EXPRESSIVITY OF EMERGENT LANGUAGES IS A TRADE-OFF BETWEEN CONTEXTUAL COMPLEXITY AND UNPREDICTABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

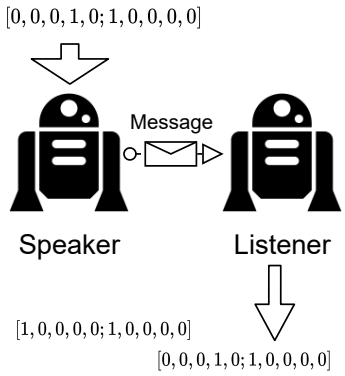
Researchers are using deep learning models to explore the emergence of language in various language games, where simulated agents interact and develop an emergent language to solve a task. We focus on factors which determine the *expressivity* of emergent languages, which reflects the amount of information about input spaces those languages are capable of encoding. We measure the expressivity of emergent languages based on their generalisation performance across different games, and demonstrate that the expressivity of emergent languages is a trade-off between the complexity and unpredictability of the context those languages are used in. Another novel contribution of this work is introducing a contrastive loss into the learning methods on referential games. We show that using our contrastive loss alleviates the collapse of message types seen using standard referential loss functions.

1 INTRODUCTION

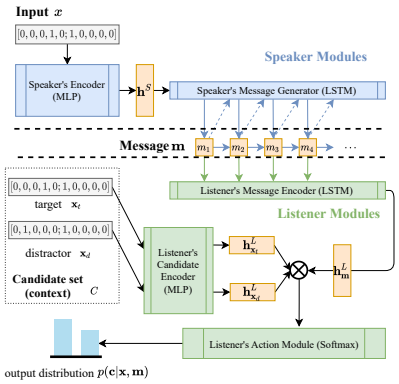
Language games were first introduced by Wittgenstein (1954) to explore the acquisition of meanings of language utterances. Instantiating this concept with the signalling game design from Lewis (1969) enables linguists to explore the emergence of linguistic structure (Kirby, 2001; Kirby & Hurford, 2002) where artificial languages are represented as symbolic systems. The success of deep learning (DL) models on complicated cognitive tasks (Krizhevsky et al., 2012; LeCun et al., 2015; Silver et al., 2016) then inspired researchers to apply DL-based models to language games for investigating agents' ability to invent communication protocols without preset rules.

In language games implemented in existing works (e.g. Lazaridou et al., 2018; Graesser et al., 2019; Zhang et al., 2019; Lee et al., 2018; Yuan et al., 2020), there are usually two types of agents, *speakers* that emit messages based on their observations (i.e. target objects) and *listeners* that receive messages and take actions accordingly, as illustrated in Figure 1. Based on the goals for listeners, we can categorise most of the games into the following three types: 1) *referential* games where listeners need to select the target object observed by the speaker among a set of candidate objects (candidates for short) (e.g. Lazaridou et al., 2017; Ren et al., 2020); 2) *reconstruction* games where listeners need to reconstruct the speaker's observation (e.g. Chaabouni et al., 2020; Kharitonov et al., 2020); and 3) *navigation* games where listeners need to go to the location specified by the speaker (e.g. Lowe et al., 2017; Kajić et al., 2020). We focus on referential games, illustrated in Figure 1a, which have been well investigated in both the linguistic community (e.g. Winters et al., 2018; Kirby et al., 2014) and emergent communication community (e.g. Lazaridou et al., 2018; Li & Bowling, 2019).

It is reasonable to assume that listeners in referential games can differentiate the target object from the other distractors in the context as long as the speaker's message encodes some unique feature of the target. Therefore, the speaker's messages must convey different amount of information for listeners to complete their task in games with candidates that are similar in different degree. For example, to distinguish two very similar objects in the candidates set, e.g. blue wide-stripped shirt and blue narrow-stripped shirt, it requires more information about the target to be encoded by the speaker's message than to distinguish a shirt from a cup. Experiments with human participants show that emergent languages are sensitive to the requirements of the communicative tasks for which they are used, with languages developing in which only the necessary discriminatory information is encoded (Winters et al., 2018). Following Kirby et al. (2015), we refer to such ability to encode information



(a) Illustration of the referential game we investigate in this work. Observations are concatenations of two 5-digits-long one-hot vectors. The objective of the listener in to select among a set of candidate inputs (the *context*).



(b) Illustration of the architecture of agents we implement for referential games. Blue modules are components of the speaker, and green modules belong to the listener. Orange and grey indicate the vectors generated by agents and input samples respectively.

Figure 1: Diagram of referential game and the architecture of our agents.

about input space (e.g. structure, dimensionality, distribution of samples) as **expressivity**. In this work, we explore the factors that influence expressivity in the framework of DL-based language games, and we argue that it is important to understand these factors such that agents could develop languages that are useful in general.

Our contribution in this work is threefold. First, we propose and verify a hypothesis about the determining factors of expressivity of emergent languages under the DL-based framework. Following Winters et al. (2018), we show that context size (i.e. the number of candidates the listener must select among when interpreting the speaker's message) influence both the *complexity* and *unpredictability* of that context. Complexity refers to the similarity between the items in the context: in order for the listener to select the target, more information needs to be encoded in the speaker's signal if the context contains more similar distractors for a given target. Unpredictability refers to the extent to which the information that needs to be encoded by the speaker is stable across contexts in successive episodes of communication: in games where the information required by the listener differs from trial (training epoch for DL agents) to trial, the speaker needs to encode more information on *every* trial in order to be sure to encode the information the listener needs.[1] As we show in Section 3.1, complexity and unpredictability are affected differently by the context size: as context size increases, complexity increases but unpredictability decreases. Therefore, we propose and verify the following hypothesis about the determining factors of expressivity:

The expressivity of emergent languages is a trade-off between the complexity and the unpredictability of context in language games. [1]

Our second contribution is a novel measure of expressivity based on partial ordering of languages in terms of their generalisation performances across tasks. Although expressivity is related to the amount of information encoded in a language, we illustrate that mutual information (MI) is not an appropriate measurement for expressivity in Section 3.2. Considering that one of our objectives is to facilitate a universally useful emergent language, we propose to measure expressivity of a language based on the generalisation performance of agents trained with it across different games. Since it is more challenging to quantitatively measure it, we focus on the *partial ordering* between languages in this first work towards understanding it.

Our final contribution is the introduction of a contrastive loss into referential games. We modify the contrastive loss function proposed by Chen et al. (2020), and show that it is equivalent to the loss function applied by most of the existing works in DL-based emergent communication, (e.g Havrylov & Titov, 2017; Li & Bowling, 2019; Ren et al., 2020; Guo et al., 2020) while its space complexity (in the algorithmic sense) is lower. Our contrastive loss enables our approach to overcome the technical limitations imposed by GPU memory size on large-scale referential games. Meanwhile, we also show that this contrastive loss could greatly alleviate the collapse of message types, i.e. the number of unique messages from speakers is less than the number of inputs observed by speakers.

[1]Context, complexity, and unpredictability are defined and discussed in depth in Section 3.

2 GAMES, AGENTS AND LEARNING METHODS

2.1 GAME ENVIRONMENT AND ARCHITECTURE OF AGENTS

Environment Components As shown in Figure 1b, the environment in our referential games have the following components: 1) *input space* $\mathcal{X}$ from which the speaker's observations are drawn, which consists of $10,000$ possible inputs[2] where each $\boldsymbol{x} \in \mathcal{X}$ is a concatenation of 4 one-hot vectors whose length is 10; 2) *message space* $\mathcal{M}$ consists of 10^6 6-tokens-long sequences $\boldsymbol{m}$ (called as a *message type*) where the size of the token inventory is 10; note that the tokens are initially meaningless, and meaning emerges through the negotiation of a communication protocol between speaker and listener; 3) *candidate set* C which contains a target input $\boldsymbol{x}_t$ and $|C|-1$ distractors $\{\boldsymbol{x}_{d_1}, \boldsymbol{x}_{d_2}, \ldots, \boldsymbol{x}_{d_{|C|-1}}\}$ sampled uniformly at random from $\mathcal{X}$ without replacement.

Agents The architecture of the agents we implemented is also illustrated in Figure 1b: 4) *speaker* S consists of both a multi layer perceptron (MLP) for encoding the 40-digits-long target input $\boldsymbol{x}_t$ onto a 256-digits-long embedding $\boldsymbol{h}^S$, and a decoder for generating a message $\boldsymbol{m}$ based on long-short term memory (LSTM) network (Hochreiter & Schmidhuber, 1997). 5) *listener* L is constituted by an LSTM-based sequence encoder which encodes $\boldsymbol{m}$ into a 256-digits-long embedding $\boldsymbol{h}^L(\boldsymbol{m})$, and an MLP for encoding the inputs which can be denoted by $f^L_{cenc}(\cdot)$.

Definition of an emergent language: following e.g. Ren et al. (2020), we can then define an emergent language $\mathcal{L}$ as a mapping function from the input space $\mathcal{X}$ to the message space $\mathcal{M}$, i.e. $\mathcal{L} : \mathcal{X} \mapsto \mathcal{M}$, thus it can be represented as $\{(\boldsymbol{x}_i, \mathcal{L}(\boldsymbol{x}_i)) | \forall \boldsymbol{x}_i \in \mathcal{X}\}$. Note that $\mathcal{L}$ could be a non-injective mapping function, i.e. it is possible for two different inputs to be mapped to an identical message thus the number of message types might be less than the number of unique inputs.

2.2 LEARNING METHODS OF AGENTS

In the previous works (e.g. Havrylov & Titov, 2017), the action module of L is usually a $\mathrm{softmax}$ function that would produce a categorical distribution over the candidates based on the distance between $\boldsymbol{h}^L(\boldsymbol{m})$ and embeddings of each candidate $f^L_{cenc}(\boldsymbol{x}_i)$ obtained through another MLP-based candidate encoder. By comparing the loss function used in Havrylov & Titov (2017) and the contrastive loss function proposed by Hadsell et al. (2006), we find that the aim of both functions is the same, i.e. to make the distance between positive pairs ($\boldsymbol{x}_t$ and $\boldsymbol{m}$) closer while keeping the embedding for negative pairs ($\boldsymbol{x}_d$ and $\boldsymbol{m}$) as far apart as possible. To effectively validate our hypothesis, we need to make the whole input space $\mathcal{X}$ as candidate set C in which case the batch is too large to be stored in GPU memory (since the space complexity is $\mathcal{O}(|\mathcal{X}|^2)$). Therefore, to overcome this issue, we introduce a contrastive learning method similar to Chen et al. (2020) to referential games. Suppose that we have a batch of input samples $B = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_i, \ldots, \boldsymbol{x}_{|B|}\}$ and the corresponding messages from the speaker are $\{\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots, \boldsymbol{m}_i, \ldots, \boldsymbol{m}_{|B|}\}$ which would be further encoded into $\{\boldsymbol{h}^L_{\boldsymbol{m}_1}, \boldsymbol{h}^L_{\boldsymbol{m}_2}, \ldots, \boldsymbol{h}^L_{\boldsymbol{m}_i}, \ldots, \boldsymbol{h}^L_{\boldsymbol{m}_{|B|}}\}$ by the listener L. Then the loss function for a sample $\boldsymbol{x}_i$ is defined on a $\mathrm{softmax}$ function:

$$\ell_i = -\log \frac{\exp\left({\boldsymbol{h}^L_{\boldsymbol{m}_i}}^\top \cdot f^L_{cenc}(\boldsymbol{x}_i)\right)}{\sum_{j=1}^{|B|} \exp\left({\boldsymbol{h}^L_{\boldsymbol{m}_i}}^\top \cdot f^L_{cenc}(\boldsymbol{x}_j)\right)}. \tag{1}$$

In this way, the number of candidates $|C|$ is exactly the batch size $|B|$ during training.

As for updating the parameters, we use the Adam algorithm introduced by Kingma & Ba (2014), and the learning rate is set to 10^{-4}. To allow the gradients being propagated through the discrete channel to overcome the sampling issue of messages, we apply the Gumbel-Softmax trick proposed by Jang et al. (2017), and the temperature hyper-parameter τ is set to 1.0.

Our implementation[3] of games bases on the framework *EGG* developed by Kharitonov et al. (2019). The experiments across 6 random seeds, 18 source games, 11 target games took $4,216$ hours in total, on Nvidia Tesla P100.

[2] According to the results from Chaabouni et al. (2020), an input space containing $10,000$ possible inputs is large enough for neural models to achieve $> 99\%$ generalisation accuracy.

[3] Codes are provided in the supplementary materials.

3 CONTEXT AND EXPRESSIVITY

3.1 CONTEXT, AND ITS COMPLEXITY AND UNPREDICTABILITY IN DIFFERENT GAMES

Context: Communication between humans always takes place w.r.t. some context (Winters et al., 2018). For humans, that context includes: 1) features of the environment that can be directly perceived by the dialogue participants and used for comprehending an utterance (e.g. "cup" is interpreted as an object in the immediate physical context); and 2) information that cannot be directly perceived but can affect the interpretation of utterances (e.g. using and interpreting "cup" assumes some shared knowledge of what counts as a "cup"; once a particular cup has been established as salient in particular dialogue, further utterances of "the cup" or "it" will be interpreted as referring to that specific salient cup).

Inspired by this conclusion from pragmatics, in the DL-based language games, we define *context* as the *supervision information* involved in the calculation of losses, i.e. the context of a specific input $\boldsymbol{x}_t$, denoted as $\mathcal{C}(\boldsymbol{x}_t)$, is the space of samples involved in the calculation of loss. As for in referential games, the cross entropy loss is calculated based on the distance between the message embedding $\boldsymbol{h}^L(\boldsymbol{m})$ and candidate embedding $f^L_{cenc}(\boldsymbol{x}_i)$ where $\boldsymbol{x}_i \in C$, thus $\mathcal{C}(\boldsymbol{x}_t) = C \subseteq \mathcal{X}$ in referential games.

Complexity of context: We assume that the goal of communication is to distinguish a target object from other possibilities in the context (as defined above). It therefore follows that the similarity of distractors in the context to the target influences the communicative precision required, and that greater precision is required to distinguish the target from a more similar distractor. For example, it is relatively easy in natural language to distinguish e.g. a cat from a table (a relatively general label like "cat" or "animal" would suffice), but harder to make fine-grained distinctions between very similar objects e.g. a Scottish Fold Cat and an American Shorthair Cat (a specialist vocabulary or a lengthy description is necessary).

Following the assumption (described verbally above) that a context which contains more similar objects makes the game harder because there are fewer unique features which distinguish the target from the distractors, we first define a neighbour set in k-th degree of a given target $\boldsymbol{x}_t$ as $\mathcal{N}_k(\boldsymbol{x}_t) = \{\boldsymbol{x}_j : d(\boldsymbol{x}_t, \boldsymbol{x}_j) \leqslant k\}$ where d is a distance function that can properly capture the similarity between inputs, e.g Hamming distance in our setting. The complexity of $\mathcal{C}(\boldsymbol{x}_t)$ is then defined as the expectation of the probability that $\mathcal{C}(\boldsymbol{x}_t)$ contains an object from $\mathcal{N}_k(\boldsymbol{x}_t)$, i.e.

$$\mathbb{E}_{\boldsymbol{x}_t}\left[\frac{1}{|\mathcal{C}(\boldsymbol{x}_t)|}\sum_{\boldsymbol{x}_j}\mathbb{I}\left(x_j \in \mathcal{N}_k(\boldsymbol{x}_t)\right)\right]. \tag{2}$$

In our referential games, since the sampling procedure is independent Bernoulli without replacement, the value of the above expectation is then $1 - \left(\frac{|\mathcal{X}|-|\mathcal{N}_k(\boldsymbol{x}_t)|}{|\mathcal{X}|}\right)^{|C|}$ which is a monotonically increasing function w.r.t $|C|$ and a fixed k (i.e. larger contexts are more complex since they are more likely to include items which are more similar to the target).

Unpredictability of context: Our notion of unpredictability comes from experimental work with human participants, e.g. Winters et al. (2018). Suppose the aim is to distinguish a striped t-shirt from a number of distractors, and there are two sequences of context: i) three runs of games where distractors are all cups; ii) three runs of games where distractors are a cup, a plain t-shirt, and a pencil. In the first sequence of games, participants would become certain that "t-shirt" is enough for distinguishing the target, whereas in the second sequence participants would learn that a more overspecified utterance (e.g. "striped t-shirt") is required to guarantee comprehension after a failure on the trial involving the plain t-shirt distractor. That is, the context in the first sequence is more predictable than the second. Winters et al. (2018) show that human participants are sensitive to this kind of unpredictability, and adapt their communicative strategies accordingly.

In DL-based games, we refer to the context at the e-th trial, i.e. the e-th training epoch, of a target $\boldsymbol{x}_t$ as $\mathcal{C}^e(\boldsymbol{x}_t)$. Following the above example, the *unpredictability* of context is then defined as the

expectation of the probability that $\mathcal{C}^{e+1}(\boldsymbol{x}_t)$ contains an object that is not from $\mathcal{C}^{e}(\boldsymbol{x}_t)$, i.e.

$$\mathbb{E}_{\boldsymbol{x}_t}\left[\frac{1}{|\mathcal{C}^{e+1}(\boldsymbol{x}_t)|}\sum_{\boldsymbol{x}_j\in\mathcal{C}^{e+1}(\boldsymbol{x}_t)}\mathbb{I}\left(\boldsymbol{x}_j\notin\mathcal{C}^{e}(\boldsymbol{x}_t)\right)\right] \tag{3}$$

In referential game, since the sampling procedure is independent Bernoulli without replacement, the probability of having an object not from $\mathcal{C}^{e}(\boldsymbol{x}_t)$ in $\mathcal{C}^{e+1}(\boldsymbol{x}_t)$ is then simply $1-\frac{|\mathcal{C}^{e}(\boldsymbol{x}_t)|}{|\mathcal{X}|}=1-\frac{|C|}{|\mathcal{X}|}$ (i.e. smaller contexts are more unpredictable, since contexts on successive trials are more likely to differ in their composition).

3.2 Measuring Expressivity in Language Games

Following the intuition that the different types of games require messages to convey different amount of information about input space $\mathcal{X}$, Mutual Information may seem to be a good candidate measure for expressivity. However, we found that even for two emergent languages that are both deterministic and bijective mapping functions, i.e. they have the same MI value $\log(|\mathcal{X}|)$, their expressivity can still be different. Further discussion about MI can be found in Appendix B.

Since it is challenging to directly quantify the expressivity of emergent languages, we instead focus on the partial orders between them in this work. Inspired by a common practice in the transfer learning community (e.g. Pan & Yang, 2009), where models are trained on a source domain but evaluated on different domains, here we propose a definition of the partial order between the expressivity of two emergent languages $\mathcal{L}_A$ and $\mathcal{L}_B$, i.e. $\mathfrak{E}_{\mathcal{L}_A}$ and $\mathfrak{E}_{\mathcal{L}_B}$:

Definition 3.1 ($\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_A} > \mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_B}$)**.** Let g be an instance of language game, and $\mathcal{G}$ be a set of language game instances. Then, given $\mathcal{G}$, if the generalisation performance of $\mathcal{L}_A$ is better than $\mathcal{L}_B$ on some language games $\mathcal{G}'$, i.e. $\forall g\in\mathcal{G}',\ \mathfrak{P}^{g}_{\mathcal{L}_A}>\mathfrak{P}^{g}_{\mathcal{L}_B}$, where $\varnothing\subset\mathcal{G}'\subseteq\mathcal{G}$ and $\mathfrak{P}$ denotes the generalisation performance, while there is no statistically significant difference between the converged generalisation performance of $\mathcal{L}_A$ and $\mathcal{L}_B$ on the remaining games, i.e. $\mathfrak{P}^{g'}_{\mathcal{L}_A}\approx\mathfrak{P}^{g'}_{\mathcal{L}_B}\forall g'\in\mathcal{G}\backslash\mathcal{G}'$ [4], we then say that expressivity of $\mathcal{L}_A$ is higher than $\mathcal{L}_B$ on $\mathcal{G}$, i.e. $\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_A}>\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_B}$. The metric of generalisation performance could be varied according to different types of games.

Briefly speaking, if one language performs better than the other on some games while they perform approximately the same on the remaining games, we then say the expressivity of that language is higher than the other. Following this definition, there are several possible relationships between the expressivity of two languages $\mathcal{L}_A$ and $\mathcal{L}_B$: 1) $\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_A}>\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_B}$, defined as above; 2) $\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_A}=\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_B}$, if $\mathfrak{P}^{g}_{\mathcal{L}_A}\approx\mathfrak{P}^{g}_{\mathcal{L}_B},\ \forall g\in\mathcal{G}$; 3) $\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_B}>\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_A}$, defined as above; 4) $\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_A}$ is *incomparable* to $\mathfrak{E}^{\mathcal{G}}_{\mathcal{L}_B}$ ($\mathfrak{E}_{\mathcal{L}_A}\not\gtrless\mathfrak{E}_{\mathcal{L}_A}$), if propositions $\exists g'\in\mathcal{G},\ s.t.\ \mathfrak{P}^{g'}_{\mathcal{L}_A}>\mathfrak{P}^{g'}_{\mathcal{L}_B}$ and $\exists g''\in\mathcal{G},\ s.t.\ \mathfrak{P}^{g''}_{\mathcal{L}_B}>\mathfrak{P}^{g''}_{\mathcal{L}_A}$ are both true.

4 Predictions and Experiment Designs

4.1 Predictions From Our Hypothesis

Our hypothesis states that there is a trade-off between complexity and unpredictability that determines the expressivity. We also illustrate above that complexity is a monotonically *increasing* function of $|\mathcal{C}(\boldsymbol{x}_t)|=|C|$ while unpredictability is a monotonically *decreasing* function of it. That is, we cannot increase both of them at the same time with a fixed-sized input space $\mathcal{X}$. Therefore, we can expect the following phenomena about the relations between expressivity and the corresponding contextual complexity and unpredictability:

1. higher complexity in games having same contextual unpredictability can facilitate emergent languages having higher expressivity;

[4] Note that $\approx$ in this work means that there is no statistically significant difference between means of two sequences.

2. higher unpredictability in games having same contextual complexity can facilitate emergent languages having higher expressivity;
3. if the size of $\mathcal{X}$ is fixed, a language which emerges in a game with moderate complexity and unpredictability should have the highest expressivity.

4.2 Language Transfer Experiment

To verify the above three predictions and thus our hypothesis, we evaluate the generalisation performance of languages emerging in various games following the procedure outlined in Procedure 1. Briefly speaking, the language transfer procedure is to train new listener to select the inputs among the candidates sets given *only* the messages.

Procedure 1: Procedure for the language emergence and transfer experiment

Input: A set of source game $\mathcal{G}_s$, a set of target game $\mathcal{G}_t$
for *every game* g_s^i *in* $\mathcal{G}_s$ **do**
  1. initialise a new speaker and listener for g_s^i, and train them to play g_s^i with the whole $\mathcal{X}$;
  2. after the agents converge on g_s^i, record $\mathcal{L} = \{(\boldsymbol{x}, \boldsymbol{m}) | \boldsymbol{x} \in \mathcal{X}\}$;
  3. randomly shuffle and split $\mathcal{L}$ into 2 disjoint sets $\mathcal{L}_{train}$ and $\mathcal{L}_{test}$ where $|\mathcal{L}_{train}| = 90\% \cdot |\mathcal{L}|$;
  4. **for** *every game* g_t^j *in* $\mathcal{G}_t$ **do**
    1. initialise a new listener for g_t^j;
    2. train the listener with $\mathcal{L}_{train}$ to complete g_t^j;
    3. record the *accuracy* of listener on $\mathcal{L}_{test}$ as the **generalisation performance of** g_s^i **on** g_t^j;
  end
end

Since our hypothesis concerns the size of C, we propose to experiment on a series of referential games with varying candidate set size $|C|$ ("referX" where X= $|C|$). Meanwhile, as these referential games have both different complexity and unpredictability, we also implement a series of referential games with different $|C|$ but the batches are fixed during training such that there is no unpredictability of the context ("referXf" where X= $|C|$). To sum up, in our experiment, $\mathcal{G}_s$ = {"refer2/f", "refer10/f", "refer100/f", "refer1000/f", "refer2500/f", "refer5000/f", "refer7500/f", "refer10000"}[5]. Since language games with fixed context are not good simulation of human communication , we keep only the ones having varying batches from $\mathcal{G}_s$ in $\mathcal{G}_t$.

As a naming convention to allow us to refer easily to the factors that influence a language's expressivity, we refer to emergent languages by their parameters of their source game, e.g. "refer10 language" represents the language which emerges in referential games whose $|C| = 10$ (with context varying across epochs).

5 Results & Discussion

To alleviate the effect from randomness as much as possible, all the following results are obtained with 6 different random seeds. All values are averaged across the seeds, and the corresponding standard deviation (SD) are also provided (either in this section or appendices).

5.1 Under Fixed Unpredictability, Higher Complexity Leads to Higher Expressivity

To verify the Prediction 1, we compare the generalisation performance of games with fixed context, i.e. the unpredictability of their context are the same, on $\mathcal{G}_t$, and the results are shown in Figure 2. Languages which emerge in source games whose contexts have higher complexity consistently perform better on tasks from $\mathcal{G}_t$ than those which emerge in source games with lower-complexity

[5]There is no "refer10000f" since no unpredictability is possible when $|B| = |C| = 10,000 = |\mathcal{X}|$

contexts, since all lines keep growing with the increase of values on x-axis (i.e. complexity). That is, the generalisation performance is a *monotonically increasing* function of $|C|$ if there is no unpredictability in the context, which matches with the prediction of our hypothesis.

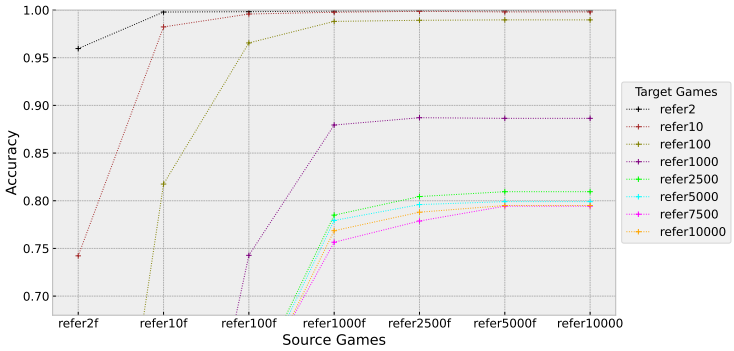


Figure 2: Generalisation performance of languages from fixed context with varying complexity. The x-axis indicates the source games, and y is the accuracy. Lines in different colours represent the generalisation performance on different target games. We plot only the means of the converged performance to keep the figure readable (SD values are provided in Table 3 in Appendix C.1), and remove 'refer7500f' from the x-axis since it is a mix of $|C| = 7,500$ and $|C| = 2,500$.

5.2 Under Fixed Complexity, Higher Unpredictability Leads to Higher Expressivity

To verify the Prediction 2, we compare the difference between generalisation performance of all games in $\mathcal{G}_s$ ("referX") and their corresponding fixed-batch version ("referXf"). As indicated by that all points in Figure 3 are positive, languages from context with varying objects always perform better than the corresponding fixed-context version on all target games. Since all "referXf" has lower unpredictability of context than the corresponding "referX" games, this indicates that higher unpredictability can unilaterally lead to better generalisation performance of emergent language across various target games, which matches with the prediction of our hypothesis.

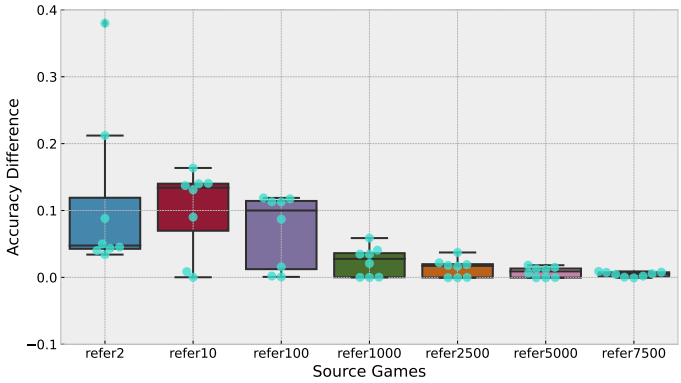


Figure 3: Generalisation performance gaps between languages from varying context and their corresponding fixed-context version on all target games. The x-axis indicates the types of source games, and the y-axis indicates the distribution of difference between the generalisation performance of "referX" and "referXf" on all target games (i.e. "referX - referXf"). Each point represent the difference on a specific kind of target game, and the boxes show the corresponding values of medians, 25th-percentile, and 75-th percentile. "refer10000" is not plotted since it doesn't have the fixed-context twin game.

5.3 EXPRESSIVITY IS A TRADE-OFF BETWEEN CONTEXTUAL COMPLEXITY AND UNPREDICTABILITY

To verify the Prediction 3, we compare performance for all source games with varying context on all target games, and the results are shown in Figure 4 (see also Appendix C.2 for more details).

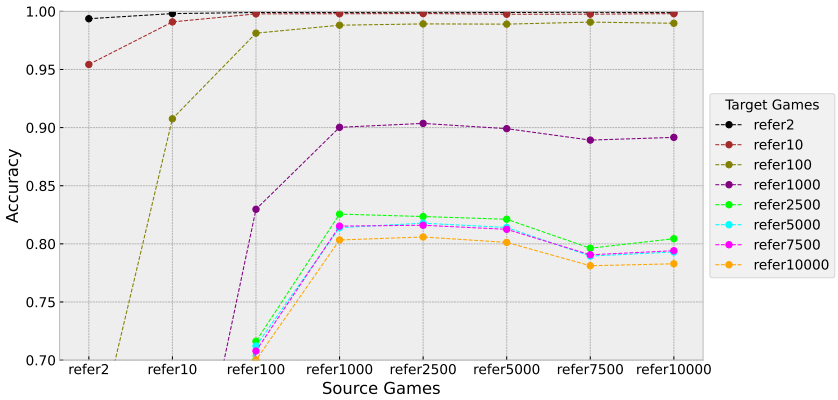


Figure 4: Generalisation performance of language from different source games on the various target games. Each line represents a specific kind of *target game*. The x-axis indicates the type of source games, and the y-axis is the accuracy on target games. To keep the figure readable, we remove the error bars, and instead provide the corresponding SD values in Table 4. We limit the range of y-axis to $[0.7, 1.0]$ to emphasise the difference between languages on complex games, more values are provided in Figure 8.

As shown in Figure 4, on whichever type of game, the peak of the curves is always among 'refer1000', 'refer2500', or 'refer5000', which means that the best expressivity always emerges in these 3 types of games. Considering that the complexity of context in referential game is a *monotonically increasing* function of the $\frac{|C|}{|\mathcal{X}|}$ while the contextual unpredictability is a *monotonically decreasing* of it, the results shown in Figure 4 support our hypothesis, i.e. the expressivity is a trade-off between the contextual complexity and unpredictability.

To be specific, although context of 'refer2', 'refer10', and 'refer100' have high unpredictability, their complexity is low thus the languages have lower expressivity. On the other hand, context of 'refer7500' and 'refer10000' have high complexity, but expressivity of neither language is the highest due to the high predictability of the context in these games. Unpredictability of context incentivises the agents to encode more information into the messages such that the messages could overcome the uncertainty of contexts and be interpreted in various contexts. Therefore, the three languages ('refer1000', 'refer2500', and 'refer5000') from context with both moderate complexity and unpredictability perform better than the others. To make it more straightforward to compare the expressivity between these languages, we also provide another view of the above results in Figure 9 in Appendix C.2, which shows that 'refer1000' $\gtrneqq$ 'refer2500' $\approx$ 'refer5000' $>$ 'refer7500' $\approx$ 'refer10000' $>$ 'refer100' $>$ 'refer10' $>$ 'refer2'.

To eliminate the possibility that the above phenomena are caused by the hyperparameters we used in our games, we also run the language transfer experiment with varying communication channel capacity and agent capacity (i.e. the size of hidden layers). As shown in the figures in Appendix D, both 'refer1000' and 'refer2500' perform better than 'refer10000' across varying hyperparameters, which shows that the trade-off phenomenon is consistent over game settings.

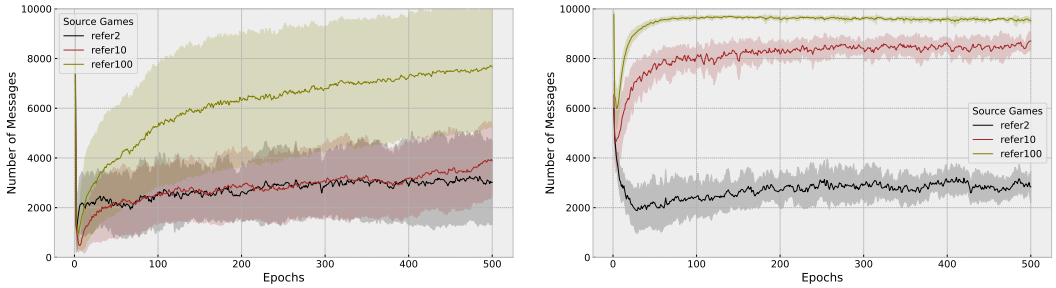
5.4 CONTRASTIVE LOSS ALLEVIATES THE COLLAPSE OF MESSAGE TYPES

As we use our contrastive loss instead of the conventional referential loss applied by (e.g. Havrylov & Titov, 2017; Ren et al., 2020), we also compare the behaviours of both loss functions on $\mathcal{G}'_s = \mathcal{G}'_t =$ {"refer2", "refer10", "refer100"}.

We refer to the phenomenon that the number of message types is greatly less than the size of input space as "message type collapse", and the mappings that maps several different inputs to a same message type as "degeneracy components". In the comparison between the two loss functions, we observe that contrastive loss is more efficient on removing degenerate components in large-scale games (where $|C| \geqslant 100$), and can better avoid the collapse of the number of message types. To be specific, as shown in Figure 5a, the numbers of message types collapses to less than 500 on games with conventional referential loss. They, however, would only decrease to roughly 5000 on 'refer10' and 'refer1000' with contrastive loss as shown in Figure 5b. The contrastive loss can also make the training of referential game more stable, as illustrated by the smaller variance in Figure 5b.

Our results also indicate that it is necessary to establish a referential game with candidate sets containing enough many distractors to guarantee that the emergent language will be relatively unambiguous. As far as we know, the existing works do not investigate the effects of the size of candidate sets, e.g. Lee et al. (2018) set $|C| = 2$, Li & Bowling (2019) set $|C| = 5$, and $|C| \leqslant 20$ is used by Lazaridou et al. (2018). Given our results, our heuristic is that $|C|$ should be higher than either $1,000$ for large input spaces or $\frac{1}{10}|\mathcal{X}|$ for smaller ones.

We also explore more about the degeneracy degree (as well as structure degree) of the mappings constituting different languages, and the results and discussion are provided in Appendix A.



(a) With conventional referential loss

(b) With contrastive loss

Figure 5: How number of message types change over training epochs with different loss functions. The lines are the means over 6 runs and shadow areas indicate the corresponding variance.

6 Conclusion & Future Work

To our knowledge, this is the first work exploring the *expressivity* of emergent languages from language games involving DL-based agents that have different context size. We first proposed a hypothesis about the factors influencing the expressivity of emergent languages, and defined a partial ordering between expressivity of different emergent languages based on the generalisation performance across a set of games. Through a series of experiments on referential games with varying numbers of candidates, we then validated the three predictions from our hypothesis, which furthers indicates that expressivity is indeed a trade-off between the complexity and unpredictability of context in language games. To overcome the memory inefficiency issue caused by the conventional referential loss function, we introduced a contrastive loss into our implementation, and showed that it can also alleviate the collapse of message types.

Since this is a first step towards defining and understanding the expressivity of emergent languages, there are still some questions unanswered by this work. For example, some properties of our emergent languages are surprising given the simplicity of some of our referential games. In the simplest setting ($|C| = 2$), we see emergent languages with more than $2,000$ message types, whereas an optimal strategy requires a language consists of only 40 unique messages (since there are only 40 different values for all properties in $\mathcal{X}$ and two items in the context always differ on at least one property). Thus, it needs further investigation of how redundancy helps during the communication between agents. More importantly, we lack understanding of the structure of mappings constituting a language, especially when the languages are generated or encoded by neural networks. We believe that the expressivity of languages still needs further exploration in order to maximise the generalisation performance of emergent languages across various tasks.

REFERENCES

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

Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. *arXiv preprint arXiv:2004.09124*, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Laura Graesser, Kyunghyun Cho, and Douwe Kiela. Emergent linguistic phenomena in multi-agent communication games. *arXiv preprint arXiv:1901.08706*, 2019.

Shangmin Guo, Yi Ren, Serhii Havrylov, Stella Frank, Ivan Titov, and Kenny Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *EvoLang*, 2020.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

Serhii Havrylov and Ivan Titov. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems*, 2017.

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

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. 2017. URL `https://arxiv.org/abs/1611.01144`.

Ivana Kajić, Eser Aygün, and Doina Precup. Learning to cooperate: Emergent communication in multi-agent navigation. *arXiv preprint arXiv:2004.01097*, 2020.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 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*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-3010. URL `https://www.aclweb.org/anthology/D19-3010`.

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. Entropy minimization in emergent languages. In *International Conference on Machine Learning*, pp. 5220–5230. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Simon Kirby. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2): 102–110, 2001.

Simon Kirby and James R Hurford. The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language*, pp. 121–147. Springer, 2002.

Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current opinion in neurobiology*, 28:108–114, 2014.

Simon Kirby, Monica Tamariz, Hannah Cornish, and Kenny Smith. Compression and communication in the cultural evolution of linguistic structure. *Cognition*, 141:87–102, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

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

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input. *arXiv preprint arXiv:1804.03984*, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. Emergent translation in multi-agent communication. In *International Conference on Learning Representations*, 2018.

David Lewis. *Convention: A philosophical study*. John Wiley & Sons, 1969.

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

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *arXiv preprint arXiv:1706.02275*, 2017.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model. *ICLR*, 2020.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

James Winters, Simon Kirby, and Kenny Smith. Contextual predictability shapes signal autonomy. *Cognition*, 176:15–30, 2018.

Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 1954.

Luyao Yuan, Zipeng Fu, Jingyue Shen, Lu Xu, Junhong Shen, and Song-Chun Zhu. Emergence of pragmatics from referential game between theory of mind agents. *arXiv preprint arXiv:2001.07752*, 2020.

Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Efficient communication in multi-agent reinforcement learning via variance based control. *arXiv preprint arXiv:1909.02682*, 2019.

A MORE RESULTS AND DISCUSSION ON EXPRESSIVITY

In this section, we provide two more intermediate properties of emergent languages that are also decided by complexity and unpredictability of context, and may influence the expressivity.

A.1 DEGREE OF DEGENERACY AFFECTS BUT DOES NOT FULLY DETERMINE THE EXPRESSIVITY

Inspired by Kirby et al. (2015), the first intermediate factor we explore is the *degree of degeneracy* (i.e. the degree of ambiguity) of the languages. As a language is defined as a mapping function $\mathcal{L} : \mathcal{X} \mapsto \mathcal{M}$, the degree of its degeneracy could be measured by the number of distinct message types.

As shown in the Table 1, in referential languages, the degeneracy degree decreases rapidly with the increase of the contextual complexity of the source games, and reaches a ceiling around $|D| \geqslant 1,000$. There is a positive correlation between the number of message types and expressivity for the languages with low complexity, which indicates that degree of degeneracy could be an intermediate factor that is correlated with the expressivity of these languages. However, the fact that the average number of message types on refer10000 is higher than refer1000, while the expressivity of 'refer1000' is higher than 'refer10000', i.e. emergent languages having lower level of degeneracy does not necessarily have higher expressivity, indicates that degeneracy degree of one language cannot fully determine its expressivity.

Game	refer2	refer10	refer100	refer1000	refer2500	refer5000	refer7500	refer10000
Mean	2872.18	8565.69	9547.58	9801.60	9839.33	9841.27	9840.34	9856.96
25th-percentile	2629.0	8325.0	9501.5	9778.75	9825.0	9837.5	9819.0	9845.0
75th-percentile	3195.0	8795.5	9639.25	9823.25	9854.25	9864.25	9866.0	9871.0

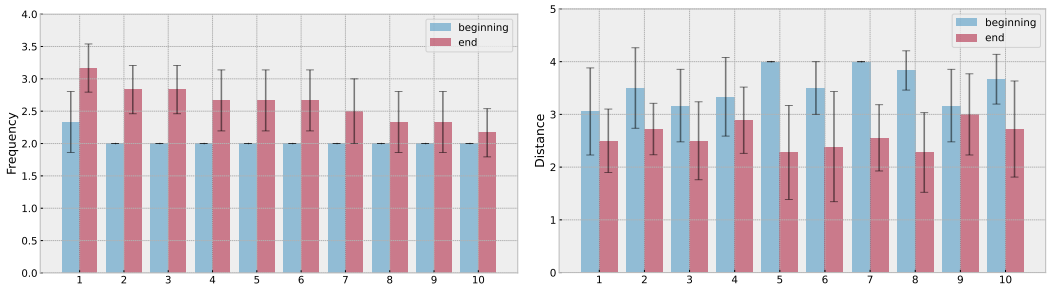
Table 1: Mean, 25th-percentile, and 75th-percentile of the distributions of message types numbers from different source game.

A.2 STRUCTURE DEGREE OF MAPPINGS OF DIFFERENT LANGUAGES

Since the degenerate degree is not the only factor determining expressivity, another factor to be considered is how the mappings of a language $\mathcal{L}$ are structured after the training procedure of agents on source games. According to the cardinality of the mappings of $\mathcal{L}$, we can divide them into two disjoint sets: i) *one-to-one* mappings, i.e. an input is mapped to a unique message; and ii) *many-to-one* mappings (a.k.a degenerate components), i.e. several inputs are mapped to the same message instance. Between this two types, the structure of one-to-one mappings has been explored under the view of compositionality in the previous works, e.g. (Brighton & Kirby, 2006; Andreas, 2019). However, as illustrated by Chaabouni et al. (2020), a good generalisability does *not* require a good compositionality, which means that measurements for compositionality are not good predictors for expressivity either. In this work, we only focus on the *many-to-one* mappings.

Given the objectives of games, for the degenerate components in an emergent language, a message $\boldsymbol{m}$ being mapped to several closely-related inputs could imply that it captures common features among these inputs. Suppose that several inputs $X_K = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K\}$ are mapped to the same message $\boldsymbol{m}$ in a converged language $\mathcal{L}$, we hypothesise that *the average distances between every pair of messages*, e.g. $\frac{1}{|X_K|}\sum_{i,j\in\{1,\ldots,K\}} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$, would become smaller after training, compared with their random initial values. Therefore, we compare the averaged distances between meanings of degenerate components as well as the frequency of degenerate components at the beginning and end of training. To make the plots more readable, we only plot the statistics about the top-10 most frequent ambiguous message types, where the frequency of a message type is defined as the number of inputs that are mapped to it in a languages. Results on 'refer1000' (one of the most expressive language in our experiments) are shown in Figure 6.

By comparing the data of beginning and end of training in Figure 6b, we can easily observe that *the averaged distance* decreased throughout the training, which indicates that the degenerate components actually became more although not fully structured. Another observation we have limited understanding of is that 'refer1000' becomes more ambiguous after training (shown by that the red bars



(a) Frequencies of degeneracy components (b) Distance between corresponding inputs

Figure 6: Statistics about frequency of *Top-10* most frequent message type and the average Hamming distances between their meanings at the beginning and end of training on 'refer1000'. The bars indicate the frequency and the Hamming distance, and the black lines indicate the corresponding SD.

representing the statistics at the end of training are lower than the bars representing the statistics at the beginning of training), although its accuracy does improve to 100%. However, limited by the scope of this work, we will leave further exploration of it to future works.

B MUTUAL INFORMATION IS INSUFFICIENT

In Section 3.2, we mentioned that MI cannot completely reflect the expressivity of an emergent language. Here, we provide more detailed explanations. On one hand, a higher MI value does not necessarily mean more task-relevant information, e.g. encoding colour in messages is unnecessary for tasks involving only the shapes of inputs although it provides more information. On the other hand, not only the determinacy of the mappings matters for generalising them, but also the structures, e.g. it would be easier for humans to extrapolate compositional languages to novel concepts than holistic ones although they are both unambiguous and therefore have the same MI. In the following paragraphs, we will give more reasoning and empirical evidence to support our claim.

Following the mapping function definition of emergent languages illustrated in Section 2.1, suppose that we get a message type collection $M \subset \mathcal{M}$ in some training epoch, the calculation of the mutual information between M and $\mathcal{X}$ is given as follows:

$$
\begin{aligned}
I(M|X) &= \sum_{\boldsymbol{x}\in X}\sum_{\boldsymbol{m}\in M} p(\boldsymbol{x},\boldsymbol{m}) \log \frac{p(\boldsymbol{x},\boldsymbol{m})}{p(\boldsymbol{x})p(\boldsymbol{m})} \overset{p(\boldsymbol{x},\boldsymbol{m})=\frac{1}{|\mathcal{X}|}}{=} \sum_{\boldsymbol{x}\in \mathcal{X}}\sum_{\boldsymbol{m}\in \mathcal{M}'} \frac{1}{|\mathcal{X}|} \log \frac{p(\boldsymbol{x},\boldsymbol{m})}{p(\boldsymbol{x})p(\boldsymbol{m})} \\
&= \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in X}\sum_{\boldsymbol{m}\in M} \log \frac{p(\boldsymbol{x},\boldsymbol{m})}{p(\boldsymbol{x})p(\boldsymbol{m})} = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in X}\sum_{\boldsymbol{m}\in M} (\log p(\boldsymbol{x},\boldsymbol{m}) - \log p(\boldsymbol{x}) - \log p(\boldsymbol{m})) \\
&\overset{p(\boldsymbol{x})=\frac{1}{|\mathcal{X}|}}{=} \frac{1}{N}\sum_{\boldsymbol{x}\in X}\sum_{\boldsymbol{m}\in M} \left(\log \frac{1}{|\mathcal{X}|} - \log \frac{1}{|\mathcal{X}|} - \log p(\boldsymbol{m})\right) = \frac{1}{|\mathcal{X}|}\sum_{\boldsymbol{x}\in \mathcal{X}}\sum_{\boldsymbol{m}\in M} (-\log p(\boldsymbol{m})) \\
&= -\sum_{\boldsymbol{m}\in M} \log p(\boldsymbol{m}) = -\sum_{\boldsymbol{m}\in M} \log \frac{|\mathcal{L}(\boldsymbol{m})|}{|M|} = -\sum_{\boldsymbol{m}\in M} (\log |\mathcal{L}(\boldsymbol{m})| - \log |M|) \\
&= \sum_{\boldsymbol{m}\in M} (\log |M| - \log |\mathcal{L}(\boldsymbol{m})|)
\end{aligned}
\tag{4}
$$

where $p(\boldsymbol{x}) = \frac{1}{|\mathcal{X}|}$ as $\boldsymbol{x}$ follows a uniform distribution over the input space, $p(\boldsymbol{x},\boldsymbol{m}) = \frac{1}{|\mathcal{X}|}$ since every $\boldsymbol{x}$ is mapped to a message, and $|\mathcal{L}(\boldsymbol{m})|$ is the appearance number of $\boldsymbol{m}$ in one language.

By observing the above equation, it is straightforward to see that the MI value is decided by the number of message types, i.e. the degeneracy degree of one language. Therefore, if two emergent languages have approximately the same number of messages types (as shown in Table 1 and degeneracy degree, the MI values for these two languages should also be the same. However, as we illustrate in

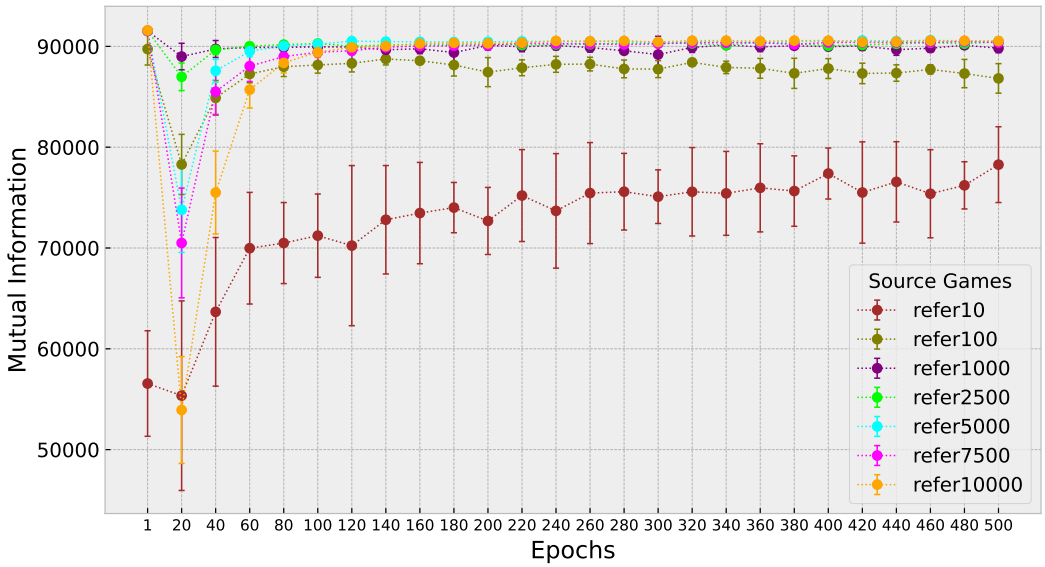


Figure 7: Mutual information curves over training epochs in different types of games. The converged values are provided in Table 2

Section A.1 that degree of degeneracy cannot fully determine the expressivity of emergent languages, we can then deduce that MI cannot fully measure the expressivity of emergent languages.

To provide more empirical evidence to support our claim, we tracked the MI values over training epochs, and the results are shown in the following Figure 7. As we can see, there is no statistically significant difference between the MI values of 'referX' (where X$\geqslant 1,000$) while we can show that their expressivity is different in Section 5.3.

Game	refer10	refer100	refer1000	refer2500	refer5000	refer7500	refer10000
Mean	78269.22	86823.66	89823.16	90447.69	90504.40	90378.89	90533.06
σ	3760.05	1464.69	443.12	213.76	412.37	306.85	139.42

Table 2: Converged MI values of different types of games. Both mean and standard deviation (σ) are given.

C Language Transfer Experiment Results

C.1 Raw Results of Complexity Experiment

In this section, we provide the raw data of Figure 2 in the following Table 3. Not that although "refer1000f" performs the best on "refer10" than all other languages, we argue this outlier doesn't influence conclusion of our prediction considering the simplicity of "refer10".

C.2 More Results of Trade-off Experiment

In this section, we first provide the complete version of Figure 4, i.e. Figure 8. Since the performance of 'refer2', 'refer10', and 'refer100' are too bad on complex target games, it is very clear that their expressivity are much lower that the rest. Therefore, to emphasise the difference between the other languages, we limit the range of y-axis to $[0.7, 1.0]$ in Figure 8.

We also illustrate the partial order between the expressivity of different languages in Figure 9. It is very clear that the 'refer2', 'refer10', and 'refer100' performs worse than the others on referential games with more candidates than their sources. Combining this fact with the fact that 'refer7500' and 'refer10000' perform worse than 'refer1000', 'refer2500', and 'refer5000' shows that increasing the

Source	Statistics	refer2	refer10	refer100	refer1000	refer2500	refer5000	refer7500	refer10000
refer2f	mean	0.9596	0.7423	0.2244	0.0322	0.0168	0.0163	0.0156	0.0163
	σ	0.0236	0.1493	0.1308	0.0203	0.0103	0.0105	0.0111	0.0119
refer10f	mean	0.9979	0.9823	0.8175	0.3259	0.2028	0.1918	0.1909	0.1879
	σ	0.0007	0.0044	0.0404	0.0518	0.0461	0.0446	0.0500	0.0418
refer100f	mean	0.9983	0.9958	0.9655	0.7428	0.5988	0.5933	0.5946	0.5871
	σ	0.0010	0.0029	0.0106	0.0573	0.0738	0.0685	0.0719	0.0663
refer1000f	mean	0.9989	0.9977	0.9881	0.8794	0.7849	0.7793	0.7565	0.7686
	σ	0.0008	0.0006	0.0032	0.0163	0.0250	0.0242	0.0664	0.0237
refer2500f	mean	0.9992	0.9987	0.9892	0.8871	0.8044	0.7960	0.7788	0.7880
	σ	0.0008	0.0005	0.0032	0.0199	0.0259	0.0239	0.0267	0.0264
refer5000f	mean	0.9988	0.9980	0.9897	0.8916	0.8094	0.7980	0.7941	0.7928
	σ	0.0007	0.0006	0.0023	0.0077	0.0099	0.0143	0.0127	0.0122
refer10000	mean	0.9993	0.9980	0.9897	0.8864	0.8095	0.7991	0.7943	0.7950
	σ	0.0005	0.0007	0.0021	0.0084	0.0142	0.0095	0.0123	0.0107

Table 3: Mean and standard deviation σ of generalisation performance of languages from fixed-context on various target games.

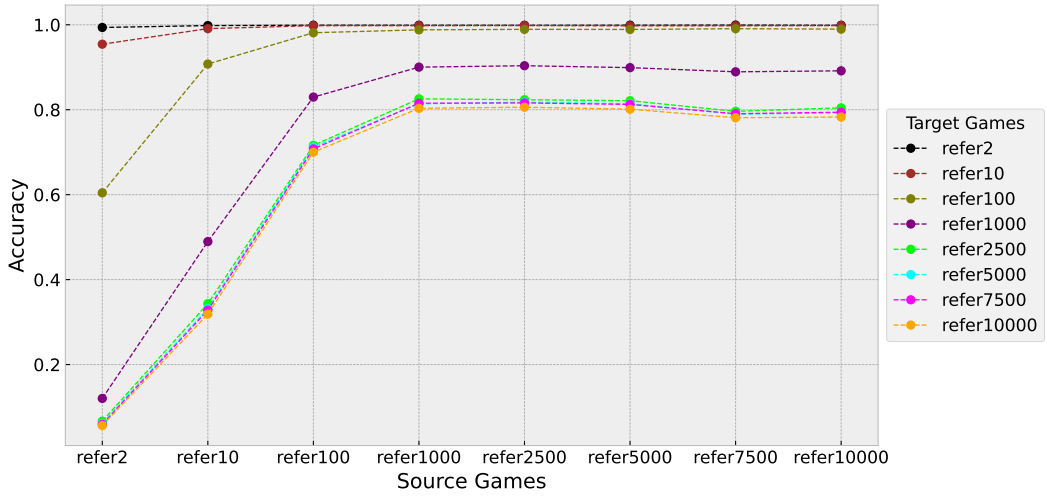


Figure 8: Same to Figure 4, but with wider range for y-axis.

size of candidate set $|C|$ would not always benefit the expressivity. As for 'refer1000', 'refer2500', and 'refer5000', there is no statistically significant difference between their performance on the simpler referential games, and they intersect on some games, thus their expressivity are approximately the same.

At last, we give the precise values of both Figure 9 and 4 in the following Table 4. All the results are obtained by multiple runs, thus both mean and standard deviation (σ) are given.

D Varying the capacity of channel and network

To verify that our findings are robust to the capacity of the communication channel and agents, we also run some extra experiments on the following different configurations of referential games:

1. *larger channel capacity*: set the length of messages to 8 and size of token inventory to 20, thus the size of message space becomes 20^8;

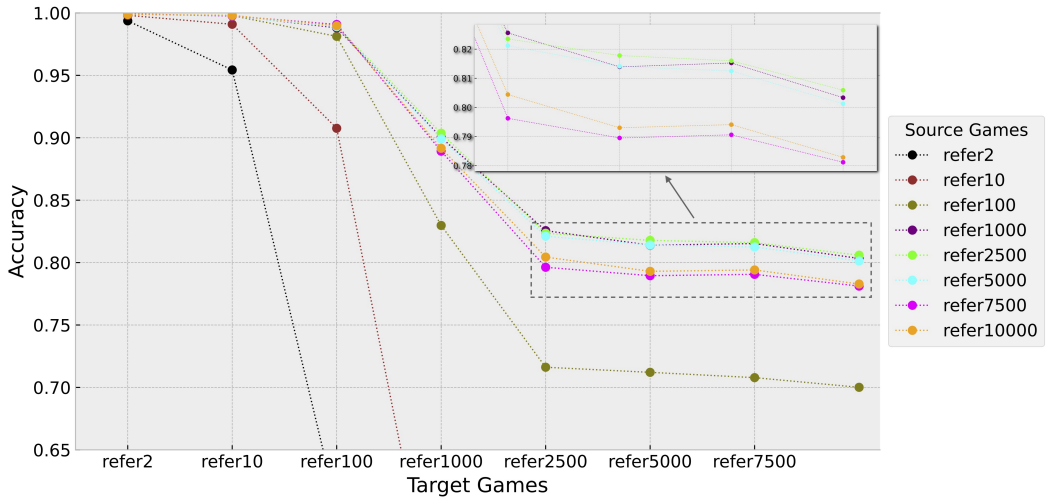


Figure 9: Generalisation performance of language from the same source game over different target games, where each line represents a *source game*. Since the plot would become not readable if we plot error bars here, we provide SD in Table 4 in Appendix C. The x-axis indicates the type of target games, and the y-axis is the accuracy on target games. If a line lies above the other over all the target games, it means that the corresponding source language has a higher expressivity than the other. On the other hand, if two lines intersect, e.g. 'refer1000' and 'refer2500' above, then the expressivity of these languages is *incomparable*.

Source	Statistics	refer2	refer10	refer100	refer1000	refer2500	refer5000	refer7500	refer10000
refer2	mean	0.9936	0.9543	0.6043	0.1203	0.0668	0.0616	0.0592	0.0565
	σ	0.0014	0.0077	0.0394	0.0146	0.0105	0.0078	0.0090	0.0047
refer10	mean	0.9980	0.9909	0.9076	0.4896	0.3434	0.3318	0.3282	0.3188
	σ	0.0007	0.0036	0.0245	0.0833	0.0809	0.0801	0.0778	0.0752
refer100	mean	0.9986	0.9978	0.9812	0.8298	0.7162	0.7121	0.7078	0.7001
	σ	0.0007	0.0009	0.0035	0.0171	0.0299	0.0312	0.0293	0.0350
refer1000	mean	0.9989	0.9979	0.9881	0.9003	0.8256	0.8139	0.8153	0.8033
	σ	0.0007	0.0009	0.0015	0.0143	0.0151	0.;0165	0.0226	0.0183
refer2500	mean	0.9988	0.9981	0.9892	0.9036	0.8235	0.8178	0.8160	0.8059
	σ	0.0006	0.0009	0.0012	0.0121	0.0099	0.0158	0.0148	0.0197
refer5000	mean	0.9990	0.9975	0.9890	0.8991	0.8212	0.8141	0.8125	0.8013
	σ	0.0004	0.0004	0.0022	0.0050	0.0140	0.0095	0.0117	0.0140
refer7500	mean	0.9990	0.9975	0.9890	0.8893	0.7963	0.7895	0.7906	0.7812
	σ	0.0005	0.0004	0.0034	0.0156	0.0206	0.0185	0.0223	0.0215
refer10000	mean	0.9988	0.9980	0.9887	0.8916	0.8044	0.7930	0.7940	0.7828
	σ	0.0007	0.0005	0.0023	0.0077	0.0099	0.0143	0.0127	0.0122

Table 4: Mean and standard deviation σ of generalisation performance of language games on each other.

2. *larger agent capacity*: set the hidden layer size to 128, thus the capacity of both speaker and listener become larger than the original setting (256);
3. *larger channel&agent capacity*: do the above two changes at the same time.

Since our key findings are about the generalisation performance between refer1000, refer2500, and refer10000, we just run these extra experiments on a game set $\mathcal{G}_{extra}$ = {'refer1000', 'refer2500', 'refer10000'} with multiple runs. The results of configuration 1, 2, and 3

are given in Figure 10, 11, and 12 respectively. It can be observed in all figures that 'refer1000' or 'refer2500' always perform better than 'refer10000' on all kinds of target games, i.e. 'refer1000' or 'refer2500' both have higher expressivity than 'refer10000' . Therefore, we believe the key observation from Figure 9 hold for varying configurations of communication channel and capacity of agents, which means that our conclusions are robust to the configuration of language games.

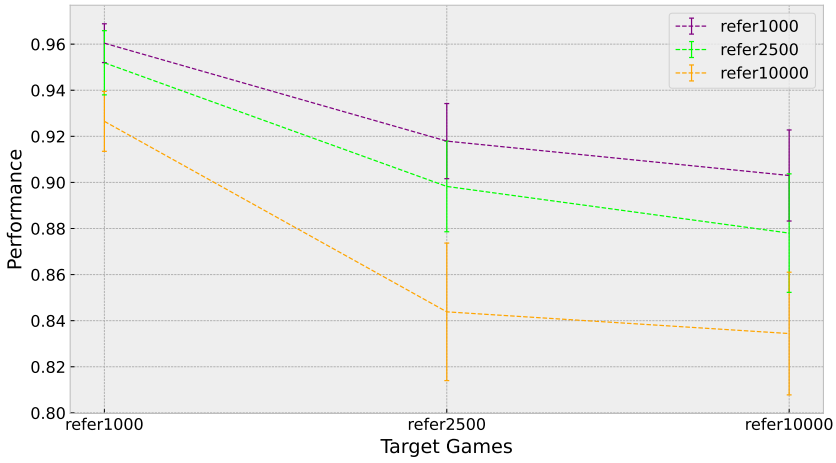


Figure 10: Results of language transfer with larger channel capacity. Error bars indicate the standard deviation of the mean values.

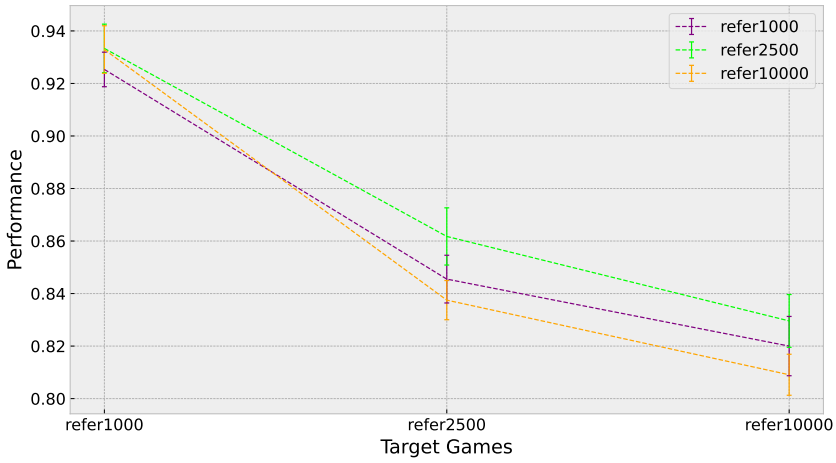


Figure 11: Results of language transfer with larger agent capacity. Error bars indicate the standard deviation of the mean values.

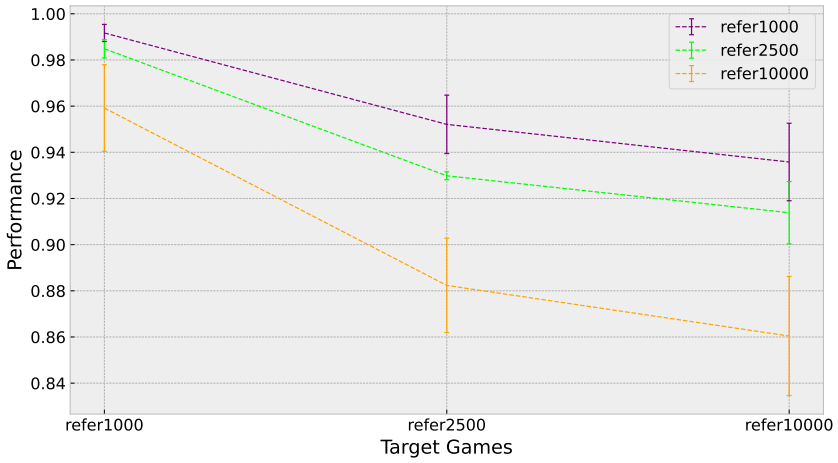


Figure 12: Results of language transfer with larger channel and agent capacity. Error bars indicate the standard deviation of the mean values.