

Enabling the Emergence of Symbolic Language without Handcrafted Inductions

Abstract

The emergence of symbolic languages with high compositionality has attracted extensive attention from a broad range of communities. Existing studies achieve high compositionality through *deliberately handcrafted* inductions (e.g., additional rewards, constructed loss functions and structural input data) in multi-agent learning, which are unnatural. Yet, few studies investigate the emergence of symbolic language with high compositionality *naturally*, i.e., without deliberately handcrafted inductions.

In this paper, we are the first to successfully achieve high compositional symbolic language in a *natural* manner without handcrafted inductions. Initially, by investigating the emergent language after removing the *deliberately handcrafted* inductions, we observe the difficulty in naturally generating high compositional language. Further, we reveal and characterize the quantitative relationship between the agent capacity and the compositionality of emergent language, with a novel mutual information-based metric for more reasonable measuring the compositionality. The experimental results lead to a counter-intuitive conclusion that lower agent capacity facilitates the emergence of language with higher compositionality. Based on our conclusion, we can get a more compositional language with a higher probability.

Introduction

The emergence of language has always been an important issue, which attracts attention from a broad range of communities, including philology, biology, and computer science. Especially in computer science, efforts in recent years trying to explore the emergent language in virtual multi-agent environments, where agents are trained to communicate with neural-network-based methods such as deep reinforcement learning (Kottur et al. 2017; Bogin, Geva, and Berant 2018; Lazaridou et al. 2018; Choi, Lazaridou, and de Freitas 2018; Jaques et al. 2019; Mul, Bouchacourt, and Bruni 2019; Kharitonov et al. 2019; Labash et al. 2020; Chaabouni et al. 2020).

The quality of emergent language is typically measured by its *compositionality*. Compositionality is a principle that determines whether the meaning of a complex expression (e.g. phrase), which is assembled out of a given set of simple

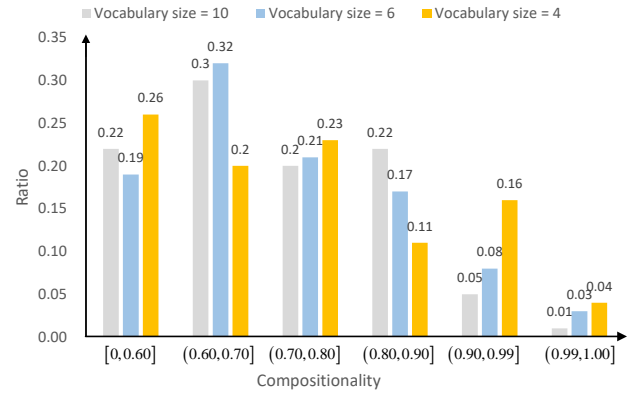


Figure 1: The distribution of compositionality for 100 emerged symbolic languages without any induction. It can be observed that high compositional symbolic language seldom emerged (e.g., < 5% for compositionality > 0.99). Moreover, varying the vocabulary size does not affect the compositionality notably.

components (e.g., symbols), can be determined by its constituent components and the rule combining them (Andreas 2018; Chaabouni et al. 2020). For example, the expression “AAAI is a conference” consists of two meaningful words “AAAI” and “conference”, and a rule for definition (“is”). Compositionality is considered to be a source of productivity, systematicity, and learnability of language, and the reason why a language with finite vocabulary can express almost infinite concepts.

Prior studies focus on achieving high compositional symbolic language through *deliberately handcrafted* inductions, e.g., additional rewards (Mordatch and Abbeel 2017), constructed loss functions (Kharitonov et al. 2019), structural input data (Lazaridou et al. 2018; Evtimova et al. 2018), memoryless (Kottur et al. 2017; Li and Bowling 2019), and ease-of-teaching (Li and Bowling 2019). Such optimization methodologies are driven by the challenges to generate high compositional symbolic without induction in an existing multi-agent environment.

Figure 1 reports the compositionality when training two agents in the widely-used listener-speaker referential

Table 1: Handcrafted inductions in related works.

Works	Handcrafted induction	Compositionality
(Kirby et al. 2015)	Expressivity and compressibility	Not quantitative, Speaker
(Kottur et al. 2017)	Listener’s memory	Not quantitative, Speaker
(Choi, Lazaridou, and de Freitas 2018)	Maximum message length	Not quantitative, Speaker+Listener
(Lazaridou et al. 2018)	Structure of input data	Quantitative, Speaker
(Evtimova et al. 2018)	Multi-modal scenarios	Quantitative, Speaker
(Li and Bowling 2019)	Population size, resetting all listeners	Quantitative, Speaker
(Chaabouni et al. 2019)	Word-order constraints	Not quantitative, Speaker
(Chaabouni et al. 2020)	Easier to decode	Quantitative, Speaker
Ours	None	Quantitative, Speaker+Listener

game (David 1969) for emerging 100 languages, and it can be observed that the compositionality of emergent language is seldom high (e.g., $< 5\%$ for compositionality > 0.99) without any induction. Moreover, varying the vocabulary size does not affect the compositionality notably. Though such unnatural inductions are useful, they prevent us from better understanding the mystery of the emergence of language and even intelligence among our pre-human ancestors. Yet, few works investigate the emergence of high compositional symbolic language *naturally*, i.e., without handcrafted inductions. In other words, it is never clear whether *natural* environment and agents are sufficient for achieving high compositionality.

This paper is the first one to achieve high compositional language without any deliberately handcrafted induction. The key observation is that the internal *agent capacity* plays a crucial role in the compositionality of emergent language. Concretely, the relationship between the agent capacity and the compositionality of emergent language is characterized, with a novel mutual information-based metric for the compositionality. Regarding the theoretical analysis, we propose a novel mutual information-based metric to measure the compositionality quantitatively. Regarding the experimental validation, we exploit the relationship between agent capacity and the compositionality of symbolic language that emerged *naturally* in our experiments. Both the theoretical analysis and experimental results lead to a counter-intuitive conclusion that *lower agent capacity facilitates the emergence of language with higher compositionality*. Therefore, by only reducing the agent capacity in such a natural environment, we can generate a more compositional language with a higher probability.

In this paper, we made the following contributions:

- To our best knowledge, we are the first work to successfully achieve high compositional symbolic language naturally, without any deliberately handcrafted induction.
- We analyze the compositionality of emerged symbolic language after removing deliberately handcrafted inductions.
- We propose a novel mutual information-based metric to measure the compositionality quantitatively, which is more reasonable.
- We experimentally exploited the relationship between agent capacity. Both theoretical analysis and experimen-

tal results lead to a counter-intuitive conclusion that lower agent capacity facilitates the emergence of symbolic language with higher compositionality.

The rest of this paper is arranged as follows. Section summarizes the related works. Section introduces the experimental setup used in this study. Section describes our proposed novel mutual-information-based metric for measuring the compositionality of symbolic language. Section gives the experimental results of the exploration for the relationship between agent capacity and compositionality. Section concludes this paper.

Related Works

Previous works focus on the *deliberately handcrafted* inductions that affect the compositionality of emergent language. Some significant works on studying the environmental inductions on the compositionality of emergent language are summarized in Table 1. For example, Kirby et al. (2015) explored how the pressures for expressivity and compressibility lead the structured language. Kottur et al. (2017) constrained the vocabulary size and whether the listener has memory to coax the compositionality of the emergent language. Lazaridou et al. (2018) showed that the degree of structure found in the input data affects the emergence of the symbolic language. Li and Bowling (2019) studied how the pressure, ease of teaching, impact on the iterative language of the population regime. Evtimova et al. (2018) designed novel multi-modal scenarios, which the speaker and the listener should access to different modalities of the input object, to explore the language emergence. These inductions are deliberately designed, which are too ideal to be true in the real world. In this paper, these handcrafted inductions above are all removed, and the high compositional language is learned only by the agent capacity.

To measure the compositionality of emergent language, metrics are proposed (Kottur et al. 2017; Choi, Lazaridou, and de Freitas 2018; Lazaridou et al. 2018; Evtimova et al. 2018; Chaabouni et al. 2020). At the initial stage, many studies only analyzed the language compositionality qualitatively (i.e. not quantitatively). For example, Choi, Lazaridou, and de Freitas (2018) printed the agent messages with the letter ‘abcd’ at some training round, and directly analyzed the compositionality on these messages. Kottur et al. (2017) introduced the dialog tree to show the evolution of language compositionality during the training process. Lat-

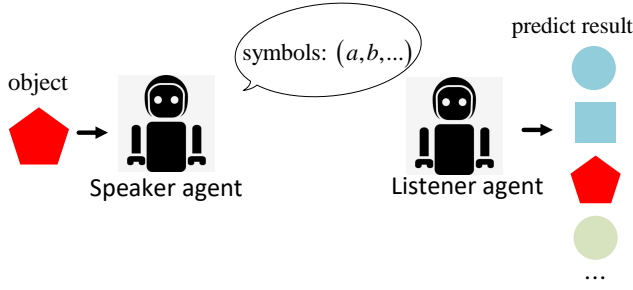


Figure 2: The referential game in this paper.

ter, some quantitative metrics are explored. The topographic similarity (Lazaridou et al. 2018) is introduced to measure the distances between all the possible pairs of meanings and the corresponding pairs of signals. Chaabouni et al. (2020) proposed the positional disentanglement, which measures whether symbols in a specific position relate to the specific attribute of the input object. From Table 1, most metrics are proposed on the sight of speaker. In our view, human beings developed the language based on a bilateral communication between the speaker and the listener. One research (Choi, Lazaridou, and de Freitas 2018) considered the metric bilaterally, but it is not a quantitative metric. In this paper, we propose a novel quantitative metric from both the speaker and the listener’s perspective.

In conclusion, the previous works induced the compositional language based on some deliberately handcrafted inductions, and the quantitative metric from the sight of both the speaker and the listener is still lacking. In this paper, we remove all the handcrafted inductions as shown in Table 1 and get a high compositional language through the internal agent capacity. Moreover, we propose a quantitative metric which take both the speaker and the listener into account.

Framework of Language Emerging

Before going to the detail of the training algorithms, we first introduce the environment, gaming rules, and agent architecture for enabling the emergence of symbolic language.

Environment setup

Figure 2 shows the entire environment used in this study, i.e., a commonly used referential game. Roughly, the referential game requires the speaker and listener to work cooperatively to accomplish a certain task. In this paper, the task is to have the listener agent reconstruct the object what the speaker claims it has seen, only through their emerged communication protocol. The consistent success in this game indicates that language has emerged between speaker and listener.

Game rules In our referential game, agents follow the following rules to finish the game in a cooperative manner. In each round, once received an input object t , Speaker S speaks symbols s to Listener L ; Listener L reconstruct the predicted result \hat{t} based on the listened symbols s ; if $t = \hat{t}$, agents win this game and receive positive rewards ($r(t, \hat{t}) = 1$); otherwise agents fail this game and receive

Algorithm 1 Learning Algorithm(t, \hat{t})

```

1: if Training the speaker agent  $S$  then
2:   for Batch  $T$  randomly selected from  $M_0 \times M_1$  do
3:     for  $t = (c_0, c_1)$  in  $T$  do
4:        $P(s_0|t), P(s_1|t) = \pi_{old}^S(s = (s_0, s_1)|t)$ 
5:       Sample  $s_0$  with  $P(s_0|t)$ ,  $s_1$  with  $P(s_1|t)$ 
6:        $P(\hat{t}|s) = \pi_{old}^L(\hat{t}|s)$ 
7:       Sample  $\hat{t}$  with  $P(\hat{t}|s)$ 
8:       Get reward  $r(\hat{t}, t)$ 
9:        $J(\theta^S, \theta^L) = E_{\pi_{old}^S, \pi_{old}^L} [r(\hat{t}, t) \cdot \frac{\pi_{old}^S(s|t)}{\pi_{old}^S(s|\hat{t})}]$ 
10:      Update  $\theta^S$  by  $\nabla_{\theta^S} J$ 
11:    end for
12:     $\pi_{old}^S \leftarrow \pi^S$ 
13:  end for
14: end if
15: if Training the listener agent  $L$  then
16:   for Batch  $T$  randomly selected from  $M_0 \times M_1$  do
17:     for  $t = (c_0, c_1)$  in  $T$  do
18:        $P(s_0|t), P(s_1|t) = \pi_{old}^S(s = (s_0, s_1)|t)$ 
19:       Sample  $s_0$  with  $P(s_0|t)$ ,  $s_1$  with  $P(s_1|t)$ 
20:        $P(\hat{t}|s) = \pi_{old}^L(\hat{t}|s)$ 
21:       Sample  $\hat{t}$  with  $P(\hat{t}|s)$ 
22:       Get reward  $r(\hat{t}, t)$ 
23:        $J(\theta^S, \theta^L) = E_{\pi_{old}^S, \pi_{old}^L} [r(\hat{t}, t) \cdot \frac{\pi_{old}^L(s|\hat{t})}{\pi_{old}^L(s|t)}]$ 
24:      Update  $\theta^L$  by  $\nabla_{\theta^L} J$ 
25:    end for
26:     $\pi_{old}^L \leftarrow \pi^L$ 
27:  end for
28: end if

```

negative rewards ($r(t, \hat{t}) = -1$).

Precisely, during the game, Speaker S receives an input object t , which is a concept-pair with two concepts from the concept set M_0 and M_1 , i.e., two one-hot vectors representing shape and color, respectively. Based on the t , Speaker S speaks a symbol sequence s , which similarly contains two words from V . The Listener L receives s and output predicted result \hat{t} , a single word (one-hot vector) corresponded with a concept-pair from the Cartesian product of $M_0 \times M_1$, which represents all the meanings of two combined words from M_0 and M_1 . Please note that since t and \hat{t} have different length, we say $t = \hat{t}$ if t expresses the same concept-pair as \hat{t} , e.g., “red circle”.

Agent architecture

Figure 3 shows the architecture of the constructed agents, including the Speaker S and Listener L .

Speaker. Regarding the Speaker S , it is constructed as a three-layer neural network. The Speaker S processes the input object t with a fully-connected layer to obtain the hidden layer h^s , which is further processed with fully-connected layers to obtain the output layer. The output layer results indicate the probability distribution of symbols with given input object t , i.e., $o_i^s = P(s_i|t)$ $i \in 0, 1$. The final readout symbols are sampled based on such probability distribution.

Listener. Regarding the Listener L , it is constructed as a three-layer neural network, too. Different from Speaker

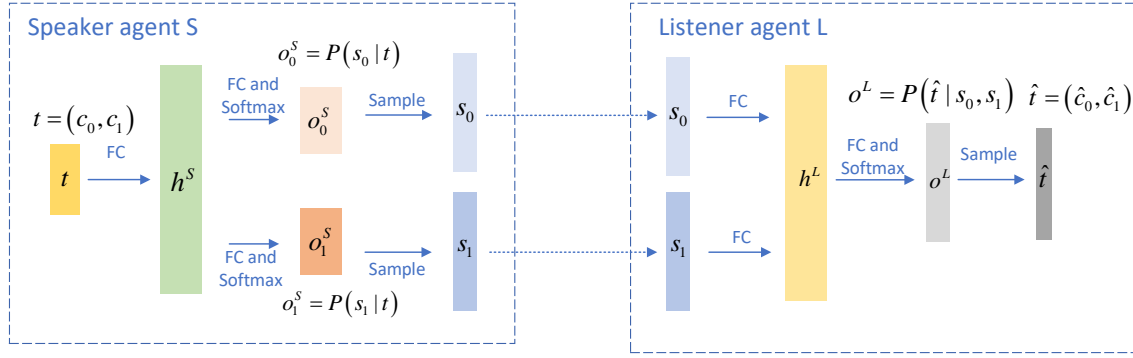


Figure 3: The architecture of agents. *Left:* speaker. *Right:* listener.

S that tries to separate input object into words, L tries to concatenates words to understand the combined meaning. The output layer results are also the probability distribution of symbols \hat{t} with given input sequence s , i.e., $o^L = P(\hat{t} | s_0, s_1)$.

Learning algorithm

To remove all the handcrafted induction as well as for a more realistic scenario, agents for this referential game are independent of each other, with no shared model parameters or architectural connections. As shown in Algorithm 1, we train the separate Speaker S and Listener L with Stochastic Policy Gradient methodology in a tick-tock manner, i.e., training one agent while keeping the other one. Roughly, when training the Speaker, the target is set to maximize the expected reward $J(\theta_S, \theta_L) = E_{\pi_S, \pi_L}[r(t, \hat{t})]$ by adjusting the parameter θ_S , where θ_S is the neural network parameters of Speaker S with learned output probability distribution π_S , and θ_L is the neural network parameters of Listener with learned probability distribution π_L . Similarly, when training the Listener, the target is set to maximize the expected reward $J(\theta_S, \theta_L)$ by fixing the parameter θ_S and adjusting the parameter θ_L .

Additionally, to avoid the handcrafted induction on emergent language, we only use the predicted result \hat{t} of the listener agent as the evidence of whether giving positive rewards. Then, the gradients of the expected reward $J(\theta_S, \theta_L)$ can be calculated as follows:

$$\nabla_{\theta_S} J = \mathbb{E}_{\pi_S, \pi_L} \left[r(\hat{t}, t) \cdot \frac{\nabla_{\theta_S} \pi^S(s_0, s_1 | t)}{\pi_{old}^S(s_0, s_1 | t)} \right] \quad (1)$$

$$\nabla_{\theta_L} J = \mathbb{E}_{\pi_S, \pi_L} \left[r(\hat{t}, t) \cdot \frac{\nabla_{\theta_L} \pi^L(\hat{t} | s_0, s_1)}{\pi_{old}^L(\hat{t} | s_0, s_1)} \right] \quad (2)$$

Mutual Information Similarity (MIS)

In this section, we propose the *Mutual Information Similarity (MIS)* as a metric of compositionality and give a thorough theoretical analysis. MIS is the similarity between an identity matrix and the mutual information matrix of concepts and symbols.

Before giving the definition of MIS, we first model the agents in the referential games. As shown in Figure 4, the

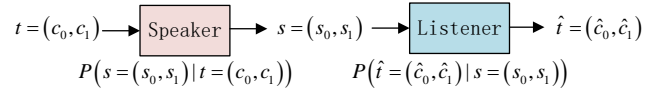


Figure 4: The information channel modeling of the agents in the referential game.

Speaker		Listener			
$(c_0, c_1) \backslash (s_0, s_1)$		$s_1 \backslash s_0$	a	b	c
a	ac	a	■	■	●
b	bc	b	■	■	●
c	aa	c	■	■	■
	ba				

Figure 5: An emergent language that the unilateral metrics cannot measure its non-compositionality. Notice that given $s_1 = a$, the listener can neither determine the shape nor the color without the knowledge about s_0 .

listener and speaker in the referential game are connected in tandem. The speaker agent can be regard as a channel, whose input is a concept $c = (c_0, c_1)$ and output is a symbol $s = (s_0, s_1)$. The listener agent can be regard as another channel, whose input is a symbol $s = (s_0, s_1)$ and output is a predict result $\hat{t} = (\hat{c}_0, \hat{c}_1)$. Since the output of the listener only depends on the symbol s , we can model the policy of the speaker agent and the listener agent by the probability distribution $P(s = (s_0, s_1) | t = (c_0, c_1))$ and $P(\hat{t} = (\hat{c}_0, \hat{c}_1) | s_0, s_1)$, respectively.

Now we can analyse the information of the concepts preserved in the transmission process given the symbol transmitted, i.e. the conditional mutual information $I(t, \hat{t} | s)$. Whenever a stable language emerged, the speaker and the listener consistently use a specific symbol s to refer to a

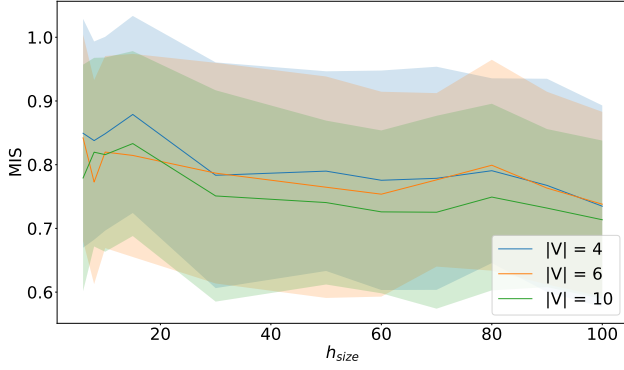


Figure 6: Compositionality of symbolic language under different parameters ($[\mu - \sigma, \mu + \sigma]$, where μ is the mean value and σ is the standard deviation).

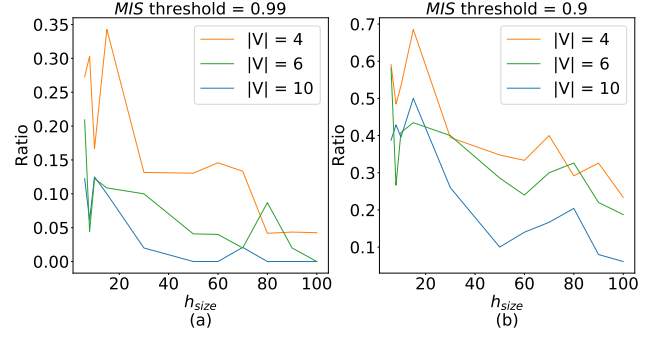


Figure 7: The ratio of high compositional language. (a) $MIS > 0.99$. (b) $MIS > 0.9$.

puted with following formula:

$$MIS_0 = \frac{1}{m} \sum_{j=0}^{m-1} \frac{\max_{i \in [0, n-1]} R(c_i, s_j)}{\epsilon + \sqrt{\sum_{i=0}^{n-1} R^2(c_i, s_j)}}, \epsilon > 0$$

$$MIS = \frac{n \cdot MIS_0 - 1}{n - 1}$$

MIS is a bilateral metric. Unilateral metrics, e.g. *topographic similarity* (*topo*()) and *posdis*(), only take the policy of the speaker into consideration. We provide an example to illustrate the inadequacy of unilateral metrics, shown in Figure 5. In this example, the speaker only uses s_1 to represent the shape. From the perspective of the speaker, the language is perfectly compositional (i.e. both *topo* and *posdis* are 1). However, the listener cannot distinguish the shape depend only on s_1 , showing the non-compositionality in this language. The bilateral metric MIS addresses such defects by taking the policy of the listener into account, thus $MIS < 1$.

Experiments

We exploit the relationship between agent capacity and the compositionality of symbolic language that emerged in our

Table 2: The Chi-square test between high-compositionality and agent capacity.

$H_0: MIS > 0.90$ is independent with h_{size}				
Vocabulary size	χ^2	df	p -value	
4	22.20	10	1.41×10^{-2}	
6	27.52	10	2.16×10^{-3}	
10	64.46	10	5.14×10^{-10}	
$H_0: MIS > 0.99$ is independent with h_{size}				
Vocabulary size	χ^2	df	p -value	
4	30.19	10	7.97×10^{-4}	
6	25.96	10	3.80×10^{-3}	
10	33.80	10	2.00×10^{-4}	

specific object t . Therefore we can safely say $I(t, \hat{t}|s) = I(t, \hat{t}|s_{t, \hat{t}})$ where $s_{t, \hat{t}} = \max_s \{P(\hat{t}|s)P(s|t)\}$. This conditional mutual information can be obtained by Equation 3.

$$I(t, \hat{t}|s_{t, \hat{t}}) = \sum_t \sum_{\hat{t}} P(t, \hat{t}|s_{t, \hat{t}}) \log \frac{P(t, \hat{t}|s_{t, \hat{t}})}{P(t)P(\hat{t}|s_{t, \hat{t}})} \quad (3)$$

We define the ratio of preserved information $R(t, s)$ as Equation 4, where $H(t)$ denotes the information entropy of t . $R(t, s)$ measures the degree of alignment between symbols and objects.

$$R(t, s) = \frac{I(t, \hat{t}|s = s_{t, \hat{t}})}{H(t)} \quad (4)$$

Following the Equation 4 we can obtain the normalized mutual information matrix M by collecting $R(c_i, s_j)$ for all i, j , as Equation 5.

$$M = \begin{pmatrix} R(c_0, s_0) & R(c_0, s_1) \\ R(c_1, s_0) & R(c_1, s_1) \end{pmatrix} \quad (5)$$

Each column of M corresponds to the semantic information carried by one symbol. In a perfectly compositional language, each symbol represents one specific concept exclusively. Therefore, the similarity between the columns of M and a one-hot vector is aligned with the compositionality of the emergent language.

Finally, we define *raw mutual information similarity* (MIS_0) as the average cosine similarity of M columns and one-hot vectors, as Equation 7. Furthermore, MIS is the normalized mutual information similarity into the $[0, 1]$ value range, which can be computed with following formula:

$$MIS_0 = \frac{1}{2} \sum_{j=0}^1 \frac{\max_{i=0,1} R(c_i, s_j)}{\epsilon + \sqrt{\sum_{i=0}^1 R^2(c_i, s_j)}}, \epsilon > 0$$

$$MIS = 2MIS_0 - 1$$

Generalized to m symbols and n objects, MIS can be com-

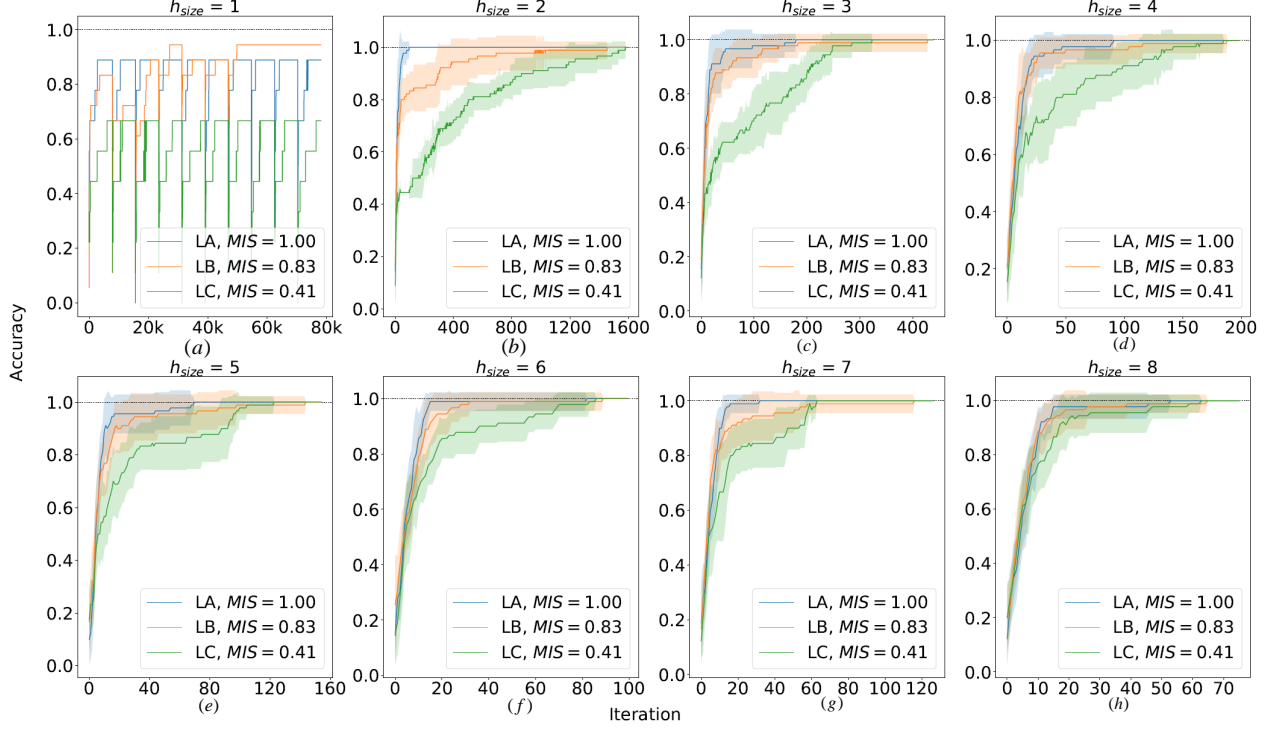


Figure 8: Accuracy of Listeners when varying h_{size} from 1 to 8. Each curve represents an average accuracy trend from 50 repeated training, with the range of $[\mu - \sigma, \mu + \sigma]$, where μ is the average accuracy and σ is the standard deviation.

natural referential game. For various configuration of vocabulary size, we fix $|M_0| = |M_1| = 3$ and train the speaker-listener agents to emerge symbolic language when varying the agent capacities, i.e., hidden layer size (h_{size}), from 6 to 100.

Figure 6 reports the experimental results. It can be observed that the mean value of MIS decreases as the value of h_{size} increases. Taking the configuration of vocabulary size $|V| = 10$ as an example, the mean value of MIS is around 0.8 when $h_{size} \leq 20$; MIS significantly decreases to 0.75 when h_{size} increases from 20 to 40; MIS further reduces to 0.7 when h_{size} increases from 40 to 100. For different vocabulary sizes, the MIS shares the similar behavior. It is because symbols in low-compositional languages carry semantic information about more concepts. As a result, higher capacity is required to characterize the complex semantic information for low-compositional language to emerge. In summary, lower agent capacity improves the possibility of emerging high compositional symbolic language.

Ratio of high compositional language.

We further breakdown our results to investigate the importance of agent capacity to the compositionality of symbolic language. Figure 7 reports the ratio of high compositional symbolic language in all emerged languages, Figure 7 (a) and (b) for $MIS > 0.99$ and $MIS > 0.9$, respectively. It can be observed that the ratio of high compositional sym-

bolic languages decreases drastically with the increase of h_{size} . Taking vocabulary size $|V| = 4$ as an example, symbolic languages with compositionality $MIS > 0.99$ take $>10\%$ mainly over all the emerged symbolic languages, when $h_{size} < 20$; the ratio reduces to $0\% \sim 5\%$ when h_{size} increases to 40; the ratio reduces around 3% when h_{size} goes beyond 40. $MIS > 0.9$ reports similar results. Notably, when h_{size} is large enough (e.g., > 40), high compositional symbolic language is hard to emerge in a natural referential game, for easy-to-emerge low compositional symbolic language is sufficient in scenarios of referential game. On the other side, agents are enforced to use compositionality to express more meanings, for the constraint from low capacity.

Additionally, we also perform χ^2 test to check the statistical significance between the high compositionality and agent capacity. Table 2 reports the χ^2 test results for $MIS > 0.99$ and $MIS > 0.9$, respectively. It can be observed that for different vocabulary sizes, the p-value is always less than 0.05, which means the high compositionality has a statistical significance related to agent capacity.

Breakdown into language teaching.

We further breakdown the learning process to investigate the language teaching scenario, where the Speaker teaches the Listener its fixed symbolic language. We define three symbolic languages in different compositionality for Speaker to teach, i.e., high (LA, $MIS = 1$), mediate (LB, $MIS = 0.83$),

LA	circle	square	triangle
red	(a, a)	(a, b)	(a, c)
blue	(b, a)	(b, b)	(b, c)
green	(c, a)	(c, b)	(c, c)

LB	circle	square	triangle
red	(a, a)	(a, b)	(a, c)
blue	(a, d)	(a, e)	(a, f)
green	(b, a)	(b, b)	(b, c)

LC	circle	square	triangle
red	(a, a)	(b, a)	(c, a)
blue	(d, a)	(e, a)	(f, a)
green	(g, a)	(h, a)	(i, a)

Figure 9: Three pre-defined language for teaching. (a) LA: high compositionality ($MIS = 1$). (b) LB: mediate compositionality ($MIS = 0.83$). (c) LC: low compositionality ($MIS = 0.41$).

low (LC, $MIS = 0.41$), see Figure 9.

Figure 8 reports the accuracy of Listener, i.e., the ratio of the correctly predicted symbols spoke by Speaker ($t = \hat{t}$), which varies with the training iterations under different agent capacities. Figure 8 (a) shows that when h_{size} equals to 1, the agent capacity is too low to handle languages. Figure 8 (b) shows that when h_{size} equals to 2, agent can only learn LA whose compositionality (i.e. MIS) is highest in all three languages. Combining these two observations, we can infer that language with lower compositionality requires higher agent capacity to ensure communicating successfully (i.e., h_{size}). Additionally, Figure 8 (c)~(h) shows that the higher agent capacity causes a faster training process for all three languages, but the improvement for different languages is quite different. It is obvious that language with lower compositionality also requires higher agent capacity to train faster.

Conclusion

In this paper, we are the first work to achieve high compositional symbolic language without any deliberately hand-crafted induction. We made the key observation that the internal *agent capacity* plays a crucial role in the compositionality of symbolic language. Together with the theoretical analysis, experimental results led to a counter-intuitive conclusion that *lower agent capacity facilitates the emergence of symbolic language with higher compositionality*. Therefore, by only reducing the agent capacity in such a natural environment, we generated a higher compositional symbolic language with a higher probability.

References

Andreas, J. 2018. Measuring Compositionality in Representation Learning. In *International Conference on Learning Representations*.

Bogin, B.; Geva, M.; and Berant, J. 2018. Emergence of Communication in an Interactive World with Consistent Speakers. *arXiv arXiv-1809*.

Chaabouni, R.; Kharitonov, E.; Bouchacourt, D.; Dupoux, E.; and Baroni, M. 2020. Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*.

Chaabouni, R.; Kharitonov, E.; Lazaric, A.; Dupoux, E.; and Baroni, M. 2019. Word-order Biases in Deep-agent Emergent Communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5166–5175. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1509. URL <https://www.aclweb.org/anthology/P19-1509>.

Choi, E.; Lazaridou, A.; and de Freitas, N. 2018. Compositional Obverter Communication Learning from Raw Visual Input. In *International Conference on Learning Representations*.

David, L. 1969. Convention: a philosophical study.

Evtimova, K.; Drozdov, A.; Kiela, D.; and Cho, K. 2018. Emergent Communication in a Multi-Modal, Multi-Step Referential Game. In *International Conference on Learning Representations*.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 3040–3049. PMLR.

Kharitonov, E.; Chaabouni, R.; Bouchacourt, D.; and Baroni, M. 2019. EGG: a toolkit for research on Emergence of lanGuage in Games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 55–60.

Kirby, S.; Tamariz, M.; Cornish, H.; and Smith, K. 2015. Compression and communication in the cultural evolution of linguistic structure. *Cognition* 141: 87–102.

Kottur, S.; Moura, J.; Lee, S.; and Batra, D. 2017. Natural Language Does Not Emerge ‘Naturally’ in Multi-Agent Dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2962–2967. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/D17-1321. URL <https://www.aclweb.org/anthology/D17-1321>.

Labash, A.; Aru, J.; Matiisen, T.; Tampuu, A.; and Vicente, R. 2020. Perspective taking in deep reinforcement learning agents. *Frontiers in Computational Neuroscience* 14.

Lazaridou, A.; Hermann, K. M.; Tuyls, K.; and Clark, S. 2018. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. In *International Conference on Learning Representations*.

Li, F.; and Bowling, M. 2019. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems*, 15851–15861.

- 440 Mordatch, I.; and Abbeel, P. 2017. Emergence of grounded
441 compositional language in multi-agent populations. *arXiv*
442 *preprint arXiv:1703.04908* .
- 443 Mul, M.; Bouchacourt, D.; and Bruni, E. 2019. Mastering
444 emergent language: learning to guide in simulated naviga-
445 tion. *arXiv preprint arXiv:1908.05135* .