One-to-Many Communication and Compositionality in Emergent Communication

Anonymous ACL submission

Abstract

Compositional languages leverage rules that derive meaning from combinations of simpler constituents. This property is considered to be the hallmark of human language as it enables the ability to express novel concepts and ease of learning. As such, numerous studies in the emergent communication field explore the prerequisite conditions for emergence of compositionality. Most of these studies set out one-to-one communication environment wherein a speaker interacts with one listener during a single round of communication game. However, real-world communications often involve multiple listeners; their interests may vary and they may even need to coordinate among themselves to be successful at a given task. This work investigates the effects of one-to-many communication environment on emergent languages where a single speaker broadcasts its message to multiple listeners to cooperatively solve a task. We observe that simply broadcasting the speaker's message to multiple listeners does not induce more compositional languages. We then analyze two axes of environmental pressures that facilitate emergence of compositionality: listeners of *different interests* and *coordination* among listeners.

1 Introduction

The field of emergent communication studies the core environmental factors in language emergence and the characteristics of emergent languages in relation to that of the human's. The recent developments in artificial neural networks have spurred research on the field utilizing communication simulations of neural agents (Lazaridou and Baroni, 2020). This has served as a crucial testbed for studying evolution of language (Briscoe, 2002), which often lacks concrete physical trace. The field has also demonstrated promising application possibilities in numerous domains leveraging language's desirable properties (Mu et al., 2023; Yao et al., 2022; Xu et al., 2022).

Compositionality (Janssen and Partee, 1997) is one of the most prominent features of human languages. Compositional languages can express complex meaning with combinations of simpler attributes leveraging systematic rule structures. This enables the ability to express novel concepts by combining familiar attributes. Compositionality is also attributed to enhancing languages' learnability (Ren et al., 2020; Davidson, 1965) and gives rise to robustness to noisy communication channel (Kuciński et al., 2021).

Determining the prerequisite environmental pressures for emergence of compositionality has been extensively studied in the field. These factors include language's learnability (Ren et al., 2020; Chaabouni et al., 2020; Smith et al., 2003; Li and Bowling, 2019), agents' capacity (Resnick et al., 2020), reliability of communication channel (Kuciński et al., 2021), task difficulty (Chaabouni et al., 2022; Choi et al., 2018; Mu and Goodman, 2021; Bouchacourt and Baroni, 2018; Lazaridou et al., 2017), and communication channel capacity (Lazaridou et al., 2018; Chaabouni et al., 2020). Recently, populations of agents have been investigated as a driving force for emergence of compositionality (Rita et al., 2022a; Michel et al., 2023) following prior sociolinguistic findings that larger population size tends to derive more structured languages (Raviv et al., 2019).

Most of these studies take one-to-one communication regime where only a single speaker-listener pair interacts with each other during an instance of game play. Even when there are multiple listeners in the system, a speaker's message is only sent to a single listener (Chaabouni et al., 2022; Michel et al., 2023; Rita et al., 2022a; Kim and Oh, 2021; Tieleman et al., 2019). Consequently, they fail to model the effects of one-to-many communication in emergent languages.

This work investigates the effects of one-to-many communication regime on the compositionality of emergent languages. In real-world communications, a single message often concerns multiple parties: an advertisement of a product, a sergeant's command to a squad, etc. In these scenarios, there are more than one interested entity for a given message. This environment opens two interesting aspects of communication, and we find that these aspects each introduce a new environmental pressure that facilitates emergence of compositionality.

First, the listeners may not share the same interests. In the case of the advertisement of a product, some of the viewers of the advertisement may only be interested in certain characteristics of the product such as colors and sizes, while others may only care about the price and brand name. While it is still the case that the advertisement must contain all of the relevant information for the product, we argue that it introduces a new pressure that forces the message to be easier to understand for listeners that are only interested in certain parts of the attributes. We hypothesize that these listeners would prefer messages that are easily interpretable, without the need to understand other details corresponding to attributes that they are not concerned with.

Second, listeners may need to coordinate among themselves to be successful at the task at hand. In the case of the sergeant's command to a squad, coordination among the squad may be required for them to have successfully carried out the mission. Hence, a misinterpretation of the command from a single listener may result in failure for the entire squad. We argue that the pressure that the language be simultaneously understood by multiple listeners forces the language to be more compositional. Intuitively, it is plausible that one listener may develop a compositionally inferior language, but it is less likely to be shared by other listeners in the group due to its inferior compositionality.

Extensive experiments confirm the hypotheses that agents of different interests and coordination among agents are crucial environmental pressures for emergence of compositionality. We find that simply broadcasting a speaker's message to multiple listeners does not enhance compositionality of induced languages. We observe emergence of compositionality when listeners of different interests are introduced or coordination pressures are injected to the environment. We then analyze what kinds of compositionalities are derived from these pressures with various compositionality measures.

2 Related work

Emergent communication and its applications Human languages exhibit a number of universal characteristics (Greenberg, 1961). The emergent communication field strives to close the gap between the communication protocols emerged from artificial agents and the natural languages with regard to these language universals. The studied characteristics include Zipf's law of abbreviation (Zipf, 1949; Chaabouni et al., 2019; Ueda and Washio, 2021; Ueda and Taniguchi, 2024), word boundaries (Harris, 1955; Ueda et al., 2023; Ueda and Taniguchi, 2024), trade-off between word-order and case-marking (Comrie, 1989; Blake, 2001; Lian et al., 2023) and compositionality (Chaabouni et al., 2020; Rita et al., 2022b). On a more practical note, the language-like properties of induced protocols facilitate numerous applications. Mu et al. (2023) leverage emergent languages' superior functional expressivity for embodied control task. Yao et al. (2022) demonstrate the effectiveness of emergent languages in low-resource language modeling, and similar results are reported in machine translation (Li et al., 2020; Downey et al., 2023). Xu et al. (2022) show emergent languages' competitive as a representation learning method. Techniques for inducing compositionality in emergent languages (Zheng et al., 2024; Li and Bowling, 2019; Ren et al., 2020) find applications in improving generic neural networks' abilities (Ren et al., 2023; Zheng et al., 2024; Noukhovitch et al., 2023).

Environmental pressures for compositionality Prerequisite conditions for emergence of compositionality are extensively studied. Kuciński et al. (2021) theoretically prove that compositional languages are more robust to message corruption and empirically verify that noisy channels facilitate compositionality. Several studies explore how capacity of communication channel (Lazaridou et al., 2018; Chaabouni et al., 2020) or capacity of neural agents (Resnick et al., 2020) affect compositionality. Cheng et al. (2023) observe that compositional languages are easier to imitate and suggest that imitability may also be a driving force for compositionality. Chaabouni et al. (2022) emphasize the task difficulty in terms of scale. Iterated learning (Smith et al., 2003; Li and Bowling, 2019; Ren et al., 2020) framework investigates the effects of language transmission across generations and finds that languages' learnability for newly created agents provide crucial pressure for compositionality.

Community structures in emergent communication Our study on the one-to-many communication regime is closely related to a line of works that investigates the effects of community structures on emergent languages. Harding Graesser et al. (2019) explore how independently formed communities' languages evolve when these communities start to interact with each other. Kim and Oh (2021) investigate the effects of different communication graphs on the languages' properties. Several studies observe that naively increasing the population size does not yield more structured languages (Chaabouni et al., 2022; Kim and Oh, 2021). Rita et al. (2022a) argue that different learning speeds in populations facilitate language structures. Michel et al. (2023) observe that limiting the communication graph with partitioning induces compositionality and generalization to unseen partners. However, all of these studies focus on one-to-one game play; hence, does not model the effects of one-to-many communication. Chaabouni et al. (2022) consider a simple voting mechanism of listeners only at inference time. Li and Bowling (2019) utilize simple message broadcasting when studying the effects of population size in iterated learning, but do not observe substantial improvements.

3 One-to-many communication game

We analyze emergent languages of agents playing a variant of Lewis reconstruction game (Lewis, 1969). The process of the game is as follows. Speaker π_θ observes an object, $x \in \mathcal{X}^K$ and produces a message $m \sim \pi_\theta(\,\cdot \mid x)$ describing the object. An object contains K attributes and each attribute can take one of $|\mathcal{X}|$ possible values. A message $m \in \mathcal{W}^T$ is a sequence of symbols of fixed length T and each symbol belongs to vocabulary $\mathcal{W}$. The game contains a set of N listeners $\{\pi_{\phi_i}\}_{i=1}^N$. Each listener π_{ϕ_i} is concerned with K_i attributes where $1 \leq K_i \leq K$. Let $x_i \in \mathcal{X}^{K_i}$ denote the K_i attributes' values the listener π_{ϕ_i} is concerned with in object x, e.g., if $K_i = K$, then $x_i = x$.

For each round of game play, the set of listeners are randomly partitioned into M groups $\{\mathcal{G}_j\}_{j=1}^M$ such that $\cup_{j=1}^M \mathcal{G}_j = \{\pi_{\phi_i}\}_{i=1}^N$ and $\cap_{j=1}^M \mathcal{G}_j = \varnothing$. Upon receiving message m, listener π_{ϕ_i} outputs its prediction for the object as $\widehat{x_i} \sim \pi_{\phi_i}(\,\cdot \mid m)$. Let $\mathcal{G}^{(i)}$ denote the indices of listeners in the group that listener π_{ϕ_i} belongs to. Listener π_{ϕ_i} receives a reward of 1 if all of the listeners' predictions in its group are correct, i.e., $R_{L_i}(x) = 1$ if $\forall j \in \mathcal{G}^{(i)}$, $\widehat{x_j} = x_j$ and 0 otherwise. The speaker receives the average reward of all listeners as a reward, which is equal to the fraction of successful listeners: $R_S(x) = \frac{1}{N}\sum_{i=1}^N R_{L_i}(x)$. See Appendix A for graphical illustrations.

4 Experimental setup

Dataset We represent each attribute's value with one-hot encoding. The number of attributes, K, is set to 4, and the number of values, $|\mathcal{X}|$, is set to 10. We set aside 10% of all attribute-value combinations as test set and use the rest as train set.

Speaker architecture One-hot encoded object x passes through a linear layer and initializes a single-layer GRU (Chung et al., 2014) of hidden size 500. It recurrently processes the input in total of $T = 5$ time steps. In each time step, outputs from it are fed to a linear layer and then goes through Softmax activation to produce vocabulary distribution of dimension $|\mathcal{W}| = 10$.

Listener architecture A listener π_{ϕ_i} is a single-layer GRU of hidden size 500. The listener sequentially processes the speaker's message m, the last output of which is then passed to K_i linear layers corresponding to the number of attributes the listener is interested in. They each go through Softmax activation and produce distribution of size $|\mathcal{X}|$ corresponding to the number of possible values an attribute can take.

Optimization We maximize each of the listeners' and speaker's expected reward with the REINFORCE algorithm (Williams, 1992). The expected reward for listener π_{ϕ_i} is written as $J_{L_i}(\phi_i) = \mathbb{E}_{x\sim p}\mathbb{E}_{m\sim\pi_\theta(\,\cdot\,|x)}R_{L_i}(x)$ and that of the speaker's is written as $J_S(\theta) = \mathbb{E}_{x\sim p}\mathbb{E}_{m\sim\pi_\theta(\,\cdot\,|x)}R_S(x)$, where p denotes the uniform distribution over $\mathcal{X}^K$. We also utilize entropy regularization for the speaker to facilitate exploration and cross entropy loss from listeners for stable training. We stop training if the success rate on the train set reaches 99. Full description of the setup is in Appendix B. See Appendix H for source code.

Reporting We report average over 10 random seeds. Throughout the paper, we use error bars to indicate 95% confidence interval and $\pm$ to denote standard deviation. **Bold** and <u>underline</u> indicate the best and second best results.

5 Evaluation metrics

Topographic similarity (TopSim) Let $D_{\text{obj}} : \mathcal{X}^K \times \mathcal{X}^K \to \mathbb{R}^+$ and $D_{\text{msg}} : \mathcal{W}^T \times \mathcal{W}^T \to \mathbb{R}^+$ be distance measures over the objects and messages, respectively. Topographic similarity (Brighton and Kirby, 2006) is Spearman's rank correlation of D_{obj} and D_{msg} over the joint uniform object, message distribution. This measures how changes in objects correlate with changes in messages. For D_{obj} and D_{msg}, we use cosine distance and Levenshtein distance (Levenshtein, 1965), respectively.

Positional disentanglement (PosDis) Let m_i denote the i-th symbol of message m. Let a_1^i denote the attribute that has the highest mutual information with m_i, i.e., $a_1^i = \arg\max_a \mathcal{I}(m_i; a)$. Similarly, let a_2^i denote the attribute that has the second highest mutual information with m_i, i.e., $a_2^i = \arg\max_{a \neq a_1^i} \mathcal{I}(m_i; a)$. Positional disentanglement (Chaabouni et al., 2020) is equal to $\frac{1}{T}\sum_{i=1}^{T} \frac{\mathcal{I}(m_i;a_1^i)-\mathcal{I}(m_i;a_2^i)}{\mathcal{H}(m_i)}$, where $\mathcal{H}(m_i)$ denotes the entropy of the i-th symbol. This measures the degree to which a single position is responsible for conveying information about an attribute.

Bag-of-symbols disentanglement (BosDis) Let n_i denote the number of occurrences of i-th symbol in vocabulary $\mathcal{W}$. Other notations follow from positional disentanglement. Bag-of-symbols disentanglement (Chaabouni et al., 2020) is equal to $\frac{1}{|\mathcal{W}|}\sum_{i=1}^{|\mathcal{W}|} \frac{\mathcal{I}(n_i;a_1^i)-\mathcal{I}(n_i;a_2^i)}{\mathcal{H}(n_i)}$. This measures how much a symbol univocally refers to an attribute.

Compositional generalization Compositional generalization is the average task success rate on unseen attribute combinations. This is calculated using the test set without regard to the group.

6 Experiments

6.1 Does naive one-to-many communication enhance compositionality of languages?

Setup In naive one-to-many communication regime, all listeners share the same interests, and there is no coordination required among the listeners. More specifically, the number of attributes each listener is interested in is identical to the number of attributes the speaker observes ($K_i = K$), and each group contains only a single listener ($|\mathcal{G}_j| = 1$).

Naive message broadcasting does not improve compositionality Figure 1 compares languages from naive one-to-many communication regime with varying number of listeners (N) against the single-listener one-to-one communication regime ($N = 1$). While some of the cases exhibit improvements, none of the differences are statistically significant (two-tailed t-test with $p = 0.05$). The results suggest that simply broadcasting a message does not introduce a meaningful pressure on language emergence.

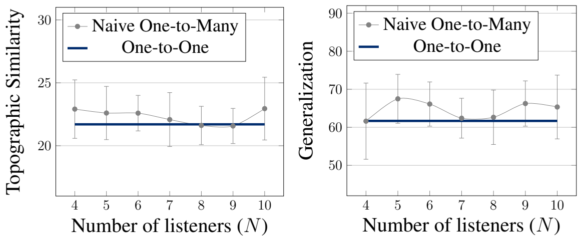


Figure 1: Language properties under varying number of listeners in naive one-to-many communication regime.

6.2 How do listeners of different interests affect language properties?

Setup We devise three kinds of listener formations for this experiment. The *partial-interest* formation contains $\binom{K}{K_i}$ listeners that are only concerned with K_i attributes. Each of $\binom{K}{K_i}$ listeners' interests are distinct attribute combinations. The *mixed-interest* formation is the same as the partial-interest formation except that it contains one additional listener that is concerned with all of the K attributes. The *full-interest* formation contains $1 + \binom{K}{K_i}$ listeners all of which are interested in all of the K attributes. As there is no coordination required, each group $\mathcal{G}_i$ contains a single listener. The test set accuracy is calculated only with the listeners that are interested in all of the attributes.

Readability pressure from different interests facilitates more structured languages In Figure 2a, we observe a trend that the more the listeners can disregard other parts of a message that they are not concerned with, the more compositional the languages tend to be. The formations with smaller number of interested attributes (K_i) exhibit higher TopSim, and the partial-interest formation's languages tend to exhibit higher TopSim compared to the mixed-interest formation. Languages from the two formations are more compositional then that of a similarly sized full-interest formation. In Figure 2d, we observe a similar trend for compositional generalization ability. We hypothesize that listeners of different interests prefer more struc-

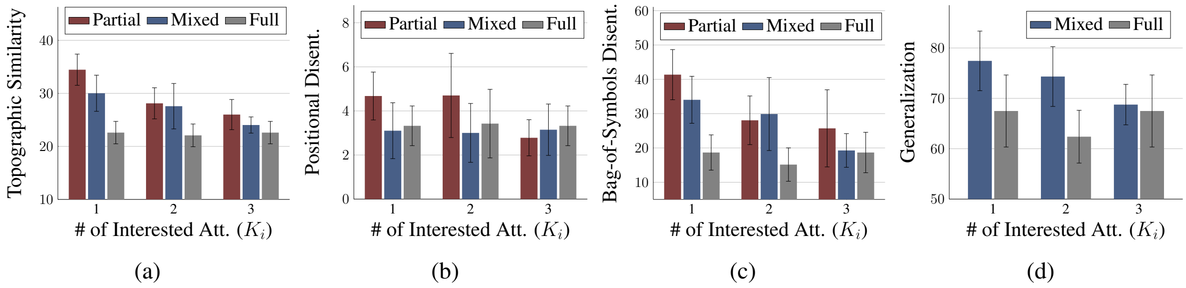


Figure 2: Comparison of language properties in listeners of different interests regime.

tured languages, so that they can more easily infer the attributes of interest from a message without needing to understand other details that are not related to their interests. In Appendix E, we find that languages from listeners of different interests are also easier to learn. We confirm that the results do not stem from the relative easiness of the task in Appendix F.

Listeners of different interests prefer symbol-wise structures rather than position We analyze what kinds of language structures are promoted by listeners of different interests. One possible structure is to denote each attribute within a certain position of a message. However, we do not observe such positional structure in regard to the number of interested attributes from Figure 2b. Another possible strategy is to associate the number of occurrences of a certain symbol to an attribute. In Figure 2c, we observe a clear trend that listeners of different interests prefer this kind of association.

6.3 How does coordination pressure affect language properties?

Setup We construct 50 listeners of the same interests ($K_i = K$). For each round of game play, the listeners are randomly split into equally sized groups. We explore the effects of coordination pressure in terms of group size ($|\mathcal{G}_j|$). A larger group size forces more listeners to be simultaneously successful at understanding the speaker's message. The test accuracy is calculated by taking average of all listeners' success rates regardless of the group.

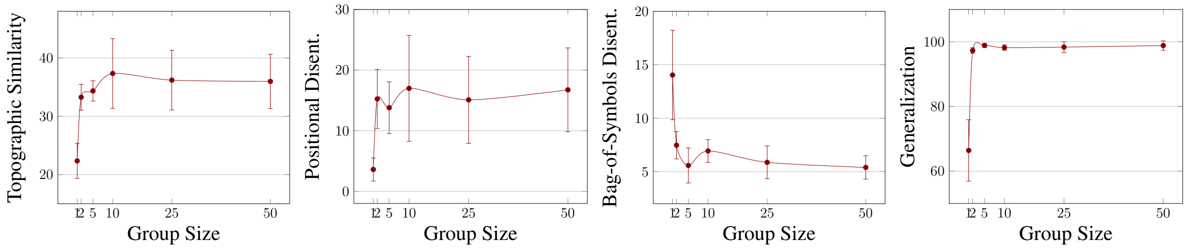
Coordination pressure amplifies preference of compositionality In Figure 3a, we observe a steep increase in TopSim as soon as a small coordination pressure is introduced to the game. TopSim steadily increases with the group size up to 10, then shows a bit of decrease at larger group sizes of 25 and 50. Similar increase is observed in compositional generalization ability in Figure 3d. We hypothesize that the coordination pressure amplifies the degree of preference for the language's compositionality from the listeners, as it requires the listeners to have a simultaneously shared understanding of a message.

Coordination pressure induces position-wise structures rather than symbols In Figure 3b, we observe increase in PosDis when coordination pressure is injected to the game, suggesting that coordination pressures induce more position-wise structures. A reverse trend is observed in BosDis in Figure 3c. The emergent languages under coordinate pressure tend to rely less on the number of occurrences of a symbol when determining an attribute's value. The results indicate that to effectively express more complicated concepts (larger number of attributes) position-wise structures are preferred.

6.4 Coordination pressure in relation to iterated learning framework

Iterated learning Iterated learning framework (Smith et al., 2003) simulates languages' transmission across generations. Li and Bowling (2019) find that periodically resetting listener's parameters forces the speaker to develop languages that are easier to teach, hence more compositional. In their experiments with populations of listeners, the authors hypothesize that resetting each listener in uniform time intervals instead of resetting them all at once could yield more structured languages as the population would contain more diverse listeners with varying degrees of experience. However, they observe that simultaneously resetting all of the listeners at the same time yield more compositional languages compared to the uniform reset regime.

Setup We conduct a small-scale experiment with two listeners to explore how coordination pressure impacts languages in iterated learning. We consider three different listener reset regimes. In simultane-



(a) (b) (c) (d)

Figure 3: Comparison of language properties under varying degrees of coordination pressure.

Metric	*Single Listener* ($N = 1$)		*Without Coordination* ($N = 2$)			*With Coordination* ($N = 2$)		
	No-Reset	**Simultaneous**	**No-Reset**	**Simultaneous**	**Uniform**	**No-Reset**	**Simultaneous**	**Uniform**
TopSim	21.72±3.3	**28.43**±2.8	22.34±2.7	**28.16**±4.0	26.16±2.1	30.91±2.3	32.37±3.4	**36.18**±4.7
Generalization	61.03±13.5	**91.34**±8.9	56.66±9.2	**92.02**±2.9	87.48±8.2	**98.27**±1.4	92.22±2.3	96.19±3.3

Table 1: Effects of coordination pressure on emergent languages in iterated learning environment.

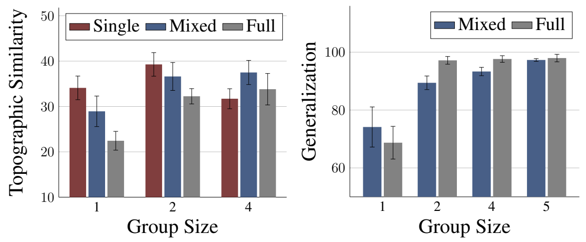
ous reset regime, we reset all listeners every 200 epochs. Uniform reset regime resets one listener at epochs $\{100, 300, 500, ...\}$, and the other listener at epochs $\{200, 400, 600, ...\}$. No-reset regime does not perform any listener reset. We also consider a single-listener system under no-reset and simultaneous reset regimes. We train the agents for 6,000 epochs.

Coordination pressure accentuates the effects of population in iterated learning In Table 1, we compare the single-listener system to the two-listener systems with and without coordination pressure (group size of 2 and 1, respectively). When there is no coordination pressure, uniform reset produces less compositional languages compared to the simultaneous reset regime, and the simultaneous reset regime in two-listener system does not show a clear improvement over the single-listener system. Under coordination pressure, uniform reset exhibits higher compositionality than the simultaneous reset regime, and simultaneous reset regime shows improvements over the single-listener system. These observations demonstrate the importance of coordination in iterated learning.

6.5 Listeners of different interests under coordination pressure

We explore how the readability pressure from listeners of different interests and coordination pressure interact with each other in language emergence.

Setup We construct three kinds of listener formations. The *single-interest* formation contains four listeners that are interested in each of the four attributes ($K_i = 1$). The *mixed-interest* formation is the same as the single-interest formation but contains one additional listener that is interested in all of the attributes. The *full-interest* formation contains five listeners that are interested in all of the four attributes ($K_i = K$). We test how these listener formations behave under varying degrees of coordination pressure expressed by group sizes of 1, 2, 4 (and 5 for the latter two formations that contain 5 listeners). We note that the latter two formations render one single-listener group at group sizes of 2 and 4 as it contains 5 listeners.



(a) (b)

Figure 4: Comparison of language properties in general one-to-many communication regime.

Readability pressure from different interests complements coordination pressure In Figure 4a, we observe that the mixed-interest formation induces more compositional languages compared to the full-interest formation in terms of TopSim under varying degrees of coordination pressure. We also observe that the single-interest formation's languages exhibit increase in TopSim as the group size is increase to 2. However, it experiences a decrease as the group size is further increased to

4. At group size of 4, the four listeners in the single-interest formation now experience success only if each listener's prediction for its attribute of interest is correct. Hence, the learning signals that shape the language may reflect less the readability pressure introduced by different interests. We further analyze derived compositionality structures in terms of PosDis and BosDis in Appendix G.

Partial interests reduce the magnitude of coordination pressure In Figure 4b, we see that the mixed-interest formation's generalization ability increases more slowly with the group size compared to the full-interest formation. We hypothesize that this stems from the fact that some of the attribute's descriptions need not be shared by multiple listeners in the mixed-interest formation. For example, at group size of 2 or 4, a description of an attribute must be agreed upon by at least 2 members of a group in the full-interest formation, in contrast to the mixed-interest formation where an attribute's description may not have to be agreed upon by multiple listeners if they happened to be interested in different sets of attributes. This results in reduction in coordination pressure for the mixed-interest formation. At group size of 5, however, coordination pressure forces any attribute's description to be agreed upon by more than one listener in the mixed-interest formation, and it exhibits similar generalization ability to the full-interest formation.

7 Experiments with raw images

We expand our study to more realistic scenarios employing datasets that consist of raw pixel images.

7.1 Listeners of different interests with raw pixel data

Experimental setup We explore the effects of readability pressures introduced by listeners of different interests in more realistic setup with 3dshapes dataset (Kim and Mnih, 2018). The dataset contains images of 3D shapes. Each image is characterized by 6 attributes such as the object's color and shape. We sample 4 values from each of these 6 attributes and perform the same experiment as in §6.2. Full description of the experimental setup is in Appendix C.

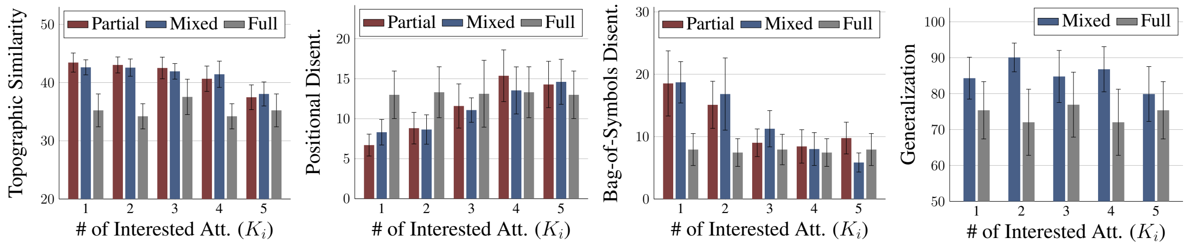
Results Overall, we observe similar trends to that of the attribute-value dataset's, suggesting that the findings in §6.2 hold in more complex environments. In Figure 5a, we find that smaller numbers of attributes of interest yield more compositional languages, and Figure 5d shows that they exhibit stronger generalization ability. We obtain more pronounced effects in terms of symbol- or position-wise structures of emergent languages. There is a clear tendency that smaller number of interested attributes produce languages that are less reliant on positional structures of messages as can be seen in Figure 5b. In Figure 5c, we also observe the tendency to denote an attribute with number of occurrences of a symbol in listeners of different interests regime.

7.2 Coordination pressure in large scale image discrimination game

Discrimination game We explore the effects of coordination pressure in a large-scale image discrimination game with ImageNet dataset (Russakovsky et al., 2015). The rules of the game are as follows. The speaker observes the target image x and sends a message m containing descriptions of the image to a set of listeners $\{\pi_{\phi_i}\}_{i=1}^{N}$. A listener π_{ϕ_i} is tasked to determine which one is the target among its context $\mathcal{C}_i$ containing other images and rewarded if all of the listeners in the group it belongs to correctly predict the target.

Scramble resistance (ScrRes) Let m' denote a randomly permuted version of message m and $\pi_{\phi_i}(x \mid m, \mathcal{C}_i)$ denote the probability assigned to the target object x by listener π_{ϕ_i} given message m and context $\mathcal{C}_i$. Scramble resistance (Bernard and Mickus, 2023) is calculated as $\frac{\min(\pi_{\phi_i}(x|m,\mathcal{C}_i),\pi_{\phi_i}(x|m',\mathcal{C}_i))}{\pi_{\phi_i}(x|m,\mathcal{C}_i)}$. A higher scramble resistance means the language is less affected by positional perturbations.

Experimental setup We use representations of images processed by a ResNet-50 (He et al., 2016) encoder pretrained on ImageNet with BYOL (Grill et al., 2020) as in Chaabouni et al. (2022); Michel et al. (2023). The context size $|\mathcal{C}_i|$ is set to 1,000 for all listeners. We use the train, validation, test splits from Chaabouni et al. (2022). We set aside 10% of the classes in the dataset as in-distribution (ID) classes and the rest as out-of-distribution (OOD) classes. We perform training and validation with ID samples in each split and evaluation with the test set containing only OOD samples. TopSim is calculated with respect to the image's representations using cosine distance. We construct 10 listeners and observe the effects of coordination pressure



(a) (b) (c) (d)

Figure 5: Comparison of language properties in different interests regime with 3dshapes dataset.

Group Size	*Task Success Rate*			*Compositionality*	
	Test (OOD)	**Val (ID)**	**Train (ID)**	**TopSim**	**ScrRes**
1	90.24±2.2	96.16±0.2	**98.45**±0.2	**22.03**±1.8	4.48±0.9
2	**94.03**±0.2	**97.59**±0.1	98.04±0.1	20.79±0.8	**6.20**±0.9
5	93.37±0.3	97.13±0.1	98.34±0.1	20.90±1.0	5.90±0.9
10	93.16±0.3	96.96±0.1	98.32±0.1	19.73±1.1	5.66±0.8

Table 2: Results on image discrimination game.

under varying group sizes. Full description of the experimental setup is in Appendix D.

Lower encounter frequency forces agents to develop more generalizable languages We report the accuracies on each split in Table 2. We observe that coordination pressure induces stronger generalization ability in both OOD and ID samples. Group size of 2 exhibits the highest generalization ability and further increase in group size results in lower generalization ability. Group size of 10 would exert the strongest coordination pressure requiring 10 listeners to simultaneously carry out the task. However, it forms only a single group throughout the training. A lower group size means that any two listeners less frequently encounter each other during training and yet need to be successful at coordination. We hypothesize that this pressure forces the agents to develop more generalizable languages.

Scramble resistance shows higher correlation with generalization than TopSim In Table 2, we observe that agents trained without any coordination exhibit the highest TopSim even though they show lowest generalization ability. Prior works (Chaabouni et al., 2022; Michel et al., 2023) also report that TopSim does not correlate with generalization ability and suggest that it may be an inadequate measure of compositionality for complex data forms. In contrast to TopSim, ScrRes shows high correlation with generalization ability, suggesting that a certain degree of positional invariance is beneficial for expressing more complex forms of data.

Coordination pressure induces languages that are easier to learn We explore how coordination pressure affects languages' learnability. To that end, we train a newly initialized listener by letting it play the discrimination game with a frozen speaker on the train set. We compare learnability of languages emerged under no coordination pressure to the languages emerged under coordination pressure from group size of 2. In Figure 6, we observe that new listeners learn languages emerged under coordination pressure faster than the ones that did not experience coordination pressure.

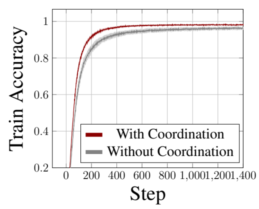


Figure 6: Learnability comparison on ImageNet.

8 Conclusion

This work investigates how one-to-many communication affects language emergence. We find that one-to-many communication introduces two complementary aspects of communication that facilitate emergence of compositionality. First, listeners of different interests exert readability pressure. This forces the language to be more structured as listeners prefer messages that do not require understanding of other aspects unrelated to the attributes of interest. Second, coordination among listeners amplify agents' preference of compositionality as the language has to be simultaneously understood by multiple listeners. Additionally, we find that coordination across different generations is an important factor in iterated learning. We verify that our findings hold in more complex environments with experiments on raw image data. Our work sheds light on the importance of one-to-many communication in the emergent communication field.

Limitations

Task complexity This work analyzes emergent languages with basic attribute-values and image datasets. While these datasets are widely employed in the emergent communication community and permit detailed analysis of compositionality, they lack the complexities of real-world environments. Recent studies propose various tasks that require more abstract reasoning (Guo et al., 2023; Zhou et al., 2024; Mihai and Hare, 2021; Patel et al., 2021). Future work may explore how our findings apply in more complex task scenarios.

Complex communication structures This study sets a basic one-to-many communication of a single speaker and the speaker's message is broadcast to all listeners in the system. However, more complex communication structures are possible. There could be multiple speakers and a speaker's message may be relayed to only certain portions of the listeners. The effects of population size (Rita et al., 2022a; Michel et al., 2023) and more complex communication graphs (Kim and Oh, 2021; Harding Graesser et al., 2019; Michel et al., 2023) could be further explored. In the coordination side, instead of forming new groups for each game play, longer listener group formation frequency could be explored. We also note that the effects of skewed interests of listeners are not explored in this work as we simply utilized all combinations of interests.

Exploration of applications Our work does not explore immediate application areas of the findings. However, the emergent communication field has demonstrated numerous application possibilities in diverse domains. Some of these find applications in improving foundation models (Noukhovitch et al., 2023; Zheng et al., 2024). It may be an interesting research direction to investigate our findings in relation to alignment of large language models (Ouyang et al., 2022; Rafailov et al., 2023) as human preferences can be decomposed into multiple attributes (Lou et al., 2024), e.g., helpfulness, politeness, etc. Our findings suggest that devising separate preference models each of which concerning a certain preference aspect could be beneficial for compositional generalization in terms of these preferences. As for the coordination pressure, multiple preference models of different value systems could be explored for simultaneously satisfying a wide range of users of varying cultural backgrounds.

Causes and implications of different compositionality structures In §6.2, we observe that listeners of different interests induce more symbol-wise structures in languages rather than position-wise structures, and we find a reverse trend when coordination pressure is exerted to the environment. We do not fully investigate the underlying mechanisms that cause these phenomena and their implications. Future work may explore how these kinds of compositionality structures affect performance in downstream tasks from the perspective of representation learning.

Theoretical analysis Through extensive experiments, we empirically verify that listeners of different interests and coordination among listeners play crucial roles in emergence of compositionality. However, more fine-grained analysis of the process would enhance the understanding these factors and facilitate applications possibilities. One could theoretically analyze the processing efforts required for listeners of different interests are indeed lower when the language is more compositional, or theoretically validate that the chances of any two listeners to stumble upon the same protocol are higher when the language is compositional.

Relationship to other environmental pressures As we discuss in §2, there are various environmental factors involved in emergence of compositionality, e.g., noisy channel (Kuciński et al., 2021). The relationship between these and the pressures investigated in this work could be further explored. For instance, we explore coordination pressure in relation to iterated learning in §6.4.

Effects of one-to-many communication on other language universals Our work focuses on one-to-many communication's effects on compositionality. However, there are other language universals that are actively studied in the emergent communication field as discussed in §2. Future work may explore how one-to-many communication affects other language universals.

Availability of attribute labels In the experiments with listeners of different interests, the listeners' interests are derived from labeled attributes. However, a dataset in question may lack such labels. Future work may investigate the ways in which interests can be formed in an unsupervised manner. One could devise information bottlenecks so that each listener would have a specialized role in the cooperative task.

References

Timothée Bernard and Timothee Mickus. 2023. So many design choices: Improving and interpreting neural agent communication in signaling games. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8399–8413, Toronto, Canada. Association for Computational Linguistics.

B.J. Blake. 2001. *Case*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

Henry Brighton and Simon Kirby. 2006. Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings. *Artificial Life*, 12(2):229–242.

T. Briscoe. 2002. *Linguistic Evolution through Language Acquisition*. Cambridge University Press.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online. Association for Computational Linguistics.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. *Anti-efficient encoding in emergent communication*. Curran Associates Inc., Red Hook, NY, USA.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. Emergent communication at scale. In *International Conference on Learning Representations*.

Emily Cheng, Mathieu Rita, and Thierry Poibeau. 2023. On the correspondence between compositionality and imitation in emergent neural communication. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12432–12447, Toronto, Canada. Association for Computational Linguistics.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. 2018. Multi-agent compositional communication learning from raw visual input. In *International Conference on Learning Representations*.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

B. Comrie. 1989. *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press.

Donald Davidson. 1965. Theories of meaning and learnable languages. In Yehoshua Bar-Hillel, editor, *Proceedings of the 1964 International Congress for Logic, Methodology, and Philosophy of Science*, pages 383–394. North-Holland.

C.m. Downey, Xuhui Zhou, Zeyu Liu, and Shane Steinert-Threlkeld. 2023. Learning to translate by learning to communicate. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 218–238, Singapore. Association for Computational Linguistics.

J.H. Greenberg. 1961. *Universals of Language*. MIT.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

Yuxuan Guo, Yifan Hao, Rui Zhang, Enshuai Zhou, Zidong Du, Xishan Zhang, Xinkai Song, Yuanbo Wen, Yongwei Zhao, Xuehai Zhou, Jiaming Guo, Qi Yi, Shaohui Peng, Di Huang, Ruizhi Chen, Qi Guo, and Yunji Chen. 2023. Emergent communication for rules reasoning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Laura Harding Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication 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)*, pages 3700–3710, Hong Kong, China. Association for Computational Linguistics.

Zellig S. Harris. 1955. From phoneme to morpheme. *Language*, 31(2):190–222.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Theo M.V. Janssen and Barbara H. Partee. 1997. Chapter 7 - compositionality. In Johan van Benthem and Alice ter Meulen, editors, *Handbook of Logic and Language*, pages 417–473. North-Holland, Amsterdam.

Eugene Kharitonov, Roberto Dessì, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2021. EGG: a toolkit for research on Emergence of lanGuage in Games. `https://github.com/facebookresearch/EGG`.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Jooyeon Kim and Alice Oh. 2021. Emergent communication under varying sizes and connectivities. In *Advances in Neural Information Processing Systems*.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Ł ukasz Kuciński, Tomasz Korbak, PawełKoł odziej, and Piotr Mił oś. 2021. Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. In *Advances in Neural Information Processing Systems*, volume 34, pages 23075–23088. Curran Associates, Inc.

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

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

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Vladimir I. Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710.

David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA.

Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yaoyiran Li, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2020. Emergent communication pretraining for few-shot machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Yuchen Lian, Arianna Bisazza, and Tessa Verhoef. 2023. Communication drives the emergence of language universals in neural agents: Evidence from the word-order/case-marking trade-off. *Transactions of the Association for Computational Linguistics*, 11:1033–1047.

Xingzhou Lou, Junge Zhang, Jian Xie, Lifeng Liu, Dong Yan, and Kaiqi Huang. 2024. Spo: Multi-dimensional preference sequential alignment with implicit reward modeling. *Preprint*, arXiv:2405.12739.

Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, and Angeliki Lazaridou. 2023. Revisiting populations in multi-agent communication. In *The Eleventh International Conference on Learning Representations*.

Daniela Mihai and Jonathon Hare. 2021. Learning to draw: Emergent communication through sketching.

Jesse Mu and Noah Goodman. 2021. Emergent communication of generalizations. In *Advances in Neural Information Processing Systems*.

Yao(Mark) Mu, Shunyu Yao, Mingyu Ding, Ping Luo, and Chuang Gan. 2023. Ec2: Emergent communication for embodied control. *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Michael Noukhovitch, Samuel Lavoie, Florian Strub, and Aaron C Courville. 2023. Language model alignment with elastic reset. In *Advances in Neural Information Processing Systems*, volume 36, pages 3439–3461. Curran Associates, Inc.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Shivansh Patel, Saim Wani, Unnat Jain, Alexander Schwing, Svetlana Lazebnik, Manolis Savva, and Angel X. Chang. 2021. Interpretation of emergent communication in heterogeneous collaborative embodied agents. In *ICCV*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, 286:20191262.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. 2020. Compositional languages emerge in a neural iterated learning model. In *International Conference on Learning Representations*.

Yi Ren, Samuel Lavoie, Mikhail Galkin, Danica J. Sutherland, and Aaron Courville. 2023. Improving compositional generalization using iterated learning and simplicial embeddings. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Cinjon Resnick, Abhinav Gupta, Jakob Foerster, Andrew M. Dai, and Kyunghyun Cho. 2020. Capacity, bandwidth, and compositionality in emergent language learning. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '20, page 1125–1133, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. 2022a. On the role of population heterogeneity in emergent communication. In *International Conference on Learning Representations*.

Mathieu Rita, Corentin Tallec, Paul Michel, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. 2022b. Emergent communication: Generalization and overfitting in lewis games. In *Advances in Neural Information Processing Systems*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Kenny Smith, Simon Kirby, and Henry Brighton. 2003. Iterated learning: A framework for the emergence of language. *Artificial life*, 9:371–86.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12. MIT Press.

Olivier Tieleman, Angeliki Lazaridou, Shibl Mourad, Charles Blundell, and Doina Precup. 2019. Shaping representations through communication: community size effect in artificial learning systems. *Visually Grounded Interaction and Language (ViGIL) Workshop*.

Ryo Ueda, Taiga Ishii, and Yusuke Miyao. 2023. On the word boundaries of emergent languages based on harris's articulation scheme. In *The Eleventh International Conference on Learning Representations*.

Ryo Ueda and Tadahiro Taniguchi. 2024. Lewis's signaling game as beta-VAE for natural word lengths and segments. In *The Twelfth International Conference on Learning Representations*.

Ryo Ueda and Koki Washio. 2021. On the relationship between Zipf's law of abbreviation and interfering noise in emergent languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 60–70, Online. Association for Computational Linguistics.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Zhenlin Xu, Marc Niethammer, and Colin Raffel. 2022. Compositional generalization in unsupervised compositional representation learning: A study on disentanglement and emergent language. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shunyu Yao, Mo Yu, Yang Zhang, Karthik Narasimhan, Joshua Tenenbaum, and Chuang Gan. 2022. Linking emergent and natural languages via corpus transfer. In *International Conference on Learning Representations (ICLR)*.

Chenhao Zheng, Jieyu Zhang, Aniruddha Kembhavi, and Ranjay Krishna. 2024. Iterated learning improves compositionality in large vision-language models. *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Enshuai Zhou, Yifan Hao, Rui Zhang, Yuxuan Guo, Zidong Du, Xishan Zhang, Xinkai Song, Chao Wang, Xuehai Zhou, Jiaming Guo, Qi Yi, Shaohui Peng, Di Huang, Ruizhi Chen, Qi Guo, and Yunji Chen. 2024. Emergent communication for numerical concepts generalization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17609–17617.

George K. Zipf. 1949. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley.

A Graphical illustration of one-to-many communication game

Figure 7 illustrates listeners of different interests in one-to-many communication game. The speaker's message is broadcast to three listeners. These listeners each have their own distinct interests. The first listener is only interested in the color of the object, while the second listener is only interested in the shape of the object. The third listener is interested in both the color and the shape of the object. The predictions of these listeners reflect their interests, hence exclusively pertain to the attributes of interest.

Figure 8 illustrates coordination among four listeners. Each of the four listeners are assigned to a group of size 2. The speaker's message is broadcast to the listeners, and each listener predicts the object's attributes. Both listeners in the first group correctly predict the object's attributes and the group is considered to be successful at the task. One of the listeners in the second group produces an incorrect prediction and this results in a failure of the task for the entire group.

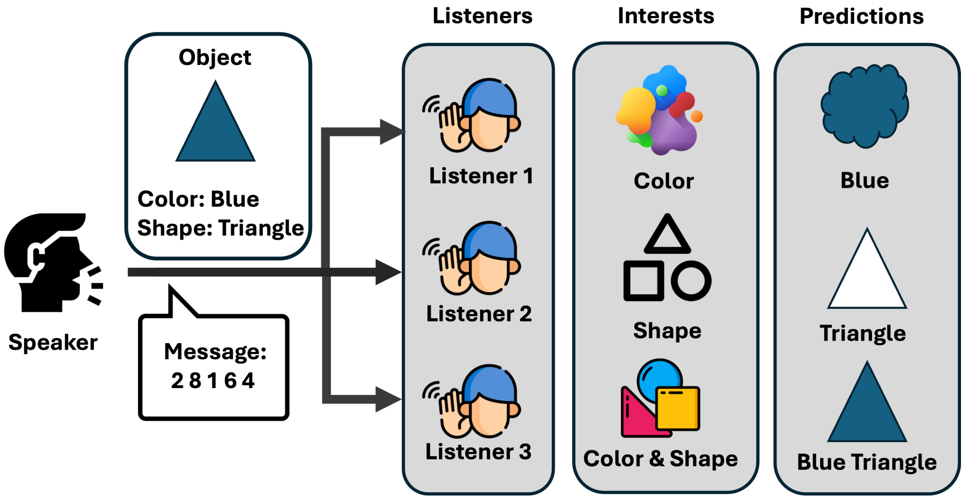


Figure 7: Illustration of listeners of different interests in one-to-many communication game. Each listener is interested in different set of attributes and its predictions only pertain to the attributes of interest.

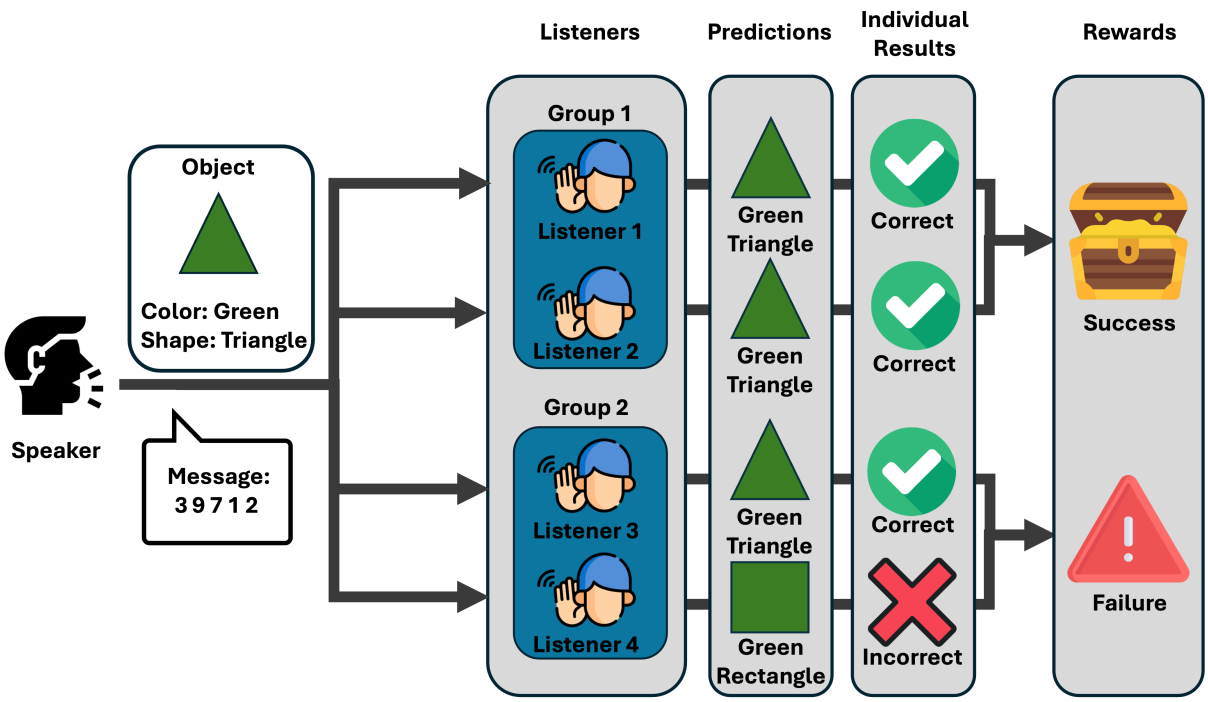


Figure 8: Illustration of coordination among listeners in one-to-many communication game. Listeners are split into groups and each listener is rewarded only if all of the listeners in the same group correctly predict the attributes.

B Experimental details

We utilize EGG framework (Kharitonov et al., 2021) which is available under MIT license. Speaker's symbol embedding size is 5 and listeners' symbol embedding size is 30. We use Adam optimizer (Kingma and Ba, 2017) with learning rate of 0.001. The batch size is set to 5120. We utilize REINFORCE with baseline (Sutton et al., 1999) where the baseline function is the average of the past rewards for the corresponding speaker or listener agent. We report compositionality metrics from the full dataset. We exclude a few runs that did not reach train accuracy of 99. At inference time, messages are constructed by selecting the symbol that has been assigned the highest probability by the sender at each time step. Experiments on raw pixel datasets follow the same setup unless otherwise specified.

Entropy regularization We add entropy regularization term in the speaker's symbol distribution to promote exploration. The magnitude of the regularization is controlled by a scaling hyperparameter γ which is multiplied to the entropy term. γ is set to induce successful language emergence on the train set of each dataset. For the experiments with attribute-values dataset, the value is set to 0.5. In the experiments with 3dshapes dataset, the value is set to 1.0. In the image discrimination experiments, the value is set to 0.1.

Cross entropy loss The training objective contains cross entropy loss from listener to stabilize training process. The cross entropy loss for listener π_{ϕ_i} is written as $-\frac{1}{K_i}\sum_{k=1}^{K_i}\log \pi_{\phi_i}(x_i^{(k)} \mid m)$, where $x_i^{(k)}$ refers to the k-th attribute in the object of interest x_i for the listener. For the speaker, listeners' average cross entropy loss is added to the reward after taking negative of it. For the listeners, each listener's own cross entropy loss is added to the reward in a similar manner. In addition to that, we directly backpropagate the cross entropy loss for each listener. Each cross entropy loss term is multiplied by a scaling hyperparameter λ. We use minimal value of λ for each dataset required for successful language emergence. In experiments with attribute-values dataset, the value is set to 0.4. For experiments with 3dshapes dataset, the value is set to 0.0. For the image discrimination experiments with ImageNet, the value is set to 0.2.

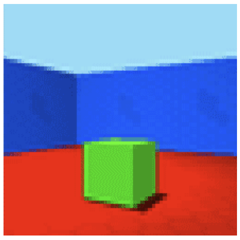


Figure 9: A sample of 3dshapes dataset.

C Experimental details on 3dshapes

We set the vocabulary size $|\mathcal{W}|$ to 6 and the length of messages T to 6. The batch size is set to 5,120. We stop training when the train accuracy reaches 99. We run each experiment with 20 random seeds and report the average. The dataset is available under Apache-2.0 license.

Dataset construction An image is characterized by 6 attributes: object's shape, object's color, object's size, color of the wall, color of the floor, and viewing orientation. Figure 9 shows a sample of the 3dshapes dataset. The number of values these attributes can take range from 4 to 14. We take 4 values from each attribute ($|\mathcal{X}| = 4$). For the attribute that correspond to the scale of the object, we choose values 0, 2, 4, 7 out of all the available values which range from 0 to 7. For the viewing orientation attribute, we choose values 0, 4, 9, 14 out of all the available values which range from 0 to 14. We construct each of the other attributes' 4 values by random sampling.

Agent architecture The speaker processes the image with a two-layer convolutional neural network (CNN) each of which is accompanied by a max pooling layer. The outputs then go through a linear layer before being processed by the single-layer GRU as described in §4. This produces a message m. CNNs have kernel size of 8, stride of 1, and filter size of 8. We utilize same padding. Max pooling layer has kernel size of 2 and stride of 2. The linear layer projects activations of dimension 2,048 to 500. A listener with the same architecture as in §4 processes the message m and outputs its prediction for the values of the image's attributes.

D Experimental details on ImageNet

The speaker processes the target image's representation of dimension 2048 with a linear layer producing activations of dimension 500. They are then

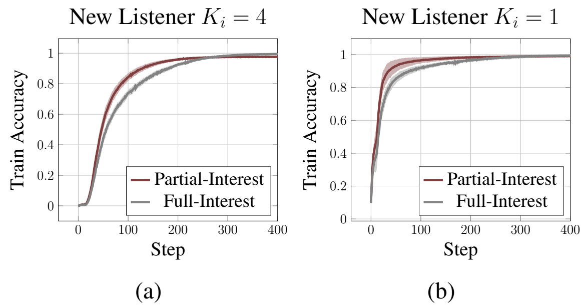


Figure 10: Learnability comparison in different interests regime. Shades indicate one standard deviation across 10 random seeds.

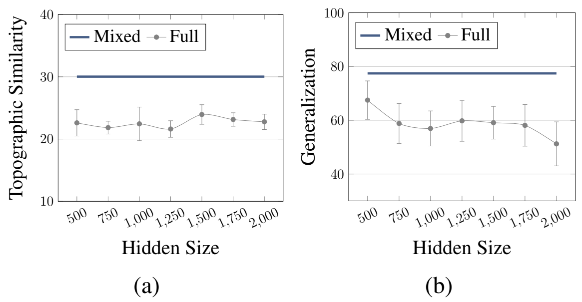


Figure 11: Language properties under varying values of listener hidden sizes in the full-interest formation in comparison with the mixed-interest formation of a fixed listener capacity.

processed by the single-layer GRU as described in §4. This produces message m containing descriptions of the target image.

A listener π_{ϕ_i} processes each of the images' representations in its context $\mathcal{C}_i$ with a linear layer then computes similarity scores of them with the message representation from the single-layer GRU described in §4. The message representation is computed from the last hidden state of the the single-layer GRU after it is passed through a linear layer. The resulting message representation has a dimension of 500. We use dot product as the similarity score function. These scores are then passed to Softmax activation to produce distribution over the context $\mathcal{C}_i$. We construct each listener's context by randomly sampling images without replacement.

The vocabulary size $|\mathcal{W}|$ and message length T are both set to 10. The batch size is set to 2048. Training is performed for 1,000 epochs and evaluation is performed with the checkpoint that exhibit the highest accuracy on the validation set. We repeat each experiment with 10 different random seeds and report the average. Scramble resistance is calculated with respect to one randomly selected listener. We report compositionality metrics from the test set. The image representations of ImageNet dataset is available under Apache-2.0 license.

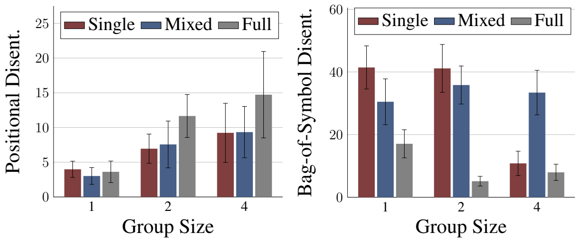
E Languages from listeners of different interests regime are easier to learn

We test if listeners of different interests in §6.2 indeed facilitate more structured, hence easier to learn languages. We take languages from the partial-interest formation with the number of interested attributes set to one ($K_i = 1$) and the full-interest formation of equal size. We randomly initialize new listeners of two different interests; one is only interested in one randomly sampled attribute ($K_i = 1$), and the other is interested in all of the four attributes ($K_i = 4$). We train these listeners by letting them play the game with the frozen senders of respective languages. In Figure 10, we observe that in both cases the languages from the partial-interest formation are easier to learn.

F Effects of relative model capacity in listeners of different interests regime

We validate that higher compositionality exhibited from listeners of different interests regime do not stem from the relative difficulty of the task as the number of attributes that need to be determined is lower in that regime. To that end, we increase the hidden size of listeners in the full-interest formation from 500 to larger values and compare them with the mixed-interest formation with $K_i = 1$. The experimental setup follows from §4. The hidden size of listeners in the mixed-interest formation is fixed to 500. Both formations contain the same number of listeners, $N = 5$.

In Figure 11a, we observe that the values of TopSim stay almost the same as the listeners' capacity is increased in the full-interest formation. This suggests that the relative capacity of the listeners in listeners of different interests regime is not the core contributing factor for the emergence of compositionality. Similarly, in Figure 11b, we observe a decrease in generalization ability as the capacity of the listeners in the full-interest formation is increased. These observations confirm that it is not the relative easiness of the task that induced more compositional languages in the listeners of different interests regime.



(a) (b)

Figure 12: Comparison of language properties in general one-to-many communication regime.

G Trade-off in the preference of symbol-wise and position-wise structures in general one-to-many communication

We analyze how the tendency to form more position-wise language structures under coordination pressure affects the tendency to form more symbol-wise language structures in listeners of different interests regime and vice versa. The experimental setup follows from §6.5. In Figure 12a, we observe in all listener formations the preference for position-wise language structures increases along with coordination pressure but the degree is less pronounced in single-interest and mixed-interest formations compared to the full-interest formation. Interestingly, Figure 12b shows that preference for symbol-wise structures in different interests regime prevails under coordination pressure unless the four single-attribute listeners are required to be always in the same group.

H Reproducibility

For training we utilized NVIDIA RTX A6000 48GB and NVIDIA A100 80GB. The most demanding task in terms of compute required less than 24GB of VRAM and took about 2 or 3 hours to complete per random seed. The number of parameters of an agent is far less than 1M in all experiments.

We make an anonymized version of our code available at: https://anonymous.4open.science/r/onetomany/.