. 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.

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

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

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations, 2018.

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

Premack, D. and Woodruff, G. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, December 1978. ISSN 1469-1825, 0140-525X. doi: 10.1017/S0140525X00076512.

Rothwell, J. *In the Company of Others: An Introduction to Communication*. McGraw-Hill Higher Education, 1999. ISBN 9781559347389. URL https://books.google.pl/books?id=VtPZAAAAMAAJ.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 0028-0836, 1476-4687. doi: 10.1038/323533a0. URL http://www.nature.com/articles/323533a0.

Rączaszek-Leonardi, J., Nomikou, I., Rohlfing, K. J., and Deacon, T. W. 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.

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

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

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

A. Experimental setup

The architecture of a neural network used in this paper is shown in Figure 3. It consists of three main parts: the sender, the receiver and the noisy discrete channel between them.

The sender network consists of a vision and encoding module. The former consists of two convolutional layers (each having 8 filters, kernel 3×3, stride 1, and elu activation function). After each convolutional layer, there is a 2×2 max pool layer with stride 2. The output of the last max pool layer is passed through two dense layers (with 64 neurons and elu activation) and a linear classifier with softmax for each symbol.

The noisy channel layer consists of a dense layer with $|A_s|$ neurons, a fixed weights matrix, and a $\log$ activation function. This is followed by a Gumbel softmax layer.

The receiver takes two one-hot encoded symbols as input and concatenates them to obtain one input vector s. Consequently, s and $1 - s$ are passed to dense layers. The sum of those layers is processed through the elu activation function and two dense layers (similarly to Kaiser & Bengio (2018)). Finally, there are two linear classifiers - one for shape and one for color and softmax layer. Each dense layer in the receiver has a width of 64.

For training, we used $\lambda_{KL} = 0.01$, $\lambda_{l_2} = 0.001$, an Adam optimizer with learning rate 0.001, and a batch size of 64.

To speed up training, we used a pre-trained convnets and the first dense layer of the sender. During the pre-training, only part of the neural network was trained. Linear classifier with 16 classes was added after the first dense layer of the sender. This classifier was then trained to recognize classes of images from our dataset (in one class all images present the same shape and color). In the final experiments, the full model was used with the pre-trained weights, which were frozen during the training.

We used pretraining for most of our experiments. It was used in the experiment with an abundant alphabet presented in Table 2 and while exploring the relationship between compositionality and noise presented in Figure 2. However, it's not strictly necessary and language presented in Table 1 was learned in end-to-end setup (without pre-training).

B. Metrics

Conflict count and topographical similarity are two measures of compositionality. Figure 4 presents a scatter plot of them. Despite a strong correlation, there is also a difference between the two, in the way synonyms are treated. For example, Table 3 presents a language which has a conflict count equal to 0 and topographical similarity 0.87. This language uses synonyms for the gray color (symbols 0 and 1) and for an ellipsoid shape (symbols 0 and 4). In particular, there is no conflict of meaning. This is possible because the alphabet is slightly larger than the space of described features (5 symbols and only 4 features). Topographical similarity penalizes synonyms (because similar meanings are described by distinct symbols) while conflict count allows for them. We can see from the bottom part of table 3, that the receiver learned to report color as grey when the symbol on the second position is 0 or 1, regardless of the shape. The same happens with an ellipsoid shape, which is reported when there is 0 or 4 in the first position, regardless of the color.

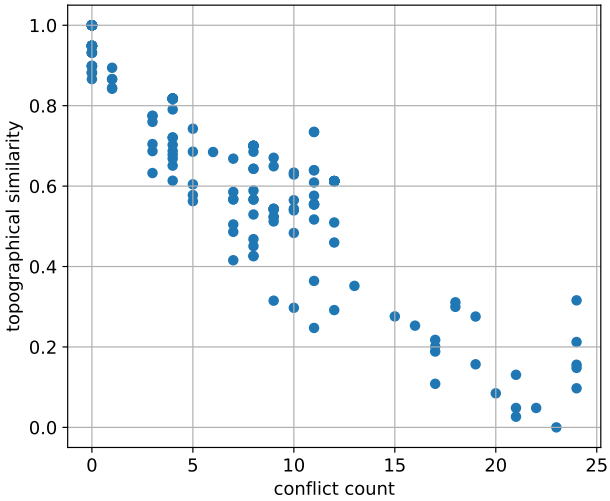

Figure 4. A scatter plot of topographical similarity and conflicts count. Linear correlation between them is -0.93.

Table 3. Language learned by the sender (top) and the receiver (bottom). $|A_s| = 5$, $\varepsilon = 0.1$. Here $\texttt{conflicts} = 0, \texttt{topo} = 0.88$.

SHAPE COLOR	BOX	CYLINDER	ELLIPSOID	SPHERE
BLUE	2,2	3,2	4,2	1,2
CYAN	2,4	3,4	4,4	1,4
GRAY	2,0	3,1	0,1	1,1
GREEN	2,3	3,3	0,3	1,3

POS1 POS0	0	1	2	3	4
0	ELL,GRA	ELL,GRA	ELL,BLU	ELL,GRE	ELL,CYA
1	SPH,GRA	SPH,GRA	SPH,BLU	SPH,GRE	SPH,CYA
2	BOX,GRA	BOX,GRA	BOX,BLU	BOX,GRE	BOX,CYA
3	CYL,GRA	CYL,GRA	CYL,BLU	CYL,GRE	CYL,CYA
4	ELL,GRA	ELL,GRA	ELL,BLU	ELL,GRE	ELL,CYA

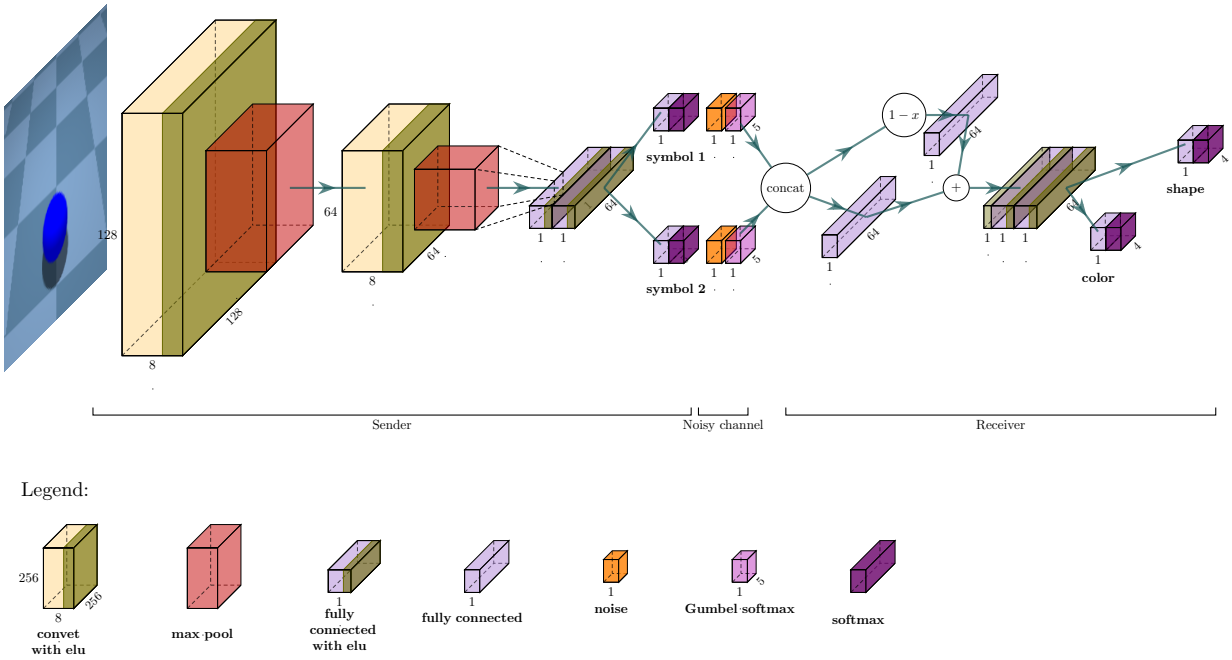

Figure 3. The architecture of the network. Here $|A_s| = 5$.

C. Optimality of compositional communication

Suppose that there are $K > 1$ features, and feature f_k takes values in a set $\mathcal{F}_k$. For simplicity, we assume that the sets $\mathcal{F}_1, \dots, \mathcal{F}_K$ are mutually disjoint, and $|\mathcal{F}_k| = F = |A_s| = d_s$, where $A_s = \{1, \dots, d_s\}$. This means that $d_f = KF$. Denote $\mathcal{F} = \prod_{i=1}^K \mathcal{F}_i$. A language is a mapping from the feature space to the set of K-symbol messages, i.e. $l \colon \mathcal{F} \to A_s^K$.

For any K dimensional vectors $\mathbf{v}, \mathbf{w}$, define $\rho(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^K \mathbf{1}(v_i = w_i)$. Let $\mathbf{s} = (s_1, \dots, s_K) \in A_s^K$ be a message. Below we assume that a noisy message, $\mathbf{s}' = (s'_1, \dots, s'_K)$, that is produced by the noisy channel, has the following conditional distribution:

$$\mathbb{P}(\mathbf{s}' = \hat{\mathbf{s}}|\mathbf{s}) = (1-\varepsilon)^{\rho(\hat{\mathbf{s}},\mathbf{s})} \varepsilon^{K-\rho(\hat{\mathbf{s}},\mathbf{s})} \left(\frac{1}{d_s - 1}\right)^{K-\rho(\hat{\mathbf{s}},\mathbf{s})},$$

for any $\hat{\mathbf{s}} \in A_s^K$, and a fixed $\varepsilon \in (0, 1)$.

Suppose that l is a one-to-one mapping, and that the sender uses l to generate messages, while the receiver uses l^{-1} to decode them. For any given symbol $\mathbf{s}$, and a corrupted message $\mathbf{s}'$, the reward they get is the amount of correctly encoded features:

$$R(\mathbf{s}, \mathbf{s}', l) = \rho(l^{-1}(\mathbf{s}), l^{-1}(\mathbf{s}')).$$

Lemma 1. *Assume that $F = d_s \geq 2$ and $\varepsilon < (d-1)/d$. Then for any distribution $\mu \in \mathcal{P}(A_s^K)$, a compositional language is optimal, in the sense that it maximizes the expected reward*

$$\max_l \mathbb{E}_{\mathbf{s}\sim\mu}[R(\mathbf{s}, \mathbf{s}', l)] = K(1-\varepsilon),$$

where the max *is taken over all one-to-one mappings l.*

Proof. We start by observing that

$$\mathbb{E}_{\mathbf{s}\sim U(A_s^K)}[R(\mathbf{s}, \mathbf{s}', l)] = \mathbb{E}_{\mathbf{s}\sim U(A_s^K)}\left[\mathbb{E}\left[\rho(l^{-1}(\mathbf{s}), l^{-1}(\mathbf{s}')|\mathbf{s}\right]\right]$$
$$= \mathbb{E}_{\mathbf{s}\sim U(A_s^K)}\left[\sum_{k=0}^K k\mathbb{P}\left(\rho(l^{-1}(\mathbf{s}), l^{-1}(\mathbf{s}')) = k|\mathbf{s}\right)\right].$$

Define $\Lambda_k(\mathbf{s}) = \{\alpha \in \mathcal{F} : \rho(\alpha, l^{-1}(\mathbf{s})) = k\}$ and $\Lambda_{k,j}(\mathbf{s}) = \{\alpha \in \mathcal{F} : \rho(\alpha, l^{-1}(\mathbf{s})) = k, \rho(\mathbf{s}, l(\alpha)) = j\}$. Then

$$\mathbb{P}\left(\rho(l^{-1}(\mathbf{s}), l^{-1}(\mathbf{s}')) = k|\mathbf{s}\right) = \sum_{\alpha\in\Lambda_k(\mathbf{s})} \mathbb{P}(\mathbf{s}' = l(\alpha)|\mathbf{s})$$
$$= \sum_{j=1}^K \sum_{\alpha\in\Lambda_{k,j}(\mathbf{s})} (1-\varepsilon)^j \varepsilon^{K-j} \left(\frac{1}{d_s-1}\right)^{K-j}$$
$$= \left(\frac{\varepsilon}{d_s-1}\right)^K \sum_{j=1}^K \sum_{\alpha\in\Lambda_{k,j}(\mathbf{s})} \left(\frac{1-\varepsilon}{\varepsilon}(d_s-1)\right)^j.$$

Suppose, that there exists $k \neq j$, such that $\Lambda_{k,j}(\mathbf{s}) \neq \emptyset$. Since, l is a one-to-one mapping and $F = d_s$, there exists a sequence of distinct numbers $k_1, k_2, \ldots, k_m$, where $m \leq F$, such that $\Lambda_{k_i,k_{i+1}}(\mathbf{s}) \neq \emptyset$, for $i \leq m$ and $k_{m+1} = k_1$. To see why, assume that such a k and j exist. Then either $\Lambda_{j,k} \neq \emptyset$ and we are done ($m = 1$), or $\Lambda_{j,k} = \emptyset$ and there exists $j_1 \notin \{j, k\}$ such that $\Lambda_{j,j_1} \neq \emptyset$. After repeating this argument at most F times, we will find the desired sequence $k_1, \ldots, k_m$.

Let $\alpha_{i,i+1} \in \Lambda_{k_i,k_{i+1}}(\mathbf{s})$ and define $\hat{l}$ which is equal to l, except for $\alpha_{1,2}, \ldots, \alpha_{m,m+1}$, where

$$\hat{l}(\alpha_{i,i+1}) = l(\alpha_{i-1,i}),$$

and $\alpha_{0,1} = \alpha_{m,m+1}$. Then

$$\begin{aligned}
&\sum_{k=1}^{K} k\mathbb{P}\left(\rho(\hat{l}^{-1}(\mathbf{s}), \hat{l}^{-1}(\mathbf{s}')) = k|\mathbf{s}\right) \\
&- \sum_{k=1}^{K} k\mathbb{P}\left(\rho(l^{-1}(\mathbf{s}), l^{-1}(\mathbf{s}')) = k|\mathbf{s}\right) \\
&= \left(\frac{\varepsilon}{d_s - 1}\right)^K \sum_{i=1}^{m} \Big[k_i \left(\frac{1-\varepsilon}{\varepsilon}(d_s - 1)\right)^{k_i} \\
&- k_i \left(\frac{1-\varepsilon}{\varepsilon}(d_s - 1)\right)^{k_{i+1}} \Big] > 0,
\end{aligned}$$

where the last line follows from the rearrangement inequality and the fact that $\frac{1-\varepsilon}{\varepsilon}(d_s - 1) > 1$ (by our assumption). This means that until there exists $k \neq j$ such that $\Lambda_{k,j}(\mathbf{s}) \neq \emptyset$, we can improve l. So suppose that $\Lambda_{k,j}(\mathbf{s}) = \emptyset$ for any $k \neq j$. Then

$$\begin{aligned}
&\mathbb{P}\left(\rho(l^{-1}(\mathbf{s}), l^{-1}(\mathbf{s}')) = k|\mathbf{s}\right) \\
&= \left(\frac{\varepsilon}{d_s - 1}\right)^K \left(\frac{1-\varepsilon}{\varepsilon}(d_s - 1)\right)^k |\Lambda_{k,k}| \\
&= \left(\varepsilon \frac{F-1}{d_s - 1}\right)^K \left(\frac{1-\varepsilon}{\varepsilon} \frac{d_s - 1}{F - 1}\right)^k \binom{K}{k},
\end{aligned}$$

and consequently,

$$\max_l \mathbb{E}_{\mathbf{s} \sim U(A_s^K)}[R(\mathbf{s}, \mathbf{s}', l)] = K(1 - \varepsilon).$$

Since a compositional language satisfies $\Lambda_{k,j}(\mathbf{s}) = \emptyset$ for any $k \neq j$, it is optimal. $\square$