Emergent Language Generalization and Acquisition Speed are not tied to Compositionality

Eugene Kharitonov
Facebook AI
kharitinov@fb.com

Marco Baroni
Facebook AI, ICREA
mbaroni@fb.com

Abstract

Studies of discrete languages emerging when neural agents communicate to solve a joint task often look for evidence of compositional structure. This stems for the expectation that such a structure would allow languages to be acquired faster by the agents and enable them to generalize better. We argue that these beneficial properties are only loosely connected to compositionality. In two experiments, we demonstrate that, depending on the task, non-compositional languages might show equal, or better, generalization performance and acquisition speed than compositional ones. Further research in the area should be clearer about what benefits are expected from compositionality, and how the latter would lead to them.

1 Introduction

There is a recent spike of interest in studying the languages that emerge when artificial neural agents communicate to solve a common task (Foerster et al., 2016; Lazaridou et al., 2016; Havrylov and Titov, 2017). A good portion of such studies looks for traces of *compositional* structure in those languages, or even tries to inject such structure into them (Kottur et al., 2017; Choi et al., 2018; Lazaridou et al., 2018; Mordatch and Abbeel, 2018; Andreas, 2019; Cogswell et al., 2019; Li and Bowling, 2019; Resnick et al., 2019; Chaabouni et al., 2020). Besides possibly providing insights on how compositionality emerged in natural language (Townsend et al., 2018), this emphasis is justified by the idea that a compositional language has various *desirable* properties. In particular, compositional languages are expected to help agents to better *generalize* to new (composite) inputs (Kottur et al., 2017; Lazaridou et al., 2018), and to be faster to acquire (Cogswell et al., 2019; Li and Bowling, 2019; Ren et al., 2019).

We engage here with this ongoing research pursuit. We step back and reflect on the benefits that compositionality can bring to the emergent languages: if there is none, then it is unlikely that agents will develop compositional languages on their own. Indeed, several studies have shown that compositionality does not emerge naturally among neural agents (e.g. Kottur et al., 2017; Lazaridou et al., 2018; Andreas, 2019). On the other hand, understanding what benefits compositionality could bring to a language would help us in establishing the conditions for its emergence.

Compositionality is typically seen as a property of a language, independent of the task being considered. However, the task will likely influence properties such as generalization and ease of acquisition, that compositionality is expected to correlate with. Our experiments show that it is easy to construct tasks for which a compositional language is equally hard, or harder, to acquire and does not generalize better than a non-compositional one. Hence, language emergence researchers need to be clear about i) which benefits they expect from compositionality and ii) in which way compositionality would lead to those benefits in their setups. Otherwise, the agents will likely develop perfectly adequate communication systems that are not compositional.

2 Operationalizing compositionality

Before we proceed, let us clarify our definition of compositionality. Linguists and philosophers have extensively studied the topic for centuries (see Pagin and Westerståhl, 2010a,b, for a thorough review). However, the standard definition that a language is compositional if the meaning of each expression in it is a function of the meaning of its parts and the rules to combine them is so general as to be vacuous for our purposes (under such definition, even the highly opaque languages

Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, pages 11–15
Online, November 20, 2020. ©2020 Association for Computational Linguistics

we will introduce below are compositional, *contra* our intuitions).

In most current language emergence research, the input to language is composite in the sense that it consists of ensembles of elements. In this context, intuitively, a language is compositional if its symbols denote input elements in a *disentangled* way, so that they can be freely juxtaposed to refer to arbitrary combinations of them. More precisely, the following property might suffice for a limited but practical characterization of compositionality. Given a set of atomic input elements (for example, a set of independent attribute values), each atomic symbol should refer to one and only one input element, *independently of the other symbols it co-occurs with*.[1] A language where all symbols respect this property is compositional in the intuitive sense that, if we know the symbols that denote a set of input elements, we can assemble them (possibly, following some syntactic rules of the language) to refer to the ensemble of those input elements, irrespective of whether we have ever observed the relevant ensemble. Consider for example a world where inputs consist of two attributes, each taking a number of values. A language licensing only two-character sequences, where the character in the first position refers to the value of the first attribute, and that in the second position independently refers to the value of the second, would be compositional in our sense. On the other hand, a language that also licenses two-character sequences, but where both characters in a sequence are needed to decode the values of both the first and the second input attribute, would not be compositional. We will refer to the lack of symbol interdependence in denoting distinct input elements as *naïve compositionality*.[2]

We believe that naïve compositionality captures the intuition behind explicit and implicit definitions of compositionality in emergent language research. For example, Kottur et al. (2017) deem non-compositional those languages that either use single symbols to refer to ensembles of input elements, or where the meaning of a symbol depends on the context in which it occurs. Havrylov and Titov (2017) looked for symbol-position combinations that encode a single concept in an image, as a sign of a compositional behavior. A naïvely compositional language will maximize the two recently proposed compositionality measures of residual entropy (Resnick et al., 2019) and positional disentanglement (Chaabouni et al., 2020).

Naïve compositionality is also closely related to the notion of *disentanglement* in representation learning (Bengio et al., 2013). Interestingly, Locatello et al. (2018) reported that disentanglement is not necessarily helpful for sample efficiency in downstream tasks, as had been previously argued. This resonates with our results below.

3 Communication Game

We base our experimental study on a one-episode one-direction communication game, as commonly done in the relevant literature (Lazaridou et al., 2016, 2018; Havrylov and Titov, 2017; Chaabouni et al., 2019). In this setup, we have two agents, Sender and Receiver. An input $\boldsymbol{i}$ is fed into Sender, in turn Sender produces a message $\boldsymbol{m}$, which is consumed by Receiver. Receiver produces its output $\hat{\boldsymbol{o}}$. Comparing the output $\hat{\boldsymbol{o}}$ with the ground-truth output $\boldsymbol{o}$ provides a loss. We used EGG (Kharitonov et al., 2019) to implement the experiments.[3]

In contrast to the language emergence scenario, we use a hard-coded Sender agent that produces a fixed, pre-defined language. This allows us to easily control the (naïve) compositionality of the language and measure how it affects Receiver's performance. This setup is akin to the motivating example of Li and Bowling (2019).

We study two Receiver's characteristics: (i) acquisition speed, measured as the number of epoch needed to achieve a fixed level of performance on training set, and (ii) generalization performance on held-out data.

4 Experimental setup

To demonstrate that compositionality of a language alone, detached from the task at hand, does not necessarily lead to higher generalization or faster acquisition speed, we design two experiments.

The first experiment (*attval*) operates in an attribute-value world, similar to those of Kottur et al. (2017); Chaabouni et al. (2019). We fix two languages, one compositional and one not,

[1] We leave the definition of what counts as an atomic symbol open: it could be a single character, a character bound to a certain position in a message string, a character sequence, etc.

[2] Naïve in the sense that it is only appropriate when complex meanings are ensembles of atomic meanings. The definition breaks down when complex meanings result from functions that merge their components in different ways than simple ensembling, as is often the case in natural language.

[3] The code is available at https://github.com/facebookresearch/EGG/tree/master/egg/zoo/compositional_efficiency.

	Acquisition speed		
	task-identity	task-linear	task-entangled
	LSTM		
lang-identity	$5.3_{\pm 0.1}$	$30.0_{\pm 1.4}$	$20.1_{\pm 0.6}$
lang-entangled	$20.1_{\pm 0.5}$	$26.6_{\pm 0.8}$	$5.5_{\pm 0.2}$
	GRU		
lang-identity	$5.6_{\pm 0.2}$	$94.0_{\pm 7.7}$	$57.4_{\pm 13.9}$
lang-entangled	$37.2_{\pm 2.7}$	$91.5_{\pm 4.7}$	$5.7_{\pm 0.2}$
	Test accuracy		
	task-identity	task-linear	task-entangled
	LSTM		
lang-identity	$0.97_{\pm 0.00}$	$0.0_{\pm 0.0}$	$0.06_{\pm 0.01}$
lang-entangled	$0.08_{\pm 0.01}$	$0.0_{\pm 0.0}$	$0.97_{\pm 0.00}$
	GRU		
lang-identity	$0.97_{\pm 0.00}$	$0.0_{\pm 0.0}$	$0.06_{\pm 0.01}$
lang-entangled	$0.10_{\pm 0.02}$	$0.0_{\pm 0.0}$	$0.97_{\pm 0.00}$

Table 1: Attval experiment. Top: epochs to achieve perfect accuracy on training set. Bottom: test accuracy after convergence. $\pm$ marks 1 standard error of the mean.

and build three tasks: (i) "easy" for compositional language and "hard" for non-compositional; (ii) equally "hard" for both; (iii) "hard" for compositional language and "easy" for non-compositional language. Informally, we control the amount of computation needed by Receiver to perform a task starting from a language, where it can be equally hard to rely on compositional or non-compositional languages, or the answers could even be readily available in a non-compositional language.

In the second experiment (*coordinates*), we design a single task that is equally "easy" for an entire family of languages (parameterized by a continuous value), including compositional and non-compositional ones. The task is to transmit points on the 2D plane (thus, the input ensembles here are pairs of point coordinates). Here, we leverage the observation that a typical neural model has a linear output layer, for which it is equally easy to learn any rotation of the ground-truth outputs. Such rotation-group-invariance could play role in games where continuous image embeddings are used as input (Lazaridou et al., 2016; Havrylov and Titov, 2017).

4.1 Attval experiment

Input Sender's input $\boldsymbol{i}$ is a two-dimensional vector; each dimension encodes one of two attributes, each having n_v values: $\boldsymbol{i} \in \{1..n_v\} \times \{1..n_v\}$.

Languages We consider two languages, with messages of length two and vocabulary size n_v. The first language, *lang-identity*, represents the inputs as-is, by putting the value of the first (second) attribute in the first (second) position: $(m_1, m_2) \leftarrow (i_1, i_2)$. In the second language, *lang-entangled*, the first and the second positions are obtained as follows:

$$m_j \leftarrow (i_1 + (-1)^j \cdot i_2) \mod n_v, \quad j \in \{1, 2\} \tag{1}$$

Lang-identity and *lang-entangled* have exactly the same n_v^2 utterances. While *lang-identity* is naïvely compositional (one symbol encodes one attribute only), *lang-entangled* is not: each symbol of an utterance encodes equal amount of information about both attributes and both symbols are equally needed for decoding each attribute.[4]

Tasks We consider three tasks. In all of them, Receiver outputs two discrete values, $\boldsymbol{o} \in \{1..n_v\} \times \{1..n_v\}$. In *task-identity*, Receiver has to recover the original input of Sender, $\boldsymbol{i}$. In the second task, *task-linear*, Receiver needs to output two values that are obtained as integer linear-modulo operations of the original input values: $\boldsymbol{o} \leftarrow A \cdot \boldsymbol{i} + \boldsymbol{b} \mod n_v$. In the third task, *task-entangled*, we require Receiver to output $o_j \leftarrow (i_1 + (-1)^j \cdot i_2) \mod n_v$. In this task, the output values derive from the same attribute transform applied in the *lang-entangled* language (Eq. 1). This language-task pair mirrors the *lang-identity*/*task-identity* pair: each symbol encodes one *output* value.

Architecture and hyperparameters Receiver is implemented as an LSTM (Hochreiter and Schmidhuber, 1997) or a GRU cell (Cho et al., 2014). Its output layer specifies two categorical distributions over n_v values, encoding two output values. As a loss, we use the sum of per-output negative log-likelihoods. We used the following hyperparameters: $n_v = 31$; hidden layer size 100; embedding size 50; batch size 32; 500 epochs training with Adam (learning rate 10^{-2}). Each configuration was run 20 times with different random seeds. A random $1/5$ of the data is used as test set.

4.2 Coordinates experiment

Input We sample points uniformly from the unit circle, centered at the origin: $\boldsymbol{i} \in \mathbb{R}^2, \ \boldsymbol{i}^T\boldsymbol{i} \leq 1$. We sample 10^3 points for training, 10^3 for testing.

[4] Note that *lang-entangled* is still (non-naïvely) compositional, in the sense that its messages can be predictably derived by applying Eq. 1 to the input pairs.

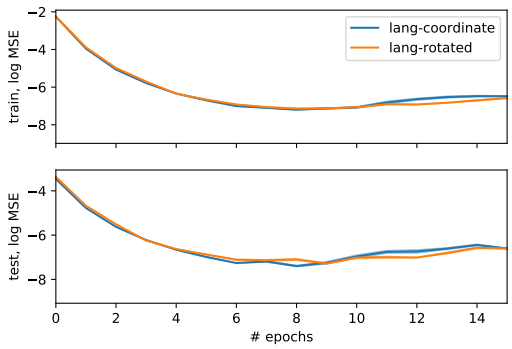

Figure 1: Coordinates experiment: log MSE vs. training epoch.

Languages We consider two languages with utterances of length two. In the first language, *lang-coordinate*, Sender sequentially transmits both coordinates of a point: $m_j \leftarrow i_j$. More precisely, the symbols refer to discretized coordinates from a $n_v \times n_v$ square grid, covering $[-1, 1] \times [-1, 1]$. This language is naïvely compositional w.r.t. the coordinate-wise representation of the inputs.

We construct the second language, *lang-rotated*, in the following way. We start with *lang-coordinate*, but apply a rotation of the plane by $\pi/4$ before feeding a point into Sender.[5] Effectively, this makes Sender "use" a rotated coordinate grid for encoding the coordinates. As a result of the rotation, *lang-rotated* ceases to be naïvely compositional in the original (non-rotated) world. Each symbol of *lang-rotated* carries equal amounts of information about both coordinates of $\boldsymbol{i}$.

Task Receiver has to recover the original (non-rotated) coordinates $\boldsymbol{i}$ of a point.

Architecture and loss Receiver is an LSTM with hidden size 100 and embedding size 50; n_v is 100; batch size is 32; we use Adam with learning rate 10^{-3}. As a loss, we use MSE. We run each configuration with 10 random seeds.

5 Results

Attval experiment In Table 1 we provide the results of the *attrval* experiment, depending on language, task, and Receiver architecture. We report the number of epochs to achieve perfect accuracy on training set (top) and the accuracy on the hold-out set after training (bottom).

[5]Rotating by any angle $(0, \pi/2)$ makes the language non-compositional; $\pi/4$ maximally entangles it.

Consider the convergence speed first. For both Receiver architectures *lang-identity* converges considerably faster than *lang-entangled*. This agrees with the findings of Li and Bowling (2019). However, in *task-linear* both languages demonstrate roughly the same convergence speed (the difference is not stat. sig.). In *task-entangled*, *lang-entangled* becomes more efficient to acquire than the naïvely compositional *lang-identity*. Interestingly, the acquisition times of *task-identity*/*lang-identity* and *task-entangled*/*lang-entangled* are symmetrical.

Next, consider the test accuracy of the same runs as above, measuring generalization to new attval combinations. We observe the same patterns: *task-linear* is equally hard to generalize from both languages; *lang-identity* reaches high test accuracy in *task-identity*, while *lang-entangled* leads to equally high accuracy on *task-entangled*. In contrast, *lang-identity* performs very poorly on *task-entangled*, just as *lang-entangled* does on *task-identity*.

Coordinates experiment Figure 1 reports learning curves for train and test sets (cueing acquisition speed and generalization, respectively). There is little difference between the compositional and non-compositional languages, in either training or held-out loss trajectories. Note also that, evidently, the linear mapping required here to undo the non-naïvely compositional transformation is easier for the networks than the non-linear operation we applied in the Attval experiment, pointing to the importance of taking the intrinsic biases of neural networks into account when designing language emergence experiments.

6 Discussion and Conclusion

Our toy experiments with hand-coded languages make the possibly obvious but currently overlooked point that, in isolation from the target task, there is nothing special about a language being (naïvely) compositional. A non-compositional language can be equally or faster to acquire than a compositional one, and it can provide the same or better generalization capabilities. Thus, if our goal is to let compositional languages emerge, we should be very clear about which characteristics of our setup should lead to its emergence.

Our concern is illustrated by the recent findings of Chaabouni et al. (2020), who observed that the degree of compositionality of emergent languages is not correlated with the generalization capabilities of the agents that rely on them to solve a task.

Indeed, lacking any specific pressure towards developing a (naïvely) compositional language, their agents were perfectly capable of developing generalizable but non-compositional communication systems. Our experiments provide a plausible explanation of their findings.

A stronger conclusion is that perhaps we should altogether forget about compositionality as an end goal. The current emphasis on it might just be a misguided effect of our human-centric bias. We should instead directly concentrate on the properties we want agent languages to have, such as fast learning, transmission and generalization.

References

Jacob Andreas. 2019. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8).

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of ACL*.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Advances in Neural Information Processing Systems*, pages 6290–6300.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. 2018. Compositional obverter communication learning from raw visual input. In *ICLP*.

Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. 2019. Emergence of compositional language with deep generational transmission. *arXiv preprint arXiv:1904.09067*.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *NIPS*, Barcelona, Spain.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *NeurIPS*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: a toolkit for research on emergence of lanGuage in games. In *EMNLP*.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of EMNLP*, pages 2962–2967, Copenhagen, Denmark.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *ICLR*.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. *arXiv preprint arXiv:1906.02403*.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. 2018. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *AAAI*.

Peter Pagin and Dag Westerståhl. 2010a. Compositionality I: Definitions and variants. *Philosophy Compass*, 5(3):250–264.

Peter Pagin and Dag Westerståhl. 2010b. Compositionality II: Arguments and problems. *Philosophy Compass*, 5(3):265–282.

Yi Ren, Shangmin Guo, Serhii Havrylov, Shay Cohen, and Simon Kirby. 2019. Enhance the compositionality of emergent language by iterated learning. In *Proceedings of the NeurIPS Emergent Communication Workshop*.

Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew M Dai, and Kyunghyun Cho. 2019. Capacity, bandwidth, and compositionality in emergent language learning. *arXiv preprint arXiv:1910.11424*.

Simon Townsend, Sabrina Engesser, Sabine Stoll, Klaus Zuberbühler, and Balthasar Bickel. 2018. Compositionality in animals and humans. *PLOS Biology*, 16(8):1–7.