Emergent Communication for Numerical Concepts Generalization

Enshuai Zhou[1, 2, 3], Yifan Hao[2], Rui Zhang[2], Yuxuan Guo[1, 2, 3], Zidong Du[2, 5],
Xishan Zhang[2, 3], Xinkai Song[2], Chao Wang[1], Xuehai Zhou[1], Jiaming Guo[2],
Qi Yi[1, 2, 3], Shaohui Peng[6], Di Huang[2], Ruizhi Chen[6], Qi Guo[2], Yunji Chen[2, 4]*
[1]University of Science and Technology of China
[2]State Key Lab of Processors, Institute of Computing Technology, CAS
[3]Cambricon Technologies
[4]University of Chinese Academy of Sciences
[5]Shanghai Innovation Center for Processor Technologies
[6]Intelligent Software Research Center, Institute of Software, CAS
enszhou@mail.ustc.edu.cn, {haoyifan, cyj}@ict.ac.cn

Abstract

Research on emergent communication has recently gained significant traction as a promising avenue for the linguistic community to unravel human language's origins and explore artificial intelligence's generalization capabilities. Current research has predominantly concentrated on recognizing **qualitative** patterns of object attributes(e.g., shape and color) and paid little attention to the **quantitative** relationship among object quantities which is known as the part of numerical concepts. The ability to generalize numerical concepts, i.e., counting and calculations with unseen quantities, is essential, as it mirrors humans' foundational abstract reasoning abilities. In this work, we introduce the NumGame, leveraging the referential game framework, forcing agents to communicate and generalize the numerical concepts effectively. Inspired by the human learning process of numbers, we present a two-stage training approach that sequentially fosters a rudimentary numerical sense followed by the ability of arithmetic calculation, ultimately aiding agents in generating semantically stable and unambiguous language for numerical concepts. The experimental results indicate the impressive generalization capabilities to **unseen** quantities and regularity of the language emergence from communication.

1 Introduction

Research on emergent communication has gained widespread attention in recent years (Lazaridou, Peysakhovich, and Baroni 2016; Choi, Lazaridou, and de Freitas 2018; Conklin and Smith 2023). It primarily involves using deep neural networks to simulate communication among multiple agents to complete collaborative tasks. From linguistics and cognitive psychology perspectives, studying emergent communication can provide a new experimental method and may validate specific linguistic and cognitive hypotheses quickly (Chaabouni et al. 2019; Rita, Chaabouni, and Dupoux 2020). From the standpoint of artificial intelligence, the language emergence from communication can help agents generalize on cooperative tasks better (Mu and Goodman 2021; Xu, Niethammer, and Raffel 2022).

The ability to generalize numerical concepts, i.e., counting and calculating on unseen quantities, is essential. According to linguistics and cognitive psychology, this ability is considered foundational for human abstract reasoning (Gelman and Gallistel 1986; Wiese 2003). Natural language possesses a comprehensive numerical system that allows humans to describe the number of objects accurately, concisely, and efficiently (Hiraiwa 2017). Furthermore, humans can perform more complex mathematical operations based on numerical concepts and digits, constructing a complete arithmetic system (Dehaene 2011). However, previous research has focused chiefly on recognizing **qualitative** patterns of object attributes(e.g., shape and color) (Kottur et al. 2017; Kuciński et al. 2021) and paid little attention to the agents' numerical concepts. It remains challenging to help agents understand the **quantitative** relations between numbers (i.e., quantities) through emergent communication.

In this work, we introduce the NumGame, leveraging the referential game framework (Lazaridou, Peysakhovich, and Baroni 2016), where agents are mandated to communicate and generalize their comprehension of numerical concepts proficiently. Specifically, agents are tasked with generalizing (in a few-shot learning manner) over unseen quantities via emergent communication in NumGame, encompassing two core tasks: Counting and Calculating. In the Counting task, agents must precisely evaluate unseen quantities of objects. In the Calculating task, agents face the challenge of deducing arithmetic relations (including addition, subtraction, and maximization) among unseen quantities. Both tasks in NumGame require the agents to understand rather than mechanically memorize quantities and their relations, making the agents' training difficult to converge effectively. Drawing inspiration from the human learning process of numbers(Wiese 2003; Hiraiwa 2017), we present a two-stage training approach comprising NumSen and NumRel. In this approach, we first employ the NumSen method to foster a rudimentary numerical sense of the agents. Then, we guide the agents to gain a foundational understanding of basic arithmetic relations between numbers within a specified

*Corresponding author.

range by the NumRel method. This progression ultimately enables the agents to generate language that expresses numerical concepts semantically stable and unambiguously and facilitates the generalization over unseen quantities and the arithmetic relations among them.

To quantitatively evaluate the effectiveness of our methods, we focus on natural language as the target and use generalization ability and regularity of the emergent language as two metrics to assess the agents' understanding of numerical concepts. We also visualize the language distribution after convergence to help readers better understand its structure. Ultimately, the experimental results demonstrate that by using the two-stage (i.e., NumSen and NumRel) training approach:

- (Section 7.2) The agents can accurately generalize over unseen quantities in the Counting task.
- (Section 7.3) The agents can perform basic calculations on unseen quantities in the Calculating task.
- (Section 7.4) Furthermore, the emerged messages between agents exhibit a solid order relation.

2 Related Work

Human numerical Concepts. Compared to other animals, humans have a remarkable grasp of numerical concepts(Hauser, Carey, and Hauser 2000; Drucker and Brannon 2014). Only humans possess the ability to use a finite set of numerical symbols to precisely describe quantities of objects and perform calculations using numbers(Pica et al. 2004; Butterworth 2005; Dehaene 2011). The concept of number is highly significant for humans (Conant 1896; Dehaene 2011) and is considered the foundation of human abstract reasoning and symbolic thinking ability (Gelman and Gallistel 1986; Wiese 2003; Feigenson, Dehaene, and Spelke 2004; Coolidge and Overmann 2012).

Many research works suggest that humans' precise grasp of the number concept arises from two main factors: number sense (Wiese 2003; Pica et al. 2004; Dehaene 2011) and human language (Hauser, Chomsky, and Fitch 2002; Wiese 2007; Hiraiwa 2017). Number sense can be divided into two parts: (1) the ability to recognize small quantities exactly, and (2) the ability to approximately recognize the magnitudes of larger quantities (Dehaene 2011). Even in prelinguistic eras, humans possessed number sense, and many animals also exhibited similar numerical abilities (Wiese 2003; Dehaene 2011). However, no animal possesses numerical abilities as powerful as humans do. Therefore, having number sense alone is insufficient; human language also plays a crucial role in the development of numerical concepts(Hiraiwa 2017). Human language is a unique communication system based on the recursive combination of a finite set of symbols (Berwick and Chomsky 2016). This unique property of the language may be another fundamental basis for the infinite expressive capacity of the human numerical system (Dehaene 2011).

Inspired by these insights, we incorporate number sense and language into the process of intelligent agents learning numerical concepts.

Emergent Communication. Using the Lewis signaling Game to research communication emergence in multi-agent systems has recently drawn more interest (Lewis 1969). Classified by motivation, some previous studies focus on how cognitive or social science views shape emergent communication, such as population heterogeneity(Chaabouni et al. 2019), linguistic complexity (Tucker et al. 2022), and efficiency of language(Chaabouni et al. 2019). Other previous studies focus on how to improve the quality of the emerged languages, such as compositionality (Conklin and Smith 2023), generalization (Xu, Niethammer, and Raffel 2022; Mu and Goodman 2021), and transferability on downstream tasks(Chaabouni et al. 2022).

In these works, the agents are required to extract and convey qualitative concepts, such as the object's shape, color, or location in the image. However, these works missed the language emergence of quantitative numerical concepts. For example, (Feng, An, and Lu 2023) constructs a multi-object environment that primarily centers on the positional relations among objects yet maintains a qualitative perspective. (Guo et al. 2019) differentiates the target and distractors based on the number of objects. Yet, in that approach, numbers are merely treated as classification labels and do not capture the intrinsic relations among them.

In this work, we delve into the quantitative concepts—the numerical concepts—and explore arithmetic relations among numbers. We propose a scenario in which agents are required to count and calculate quantities, which will compel them to comprehend the internal relations between quantities. We also propose the two-stage training method to facilitate their understanding of the numerical concepts.

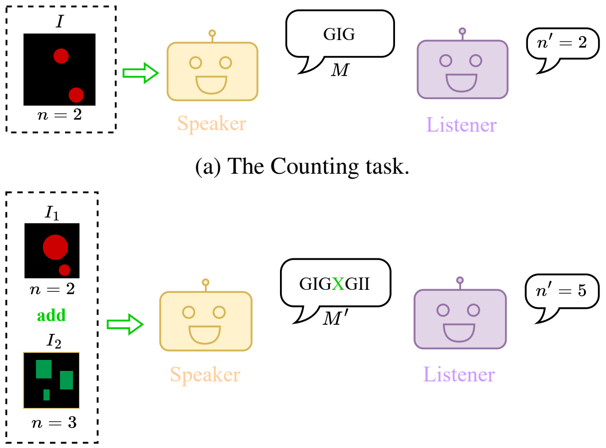
3 Environment

Based on the referential game, we propose a new game called NumGame, where the agents are required to communicate the number concept to complement the game. Additionally, we have developed a new dataset called NumWorld Dataset to evaluate the agents' performance. In this section, we will introduce the NumGame and the NumWorld Dataset.

3.1 NumGame

Figure 1 illustrates the basic setup of NumGame. In the NumGame, there are two agents involved: a speaker S and a listener L. The objective of the game is Counting or Calculating the number of objects in the images through communication and cooperation between agents.

Figure 1a shows the Counting task, where the speaker is presented with an image denoted as I, containing n objects of the same category, where $n \in N$. Subsequently, the speaker generates a message $M = \{m_1, ..., m_l\}$ based on the quantity of the objects in the image. Specifically, M is a sequence of l discrete symbols, where each symbol m_i is a one-hot vector of size v, and v is the size of vocabulary V. We regard the message M as an emergent number like natural numbers in human language. The listener, on the other hand, receives the message M and uses it to make an



(a) The Counting task.

(b) The Calculating task.

Figure 1: The NumGame tasks, requiring the agents Counting or Calculating the quantities of objects.

informed guess denoted as n' about the number of objects in the image I. If the listener's guess aligns with the actual quantity, i.e., $n = n'$, the game is considered successful; otherwise, it is deemed as failed.

The Calculating task is similar to the Counting task, with the main difference being that it requires the agents to perform arithmetic calculations on the quantities represented by two images. Figure 1b illustrates the process of the agents collaborating to complete an *add* task. The speaker is presented with two images and an arithmetic calculations symbol, then generates a message M' describing this arithmetic expression and passes it to the listener. The structure of the listener is the same as that of the Counting task – it needs to deduce the final result of the calculation from the message.

We focus on the generalization ability of the agents in the NumGame. Specifically, we are interested in the agents' ability to generalize to unseen quantities. To this end, we will train the agents on a subset of N and evaluate their performance on the remaining unseen quantities.

3.2 NumWorld Dataset

The NumWorld Dataset is developed based on the ShapeWorld dataset (Kuhnle and Copestake 2017), which serves as a synthetic dataset for visual reasoning. Within the NumWorld dataset, each sample is represented as a tuple (I, n), where I is an image containing n objects of the same category, all set against a black background. The image's resolution is 128 × 128, and the quantity n varies from 1 to 32. Each object in the dataset possesses 2 controllable attributes: *Shape* and *Color*. Both attributes have 5 possible values, and these attribute values jointly determine the category of the object. Additionally, the objects' locations and orientations are randomly generated within each image and are non-overlapping to ensure unambiguous counting. Moreover, as the number of objects increases the size of the objects diminishes, ensuring that the total pixel area remains approximately constant across all images.

The collection of all possible values for n is denoted as $\mathbf{N} = \{1, 2, ..., 32\}$. Importantly, the quantity n encompasses distinct ranges across diverse training and testing stages. We have defined three distinct sub-datasets: **Sen**, **Lang**, and **OOD**, each encompassing distinct quantity ranges for specific purposes (refer to Section 6.1 for comprehensive information).

4 Method

Drawing inspiration from the human learning process of numbers, we propose a two-stage training approach comprising NumSen and NumRel in this section.

4.1 NumSen: Pretrain the Number Sense

Number sense is a crucial ability for humans to approximate the number of objects even before language acquisition. As a result, we believe it is essential to pretrain the speaker's number sense before initiating language training.

We formulate the number sense pretraining as a vision-only process for the speaker. The vision encoder of the speaker takes an image I as input and generates a feature vector f. Subsequently, a projection head is employed to predict the quantity n of objects present in the image. Following pretraining, the vision encoder of the speaker will be utilized in the language training phase, while the projection head will be discarded.

Based on previous linguistics research (Wiese 2003), human number sense exhibits distinct responses to smaller quantities (subitizing, typically less than or equal to 4) and larger quantities (magnitude estimation, usually greater than 4). As a result, we divide the quantity n used for pretraining into two segments: $n \leq 4$ and $n > 4$. For the $n \leq 4$ segment, we employ all possible quantities $N'_0 = 1, 2, 3, 4$ to train the speaker's subitizing ability. For the $n > 4$ segment, we use a subset $N''_0 = 8, 16, 32$ to train the ability to recognize larger quantities. The choice of using only powers of 2 for training the magnitude estimation ability is motivated by our desire for the agent's number sense to closely resemble that of humans, which typically cannot precisely recognize all larger numbers. Consequently, the quantity n used for pretraining is $N_0 = N'_0 \cup N''_0 = 1, 2, 3, 4, 8, 16, 32$.

4.2 NumRel: Learn Relations between Numbers

Language is a powerful tool for humans to communicate about numbers. Based on this, we also train the agents to use language (emerging from communication) to communicate numerical concepts. If we only train the agents to communicate a single quantity, then each quantity would be essentially treated as a classification label, and the speaker does not need to understand the actual meaning of the numbers or the relations between them. Consequently, training the agents in this manner would not lead to a genuine understanding of the numerical concepts.

Considering how humans learn numbers, simple calculations (e.g., addition and subtraction) play a crucial role in fostering a better understanding of numerical concepts. To address this, we propose a novel approach called **NumRel**

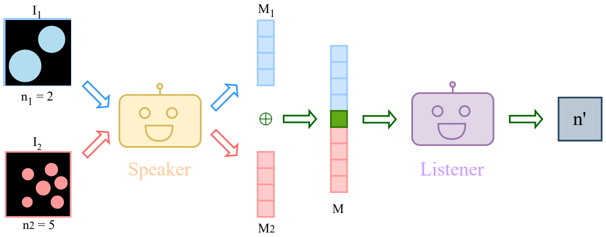


Figure 2: The NumRel training process. The speaker receives two images I_1 and I_2 and generates corresponding M_1 and M_2. These messages and an operator o are then concatenated to form the final message M. The listener receives this message and outputs a guess, n', representing the result of the operation $n_1\ o\ n_2$.

to train the agents. The NumRel method involves performing simple arithmetic calculations within a specified quantity range during training, which aids the agents in understanding the relations between numbers and then grasping the numerical concepts.

Figure 2 illustrates the NumRel setup. In NumRel, two samples, (I_1, n_1) and (I_2, n_2), are randomly selected from the original dataset and combined to form a single sample denoted as (I_1, I_2, n_1, n_2). A random operator o is then chosen from a predefined set of operators, and the target number n is calculated as the result of the operation $n_1\ o\ n_2$. Consequently, the sample is further represented as $(I_{1,2}, n_1\ o\ n_2)$. The speaker generates two distinct messages, M_1 and M_2, corresponding to I_1 and I_2, respectively. These messages and the operator o are concatenated to form the final message M. Subsequently, the listener receives the message M and outputs a guess denoted as n' concerning the result of the operation $n_1\ o\ n_2$.

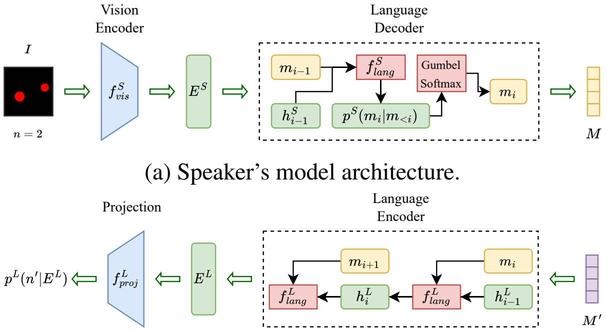
It is essential to note that the result n of the operation $n_1\ o\ n_2$ shares the same range as the original quantities n_1 and n_2. This design ensures that the Out-of-Distribution (OOD) test remains equitable and fair.

5 Model

As shown in Figure 3, the entire model consists of two components: speaker and listener. In NumGame G, the speaker takes an image I as input and generates a conditional distribution over messages $p^S(M|I)$, and the listener takes the message M and outputs a distribution over quantities $p^L(n'|M)$. In the following, we will introduce the architecture of each component and the optimization method.

Speaker. The speaker takes an image I as input and encodes it into an embedding E^S using a ResNet-50 (He et al. 2016) vision encoder f^S_{vis} , i.e., $E^S = f^S_{vis}(I)$. Then, a GRU (Chung et al. 2014) message decoder f^S_{lang} takes embedding E^S as initial hidden state h^S_0 to generate a sequence of distribution over tokens $p^S(M|h^S_0) = \prod_i p^S(m_i|m_{<i})$. The discrete message M is sampled from the distribution.

Listener. The listener takes the message M as input and encodes it into a message embedding E^L using a GRU message encoder f^L_{lang}, where E^L is the last hidden state $h^L_{|M|}$.



(a) Speaker's model architecture.

(b) Listener's model architecture.

Figure 3: The model architecture of the speaker and listener. M is a message corresponding to the image input I. M' can be a single message or a message-operator-message concatenation generated by the speaker.

The embedding E^L is then fed into a projection module f^L_{proj} to generate a distribution over quantities $p^L(n'|E^L) = Softmax(f^L_{proj}(E^L))$.

Optimization. Because the number of objects n is discrete, we model the task of predicting the quantity n as a classification problem. The listener's output $p^L(n'|M)$ is a distribution over all possible quantities $n' \in \{0, 1, 2, ..., N\}$, where N is the maximum number of objects. The model's parameters are optimized by maximizing the likelihood of the correct quantity n given the image I.

However, if we treat n merely as a one-hot categorical label, just as it's done in standard classification tasks, the agents would not understand the numerical concepts effectively. The reason is that the one-hot label cannot reflect the semantic similarity between different numbers which is a core aspect of numerical concepts. For example, the label $n = 1$ is semantically more similar to $n = 2$ than to $n = 10$. To address this problem, we use the **soft label** technique (Diaz and Marathe 2019) to train the listener. It converts the original label n into a soft probability distribution $\widetilde{\mathbf{n}} = SoftLabel(n) = (s_0, s_1, .., s_N)$, where N is the maximum number and

$$s_i = \frac{e^{-\phi(i,n)}}{\sum_{j=0}^{N} e^{-\phi(j,n)}} \tag{1}$$

$\phi(i, n)$ is a distance function between the number i and n, which could be absolute or squared distance. In conclusion, the final form of the loss is:

$$\mathcal{L}(G) = -E[\log p^L(\widetilde{\mathbf{n}}|M)], m \sim p^S(M|I) \tag{2}$$

Now, this formulation could be regarded as an ordered discrete regression problem, which is consistent with our quantity prediction task.

To make the training process end-to-end differentiable, we use the Gumbel-Softmax relaxation (Jang, Gu, and Poole 2017) with a temperature τ, which is a continuous relaxation of the discrete distribution during training. When testing, we use the $argmax$ function to get the discrete distribution from the continuous distribution.

6 Experimental Setup

Our model implementation and training process are based on the PyTorch (Paszke et al. 2019) framework and partially adapted from the EGG (Kharitonov et al. 2019) toolkit.

6.1 Dataset Preparation

Based on the ShapeWorld dataset (Kuhnle and Copestake 2017), we generate the NumWorld dataset, which consists of three sub-datasets with different ranges of quantities for different training stages:

- **Sen.** $N_0 = \{1, 2, 3, 4, 8, 16, 32\}$. The Sen dataset is used to train the number sense of the speaker. The range of n is N_0. The reason for setting N_0 in this way is explained in Section 4.1. Because the Sen is used to pretrain the vision module to recognize each distinct quantity, there are only images and the corresponding labels in it.
- **Lang.**. $N_1 = N \setminus \{10, 13, 15, 18, 20, 23, 25, 28, 30\}$. During the stage of using language, we only use the subset N_1 of N to train agents to communicate numerical concepts through discrete language. There are images, labels, and additional calculation expressions in it.
- **OOD.** $N_2 = \{10, 13, 15, 18, 20, 23, 25, 28, 30\}$. We use the remaining subset N_2 of N to evaluate the agents' generalization ability on OOD tasks. Notably, the quantities in N_2 are not seen in the two training stages, i.e., $N_2 \cap N_0 = \emptyset$ and $N_2 \cap N_1 = \emptyset$. Considering zero-shot learning may be more challenging, we give the agents fewer samples with $n \in N_2$ to train on before evaluation, i.e., few-shot learning. There are also images, labels, and calculation expressions in it.

Each sub-dataset contains both a training set and a validation set, which is used to select the best model for the next training stage or evaluate the model's generalization performance. The calculation expressions are symbolic, with a format of $(n_1, n_2, o, n_1\ o\ n_2)$, where o is randomly selected from the set $\{add, sub, max\}$. Each expression $n = n_1\ o\ n_2$ represents an arithmetic relation o between two quantities n_1 and n_2 just like the formulation in (Bahdanau et al. 2018). The agent will randomly select images within the dataset based on the expression. Therefore, the total number of images the agents see remains constant regardless of whether calculations are performed. In addition, the calculations are closed, meaning the range of results matches the range of corresponding operands to ensure genuine testing of the agents' out-of-distribution generalization ability.

In each dataset, the frequency of (1) each attribute value of the objects, (2) each quantity within the given range, and (3) each possible calculation expression are uniform, ensuring that the agents are trained and tested on a balanced dataset. Table 1 shows the detailed statistics of the three sub-datasets.

6.2 Evaluation Metrics

We focus on agents' ability to complete different tasks through their language and the regularity exhibited in the emergent language. These two aspects correspond to the functionality and semantics of the emergent language.

Sub-dataset	#Images		#Exprs	
	Train	Test	Train	Test
Sen	20k	5k	-	-
Lang	20k	5k	80k	20k
OOD	1k	5k	0.3k	20k

Table 1: Sizes of each dataset. #Exprs is the number of calculation expressions.

Generalization ability. We use the classification accuracy on the test set of each sub-dataset to evaluate the agents' generalization ability, which is divided into two aspects: in-distribution(ID) generalization and out-of-distribution(OOD) generalization. The ID generalization refers to the agents' ability to generalize over unseen images containing **seen** quantities, where the accuracy is reported on Lang's validation dataset during the training process. The OOD generalization, which is the focus of our attention, refers to the ability to generalize over unseen images containing **unseen** quantities, where the accuracy is evaluated on the OOD test dataset after a few-shot learning.

Regularity. Regularities play a pivotal role in language. According to (Smith and Wonnacott 2010), regularity pertains to the level of certainty of the statement given a meaning. In this context, stronger language regularity signifies a more direct association between meanings and statements. For our study, each image yields a (n, M) pair, where n serves as the meaning, denoting the object quantity, and M represents the generated message by the speaker, treated as the statement. This yields paired sets, $(\mathbb{N}, \mathbb{L})$, where $\mathbb{N}$ encompasses all conceivable meanings, and $\mathbb{L}$ encompasses the entire emergent language. We ignore the internal message structure and consider it a distinct symbol when evaluating regularity between $\mathbb{N}$ and $\mathbb{L}$. Similar to the approach employed in (Mu and Goodman 2021), we utilize the Adjusted Mutual Information (AMI) (Vinh, Epps, and Bailey 2010) between $\mathbb{N}$ and $\mathbb{L}$

$$\text{AMI}(\mathbb{N}, \mathbb{L}) = \frac{I(\mathbb{N}, \mathbb{L}) - \mathbb{E}[I(\mathbb{N}, \mathbb{L})]}{\max(H(\mathbb{N}), H(\mathbb{L})) - \mathbb{E}[I(\mathbb{N}, \mathbb{L})]}, \qquad (3)$$

to measure the regularity of the emergent language, where $I(\mathbb{N}, \mathbb{L})$ is the mutual information between $\mathbb{N}$ and $\mathbb{L}$, $H(\mathbb{N})$ and $H(\mathbb{L})$ are the entropy of $\mathbb{N}$ and $\mathbb{L}$, respectively, and $\mathbb{E}[I(\mathbb{N}, \mathbb{L})]$ is the expected mutual information between $\mathbb{N}$ and $\mathbb{L}$. In contrast to mutual information, AMI offers greater resilience against the influence of the number of unique messages and corrects the agreement's effect solely due to chance between $\mathbb{N}$ and $\mathbb{L}$. The AMI ranges from 0 to 1, with 1 indicating a perfect match between $\mathbb{N}$ and $\mathbb{L}$.

6.3 Training Details

We perform hyperparameter tuning on a small validation set to select the best model for each training stage. Regarding the discrete communication channel connecting the speaker and listener, we set the maximum message length $|M| = 3$ and the vocabulary size $|V| = 16$, sufficient for the speaker

Method		Acc	AMI
NoRel	w/o NumSen	0.82(0.05)	0.90(0.01)
	w/ NumSen	**0.84(0.04)**	**0.92(0.01)**
NumRel	w/o NumSen	0.78(0.04)	0.89(0.02)
	w/ NumSen	**0.81(0.05)**	**0.91(0.02)**

Table 2: Accuracy and AMI on the validation set in the second stage with different training methods.

to express all the possible quantities. The speaker and listener are trained with the AdamW optimizer (Loshchilov and Hutter 2018). The learning rates vary across different training stages, and simultaneously, different sub-modules within the model also have distinct learning rates. To prevent overfitting, we employ various techniques such as learning rate warmup, learning rate decay, and weight decay (details in Appendix B). The temperature τ in the Gumbel-Softmax also decays from 2.0 to 0.1 with a rate of 0.9.

7 Results

In this section, we demonstrate by experimental results that the proposed two-stage training method can help agents: 1) accurately represent seen (ID) quantities (Sec 7.1); 2) count unseen (OOD) quantities (Sec 7.2); 3) calculate on unseen (OOD) quantities (Sec 7.3); 4) emerge languages to capture the order relation between quantities (Sec 7.4). All the results, `mean(std)`, are derived with 6 random seeds.

7.1 Two-Stage Training

NumSen. The first stage involves number sense pretraining, which trains the speaker's visual module using the NumSen method. As mentioned earlier, this stage employs the numerical range N_0, but in fact, we also experimented with other numerical ranges N_0', e.g., Fibonacci sequence and some smaller ranges. We ultimately choose N_0 because it offers the best performance in the final OOD test (see Appendix A.1). In addition, pretraining can also be skipped in this stage, in which case the subsequent training stages would start from scratch and be unrelated to NumSen.

NumRel. In the second stage, we train the two agents to communicate quantities within a limited range on the Lang dataset. There are two training paradigms: (1) 'NoRel', where agents communicate only a single quantity without any calculation; (2) 'NumRel', where agents, in addition to recognizing quantities, also perform arithmetic calculations. Both training paradigms can use the previously pretrained parameters of NumSen or start training from scratch. Table 2 presents the performance of these two training paradigms on their respective validation sets. It's evident that using NumSen pretraining yields better results in ID generalization. It's important to note that a direct comparison between these two paradigms through the validation accuracy in Table 2 is infeasible because they are different pretext tasks. In subsequent experiments, we will compare the performance of these two methods through the OOD test.

Method	Acc	AMI
w/o training	0.14 (0.02)	0.06 (0.01)
Base	0.71 (0.05)	0.70 (0.01)
+NumSen	0.72 (0.07)	0.76 (0.01)
+NumRel	0.94 (0.03)	0.82 (0.03)
+NumSenRel	**0.97 (0.01)**	**0.86 (0.03)**

Table 3: Agents' generalization ability (accuracy, higher is better) and the emergent language's regularity (adjusted mutual information score, higher is better) on unseen (OOD) quantities on the counting task.

In this training stage, we also experimented with variations of the two paradigms above as baselines for subsequent OOD tasks, such as removing compositional inductive bias in NumRel (Appendix A.2) and using different label forms (Appendix A.3).

7.2 Counting

Table 3 illustrates the influence of our approach on the agents' capacity to identify and communicate unseen quantities. Here, 'w/o training' signifies the absence of agent training (representing the performance lower bound), 'Base' signifies the speaker conveying one quantity at a time without calculation when training, '+NumSen' or '+NumRel' signifies the utilization of the respective stage from the two-stage method for agent training, and '+NumSenRel' signifies the complete two-stage training method. From top to bottom of the table, both Acc and AMI exhibit a monotonically increasing trend, which aligns with our expectations and validates the effectiveness of our approach.

From a holistic perspective, our complete two-stage method achieves the best results in terms of Acc and AMI. It can generalize to unseen quantities with both stable ($AMI = 0.86$) semantics and high ($Acc = 0.97$) accuracy, significantly surpassing the original 'Base' training method. It is noteworthy that even though we allow the agents to learn from a few samples before OOD testing (few-shot), without any prior training on the Lang dataset, the agents cannot establish successful communication ($Acc = 0.14$).

From a disaggregated perspective, the use of NumRel (last two rows of Table 3) yields significantly better results compared to not using it (rows 2 and 3 of Table 3), with accuracy ~ 0.9 vs. ~ 0.7. This strongly indicates that compelling agents to learn calculations of numbers contributes to a better understanding and generalization of numerical concepts. Furthermore, regardless of whether calculations are involved, using NumSen pretraining improves performance (Acc 0.97 vs. 0.94, 0.72 vs. 0.71). Although the improvement from NumSen in accuracy is not substantial, it's understandable since the number sense in the early stage itself is quite vague and limited. Similar to humans, who primarily rely on language and calculation after birth to continuously enhance their cognitive capacity in numerical concepts, the early-stage number sense is useful but insufficient.

In addition, to compare different methods in a statistical sense, we also conduct Kolmogorov-Smirnov tests (Massey

Method	Acc	AMI
w/o training	0.27 (0.02)	0.02 (0.01)
Base	0.18 (0.07)	0.64 (0.06)
+NumSen	0.20 (0.07)	0.70 (0.03)
+NumRel	0.79 (0.05)	0.81 (0.04)
+NumSenRel	**0.83 (0.04)**	**0.83 (0.04)**

Table 4: Agents' generalization ability and the language's regularity on unseen quantities on the calculating task.

1951) on the Acc and AMI results in Table 3, representing the p-value as $KS(\text{F} \leq \text{G}) = (p_{Acc}, p_{AMI})$. We get

- $KS(\text{Base} \leq \text{+NumRel}) = (0.001, 0.07)$
- $KS(\text{Base} \leq \text{+NumSen}) = (0.54, 0.01)$
- $KS(\text{+NumRel} \leq \text{+NumSenRel}) = (0.07, 0.07)$

It also shows that (1) NumRel has a significant impact on Acc and AMI, (2) NumSen is more helpful in generating clear semantics than improving accuracy.

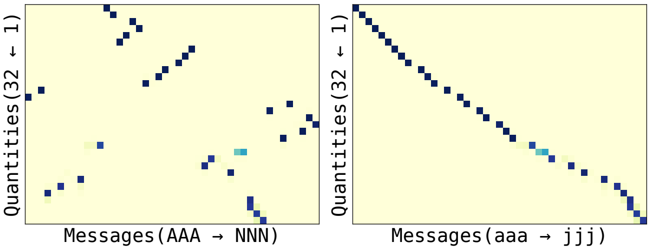
7.3 Calculating

Table 4 presents the performance of our approach on the calculating task. As mentioned earlier, each table row corresponds to different training methods. Similar to Table 3, the increasing trend from top to bottom in Table 4 also demonstrates that our approach significantly aids agents in performing calculations on unseen quantities. It's reasonable that the task performance on calculating is lower than on counting, as performing calculations is inherently more challenging than simple counting.

In line with the counting task, our two-stage method also demonstrates superior performance in the calculating task, achieving the highest accuracy and AMI. While the training methods with NumRel don't help the agents perform the calculating task perfectly, the $Acc \sim 0.8$ and $AMI \sim 0.8$ (further error analysis can be found in Appendix C) are sufficient to indicate that effective communication concerning arithmetic calculations has been established. After all, calculating is more challenging. In contrast, when NumRel is omitted, agents struggle with effective communication ($Acc \sim 0.2$) as they hadn't learned how to perform calculations during previous training, even with the few-shot learning of a small number of calculation samples before OOD testing. Although the AMI without NumRel training might exceed 0.6, it merely indicates the agents can articulate individual numbers without comprehending their interrelations. Notably, the accuracy of 'Base' is lower than 'w/o training'(0.18 vs. 0.27), which also shows that 'Base' might lead the agents to overfit individual numbers while neglecting the relations between numbers. NumSen still plays a role, but without NumRel, it cannot facilitate effective generalization on its own. This reiterates the limitations of number sense and the necessity of calculations.

7.4 Semantic Analysis

We randomly selected a seed and fine-tuned the '+NumSenRel' pre-trained model on the entire numerical range $(1 \sim 32)$ for the counting task. Then, we visualize the distributions $P(M|n)$ of agents' messages given quantities with a heatmap to analyze the structure and semantics of the emerged language.



(a) Original order. (b) Remapped order.

Figure 4: Visualization of the distributions $P(M|n)$ of a randomly selected seed. The y-axis represents the quantity n, and the x-axis represents the message M. The darker the blue color means the higher the probability.

Original	G	I	H	A	N	D	P	L	C	M
Remapped	a	b	c	d	e	f	g	h	i	j

Table 5: The mapping between two orders of the tokens.

As shown in Figure 4, there is almost a one-to-one mapping between quantities and messages, which indicates that each quantity can be accurately identified and represented. Furthermore, we find that messages near numbers often have similar messages under the original arbitrary lexicographic order (Figure 4a, in capital letters). To better understand the intrinsic structure of the agent's language, we infer the agents' lexicographic order (with Table 5) and remap the messages (details about the inference process in Appendix D). As shown in Figure 4b, under the agents' order, the messages closely mirror the quantity order. The results indicate that the language captures the order relation between quantities well. Such orderly encoding is likely one of the sources for the agents' ability to generalize the numerical concepts.

8 Conclusion

In this work, we focus on the emergence of numerical concepts and propose a novel two-stage training approach to facilitate the agents' ability to generalize numerical concepts in the multi-agent communication setting. We demonstrate that the proposed method enables agents to achieve high accuracy and AMI in communicating about unseen quantities and performing arithmetic calculations. Furthermore, the language emergence from the communication exhibits a solid order relation. Our work provides a new perspective on the study of numerical concepts. In future work, we will further enhance the agents' ability to perform more complex calculations and delve deeper into studying the encoding patterns of numbers within the emergent language.

Acknowledgments

This work is partially supported by the National Key R&D Program of China(under Grant 2021ZD0110102), the NSF of China(under Grants 61925208, 62102399, 62222214, U22A2028), CAS Project for Young Scientists in Basic Research(YSBR-029), Youth Innovation Promotion Association CAS and Xplore Prize.

References

Bahdanau, D.; Murty, S.; Noukhovitch, M.; Nguyen, T. H.; de Vries, H.; and Courville, A. 2018. Systematic Generalization: What Is Required and Can It Be Learned? In *International Conference on Learning Representations*.

Berwick, R. C.; and Chomsky, N. 2016. *Why Only Us: Language and Evolution*. The MIT Press. ISBN 978-0-262-53349-2.

Butterworth, B. 2005. The Development of Arithmetical Abilities. *J Child Psychol Psychiatry*, 46(1): 3–18.

Chaabouni, R.; Kharitonov, E.; Dupoux, E.; and Baroni, M. 2019. Anti-Efficient Encoding in Emergent Communication. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chaabouni, R.; Strub, F.; Altché, F.; Tarassov, E.; Tallec, C.; Davoodi, E.; Mathewson, K. W.; Tieleman, O.; Lazaridou, A.; and Piot, B. 2022. Emergent Communication at Scale. In *International Conference on Learning Representations*.

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

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. arxiv:1412.3555.

Conant, L. L. 1896. *The Number Concept: Its Origin and Development*. Macmillan and Company.

Conklin, H.; and Smith, K. 2023. Variable Compositionality Reliably Emerges in Neural Networks. In *International Conference on Learning Representations*.

Coolidge, F. L.; and Overmann, K. A. 2012. Numerosity, Abstraction, and the Emergence of Symbolic Thinking. *Current Anthropology*, 53(2): 204–225.

Dehaene, S. 2011. *The Number Sense: How the Mind Creates Mathematics, Revised and Updated Edition*. New York: Oxford University Press, updated edition edition. ISBN 978-0-19-975387-1.

Diaz, R.; and Marathe, A. 2019. Soft Labels for Ordinal Regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4733–4742. Long Beach, CA, USA: IEEE. ISBN 978-1-72813-293-8.

Drucker, C. B.; and Brannon, E. M. 2014. Rhesus Monkeys (Macaca Mulatta) Map Number onto Space. *Cognition*, 132(1): 57–67.

Feigenson, L.; Dehaene, S.; and Spelke, E. 2004. Core Systems of Number. *Trends Cogn Sci*, 8(7): 307–314.

Feng, Y.; An, B.; and Lu, Z. 2023. Learning Multi-Object Positional Relationships via Emergent Communication. arxiv:2302.08084.

Gelman, R.; and Gallistel, C. R. 1986. *The Child's Understanding of Number:*. Cambridge, MA: Harvard University Press. ISBN 978-0-674-11637-5.

Guo, S.; Ren, Y.; Havrylov, S.; Frank, S.; Titov, I.; and Smith, K. 2019. The Emergence of Compositional Languages for Numeric Concepts Through Iterated Learning in Neural Agents. arXiv.

Hauser, M. D.; Carey, S.; and Hauser, L. B. 2000. Spontaneous Number Representation in Semi-Free-Ranging Rhesus Monkeys. *Proc Biol Sci*, 267(1445): 829–833.

Hauser, M. D.; Chomsky, N.; and Fitch, W. T. 2002. The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, 298(5598): 1569–1579.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hiraiwa, K. 2017. The Faculty of Language Integrates the Two Core Systems of Number. *Frontiers in Psychology*, 8.

Jang, E.; Gu, S.; and Poole, B. 2017. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*.

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. Hong Kong, China: Association for Computational Linguistics.

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.

Kuciński, Ł.; Korbak, T.; Kołodziej, P.; and Miłoś, P. 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, 23075–23088. Curran Associates, Inc.

Kuhnle, A.; and Copestake, A. 2017. ShapeWorld - A New Test Methodology for Multimodal Language Understanding. arxiv:1704.04517.

Lazaridou, A.; Peysakhovich, A.; and Baroni, M. 2016. Multi-Agent Cooperation and the Emergence of (Natural) Language.

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

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Massey, F. J. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, 46(253): 68–78.

Mu, J.; and Goodman, N. D. 2021. Emergent Communication of Generalizations. In *NeurIPS*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Pica, P.; Lemer, C.; Izard, V.; and Dehaene, S. 2004. Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science*, 306(5695): 499–503.

Rita, M.; Chaabouni, R.; and Dupoux, E. 2020. “LazImpa”: Lazy and Impatient Neural Agents Learn to Communicate Efficiently. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 335–343. Online: Association for Computational Linguistics.

Smith, K.; and Wonnacott, E. 2010. Eliminating Unpredictable Variation through Iterated Learning. *Cognition*, 116(3): 444–449.

Tucker, M.; Levy, R. P.; Shah, J.; and Zaslavsky, N. 2022. Trading off Utility, Informativeness, and Complexity in Emergent Communication. In *Advances in Neural Information Processing Systems*.

Vinh, N. X.; Epps, J.; and Bailey, J. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11(95): 2837–2854.

Wiese, H. 2003. *Numbers, Language, and the Human Mind*. Cambridge University Press. ISBN 978-0-521-83182-6.

Wiese, H. 2007. The Co-Evolution of Number Concepts and Counting Words. *Lingua*, 117(5): 758–772.

Xu, Z.; Niethammer, M.; and Raffel, C. 2022. Compositional Generalization in Unsupervised Compositional Representation Learning: A Study on Disentanglement and Emergent Language. arXiv.