Internal and External Pressures on Language Emergence: Least Effort, Object Constancy and Frequency

Anonymous CoNLL submission

Abstract

In previous work, artificial agents were shown to achieve almost perfect accuracy in referential games where they have to communicate to identify images. Nevertheless, the resulting communication protocols rarely display salient features of natural languages, such as compositionality. In this paper, we propose some realistic sources of pressure on communication that avert this outcome. More specifically, we formalise the principle of *least effort* through an auxiliary objective. Moreover, we explore several game variants, inspired by the principle of *object constancy*, in which we alter the frequency, position, and luminosity of the objects in the images. We perform an extensive analysis on their effect through compositionality metrics, diagnostic classifiers, and zero-shot evaluation. Our findings reveal that the proposed sources of pressure result in emerging languages with less redundancy, more focus on high-level conceptual information, and better abilities of generalisation. Overall, our contributions reduce the gap between emergent and natural languages.

1 Introduction

One of the key requirements for a machine to be intelligent is its ability to communicate in natural language (Mikolov et al., 2018). While supervised approaches with labelled texts have recently achieved unprecedented performances in several applications (Chen et al., 2017, *inter alia*), they still neglect fundamental components of natural communication, such as the speakers' intention and the function of their utterances (Clark, 1996).

This functional aspect of language instead is captured by multi-agent games (Kirby, 2002), in which agents have to communicate about some shared input space (e.g. images). Agents usually manage to communicate with success, measured in terms of *task accuracy* (Mordatch and Abbeel, 2018; Cohoi et al., 2018, *inter alia*), if the setting is fully cooperative (Cao et al., 2018). However, Kottur et al. (2017) have shown that the emerged languages rarely display features inherent to natural languages, such as compositionality of meaning and generalisation to novel objects. For instance, agents might develop protocols to refer to specific pixel values, rather than concept-level information (Bouchacourt and Baroni, 2018).

Referential games are a perfect controlled environment to study how *sources of pressure* on the agents affect the 'naturalness' of emergent languages. Previous work has proposed to limit the memory of neural agents across turns of dialogue (Kottur et al., 2017) or to soft-constrain the active vocabulary size (Mordatch and Abbeel, 2018). However, these constraints seem at odds with the capacity of human memory. In this work, we propose a set of yet unexplored but more realistic sources of pressure, either internal to the agents or external, pertaining to the input space.

An internal source of pressure, inspired by the principle of least effort (Zipf, 1935; Haiman, 1983, see § 2.1), compels the agents to keep the length of sentences to the bare minimum. We implement this pressure through an auxiliary loss that incentivises the generation of the end-of-sentence token as early as possible. Several external pressures instead are implemented as game variants, where we control for the frequency, the position, and the illumination of objects in images. These game variants are again motivated by principles governing human perception, such as object constancy (Lorenz, 1977; Gillam, 2000, see § 2.2).

Our results demonstrate that the internal pressure efficiently compresses the sentence lengths and the vocabulary size without loss of accuracy. Moreover, based on established metrics of compositionality (Cohoi et al., 2018; Lazaridou et al., 2018; Bouchacourt and Baroni, 2018) and zero-shot evaluation, we show that agents with pressure towards object constancy achieve the highest scores. Finally, diagnostic classification reveals how the external pressures make agents sensitive to higher-level object properties.

In general, we offer a series of contributions. In addition to a novel model objective and game variants, we establish a methodology to adapt the communication hyper-parameters automatically. Moreover, we draw connections to principles of human cognition, thus aligning the multi-agent game to hypotheses on natural language evolution (Nowak and Krakauer, 1999). We elaborate on such principles in § 2. In § 3, we outline the basic setup for the referential game and the dataset. The auxiliary loss and game variants that operationalise the cognitive principles are described in § 4. We discuss the metrics for evaluation in § 5 and provide the results in § 6. The main conclusions to our work are summarised in § 7.

2 Motivation

The proposed sources of pressure on emergent languages are motivated on the basis of general principles of human communication and perception. In this section, we outline the principles of least effort, object constancy, and object frequency.

2.1 Least Effort

While human speakers try to maximise the distinctiveness of the information conveyed, they also minimise the effort involved. A version of this principle was originally formulated by Zipf (1935, p. 29) – who pointed out the correlation between word frequency rank and word length – and was later generalised to every reduction of linguistic expressions by Haiman (1983) under the name 'Principle of Economy.' This principle is also reminiscent of the maxim of quantity in pragmatics (Grice et al., 1975; Levinson, 2000), which requires to give no more information than needed.

As this principle is a key factor in explaining the variation of natural languages (Haiman, 1983), it posits a realistic constraint on language emergence. Moreover, our operationalisation of the principle allows the model to determine automatically the maximum length of sequences and the number of symbols in its vocabulary. In doing so, the complexity of the emergent language is gauged according to the data and task at hand. This has the methodological advantage of not requiring to preset these hyper-parameters arbitrarily or performing grid search on task accuracy (which rarely corresponds to natural language properties).

2.2 Subjective Constancy

Reality as it is immediately sensed is shapeless and ever-changing. However, animals evolved to various degrees the ability to perceive a reality of objects, namely constant and discrete entities lying behind the tangle of sensation (Lorenz, 1977). The same ability is connected with abstraction: over repeated impressions, animals learn to neglect what is contingent (due to the environment or their internal disposition), and group instances of objects with recurring patterns into the same conceptual class (Gillam, 2000).

Object constancy involves several different and independent mechanisms, regarding, among others: i) the **colour** of the object, under different light conditions (Holst, 1957); ii) the **position** of the object, under different perspectives (Holst, 1969–1970). For instance, bees have to identify flowers by their colour independently from the time of the day (red of dusk or gold of dawn). In our implementation, we manipulate the images in such a way that agents are exposed to the same object with different position or luminosity. As a consequence, we expect the agents to acquire some sort of constancy mechanisms.

2.3 Object Frequency

The distribution of objects and features in the real world are highly non-uniform. Agents encounter objects in the environment with different frequencies. Furthermore, the degree of association between features and objects can vary: for instance, berries evoke the colour blue less vividly than the sky. Frequency facilitates the correct classification of object instances (Nosofsky, 1988). Moreover, Medin and Schaffer (1978) have shown that more frequent stimuli lead to an increasing perceptual differentiation in the region of their features. As a consequence, agents are imprinted with respect to specific features rather than the stimulus as a whole, and stimuli become decomposable into their 'building blocks' (Schyns et al., 1998). Recently, Hendrickson and Perfors (2019) have also shown how a Zipfian distribution of words and referents can accelerate word meaning acquisition compared to a uniform one.

3 Setup

We study language emergence in the context of task-oriented multi-agent games. In the current section, we present our baseline setup (§ 3.1) and the dataset that we use (§ 3.2).

3.1 Game definition

In the game we study, two agents play a referential game. One agent, the Sender, has to describe an image; the other agent, the Receiver, has to pick the correct image out of a line-up of confounders. We follow the setup of Havrylov and Titov (2017):

1. There are N images represented by z-dimensional feature vectors $f_n = \{i_1, ..., i_z\}$. A target image t is sampled and shown to the Sender.
2. The Sender generates a message m with a maximum length L that consists of a sequence of words $\{w_i, \ldots, w_{\leq L}\}$ from a vocabulary of size $|V|$.
3. The Receiver uses m to identify t in a set of images that contains k distracting images and t in random order.

We implement both the Sender and Receiver agents as LSTM networks. Unless otherwise specified, we follow again the training procedures and error definitions of Havrylov and Titov (2017).[1] A scheme of this setup is shown in Figure 1.

3.2 Data

We use images from a modified version of the SHAPES dataset (Andreas et al., 2016). This dataset consists of 30 x 30 pixel images. Each image contains exactly one 2D object, which is characterised by a shape (circle, square, triangle), a colour (red, green, blue), and a size (small, big). The objects are positioned in a logical grid of three rows and three columns. In the baseline setting, we sample both images and distractors uniformly from the images in this space. In the next section, we introduce three alternative versions of the game, in which images are selected following more naturalistic procedures.[2]

4 Formalisation of pressures

As the core contribution of this paper, we propose a series of changes to the baseline setup in order to incorporate model internal and external pressures related to concept learning. In the current section, we describe these alternative setups.

4.1 Least effort pressure

Arguably, communicative success is not the only factor that comes into play in natural interactions. In fact, agents should also abide by the principle of least effort. We formalise this idea with a *vocabulary loss*, that encourages the agents to use shorter messages and fewer words. For each time step *t*, the logits s of the Sender over the vocabulary are discounted by C, the normalised count of distinct words in the vocabulary that have been produced so far. After squashing these values into a probability distribution with Softmax, we estimate its negative log-likelihood. The formula can be written as:

$$\mathcal{L}_v = \sum_{1 \leq t \leq L} - \log \frac{\exp(s_w^{(t)} - C^{(t)})}{\sum_j \exp(s_j^{(t)} - C^{(t)})}$$

$$= \sum_{1 \leq t \leq L} C^{(t)} - s_w^{(t)} + \log \left(\sum_j \exp(s_j^{(t)} - C^{(t)}) \right)$$

where w is the word generated at time step t. Due to the term C, this vocabulary loss is lower when the model uses fewer words. As the end of sequence symbol $<S>$ is part of the vocabulary, the loss is also implicitly encouraging shorter sentences.[3] This auxiliary loss is added to the main loss of the system, with a weighting factor λ, and minimised during optimisation.

4.2 Location invariance

Among the cognitive mechanisms governing object constancy and abstraction, a key role is played by location invariance. This mechanism allows animals to conceive objects as identical even when they move, and the object reflection on their retina has shifted (see § 2.2). We formalise the pressure to develop location invariant concepts by introducing a mismatch between the exact object instance shown to the Sender and Receiver. More precisely,

[1] For brevity, we omit these details from the full paper, but report them in Appendix A.

[2] While we work on synthetic data, the same expedients can be easily applied to natural datasets like COCO (Lin et al., 2014).

[3] We experimented with adding an additional parameter α to explicitly scale the counts of $<S>$and modulate its emission. However, we found the best value of α to be 1.

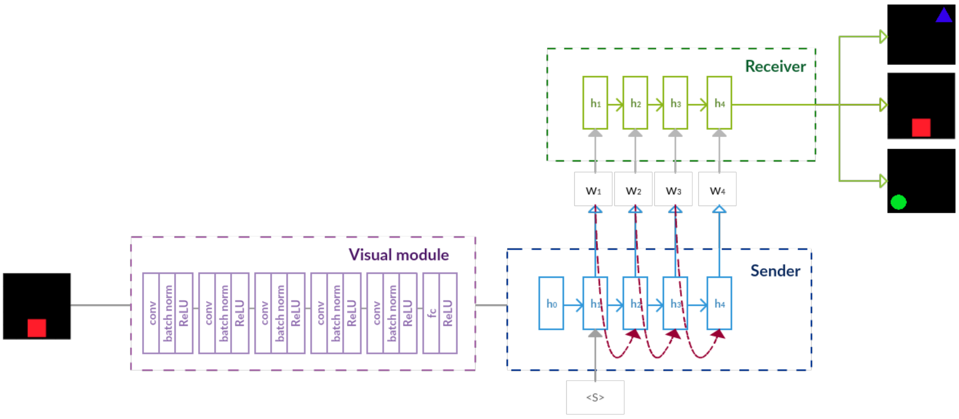

Figure 1: Architecture overview of the Sender and Receiver. The visual module shows the CNN architecture used for extracting features from the input.

when the Sender is shown a target image t, characterised by the quintuple (color, size, shape, horizontal position, vertical position) – the target image t' of the Receiver contains an object with the same shape, colour, and size, but a different position. We hypothesise that this setup will encourage the emergence of general purpose symbols for colour, shape, and size, since agents are pressured to refer to these concepts consistently across different perspectives.

4.3 Colour constancy

Another object-constancy mechanism allows animals to identify objects with altered hues, when the conditions of illumination change (see § 2.2). To introduce this pressure in our game setup, we follow a similar protocol as in § 4.2: we encourage colour constancy by slightly perturbing the Sender's target image t into the target image of the Receiver t'. More specifically, the target image of Sender and Receiver are identical in all dimensions, except their overall *brightness*. Therefore, two different *brightness* values b_1 and b_2 are assigned to each image, so that:

$$0.2 < b_1 < 0.8, \quad 0.2 < b_2 < 0.8, \\ \text{and} \quad |b_1 - b_2| \geq 0.2$$

4.4 World distribution

Finally, we consider the effect of frequency on concept memorisation, by exposing agents to non-uniform distributions of objects (see § 2.3). To obtain such a non-uniform world, we skew the distributions of different shapes $p(shape)$, as well as the conditional probability distribution $p(colour|shape)$. In particular, we sample the probabilities such that for all pairs of distinct shapes $s_1, s_2 \in shapes$, it holds that:

$$|\ (p(s_1) - p(s_2))\ | \leq 0.2$$

And for all pairs of distinct colours $c_1, c_2 \in colours$ and all $s \in shapes$:

$$|\ (p(c_1|s) - p(c_2|s))\ | \leq 0.8$$

We sample images from these new distributions. In the resulting worlds, shapes are more likely to have some colours rather than others, and some shapes are more likely than others.

5 Experiments

We now describe the procedures that we use for training and evaluation, and provide details on the (hyper)parameter settings.

5.1 Architecture and hyperparameters

The Sender and Receiver LSTMs have an embedding size of 256 and a hidden layer size of 512. The number of distracting images k for the Receiver is 3 for all experiments. The initial vocabulary size V of the Sender is 25 and the maximum message length L is 10. We empirically set the weighting factor λ for the vocabulary loss to 0.1.[4]

To pre-process the images before they are given to the agents, we use a visual module inspired by the CNN architecture of Cohoi et al. (2018), consisting of five convolutional layers followed by batch normalisation with ReLU as activation function. Each layer has 20 filters with a kernel size 3 and stride 2 with no padding. This block is followed by a fully connected layer that produces

[4] We found this to be the most stable value in terms of accuracy and emerging language properties.

2048-dimensional feature vectors activated by a ReLU function. The visual model is *pretrained* separately for each game variant, by playing that variant of the game once, resulting in four visual modules that are specialised for a particular game variant. The task accuracies during pretraining were all between 0.9 and 1.[5]

5.2 Training

In all games, we use 75k, 8k, 40k samples for the train, validation and test sets, respectively. We train the agents using Gumbel-Softmax with a temperature of 1.2, a batch size of 128, and the Adam optimiser with an initial learning rate of 0.0001. We use early stopping with a patience of 30 to avoid overfitting. We run every experiment 3 times and report the average results.

5.3 Evaluation

In addition to monitoring game *accuracy*, defined as the ratio of games in which the Receiver correctly identifies the target image, we evaluate the characteristics of the emerging communication with a range of different metrics that have been established in previous work.

Average message length In order to understand to which extent agents manage to compress their communication, we keep track of the average number of tokens in the messages produced by the Sender.

Active symbols The counterpart of the average message length is the *active symbols* metric, which expresses how many symbol types from its vocabulary are used by the Sender.

Message distinctness To succesfully complete the game, it may not be necessary to refer to all features of the input image. Following Cohoi et al. (2018), we used *message distinctness* as an estimate of how much of the image features is captured in a message. Message distinctness is defined as the count of unique messages per batch divided by the batch size. As a reference point for this metric, we compute also the number of distinct images. Generating more messages than the reference point suggests that agents are using multiple messages to refer to the same picture. Conversely, generating fewer messages than the reference point indicates that agents use a shallower language, not covering all aspects of the image.

Perplexity per symbol. As in Havrylov and Titov (2017), we used the perplexity per symbol metric to measure how often a symbol was used in a message to describe the same object:

$$Ppl = \exp(-\sum[s^{(t)} \cdot \log(s^{(t)})])$$

where $s^{(t)}$ are the vocabulary scores (given by an affine transformation of the Sender's hidden state at timestep t) for all symbols in the vocabulary. A lower perplexity shows that the same symbols are consistently used to describe the same objects.

Topographic similarity We study the topographic similarity (TS) between the message and input space, defined as the pairwise Pearson correlation between points in those spaces (Brighton and Kirby, 2006). As in Lazaridou et al. (2018), we use this metric to measure the extent to which similar objects receive similar messages.

Language entropy The language entropy S denotes the variability of the number of symbols in the language. It is given by the formula

$$S = -\sum_{w \in V} [c_w \cdot \log(c_w)]$$

where c_w is the count of the occurrences in the produced messages of each symbol w for all symbols in the vocabulary V.

Representation Similarity Analysis (RSA) is defined analogously to TS, but is computed on the continuous hidden representations of the Sender and Receiver (Kriegeskorte et al., 2008). As in Bouchacourt and Baroni (2018), we use this metric to measure the distance between two points in different embedding spaces. Sender-Receiver RSA indicates the RSA between the Sender's and Receiver's embedding spaces. Sender-Input RSA and Receiver-Input RSA indicate the correlation between the agents and the input space.

6 Results

In this section, we provide the experimental results and discuss them critically. In § 6.1, we establish the effect of the vocabulary penalty on the sequence lengths and vocabulary sizes in the emerging languages. In § 6.2, we compare the impact of different external pressures on the model's ability to generalise in *zero-shot* evaluation. In § 6.3, we consider the compositionality of the protocols evolved in each game variant in the light of the

[5] For the specific results, see Appendix B.

	Acc	$\mu(\ell)$	$\sigma^{\mathbf{2}}(\ell)$	$\|\mathbf{\Sigma}\|$	$\|\mathbf{M}\|$
baseline	0.99	11.0	0.0	20.67	4.81
penalty	0.98	6.10	0.87	13.0	3.54

Table 1: Accuracy, average message length, variance of the message length, number of active symbols Σ, and average number of unique symbols per message M on the test set when playing the baseline and penalised games. All values are averaged over three runs.

metrics described in § 5.3. Finally, we investigate which image features can be *decoded* from the emitted messages in § 6.4.

6.1 Least-effort pressure

Maintaining a maximum vocabulary size and message length of 25 and 10, respectively, we train Sender-Receiver pairs with and without the penalty loss. Results are shown in Table 1 for the baseline setup, and Table 2 for the game variants.

Baseline setup Based on Table 1, 37% fewer symbols were used in games trained with the penalty loss. The average length of the messages decreased in 45%. Additionally, the variance in the message length increased from 0 to 0.87, showing the variability of the sequence lengths needed to play the game, as opposed to always using the maximum allowed length. Moreover, there is 26% more symbol reuse within the sequences in the penalty case, as shown by the lower number of unique symbols per message. In terms of accuracy, however, there is no clear difference between games with and without the vocabulary loss. Using fewer words and shorter messages does not, at least in this case, hamper communication success. This indicates that the original models used unnecessarily many symbols.

Game variants For the different game variants, the penalty has a similar effect on the language statistics (shown in Table 2): fewer words are used, the average message length is shorter, and there is more word reuse per generated message. The language compression is most evident in the location invariance setups, where fewer messages are required to fully describe the input space: two objects are considered identical if they share colour, shape and size, regardless of their position in the grid. The models trained without penalty do not reflect this difference, and use the maximum message length they are allowed.

These results show that the use of the vocabulary loss gives rise to languages with symbol reuse. It allows the model to dynamically adjust the vocabulary size and sequence lengths while still being able to successfully solve the game. Given this positive result, we use the vocabulary penalty with a $\lambda = 0.1$ in all subsequent experiments.

6.2 Zero shot evaluation

To assess how well the agents learned to generalise in the different setups, we run a zero-shot evaluation experiment where agents have to communicate about unseen objects. Following the approach of Cohoi et al. (2018), we retrain a model for each game variant, this time removing three objects from the training set images: red triangle, blue square, and green circle. We then test these the retrained models on 40504 rounds of the game, where in each round the target is one of the held-out objects. The distractors are uniformly sampled from a set of objects containing both the training and held-out objects. The prediction accuracies are reported in Table 3.

All results are above chance level (0.25), which would be the average accuracy if the Receiver chose a random image every time out of the four candidates. The highest communication success was obtained in the colour constancy (without penalty) and world distribution (with penalty) experiments. Interestingly, the models are not similarly *ranked* in the penalty and no penalty conditions, pointing to an interaction between the two different pressures that we do not yet understand.

6.3 Metrics

We report the values for the metrics outlined in § 5.3 for all game variants in Table 4.

Message distinctness The number of distinct images (our reference point, as mentioned in Section 5.3) for the baseline game, the colour constancy game, and the world distribution game, is 162 (3 shapes × 3 colours × 2 sizes × 3 rows × 3 columns). Since this number is larger than the batch size, the expected message distinctness is 1. The baseline model averaged a message distinctness of 0.7880, the colour constancy model 0.4921, and the world distribution model 0.8396. Thus, the world distribution game brings agents the closest to capturing the entirety of the image representation, a finding which will be further confirmed in § 6.4.

Game	Penalty	Acc	$\mu(\ell)$	$\lvert\Sigma\rvert$	$\lvert\mathbf{M}\rvert$
Location invariance	Off	0.91	11.00	12.33	2.85
Colour constancy	Off	0.99	11.00	21.67	3.25
World distribution	Off	0.99	11.00	25.00	4.38
Location invariance	On	0.90	6.66	5.33	2.36
Colour constancy	On	0.99	7.49	10.0	2.64
World distribution	On	0.98	7.04	13.33	3.54

Table 2: Statistics for the game variant models calculated on the test set: accuracy, average message length, number of active symbols Σ, and average number of unique symbols per message M. Averages over 3 runs.

Game	Penalty	Acc
Baseline	Off	0.60
Location invariance	Off	0.33
Colour constancy	Off	**0.71**
World distribution	Off	0.46
Baseline	On	0.40
Location invariance	On	0.36
Colour constancy	On	0.33
World distribution	On	**0.52**

Table 3: Zero-shot accuracy on the four game variants. Average over three runs.

In the location invariance experiment there are only 18 symbolically different images, since two objects are considered the same irrespective of their horizontal and vertical position. With a batch size of 128, this gives an expected message distinctness of 18/128=0.14 per batch. The model averaged a message distinctness of 0.2287, which indicates that the same objects are sometimes referred to with different messages (in other words, contrary to evidence, the model may still consider location to be a relevant property!).

Perplexity per symbol The colour constancy game achieved the lowest perplexity per symbol, both with and without the vocabulary penalty. This means that, on average, 1.3 and 2.2 symbols (respectively) were used to denote the same object, which is preferred over having many redundant symbols referring to the same object.

RSA values Even more revealing is the similarity between the representation of the objects in the agents' embedding spaces, which is what RSA depicts. There is a high RSA Sender-Receiver score in all game variants, with scores peaking when the vocabulary penalty was applied. High RSA Sender-Receiver scores are to be expected since a match on the embedding spaces of the agents is necessary for communication success. However, it is the RSA with respect to the input that indicates whether the semantics of the agents' messages reflects the input structure. Here, similarly to the perplexity per symbol metric, the colour constancy condition triggered the highest scores both for the Sender and the Receiver when the penalty is on. On the other hand, in absence of penalty, the location invariance game obtained the highest (absolute) RSA scores.

Topographic similarity A further indication that the location invariance condition has a positive effect on the semantics of the messages comes from topographic similarity: irrespective of the presence of the penalty, the highest score (i.e., the highest correlation between messages and the object space) was obtained in this game variant.[6]

Language entropy The location invariance game, with and without penalty, also achieved the lowest language entropy as it uses the least symbols of the vocabulary.

6.4 Diagnostic classification of properties

To inspect which properties of the input space are retained by the agent messages, we perform an analysis based on diagnostic classification (Hupkes et al., 2018). We train an RNN to encode the messages generated by the Sender and predict from its final hidden state the value for each symbolic property of the input image (shape, colour, size, horizontal position, vertical position). Table 5 shows the accuracy of each classifier on the test messages. The baseline model has the lowest scores for shape and colour, and is able to solve the task by mostly communicating row and column information. On the other hand, the location invariance experiment cannot rely on position informa-

[6] In the Appendix, we plot the development of agent-input RSA and topographic similarity across the training progress in the four games.

Game	Penalty	Ppl symbol	RSA S-R	RSA S-I	RSA R-I	Top. Sim.	Lang. entropy
Baseline	Off	4.19	0.91	0.71	0.63	0.31	2.73
Location inv	Off	3.11	**0.96**	0.69	0.69	**0.38**	**2.17**
Colour const	Off	**2.18**	0.91	**0.72**	**0.71**	0.35	2.82
World distrb	Off	3.17	0.89	0.66	0.62	0.28	3.00
Baseline	On	1.74	0.95	0.46	0.45	0.20	1.61
Location inv	On	1.82	0.97	**0.58**	**0.62**	**0.30**	**1.59**
Colour const	On	**1.32**	**0.98**	0.51	0.52	0.24	1.73
World distrb	On	1.38	0.96	0.36	0.38	0.11	1.63

Table 4: Metrics on the test set: perplexity per symbol, RSA Sender-Receiver, RSA Sender-Input, RSA Receiver-Input, topographic similarity, and language entropy. Showing the average of three different runs per configuration.

Game	Penalty	Shape	Colour	Size	Row	Column
Baseline	Off	0.56	0.84	0.86	0.98	0.98
Location invariance	Off	0.84	1.00	1.00	0.33	0.33
Colour constancy	Off	0.54	0.82	0.81	1.00	1.00
World distibution	Off	0.80	0.91	0.94	0.99	0.98
Baseline	On	0.53	0.45	0.60	0.93	0.96
Location invariance	On	0.65	0.99	0.91	0.33	0.34
Colour constancy	On	0.36	0.67	0.60	0.99	1.00
World distibution	On	0.68	0.73	0.88	0.97	0.97
Chance		0.33	0.33	0.50	0.33	0.33

Table 5: Test accuracy of the five diagnostic classifiers for the four different games (average of three models).

tion, thus performing at a chance level as expected. Rather, this model mostly encodes information about colour and size while playing the game, thereby supporting the hypothesis that the right environmental pressure encourages the encoding of higher-level information. The colour constancy setting seems to have some moderate impact on the colour semantics encoded by the messages. The best results come once more from the world distribution game: a non-uniform (Zipfian) distribution of the objects induces a language that encodes, with high accuracy, all different properties of the image.

7 Conclusions

While most artificial agents achieve communication success in referential games, the emerging protocols are far from natural. Therefore, we coax the agent languages into developing desirable properties through sources of pressure that are both effective and realistic in terms of human cognition. In particular, we encourage the agents to make the least effort (in terms of sentence length and active vocabulary) through an auxiliary loss. Moreover, inspired by principles of perceptual constancy and frequency, we introduce external pressure by manipulating the appearance and frequency distribution of objects within images. Firstly, we found that least effort reduces message redundancy without loss of communication accuracy. Secondly, according to a series of well established metrics, external pressures facilitate the emergence of communicative protocols with a higher degree of compositionality. Thirdly, some sources of pressure such as colour constancy increase the accuracy in zero-shot communication, hence leading to a better ability to generalise. Finally, we reveal through diagnostic classifiers that agents under external pressures retain high-level information (such as shape or color of objects) rather than local pixel features. In general, the sources of pressure we propose bring forth a series of advantages: 1) they encourage more natural communication protocols; 2) they mitigate the arbitrariness of hyper-parameter setting; 3) they are realistic and justified by general principles of human cognition. In the future, this could help shedding light on the evolution of natural languages.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 39–48.

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

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

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *6th International Conference on Learning Representations, (ICLR)*.

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35.

H.H. Clark. 1996. *Using Language*. [ACLS Humanities E-Book]. Cambridge University Press.

Edward Cohoi, Angeliki Lazaridou, and Nando de Freitas. 2018. Compositional obverter communication learning from raw visual input. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Barbara Gillam. 2000. Perceptual constancy. In A E Kazdin, editor, *Encyclopedia of psychology*, volume 6, pages 89–93. Oxford University Press.

H Paul Grice, Peter Cole, Jerry L Morgan, et al. 1975. Logic and conversation. In Peter Cole and J. Morgan, editors, *Syntax and semantics, vol. 3: Speech acts*, pages 41–58. New York: Academic Press.

John Haiman. 1983. Iconic and economic motivation. *Language*, 59:781–819.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 2146–2156.

Andrew T Hendrickson and Amy Perfors. 2019. Cross-situational learning in a zipfian environment. *Cognition*, 189:11–22.

Erich von Holst. 1957. Aktive Leistungen der menschlichen Gesichtswahrnehmung. *Studium Generale*, 10(4):232.

Erich von Holst. 1969–1970. *Zur Verhaltensphysiologie bei Tieren und Menschen*. Piper.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Simon Kirby. 2002. Natural language from artificial life. *Artificial life*, 8(2):185–215.

Satwik Kottur, Jos M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

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

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Konrad Lorenz. 1977. *Behind the mirror: A search for a natural history of human knowledge*, volume 610. Egmont Books (UK).

Douglas L Medin and Marguerite M Schaffer. 1978. Context theory of classification learning. *Psychological review*, 85(3):207.

Tomas Mikolov, Armand Joulin, and Marco Baroni. 2018. A roadmap towards machine intelligence. *Lecture Notes in Computer Science*, page 2961.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Robert M Nosofsky. 1988. Similarity, frequency, and category representations. *Journal of Experimental Psychology: learning, memory, and cognition*, 14(1):54.

Martin A Nowak and David C Krakauer. 1999. The evolution of language. *Proceedings of the National Academy of Sciences*, 96(14):8028–8033.

Philippe G Schyns, Robert L Goldstone, and Jean-Pierre Thibaut. 1998. The development of features in object concepts. *Behavioral and brain Sciences*, 21(1):1–17.

George Kingsley Zipf. 1935. *The psycho-biology of language*. Boston: Houghton Mifflin.

A Full Game Description

A.1 Agents

We implement both the Sender and Receiver as LSTM networks. The architecture as a whole is depicted in Figure 1.

Sender The inputs of the Sender are the feature representation f of the target image t (which we refer to as f_t) and a special start token $<S>$. Starting from an initial hidden state h_0^S, which is obtained by linearly transforming f_t, at each decoding step i the Sender generates a token w_i by sampling from its output distribution, until the special end of sequence token $<S>$is generated or the maximum sequence length L is reached.

Receiver The Receiver receives the message m generated by the sender as input. It encodes this message and then uses a transformation of its last hidden state h_l^R to select an image from the four images that it is given (one target + three distractors).

A.2 Training Signal

The communication loss of the system is defined by

$$\mathcal{L}_c = \sum_{k=1}^{K} [max(0, 1 - q(t) + q(d_k))]$$

where the score function $q(x) = f_x^T g(h_l^R)$. d is each distracting image, so $K = 3$.

Communication success happens when the target's score is higher than all the distractors' scores.

Additionally, we compute a vocabulary loss defined by:

$$\mathcal{L}_v = \sum_{i=1}^{E} CrossEntropy[s_i - C, w_i]$$

where E represents the effort, taken to be the length of the message uttered by the agent, s_i represents the vocabulary scores at timestep i , C is the normalised counts of all tokens in the vocabulary that have been produced so far, and w_i is the word sampled at timestep i.

The total loss is computed by the weighted sum

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_v$$

B Visual Module Training

The communication success obtained while playing the different games and training the corresponding CNN alongside is shown in Table 6.

Baseline	Location invariance	Colour constancy	World distribution
0.97	0.89	0.89	0.98

Table 6: Test accuracy on the four different games when training the visual module.

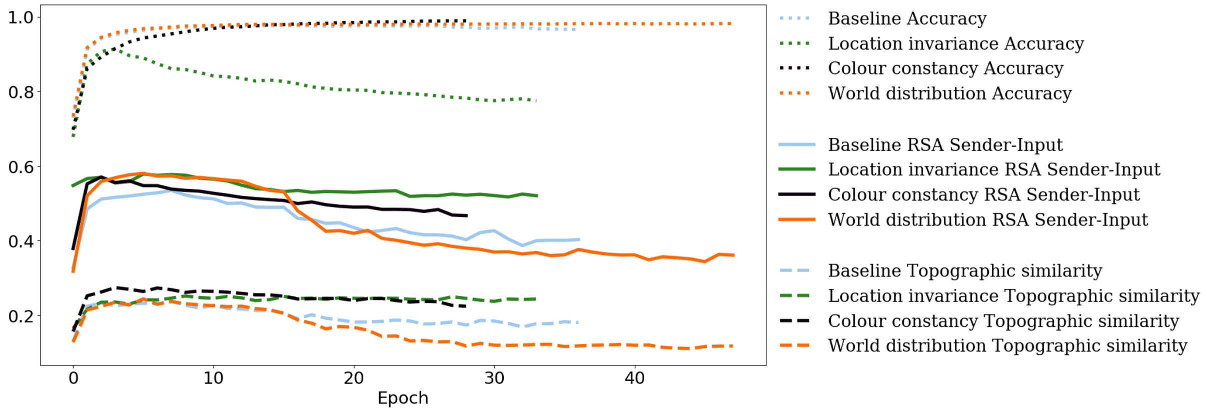

Figure 2: Development of metrics with respect to the input for the four games during training when using the penalty.

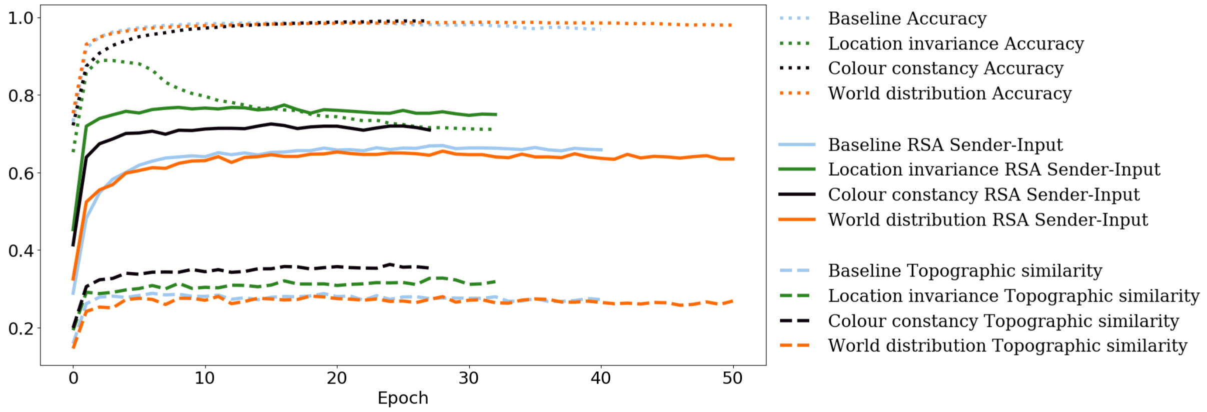

Figure 3: Development of metrics with respect to the input for the four games during training when not using the penalty.