THE EMERGENCE OF COMPOSITIONAL LANGUAGES FOR NUMERIC CONCEPTS THROUGH ITERATED LEARNING IN NEURAL AGENTS

SHANGMIN GUO *[*1], YI REN[2], SERHII HAVRYLOV[2], IVAN TITOV[2], STELLA FRANK[3], and KENNY SMITH[3]

[*]Corresponding Author: sg955@cam.ac.uk
[1]Department of Computer Science and Technology, University of Cambridge, Cambridge, UK
[2]School of Informatics, University of Edinburgh, Edinburgh, UK
[3]School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, UK

Since first introduced by Hurford (1989), computer simulation has been an increasingly important tool in evolutionary linguistics. Recently, with the development of deep learning techniques, research in grounded language learning has also started to focus on facilitating the emergence of compositional languages without pre-defined elementary linguistic knowledge. In this work, we explore the emergence of compositional languages for numeric concepts in multi-agent communication systems. We demonstrate that compositional language for encoding numeric concepts can emerge through iterated learning in populations of deep neural network agents. However, language properties greatly depend on the input representations given to agents. We found that compositional languages only emerge if they require less iterations to be fully learnt than other non-degenerate languages for agents on a given input representation.

1. Introduction

With recent advances in deep learning (DL), it has been shown that computational agents can master a variety of complex cognitive tasks (Mnih et al., 2015; Silver et al., 2017). Recent work in grounded language learning (Hermann et al., 2017; Havrylov & Titov, 2017) applied DL techniques to enable agents to discover through learning communication protocols exhibiting language-like properties, e.g. hierarchy and compositionality. Using DL methods allow us to overcome the language pre-defining issue present in current computer simulation methods in evolutionary linguistics as in Steels (2005) and Cangelosi and Parisi (2012). The issue consists in having all basic linguistic elements (such as symbols and rules of generating phrases) to be pre-specified instead of being invented from scratch. In contrast to previous works (Mordatch & Abbeel, 2018; Cao et al., 2018) which focus on the emergence of referential signalling systems, we explore the emergent compositionality of the **non-referential** numeric concepts (which

[*]Work done at University of Edinburgh.

will be explained in Section 2.2) by designing a **referential** game in which agents need to transmit numerical concepts to communicate successfully.

Inspired by Kirby, Tamariz, Cornish, and Smith (2015), we model the emergence of communication protocols in dyads (i.e. the smallest possible social group of two agents) that are nodes in iterated learning chain (Kirby, 1999). We observe that iterated learning can facilitate the emergence of compositional languages for numeric concepts. However, the emergence of languages with such properties depends on the representation of numerical concepts present in the objects observed by the agents during the training. To be specific, compositional languages emerge when numeric concepts are: i) represented as a concatenation of one-hot vectors directly representing numbers; ii) implied in images of scenes featuring different number of objects. Further, we show that input representations influence the difficulty of learning a particular language by the agents, which explains the different results in case of iterated learning. For numerical concepts, we, therefore, argue that one necessary condition for the emergence of compositional languages in iterated learning is that these languages can be fully learnt [1] with less iterations for agents (especially listeners), compared with holistic languages and emergent languages from dyads.

2. Model Methods

2.1. *The Bag-Select Game*

To test whether computational agents can learn to transmit numerical concepts, we propose a referential game called as "Bag-Select" game which is briefly illustrated in Figure 2.1.

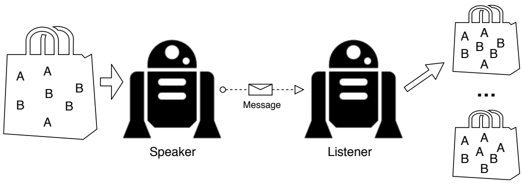


Figure 1. Sketch diagram of the Bag-Select game. The speaker observes a bag of objects of distinct types. The bag can contain a different number of objects of the specific type (here, three As and four Bs). The speaker produces a message, and the listener uses it to select the bag, that the speaker initially observed. The original bag is contained in a set among several other distinct bags, which differ only in the number of As and Bs.

In our game settings, there are two different kinds of agents: i) speaker S that observes the input b_i at the beginning of round i and then generates a message m_i; ii) listener L that receives m_i and then selects $\hat{b}_i$ among candidates c_i^k where

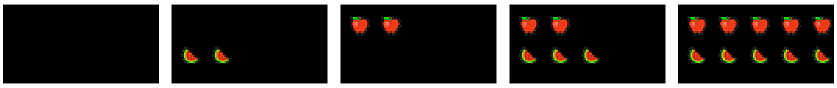
[1] A language is said to be fully learnt if: i) a speaker can always reproduce same messages as in the language given the inputs; ii) a listener could always obtain 100% accuracy given only the messages in it.

$k \in \{1, 2, \ldots, 15\}$[2] (among which one would be b_i, and the other fourteen would be uniformly sampled from the whole meaning space excluding b_i). The game only succeeds if $\hat{b_i}$ matches b_i. The speaker does not have access to the entire candidate list, only to the correct bag b_i, which implies that the number of each object type has to be encoded in the message in order to reliably succeed in the game.

2.2. *Representations of Bags*

The overall architecture of our implementation is similar to communication models proposed by Havrylov and Titov (2017). However, unlike theirs, in our game, an input b_i can be

1. **Concatenation**:a concatenation of one-hot vectors that represent numbers of each kind of objects, e.g. "2A3B" (a bag containing 2 As and 3 Bs) would be represented as $[001000; 000100]$ and "2A0B" would be represented as $[001000; 100000]$.
2. **Image**: an image containing different numbers of objects, e.g. "0A0B", "0A2B", "2A0B", "2A3B", "5A5B" would be represented as Figure 2 (a-e) respectively.
3. **Bag**: a bag of one-hot vectors that represent the quantity of different types of objects, e.g. "2A3B" and "2A0B" would be represented as $\{[01], [01], [10], [10], [10]\}$ and $\{[01], [01]\}$ respectively.



(a) "0A0B" (b) "0A2B" (c) "2A0B" (d) "2A3B" (e) "5A5B"

Figure 2. Example of an image representation of input bags that contain numerical properties. Captions under each sub-figures indicate the corresponding meaning.

As there is no specific value that can be referred to as numbers of an object in our Image and Bag representations, numeric concepts are **non-referential** in our games.

Different types of inputs require different encoders, thus we use: i) multilayer perceptron (MLP) for concatenations; ii) the convolutional neural network (CNN) which shares the same architecture of LeNet-5 proposed by LeCun, Bottou, Bengio, Haffner, et al. (1998) for images; iii) Bag-Encoder for bags.

Our bag-encoder shares almost the same architecture as the set encoder proposed by Vinyals, Bengio, and Kudlur (2015), except that we replace the softmax function in equation (5) of (Vinyals et al., 2015) with the sigmoid function. Thus, we could keep the feature representation invariant under reordering of the vectors

[2] In our experiments, there are always 15 candidates for listeners to choose from.

in bags, and avoid introducing normalizing bias (i.e. softmax output has to sum to one) which allows proper encoding of the numbers in the distributed representation of the bag.

To keep both meaning space and message space limited and thus analysable, there are only 2 different types of objects in our game and the maximum number of each kind of objects is 5. Therefore, the size of our Concatenation/Image dataset is 36, and the size of Bag dataset is 35 (excluding the empty bag). Messages are strings of characters of maximum length 2, where there is an available vocabulary of 10 characters.

2.3. *Iterated Learning for Neural Network Models*

We contrast two types of the population model. Following Havrylov and Titov (2017), we model dyads, pairs of agents who interact repeatedly and update their network parameters to maximise communicative success. Following Kirby et al. (2015), we contrast the communication systems that emerge in dyads with those that develop in iterated learning transmission chains. In the latter case, each generation in the chain consists of a pair of agents who are first trained on input-message pairs produced by the previous generation, then update their network parameters during communication with each other to maximise communicative success, before finally generating more data to pass to the next generation. In more detail, the model includes the following three steps:

1. **Learning phase (iterated learning only)**: During this phase, we train speaker S_t separately to reproduce same messages given the inputs, with the input-message pairs generated by S_{t-1}. For example, an input-message pair is "1A0B" → "yw", then we would train speakers to produce "yw" given the input "1A0B". To do so, we use stochastic gradient descent (SGD) (Robbins & Monro, 1951) to update parameters of S_t. Gradients are computed using the back-propagation (Rumelhart, Hinton, Williams, et al., 1988) algorithm with the cross entropy loss function between speaker's predictions and the messages generated by S_{t-1}. The number of training iterations is fixed such that predefined compositional language can be fully learnt (note that language produced by S_{t-1} is not necessarily compositional). There is no such phase in the first generation of iterated learning chain, as there are no input-message pairs for training S_1.
2. **Interaction phase**: During this phase, we train S_t and L_t agents to play the communication game using SGD. The reward is represented by the negative cross entropy between the probability distribution of the listener's prediction and the one-hot representation of the correct bag. Analogous to linguistic symbols, i.e. words, the messages transmitted between dyad should contain only discrete symbols. However, discrete messages would make learning prohibitively expensive from the computational perspective for computing the gradients would require enumeration of all possible messages. To overcome this

limitation, we use the Gumbel-softmax estimator proposed by Jang, Gu, and Poole (2016) to train our models. Besides, we set the number of iterations here to be fixed over generations, and number of iterations is obtained by pretraining a dyad to promise that it is long enough for a dyad to obtain 100% communication success rate.

3. **Transmission phase (iterated learning only)**: During this phase, we feed all b_i in the training set into S_t and sample messages m_i based on the generated probability distribution over vocabulary. This builds a dataset of input-message pairs for S_{t+1} to learn from. In addition, the number of sampled input-message pairs is $2,000$ so that they effectively reflect the distribution of all possible languages - note that since there are only 35-36 distinct input meanings to be communicated, there is no data bottleneck here, and learners will see signals for the entire space of possible meanings.

2.4. *Metrics and Evaluations*

Following Brighton and Kirby (2006), we take the topological similarity between meaning space and message space as the metric for measuring compositionality of languages, and we use Hamming distance and edit distance with respect to meaning space and message space. Equivalently, the topological similarity becomes the correlation coefficient between the Hamming distances between pairs of meanings and the edit distances between their corresponding messages. This measure captures the intuition that, in a compositional language, similar meanings will be conveyed using similar signals. We denote this measure of topological similarity as ρ; holistic (non-compositional) languages will have ρ scores around 0, a perfectly compositional language will have a ρ score of close to 1.

Additionally, we also need to measure the learning performance of new learners in order to compare the learnability of different languages, which will be introduced in Section 4. To do so, we use the accuracy of reproducing messages (both sequence-level and token-level) for speakers and accuracy of choosing the correct candidate for listeners respectively.

3. Emergence of Compositional Languages

In this section, we show that compositional languages can emerge under iterated learning, but only for the Concatenation and Image representations. As training iterated learning on deep learning models is extremely time-consuming, we report results for only one run per condition. During the exploratory phases of our research, we conducted multiple runs and found that the variance of resulting patterns of emergent languages is small, which gives us confidence that these results are representative.

To verify that iterated learning could successfully amplify the probability density of languages having high compositionality, we track the change of posterior probabilities of languages over generations. The results for the Concatenation, Image and Bag input representations are shown in the middle column of Figure 3.

As can be seen from the graphs, dyads do not converge on compositional languages under any input representation. However, in iterated learning models, the probability of languages with high compositionality ($\rho > 0.6$) keeps increasing over generations and gradually dominates all other languages, for the Concatenation and Image input representations; compositional languages do not develop in the Bag input representation. The compositional structure in the languages that emerge under the Concatenation input is clear from the example language given in Figure 3 (rightmost column), as is the absence of compositionality in the example language that develops under the Bag encoding; the final emergent language on Image representation is not perfectly compositional but contains a high degree of regularity.

4. Learnability of Compositional and Emergent Languages

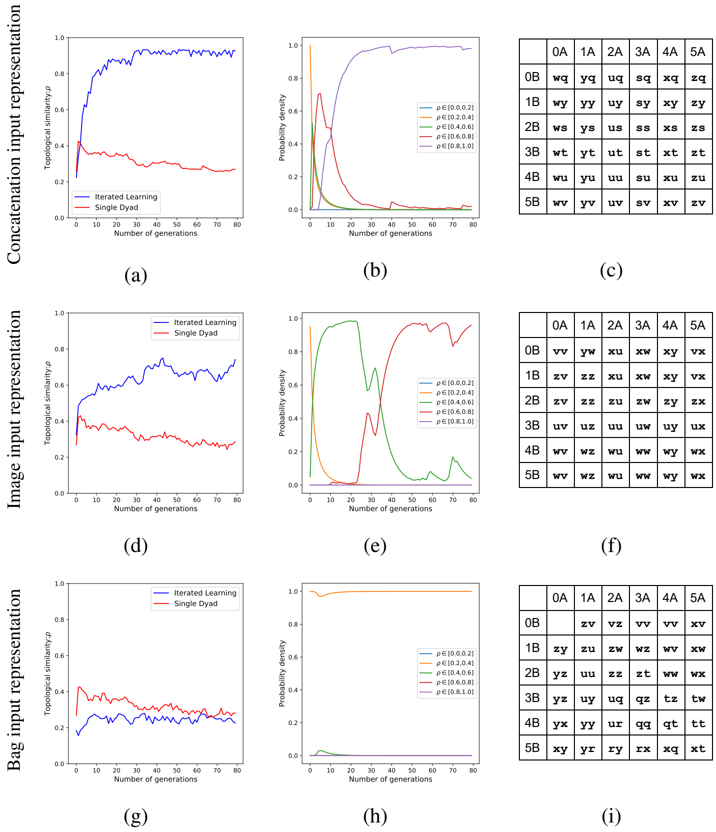
According to Kirby et al. (2015), the structure of natural languages is a trade-off between expressivity that arises during communication and compressibility that arises during learning. Meanwhile, Li and Bowling (2019) propose a hypothesis that compositional languages should be easier for listeners to learn than other less structured languages. Inspired by both of them, we hypothesise that the different effectiveness of iterated learning for different input representations observed in the above experiments is caused by different learnability of compositional languages for different input representations.

To test this hypothesis, we examine the learnability of three language types (compositional, emergent, holistic) for speakers and listeners. Our compositional test language was hand-designed and resembled the compositional languages that emerge under iterated learning in the Concatenation condition. Our holistic language was generated by randomly mapping messages from compositional languages (so that it shares same expressivity as compositional language) to inputs that constitute the whole meaning space. Our emergent test languages came from a dyad, trained to communicate as per the dyad models described above, once that dyad obtained 100% performance – as such, we would expect them to be largely holistic.

With these languages, we train speakers separately, which is illustrated in Section 2.3. At the same time, we train listeners separately to correctly complete the game with only messages in a language. For example, an input-message pair in a language is "1A0B" → "yw", then we would train listeners to select "1A0B" among the 15 candidates after taking "yw" as input. To do so, we still take the cross entropy between the correct candidate and listener's predicted probability distribution as the loss and apply SGD to update the parameters of listeners.

The learning curves of both listeners and speakers on different input representations are shown in Figure 4.

It is clear from Figure 4 that compositional languages require fewer training iterations than the other 2 kinds of languages in almost all the cases, with two exceptions: i) emergent languages has better learnability for listeners on the Bag



(a) (b) (c)

	0A	1A	2A	3A	4A	5A
0B	wq	yq	uq	sq	xq	zq
1B	wy	yy	uy	sy	xy	zy
2B	ws	ys	us	ss	xs	zs
3B	wt	yt	ut	st	xt	zt
4B	wu	yu	uu	su	xu	zu
5B	wv	yv	uv	sv	xv	zv

(d) (e) (f)

	0A	1A	2A	3A	4A	5A
0B	vv	yw	xu	xw	xy	vx
1B	zv	zz	xu	xw	xy	vx
2B	zv	zz	zu	zw	zy	zx
3B	uv	uz	uu	uw	uy	ux
4B	wv	wz	wu	ww	wy	wx
5B	wv	wz	wu	ww	wy	wx

(g) (h) (i)

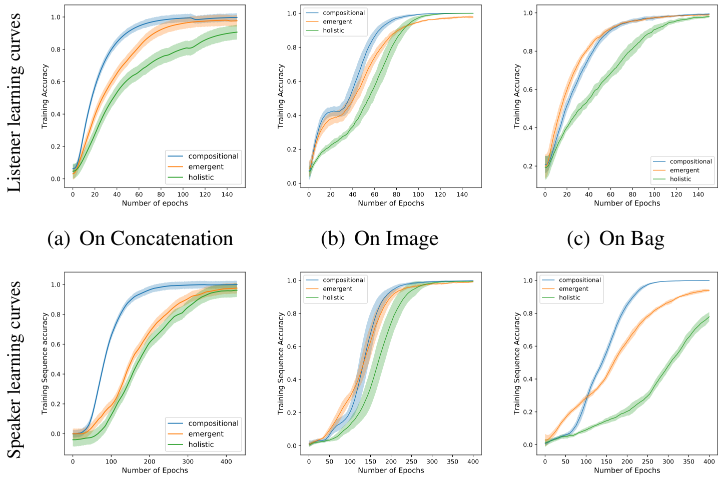
	0A	1A	2A	3A	4A	5A
0B		zv	vz	vv	vv	xv
1B	zy	zu	zw	wz	wv	xw
2B	yz	uu	zz	zt	ww	wx
3B	yz	uy	uq	qz	tz	tw
4B	yx	yy	ur	qq	qt	tt
5B	xy	yr	ry	rx	xq	xt

Figure 3. Experiments results on different input representations. The rows from top to bottom are results for Concatenation, Image and Bag representations respectively. The columns from left to right are: i) smoothed topological similarity (of language having greatest probability) over generations with different population models; ii) smoothed posterior probability of languages having different compositionality (ρ) over generations; iii) final emergent language facilitated by iterated learning, where the first row and first column are numbers of object "A" and "B" respectively.

representation; ii) compositional and emergent languages have almost the same learnability for speakers on the Image representation.

Based on the above results, considering that the topological similarity of final emergent languages given the Bag representation is much lower than Concatenation/Image representations, we argue that iterated learning will amplify the probability of compositional languages only if less training iterations are necessary for listeners to learn the compositional languages.[3] Otherwise, iterated learning

[3] As it is intuitive to show that compositional languages always have lower sample complexity than



(a) On Concatenation (b) On Image (c) On Bag

(d) On Concatenation (e) On Image (f) On Bag

Figure 4. Experiments results on learnability of different kinds of languages, the first row is for listeners and the second row is for speakers. input representations are given below each sub-figure. The lines are means of 10 runs with different random seeds, and the corresponding standard deviations are shown by the shadow area around the lines.

does not show lead to an increase in compositionality. Moreover, our results could also support the hypothesis that compositionality (which is an aspect of linguistic structure) emerges under the pressure of both expressivity and learnability (Smith, Tamariz, & Kirby, 2013), considering that emergent languages have better learnability on Bag representation than compositional languages; as such, those languages still represent a trade-off between learnability and expressivity, but under a slightly different learnability constraint. We are currently investigating why the Bag input encoding makes non-compositional languages more learnable.

5. Conclusion

We use the Bag-Select game to demonstrate that iterated learning leads to the emergence of compositional languages for transmitting numeric concepts. However, this result is dependent on the representations of inputs, and its effectiveness depends on that compositional languages have the optimal learnability for listeners in the communication game. While our findings confirm that structure of languages emerges under the pressure of both expressivity and learnability, at least for deep learning agents, the representation of the input representations affects on learnability and therefore on the structure of the emergent languages.

other non-degenerate languages and thus better learnability for speakers, we actually only need to care about learnability for listeners here, instead of both speakers and listeners as before.

References

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, *12*(2), 229–242.

Cangelosi, A., & Parisi, D. (2012). *Simulating the evolution of language*. Springer Science & Business Media.

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., & Clark, S. (2018). Emergent communication through negotiation. *arXiv preprint arXiv:1804.03980*.

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems* (pp. 2149–2159). Long Beach.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., ... others (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*.

Hurford, J. R. (1989). Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua*, *77*(2), 187–222.

Jang, E., Gu, S., & Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Kirby, S. (1999). *Function, selection, and innateness: The emergence of language universals*. OUP Oxford.

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

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, F., & Bowling, M. (2019). Ease-of-teaching and language structure from emergent communication. *arXiv preprint arXiv:1906.02403*.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... others (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529.

Mordatch, I., & Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In *Thirty-second aaai conference on artificial intelligence.*

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, 400–407.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., et al. (1988). Learning representations by back-propagating errors. *Cognitive modeling*, *5*(3), 1.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., ... others (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354.

Smith, K., Tamariz, M., & Kirby, S. (2013). Linguistic structure is an evolutionary trade-off between simplicity and expressivity. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35). Berlin, Germany.

Steels, L. (2005). The emergence and evolution of linguistic structure: from lexical to grammatical communication systems. *Connection science*, *17*(3-4), 213–230.

Vinyals, O., Bengio, S., & Kudlur, M. (2015). Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*.