Visual Referential Games Further the Emergence of Disentangled Representations

Kevin Denamganaï, Sondess Missaoui, and James Alfred Walker
Department of Computer Science
University of York
York, UK
kyd500@york.ac.uk, sondess.missaoui@york.ac.uk, james.walker@york.ac.uk

Abstract

Natural languages are powerful tools wielded by human beings to communicate information. Among their desirable properties, compositionality has been the main focus in the context of referential games and variants, as it promises to enable greater systematicity to the agents which would wield it. The concept of disentanglement has been shown to be of paramount importance to learned representations that generalise well in deep learning, and is thought to be a necessary condition to enable systematicity. Thus, this paper investigates how do compositionality at the level of the emerging languages, disentanglement at the level of the learned representations, and systematicity relate to each other in the context of visual referential games. Firstly, we find that visual referential games that are based on the Obverter architecture outperforms state-of-the-art unsupervised learning approach in terms of many major disentanglement metrics. Secondly, we expand the previously proposed Positional Disentanglement (PosDis) metric for compositionality to (re-)incorporate some concerns pertaining to informativeness and completeness features found in the Mutual Information Gap (MIG) disentanglement metric it stems from. This extension allows for further discrimination between the different kind of compositional languages that emerge in the context of Obverter-based referential games, in a way that neither the referential game accuracy nor previous metrics were able to capture. Finally we investigate whether the resulting (emergent) systematicity, as measured by zero-shot compositional learning tests, correlates with any of the disentanglement and compositionality metrics proposed so far. Throughout the training process, statically significant correlation coefficients can be found both positive and negative depending on the moment of the measure. Thus, our results on that end are inconclusive and it shows that more theoretical work is necessary.

1 Introduction

Visual referential games are at an interface between the language processing subfields of language emergence, language grounding, and the computer vision subfield of unsupervised representation learning. While language emergence raises the question of how to make artificial languages emerge with similar properties to natural languages, or at least 'natural-like' protolanguages, with compositionality at the forefront of those properties[3, 21, 43, 55], language grounding is concerned with the ability to ground the meaning of (natural) language utterances into some sensory processes, with the visual modality being the main focus of research. On one hand, emerging artificial languages' compositionality has been shown to further the learnability of said languages [35, 60, 8, 43] and, on the other hand, natural languages' compositionality promises to increase the generalisation ability of the artificial agent that would be able to rely on them as a grounding signal, as it has been found

Preprint. Under review.

to produce learned representations that generalise, when measured in terms of the data-efficiency of subsequent transfer and/or curriculum learning [26, 49, 50, 32]. More in touch with the current context of this study, Chaabouni et al. [12] showed that, when a specific kind of compositionality is found in the emerging languages (the kind that scores high on the positional disentanglement (posdis) metric for compositionality that they proposed), then it is a sufficient condition for systematicity to emerge.

Emerging languages are far from being 'natural-like' protolanguages [38, 10, 11], but sufficient conditions can be found to further the emergence of compositional languages and generalising learned representations (e.g. Kottur et al. [38], Lazaridou et al. [42], Choi et al. [16], Bogin et al. [6], Guo et al. [21], Korbak et al. [36], Chaabouni et al. [12], Denamganaï and Walker [18]). Nevertheless, the ability of neural networks to generalise in a systematic fashion has been called into question, especially when it comes to language grounding in general [28], on relational reasoning tasks [2], or on the SCAN benchmark [41, 46, 44], and more recently the gSCAN benchmark [57]. Neural networks induction biases have been investigated towards finding necessary conditions that favour the emergence of systematicity [28, 59, 37, 40, 58].

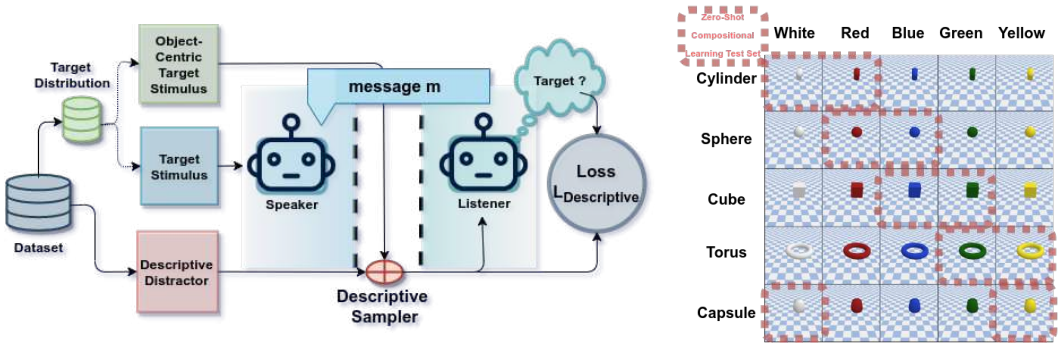
Compositionality & Systematic Generalisation/Systematicity. As a concept, compositionality has been the focus of many definition attempts. For instance, it can be defined as "the algebraic capacity to understand and produce novel combinations from known components"(Loula et al. [46] referring to Montague [47]) or as the property according to which "the meaning of a complex expression is a function of the meaning of its immediate syntactic parts and the way in which they are combined" [39]. Although difficult to define, the commmunity seem to agree on the fact that it would enable learning agents to exhibit systematic generalisation abilities (also referred to as combinatorial generalisation [5]). Some of the ambiguities that come with those loose definitions start to be better understood and explained, as in the work of Hupkes et al. [31]. In this paper, we will refer to compositionality as "the ability to entertain a given thought implies the ability to entertain thoughts with semantically related contents"[20], and thus use it interchangeably with systematicity, following the classification made by Hupkes et al. [31].

Compositionality, as a property of languages, can be difficult to measure. Brighton and Kirby [9]'s *topographic similarity* (**topsim**) which is acknowledged by the research community as the main quantitative metric for compositionality [42, 21, 59, 12, 55]. Recently, taking inspiration from disentanglement metrics, Chaabouni et al. [12] proposed two new metrics entitled **posdis** (positional disentanglement metric) and **bosdis** (bag-of-symbols disentanglement metric), that have been shown to be differently 'opinionated' in the sense that they each seem to capture different ways in which a language can be shown to be compositional.

In accordance with Hill et al. [28], Chaabouni et al. [12] also found that compositionality is not necessary but only sufficient to bring about systematic generalisation, as shown by the fact that non-compositional languages wielded by symbolic (generative) referential game players were enough to support systematic generalisation as evaluated by *zero-shot compositional learning tests* (in which a set of stimuli composed of specific attributes are held-out from the training set, while making sure that the agents are still familiarising themselves with the specific attributes in different contexts/combinations to the held-out ones). One of the necessary conditions they found to foster generalisation is the richness of stimuli (i.e. many possible values per attribute/latent dimension). Therefore, in this paper, on top of measuring the different ways in which emerging languages can be compositional, using topsim, posdis, and bosdis, we will perform zero-shot compositional learning tests in the context of a **"poorness" of stimuli**, in order to measure systematic generalisation abilities/systematicity in the most difficult context.

Disentanglement & Compositionality in Visual Referential Games. Disentanglement, as a property of learned representations, is the objective of the unsupervised representation learning field. It has been shown to be worthy of pursuit for disentangled learned representations are, at least, more sample-efficient and, at best, enabling greater performance, when considering subsequent upstream tasks [26, 24, 63]. Nevertheless, it also suffers from a definition issue [27, 19] that spawned many different metrics with different sensibility and correlations among themselves [45].

In the context of visual referential games (as opposed to symbolic ones), we are entitled to wonder **(i) in the objective of promoting compositionality in the emerging languages, how do disentangled learned representations relate to the emergence of compositionality**, **(ii) from the perspective of unsupervised representation learning, whether some induction biases of referential games**



(a) (b)

Figure 1: **(a):** Illustration of a *descriptive object-centric (partially-observable)* 2*-players/L-signal/*$N = 0$*-round/*$K = 0$*-distractor referential game*. **(b):** Example images of the dataset showing the different objects (in specific viewpoints), along with a possible zero-shot compositional learning train-test split example.

promote disentanglement, and **(iii) how do disentanglement and compositionality relate with (emergent) systematicity**.

Therefore, in order to measure disentanglement, we will make use of the **FactorVAE Score** [33], the Mutual Information Gap (**MIG**) [13], and the **Modularity Score** [56] as they have been shown to be part of the metrics that correlate the least among each other by Locatello et al. [45]. Moreover, the **posdis** and **bosdis** metrics for compositionality have been inspired by the **MIG** metric for disentanglement, therefore it is relevant to verify whether compositionality as measured by the former is correlated with disentanglement as measured by the latter. By including many metrics for both disentanglement and compositionality, we aim to make sure to capture any correlation between the many opinionated definitions of the concepts of disentanglement and compositionality.

Contributions. Firstly, following Chaabouni et al. [12]'s introduction of the posdis and bosdis metrics inspired by the MIG disentangelement metric, we expand on their work by developing a different version, more opinionated in the sense that the compositionality that it highlights is imbued with some disentanglement concerns, we discuss the advantages and disadvantages of both versions via the lenses of decoding and compression in Section 3. Our version finds its motivation in the need to discriminate further between the different kind of compositional languages that emerges in the context of obverter-based referential games, as detailed in Section 4.4.

Secondly, we show experimental results in Section 4.2 that the visual referential game paradigm furthers strong disentanglement at the level of learned representations, outperforming the state-of-the-art method FactorVAE. This paradigm is thus highlighted as an interestingly viable alternative to VAE-based [25, 33, 13] and GAN-based [14] approaches to the unsupervised learning of disentangled representations, in the context of a "poorness" of the stimuli.

Finally we investigated whether the resulting (emergent) systematicity, as measured by zero-shot compositional learning tests, correlates with any of the disentanglement and compositionality metrics proposed so far, in the context of visual referential games (see Section 4.5). Throughout the training process, statistically significant correlation coefficients can be found as both positive or negative depending on the timing of the measure. Thus, our results on that end are inconclusive, and it shows that more theoretical work is necessary.

2 Setup

2.1 Visual Referential Games

Referential/Language games emphasise the functionality of languages, namely, the ability to efficiently communicate and coordinate between agents. Following the nomenclature proposed in Denamganaï and Walker [17], we will focus primarily on a *descriptive object-centric (partially-observable)* 2*-players/*$L = 10$*-signal/*$N = 0$*-round/*$K = 0$*-distractor* referential game variant, as

illustrated in Figure 1a, with a descriptive ratio of 0.5. This ratio stands for the probability of the target stimuli to be shown to the *listener* agent, while the rest of the time it is substituted for the *Descriptive Distractor*(see Figure 1a).

As an object-centric referential game, as opposed to stimulus-centric, the listener and speaker agents are not being presented with the very same target stimuli. Rather, they are being presented with different *viewpoints* on the very same target object shown in the target stimuli, where the word *viewpoint* ought to be understood in a very large sense. Indeed, object-centrism is implemented by designing one of the latent axes of the dataset as a source of invariance for the concept of object, the **object-centric latent axis**. Thus, the listener and speaker agents would be presented with different stimuli that keeps the conceptual object being presented constant, i.e. that keeps constant the values on all the other latent axes but the object-centric latent axis value. This aspect was introduced by Choi et al. [16] (without it being of primary interest), where the pair of agents would literally be shown potentially the same 3D objects under different viewpoint, thus thinking of object-centrism as a *viewpoint* shift is historically relevant.

Concerning the communication channel, the vocabulary V is fixed with 10 ungrounded symbols, plus an eleventh grounded symbol accounting for the *end of sentence* semantic, thus $|V| = 11$. The maximum sentence length L is always equal to 10, thus placing our experiments in the context of an overcomplete communication channel whose capacity is far greater than the number of different meanings that the agents would encounter in our experiments.

We will focus exclusively on two parameterisation of the communication channel. Fisrtly, on the *Straight-Through Gumbel-Softmax* (STGS) approach proposed by Havrylov and Titov [23], as it supposedly allows a richer signal towards solving the credit assignment problem that language emergence poses since the gradient can be backpropagated from the listener agent to the speaker agent, while, in comparison, it cannot be backpropagated when using more commonly adopted approaches based on REINFORCE-like algorithms [66]. And, secondly, we investigate the *Obverter* approach proposed by Batali [4] and updated to the recent deep learning paradigm by Choi et al. [16], Bogin et al. [6], for its induction bias has proven itself powerful. As a well-discussed concept in the Theory of Mind, it posits that the speaker agent should make the assumption that the listener's mind operates similarly to its own, thus using its own understanding of a given utterance as a good enough proxy of the expected listener's understanding of the same utterance, when trying to convey a given meaning in the obvious constraint that the listener's state of mind is inaccessible to it.

Having described the referential game setup, the following section provides details on the architecture of the *speaker* and *listener* agents and the dataset used[1].

2.2 Agent Architectures

Each agent consist of, at least, a language module and a visual module. The *listener* agents also incorporates a third decision module that combines the outputs of the other two modules. In the case of the Obverter approach, both agents play the role of the *listener* from one round to another, therefore they both incorporate it. As this work focuses on learning 'good' representations, we make the architectural choice of sharing the visual module between the pairs of agents, as preliminary experiments showed it increases sample-efficiency.

In the case of the STGS approach, while the *speaker* agent is prompted to produce the output string of symbols with a *Start-of-Sentence* symbol and the visual module output as an initial hidden state, the *listener* agent consumes the string of symbols with the null vector as the initial hidden state. In the following subsections, we detail each module architecture in depth.

Visual Module. The visual module $f(\cdot)$ consists of four 3×3 convolutional layers with stride 2, followed by a fully-connected layer reducing the feature maps to flattened vectors of dimension 32. The two first convolutional layers have 32 filters, whilst the last two layers have 64. Each convolutional layer is followed by a 2D batch normalisation layer, which are found crucial in the case of the Obverter approach (without it, training does not take off), and the resulting outputs are passed through ReLU activation functions. The bias parameters of the convolutional layers are not used, as it is common when using batch normalisation layers. Inputs are resized to 64×64, thus yielding

[1]For more details, please refer to our code released at: `https://github.com/Near32/ReferentialGym/tree/develop/zoo/referential-games%2Bcompositionality%2Bdisentanglement`.

feature maps of dimension $64 \times 4 \times 4$. The input to the final fully-connected layer is a flattened representation of dimension 1024.

Language Module. The language module $g(\cdot)$ consists of a one-layer GRU network [15] in the case of the Obverter approach, and a one-layer LSTM network [29] in the case of the STGS, with 64 hidden units. In the context of the *listener* agent, the input message $m = (m_i)_{i \in [1,L]}$ (produced by the *speaker* agent) is represented as a string of one-hot encoded vectors of dimension $|V|$ and embedded in an embedding space of dimension 64 via a learned Embedding. The output of the *listener* agent's language module, $g^l(\cdot)$, is the last hidden state of the LSTM layer or GRU layer, $h^l_L = g^l(m_L, h^l_{L-1})$. In the context of the *speaker* agent's language module $g^s(\cdot)$, the output is the message $m = (m_i)_{i \in [1,L]}$ consisting of one-hot encoded vectors of dimension $|V|$, which are sampled using the STGS approach from a categorical distribution $Cat(p_i)$ where $p_i = Softmax(\nu(h^s_i))$, provided ν is an affine transformation and $h^s_i = g^s(m_{i-1}, h^s_{i-1})$. $h^s_0 = f(s_t)$ is the output of the visual module, given the target stimulus s_t.

Decision Module. Depending on the approach, the decision module can be very different. In the case of the STGS, similarly to Havrylov and Titov [23], the decision module builds a probability distribution over a set of $K+1$ stimuli/images $(s_0, ..., s_K)$, consisting of K distractor stimuli and the target stimulus (or possibly another distractor in the case of the descriptive games), given a message m:

$$p((d_i)_{i \in [0,K]} | (s_i)_{i \in [0,K]}; m) = Softmax((h^l_L \cdot f(s_i)^T)_{i \in [0,K]}). \quad (1)$$

In our case though, since there are no distractors, we set $K = 0$, and, since the agents play a descriptive game, a final category is added to encode the meaning/prediction that none of the $K+1$ stimuli is the target stimulus that the *speaker* agent was 'talking' about. The addition is made at the logit level as a learnable logit value, $logit_{no-target}$, it is an extra parameter of the model. In this case the decision module output is as follows:

$$p((d_i)_{i \in [0,K+1]} | (s_i)_{i \in [0,K]}; m) = Softmax((h^l_L \cdot f(s_i)^T)_{i \in [0,K]} \cup \{logit_{no-target}\}), \quad (2)$$

where $p(d_{K+1} | (s_i)_{i \in [0,K]}; m)$ is the predicted likelihood that none of the experienced stimuli is the target stimulus.

Similarly to Choi et al. [16], the decision module of Obverter agents is a two-layer fully-connected network $d(\cdot)$ with 128 hidden units and a ReLU activation function, taking as input the concatenation of the outputs of the visual and language modules, and outputting 2 logits for each input stimulus. The first logit $d^0(s_i)$ encodes that the listener agent is experiencing the same stimulus than the speaker agent, while the second one $d^1(s_i)$ encodes the negation. Given a set of $K+1$ stimuli, the prediction distribution of the listener agent is as follows:

$$p((d_i)_{i \in [0,K+1]} | (s_i)_{i \in [0,K]}; m) = Softmax((d^0(s_i))_{i \in [0,K]} \cup \{\frac{1}{K+1} \sum_{i=0}^{K} d^1(s_i)\}). \quad (3)$$

We also experimented with taking the minimum and maximum of the set of logits that encodes the negation, but the most efficient approach is by using an average pooling over the set of negation-encoding logits, as presented in Equation 3.

2.3 Dataset

In the following experiments, learning agents observe visual stimuli from particular train/test splits of a replication of the dataset used in the original work of Choi et al. [16], that we will refer to as 3DShapesPyBullet, and that we open-source (HIDDEN-FOR-REVIEW-PURPOSE).

Built in order to be employed as a benchmark for disentangled representation learning, the 3DShapesPyBullet dataset consists of visual representations of objects with the following attributes (illustrated in Figure 1b): a **Shape** attribute with 5 different values; a **Color** attribute with 5 different values; a **Viewpoint** attribute with 10 different values. All combinations of values along any of the generative factors/attributes/latent axes are part of the dataset, thus yielding a dataset of size 250, which fit in the context of the **"poorness" of the stimuli**. Using an object-centric referential game, we define the **Viewpoint** latent axis as our **object-centric latent axis**.

In order to build a zero-shot compositional learning test set, half of the possible values for each attribute are defined as testing-purpose values. Subsequently, for every value on the **Shape** latent axis, 2 out of the 5 available **Color** values are defined as testing-purpose values for this specific **Shape** value and the resulting stimuli are held-out, as shown in Figure 1b. Rather than sampling the held-out **Color** values to a given **Shape** value in a random fashion, we make sure that they are different from one **Shape** value to another and that all the possible **Color** values are represented at least with one **Shape** value in the training set. This ensures that the agent is familiar with all the possible values on each latent axis while remaining unaware of a subset of all the combinations possible.

The choice of defining around half of the possible values on each attribute is motivated by the results of Bahdanau et al. [2] when evaluating for emergent systematicity: a CNN+LSTM architecture, similar to that of ours here, has been found sufficiently challenged by such a train/test split.

In the original work of Choi et al. [16], a specific batch sampling scheme was used in order to guide the language emergence process towards greater completeness and informativeness (see Section 3.1), by sampling batches of stimuli with controlled amount of *Descriptive Distractors* that would differ from the target exclusively in terms of the **Shape** or the **Color** attribute value, in a form of supervised contrastive learning. We do not make use of this sampling scheme as we place ourselves in the context of unsupervised learning, where the actual latent factors should not be known in advance.

3 Metrics

In order to measure disentanglement, we adapted to our framework the implementations of the **FactorVAE Score** [33], the Mutual Information Gap (**MIG**) [13], and the **Modularity Score** [56] from the open-source work of Locatello et al. [45]. The choice of those metrics is motivated by their results showing that they correlate the least among each other, which we therefore understand as being differently opinionated about what it means for learned representations to be called disentangled.

By incorporating those different opinions/definitions about both disentanglement and compositionality, we aim to answer more thoroughly the question of finding whether disentanglement, compositionality, and systematicity are linked in the context of (discriminative) referential games. Moreover, the **posdis** and **bosdis** metrics for compositionality have been inspired by the **MIG** metric for disentanglement, therefore it is relevant to verify whether compositionality as measured by the former is correlated with disentanglement as measured by the latter. By including many metrics for both disentanglement and compositionality, we aim to make sure to capture any correlation between the many opinionated definitions of the concepts of disentanglement and compositionality.

3.1 Positional Disentanglement Metrics Formulation

As detailed in Chen et al. [13], the Mutual Information Gap (MIG) metric is defined for latent variables $(z_j)_{j\in[1,J]}$ and ground truth factors $(v_k)_{k\in[1;K]}$, where J is the dimension of the model at the level of its latent variables and K is the number of factors in the current dataset. The MIG "enforce[s] axis-alignment by measuring the difference between the top two latent variables with highest mutual information", for each ground truth factor. Summing over all ground truth factors yield the MIG score:

$$MIG(z, v) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{\mathcal{H}(v_k)}\Big(\mathcal{I}(z_{j_k^{(1)}}; v_k) - \mathcal{I}(z_{j_k^{(2)}}; v_k)\Big) \tag{4}$$

where $j_k^{(l)}$ is the index of the latent variable with the l-th highest mutual information with the k-th ground truth factor, for instance $z_{j_k^{(1)}} = argmax_{z_j}\mathcal{I}(z_j; v_k)$, and $\mathcal{H}(v_k)$ refers to the entropy of the k-th ground truth factor and is used for normalization purpose since $\forall(j,k) \in [1,J]\times[1,K]$, $0 \leq \mathcal{I}(z_j, v_k) \leq \mathcal{H}(v_k)$.

On the other hand, the Positional Disentanglement (posdis) metric is defined for symbol positions within sentences $(s_j)_{j\in[1,c_{len}]}$ and, similarly to the MIG case, ground truth factors $(v_k)_{k\in[1;K]}$, where c_{len} is the length of sentences uttered by the speaker of consideration (when omitting symbol positions with zero entropy). The posdis is defined, on the contrary to MIG, for each symbol position s_j within sentences, and summed over those possible symbol positions that are not constant (non-null entropy),

rather than over the ground truth factors:

$$posdis(s, v) = \frac{1}{c_{len}} \sum_{j=1}^{c_{len}} \frac{1}{\mathcal{H}(s_j)} \Big(\mathcal{I}(v_{k_j^{(1)}}; s_j) - \mathcal{I}(v_{k_j^{(2)}}; s_j) \Big) \tag{5}$$

where $k_j^{(l)}$ is the index of the ground truth factor with the l-th highest mutual information with the j-th symbol position within sentences, for instance $v_{k_j^{(1)}} = argmax_{v_k} \mathcal{I}(v_k; s_j)$.

Given that posdis and bosdis are defined in the context of a generative referential game where the listener agent aims to reconstruct the ground truth factors, it can be argued that these metrics take the viewpoint of the listener agent who attempts to decode the speaker's sentences in order to output predicted factors (aiming to converge on the ground truth factors). In other words, the philosophy of theses metrics is that of a decoding problem rather than that of a compression problem.

The compression approach would revolve around the speaker agent's viewpoint as it aims to compress the observed stimuli, that reflects the ground truth factors (or are the ground truth factors themselves in the case of symbolic referential games), into **informative**, **complete** and hopefully compositional sentence utterances. To further flesh out the parallel with some disentanglement definitions [19, 56], one could easily swap compositional in the last sentence for **modular**. Thus, a speaker-centred positional disentanglement metric would straightforwardly apply the formulae of Equation 4:

$$speaker - posdis(s, v) = MIG(s, v) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{\mathcal{H}(v_k)} \Big(\mathcal{I}(s_{j_k^{(1)}}; v_k) - \mathcal{I}(s_{j_k^{(2)}}; v_k) \Big) \tag{6}$$

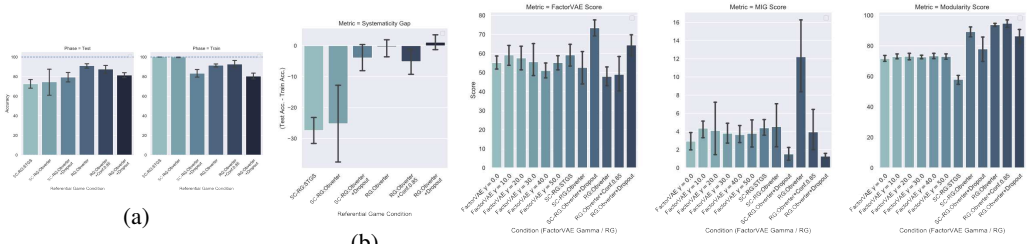
The main advantage of the speaker-centred formulation over the listener-centred one is twofold. Firstly, it is able to discriminate between complete and incomplete languages, following the definition of completeness by Eastwood and Williams [19]: "the degree to which each underlying factor is captured by a single code variable", or, in our case, a single symbol position within sentences. For instance, a language that would capture some underlying factor v_{k_1} using two or more symbol positions within sentences, for instance in a redundant fashion, would yield a lower speaker-centred posdis than languages that would capture it with only one. In the context of cheap talk where there is no cost to communication symbol usage, it might be of lesser importance, but using both metrics in concert would at least enable more insights into the kind of language that emerged.

Secondly, most importantly, the speaker-centred formulation cares about the informativeness ("the amount of information that a representation captures about the underlying factors of variation"[19]) of the language with respect to **all** the ground truth factors. Indeed, while the mutual information $\mathcal{I}(\,;\,)$ is symmetric, the posdis and MIG scores are not, and especially when it comes to informativeness: for instance, a language whose sentences would consist of only one symbol position with non-null entropy used to capture **only one of the many** ground truth factors would score perfectly on the listener-centred posdis formulation, but poorly on the speaker-centred formulation because the other underlying ground truth factor have not been captured.

In terms of disadvantage, the speaker-centred formulation of posdis loses the relaxation made by the listener-centred posdis when compared to topographic similarity: the speaker-centred posdis would penalise languages that do not exhibit a one-to-one attribute-position mapping. Borrowing from Chaabouni et al. [12], the different metrics, topsim, bosdis, listener-centred posdis, and speaker-centred posdis are differently opinionated about what it means for a language to be compositional, and we argue that the speaker-centred formulation is one of the most opinionated so far. Using them in concert is thus allowing more insights about the kind of compositionality that may emerge in each of the artificial languages we may consider.

4 Results

In this section, we detail our results when training pairs of agents for $4,000$ epochs on the 3DShapesPyBullet dataset, with 5 random seeds for the contexts with the synthetically compositional communication channels (SC-RG) and between 5 and 15 random seeds for the normal referential game contexts (RG). We used the Adam optimizer [34] with a learning rate of $6.0e^{-4}$ and PyTorch's [52] default hyperparameters. The batch size was 64, as proposed in Kim and Mnih [33]. Thus, in the case of the Obverter-based pairs of agents, the role alternation period was 2.



(a) (b) (c)

Figure 2: Zero-shot compositional learning testing and training accuracy **(a)** and Systematicity gap (**(b)**:difference between accuracy on test and train sets) for the different referential game approaches, at the end of training. **(c)**: Disentanglement scores of the different approaches. The FactorVAE approach is outperformed by the different Obverter-based referential game approaches.

4.1 Obverter vs. STGS

Figure 2a and 2b detail at the end of training the referential game accuracy against the training set and the zero-shot compositional learning test set, and the systematicity gap, respectively, for both Obverter-based and STGS-based pairs of agents. Focusing on the SC-RG contexts, we can argue that even when the language is compositional in the sense of the posdis metric, the Obverter-based approaches outperform the STGS-based approach in terms of systematicity (higher absolute accuracies shown in Figure 2a, and less negative gap shown in Figure 2b). Going forward, we focus on the Obverter-based approaches.

4.2 Referential Games further Disentanglement

Figure 2c details the disentanglement scores of the different approaches. Even when fine-tuning the FactorVAE's hyperparameter γ, the different Obverter-based referential game approaches outperform the state-of-the-art FactorVAE approach.

We observe a trade-off between the MIG and FactorVAE scores depending on whether a Dropout layer is added just before the Decision modules of the Obverter-based approaches. It highlights the Dropout mechanism as a driver for FactorVAE-centred disentanglement.

Comparing the results when using a synthetically compositional communication channel (SC-RG) from the results when using the normal referential game setting (RG), it is striking to see that disentanglement in the sense of the MIG score is maximized in the RG context, while all the other disentanglement metrics are maximised in the SC-RG contexts, when comparing to respective architecture/approach in the RG context. This result seems to imply that, in the absence of Dropout layers, there is some specific regularisation effect taking place in the Obverter-based RG setting that promotes the emergence of disentanglement in the sense of the MIG score.

4.3 Compositional Language does not further Listeners' Systematicity

In parallel to the results found in Chaabouni et al. [12] (showing that when the emerging language is compositional enough in the sense of the posdis metric then the agents wielding it are highly likely to have systematic abilities), Figure 2b shows that, in spite of learning from a posdis-biased compositional language, neither the listener from an Obverter-based approach or from a STGS-based approach are able to bridge the systematicity gap.

Of the Properties of Symbolic Stimuli. The main difference between our context and theirs is that they were using symbolic stimuli, while our results is in the context of visual stimuli. Therefore, it could be argued that a specific property found in symbolic stimuli might be a necessary condition for systematicity to emerge when the emerging languages are compositional in a posdis fashion.

Similarly to Montero et al. [48]'s interpretation of Chaabouni et al. [12]'s work, we were expecting this specific property to be captured within one of the disentanglement metrics. Yet, our disentanglement scores found in the context of synthetically compositional communication channels (see Fig. 2c) are differently high depending on the approach. Of particular interest, the presence of the Dropout layer

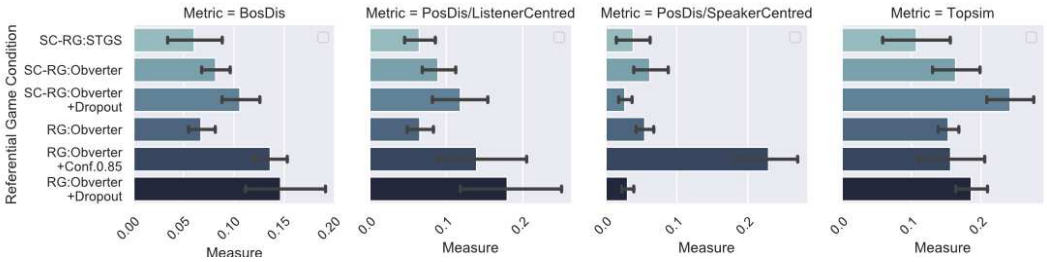


Figure 3: Compositionality metrics for the different referential game approaches, at the end of training.

just before the Decision modules of the Obverter-based agents was highlighted earlier as a mechanism furthering high disentanglement of the learned representations in the sense of the FactorVAE score, and, in the case of the systematicity gap, it induces the least negative systematicity gap among the SC-RG contexts, and the most positive systematicity gap among the RG settings.

Thus, our results would push forward the idea that, on top of promoting generalisation in the broader sense, the addition of specific Dropout layers in the architecture also brings about systematicity (/compositional or algebraic generalisation), in the context of visual referential games, and high disentanglement of learned representations in the sense of the FactorVAE score is correlated to high systematicity.

Of the Importance of the Language Emergence Process. It is important to also highlight the limitation that our SC-RG contexts do not allow for the non-stationary process through which the language used by the pair of agents would emerge as highly compositional in the sense of the posdis metric, on the contrary to Chaabouni et al. [12]'s protocol. Therefore, both the stimuli nature (symbolic or visual), on one hand, and the presence or lack there off of the language emergence process, on the other hand, could be necessary conditions for the emergence of systematicity, when the language is already compositional in the sense of the posdis metric.

4.4 Disentangling Compositionality in Emerging Languages?

Figure 3 details the different attributes of the emerging languages in the different referential game approaches. The comparison of the speaker-centred and listener-centred posdis results in the RG contexts highlights that reducing the Obverter's confidence threshold (from the default value of 0.98 to 0.85) significantly affects the informativeness and completeness of the emerging language, enabling us to discriminate between further kind of positionally disentangled compositionality.

Recalling the MIG scores in the RG contexts from Figure 2c, we can see that high (or low) disentanglement in the sense of the MIG score metric does not correlate with high (or low) compositionality in the sense of any posdis metrics. This result subsequently thickens the mystery behind disentanglement and compositionality's relationship.

4.5 Disentanglement, Compositionality, & Systematicity are not Naïvely Correlated

We compute Spearman ρ correlation matrices between the different metrics throughout the training process of Obverter-based approaches (see Appendix A), from epoch 200 to 4000, with an interval of 200 epoch between measures. While some statistically significant coefficients can be found throughout the training process, they are not consistent and rather often contradictory, irrespective of the context. From those results, the best we can conclude is that our current concepts for disentanglement, compositionality, and systematicity do not naïvely correlate. We leave it to future works to investigate analysis of a different kind, for instance as time-series variables.

5 Conclusion

Primarily, we found that Obverter-based visual referential games outperform the state-of-the-art unsupervised learning FactorVAE approach in terms of many major disentanglement metrics, and

highlighted the necessary design choices that lead to that result. Secondly, we expanded the posdis metric in a principled way with respect to disentanglement concepts of informativeness and completeness, thus allowing for further discrimination between the different kind of compositional languages that emerge in the context of referential games. Finally, investigating the relationship between disentanglement, compositionality and systematicity in this unique paradigm of visual referential games, our results provide further evidence that compositionality, disentanglement, and systematicity are not related as expected and are, thus, not fully understood yet.

Broader Impact

This work consists solely of simulations, thus alleviating some of the ethical concerns, as well as concerns regarding any consequences emerging due to the failure of the system presented. With regards to the ethical aspects related to its inclusion in the field of Artificial Intelligence, we argue that our work aims to have positive outcomes on the development of human-machine interfaces, albeit being not yet mature enough to aim for this goal. The current state of our work does not allow us to extrapolate towards negative outcomes.

This work should benefit the research community of language emergence and grounding, in its current state.

Acknowledgments and Disclosure of Funding

This work was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) [EP/L015846/1].

We gratefully acknowledge the use of Python[62], IPython[54], SciPy[64], Scikit-learn[53], Scikit-image[61], NumPy[22], Pandas[65, 51], OpenCV[7], PyTorch[52], TensorboardX[30], Tensorboard from the Tensorflow ecosystem[1], without which this work would not be possible.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

[2] D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. Courville. Systematic Generalization: What Is Required and Can It Be Learned? *International Conference on Learning Representations*, nov 2019. URL `http://arxiv.org/abs/1811.12889`.

[3] M. Baroni. Linguistic generalization and compositionality in modern artificial neural networks. mar 2019. URL `http://arxiv.org/abs/1904.00157`.

[4] J. Batali. Computational simulations of the emergence of grammar. *Approach to the Evolution of Language*, pages 405–426, 1998.

[5] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks. 2018. URL `https://arxiv.org/pdf/1806.01261.pdf`.

[6] B. Bogin, M. Geva, and J. Berant. Emergence of Communication in an Interactive World with Consistent Speakers. sep 2018. URL `http://arxiv.org/abs/1809.00549`.

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[8] H. Brighton. Compositional syntax from cultural transmission. *MIT Press*, Artificial, 2002. URL https://www.mitpressjournals.org/doi/abs/10.1162/106454602753694756.

[9] H. Brighton and S. Kirby. Understanding Linguistic Evolution by Visualizing the Emergence of Topographic Mappings. *Artificial Life*, 12(2):229–242, jan 2006. ISSN 1064-5462. doi: 10.1162/artl.2006.12.2.229. URL http://www.mitpressjournals.org/doi/10.1162/artl.2006.12.2.229.

[10] R. Chaabouni, E. Kharitonov, E. Dupoux, and M. Baroni. Anti-efficient encoding in emergent communication. *NeurIPS*, may 2019. URL http://arxiv.org/abs/1905.12561.

[11] R. Chaabouni, E. Kharitonov, A. Lazaric, E. Dupoux, and M. Baroni. Word-order biases in deep-agent emergent communication. may 2019. URL http://arxiv.org/abs/1905.12330.

[12] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, and M. Baroni. Compositionality and Generalization in Emergent Languages. apr 2020. URL http://arxiv.org/abs/2004.09124.

[13] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud. Isolating sources of disentanglement in VAEs. https://papers.nips.cc/paper/2018/file/1ee3dfcd8a0645a25a35977997223d22-Paper.pdf. Accessed: 2021-3-17.

[14] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *arXiv preprint arXiv:1606.03657*, 2016.

[15] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[16] E. Choi, A. Lazaridou, and N. de Freitas. Compositional Obverter Communication Learning From Raw Visual Input. apr 2018. URL http://arxiv.org/abs/1804.02341.

[17] K. Denamganaï and J. A. Walker. Referentialgym: A nomenclature and framework for language emergence & grounding in (visual) referential games. *4th NeurIPS Workshop on Emergent Communication*, 2020.

[18] K. Denamganaï and J. A. Walker. On (emergent) systematic generalisation and compositionality in visual referential games with straight-through gumbel-softmax estimator. 2020.

[19] C. Eastwood and C. K. I. Williams. A framework for the quantitative evaluation of disentangled representations. https://openreview.net/pdf?id=By-7dz-AZ. Accessed: 2021-4-4.

[20] J. A. Fodor, Z. W. Pylyshyn, et al. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[21] S. Guo, Y. Ren, S. Havrylov, S. Frank, I. Titov, and K. Smith. The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *arXiv preprint arXiv:1910.05291*, 2019.

[22] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

[23] S. Havrylov and I. Titov. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. may 2017. URL http://arxiv.org/abs/1705.11192.

[24] I. Higgins, A. Pal, A. Rusu, L. Matthey, C. Burgess, A. Pritzel, M. Botvinick, C. Blundell, and A. Lerchner. DARLA: Improving Zero-Shot Transfer in Reinforcement Learning. URL https://arxiv.org/pdf/1707.08475.pdf.

[25] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2 (5):6, 2017.

[26] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P. Burgess, M. Botvinick, D. Hassabis, and A. Lerchner. SCAN: Learning Abstract Hierarchical Compositional Visual Concepts. jul 2017. URL http://arxiv.org/abs/1707.03389.

[27] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner. Towards a Definition of Disentangled Representations. dec 2018. URL http://arxiv.org/abs/1812.02230.

[28] F. Hill, A. Lampinen, R. Schneider, S. Clark, M. Botvinick, J. L. McClelland, and A. Santoro. Environmental drivers of systematicity and generalization in a situated agent. Oct. 2019.

[29] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[30] T.-W. Huang. Tensorboardx, 2018. URL https://github.com/lanpa/tensorboardX.

[31] D. Hupkes, V. Dankers, M. Mul, and E. Bruni. Compositionality decomposed: how do neural networks generalise? aug 2019. URL http://arxiv.org/abs/1908.08351.

[32] Y. Jiang, S. Gu, K. Murphy, and C. Finn. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. jun 2019. URL http://arxiv.org/abs/1906.07343.

[33] H. Kim and A. Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[34] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] S. Kirby. Learning, bottlenecks and the evolution of recursive syntax. 2002.

[36] T. Korbak, J. Zubek, Ł. Kuciński, P. Miłoś, and J. Rączaszek-Leonardi. Developmentally motivated emergence of compositional communication via template transfer. oct 2019. URL http://arxiv.org/abs/1910.06079.

[37] K. Korrel, D. Hupkes, V. Dankers, and E. Bruni. Transcoding compositionally: using attention to find more generalizable solutions. jun 2019. URL http://arxiv.org/abs/1906.01234.

[38] S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. jun 2017. URL http://arxiv.org/abs/1706.08502.

[39] M. Krifka. Compositionality. *The MIT encyclopedia of the cognitive sciences*, pages 152–153, 2001.

[40] B. M. Lake. Compositional generalization through meta sequence-to-sequence learning. jun 2019. URL http://arxiv.org/abs/1906.05381.

[41] B. M. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *35th International Conference on Machine Learning, ICML 2018*, 7:4487–4499, oct 2018. URL http://arxiv.org/abs/1711.00350.

[42] A. Lazaridou, K. M. Hermann, K. Tuyls, and S. Clark. Emergence of Linguistic Communication from Referential Games with Symbolic and Pixel Input. apr 2018. URL http://arxiv.org/abs/1804.03984.

[43] F. Li and M. Bowling. Ease-of-Teaching and Language Structure from Emergent Communication. jun 2019. URL http://arxiv.org/abs/1906.02403.

[44] A. Liška, G. Kruszewski, and M. Baroni. Memorize or generalize? Searching for a compositional RNN in a haystack. feb 2018. URL http://arxiv.org/abs/1802.06467.

[45] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. Oct. 2020.

[46] J. Loula, M. Baroni, and B. M. Lake. Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks. jul 2018. URL http://arxiv.org/abs/1807.07545.

[47] R. Montague. Universal grammar. *Theoria*, 36(3):373–398, 1970.

[48] M. L. Montero, C. J. Ludwig, R. P. Costa, G. Malhotra, and J. Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qbH974jKUVy.

[49] I. Mordatch and P. Abbeel. Emergence of Grounded Compositional Language in Multi-Agent Populations. URL https://arxiv.org/pdf/1703.04908.pdf.

[50] K. Moritz Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, M. Wainwright, C. Apps, D. Hassabis, P. Blunsom, and D. London. Grounded Language Learning in a Simulated 3D World. URL https://arxiv.org/pdf/1706.06551.pdf.

[51] T. pandas development team. pandas-dev/pandas: Pandas, Feb. 2020. URL https://doi.org/10.5281/zenodo.3509134.

[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL http://jmlr.org/papers/v12/pedregosa11a.html.

[54] F. Perez and B. E. Granger. Ipython: A system for interactive scientific computing. *Computing in Science Engineering*, 9(3):21–29, 2007. doi: 10.1109/MCSE.2007.53.

[55] Y. Ren, S. Guo, M. Labeau, S. B. Cohen, and S. Kirby. Compositional Languages Emerge in a Neural Iterated Learning Model. feb 2020. URL http://arxiv.org/abs/2002.01365.

[56] K. Ridgeway and M. C. Mozer. Learning deep disentangled embeddings with the F-Statistic loss, 2018.

[57] L. Ruis, J. Andreas, M. Baroni, D. Bouchacourt, and B. M. Lake. A benchmark for systematic generalization in grounded language understanding. Mar. 2020.

[58] J. Russin, J. Jo, R. C. O'Reilly, and Y. Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics. apr 2019. URL http://arxiv.org/abs/1904.09708.

[59] A. Słowik, A. Gupta, W. L. Hamilton, M. Jamnik, S. B. Holden, and C. Pal. Exploring Structural Inductive Biases in Emergent Communication. feb 2020. URL http://arxiv.org/abs/2002.01335.

[60] K. Smith, S. Kirby, H. B. A. Life, and U. 2003. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–389, 2003. URL https://www.mitpressjournals.org/doi/abs/10.1162/106454603322694825.

[61] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, and T. Yu. scikit-image: image processing in python. *PeerJ*, 2:e453, 2014.

[62] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[63] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are disentangled representations helpful for abstract visual reasoning? May 2019.

[64] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[65] Wes McKinney. Data Structures for Statistical Computing in Python. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

[66] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] 3 main contributions summarised in the abstract and introduction and discussed in details in the result subsections.
   (b) Did you describe the limitations of your work? [Yes] Constrained to visual referential games with specific approaches.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In the broader impact section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Both in the main paper, and as an open-sourced code.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In the experimental setup section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All plots contain some measure of uncertainty.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] It will be added once the review process will have highlighted what is necessary in the paper from what is superfluous, and maybe what other experiments should be ran...

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix A : Spearman ρ Correlation Matrices

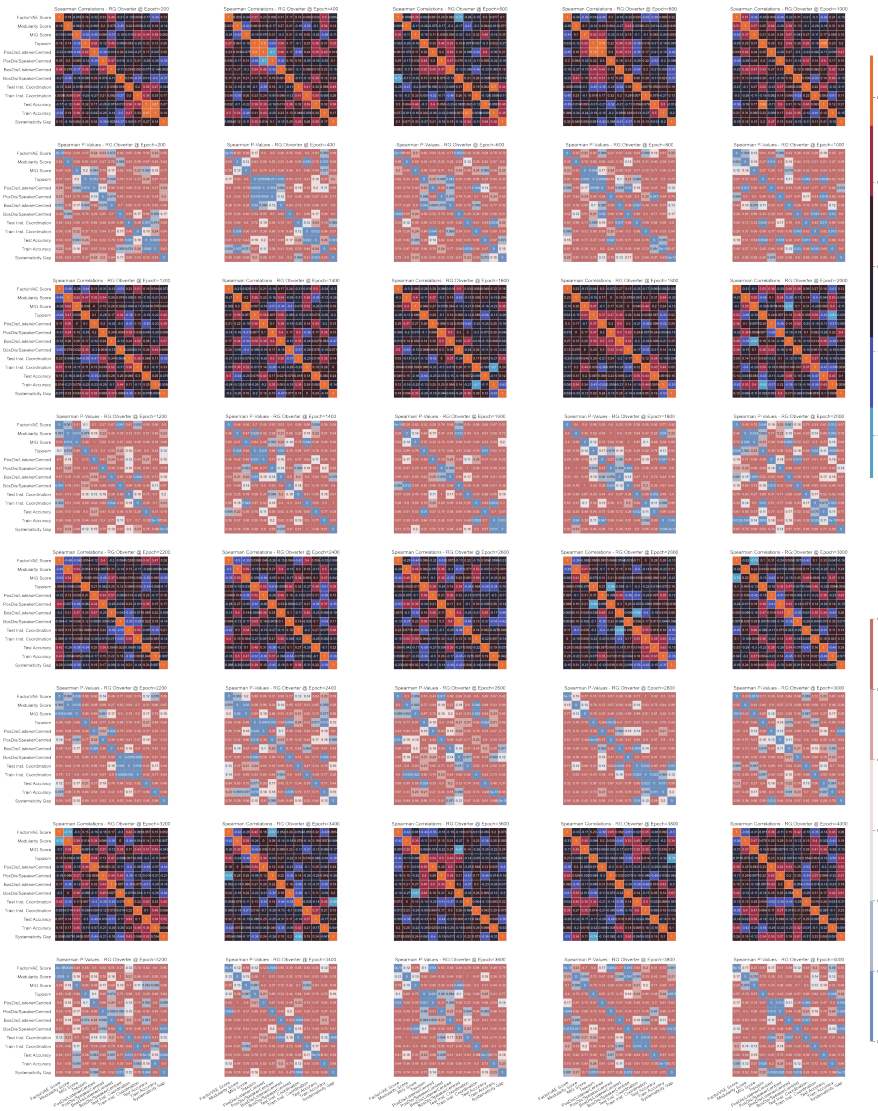


Figure 4: Spearman ρ correlation coefficients (odd rows) and p-values (even rows) matrices between the different metrics, from epoch 200 to 4000, with an interval of 200 epochs between measures, in the context of **RG:Obverter**.

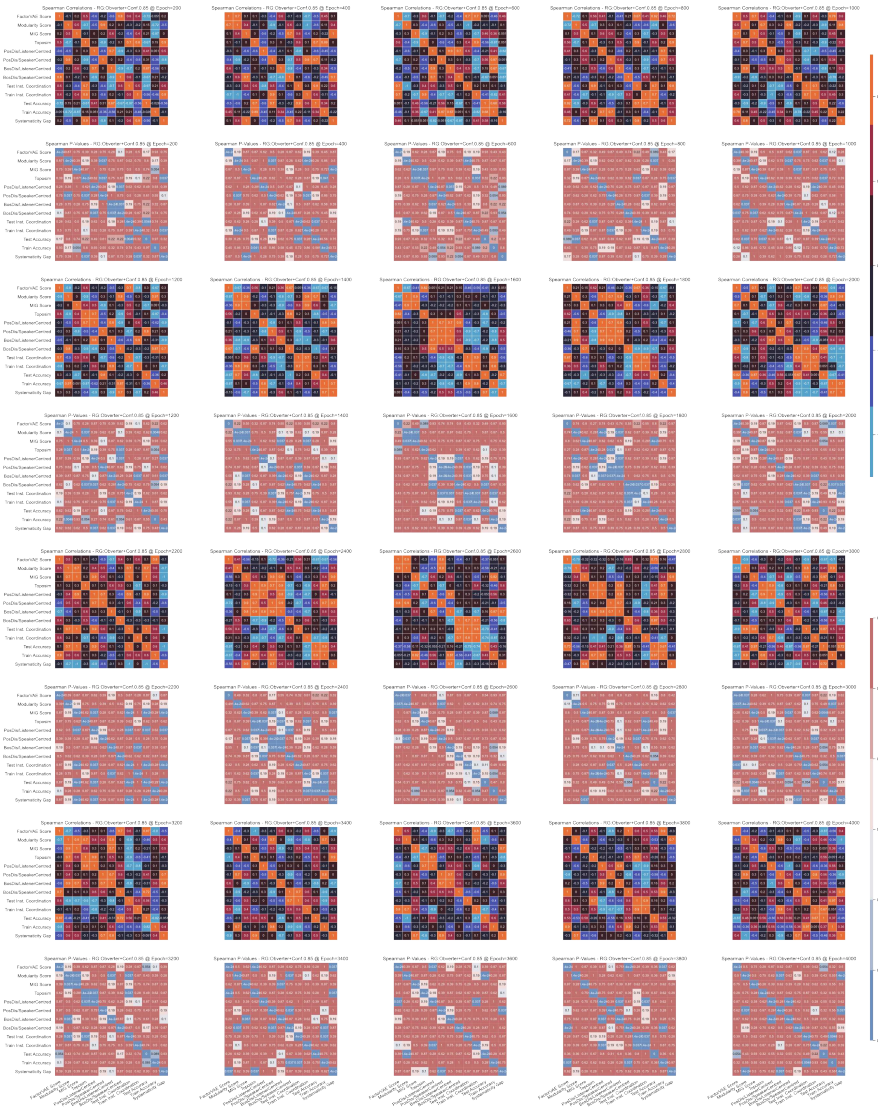


Figure 5: Spearman ρ correlation coefficients (odd rows) and p-values (even rows) matrices between the different metrics, from epoch 200 to 4000, with an interval of 200 epochs between measures, in the context of **RG:Obverter+Conf.0.85**.

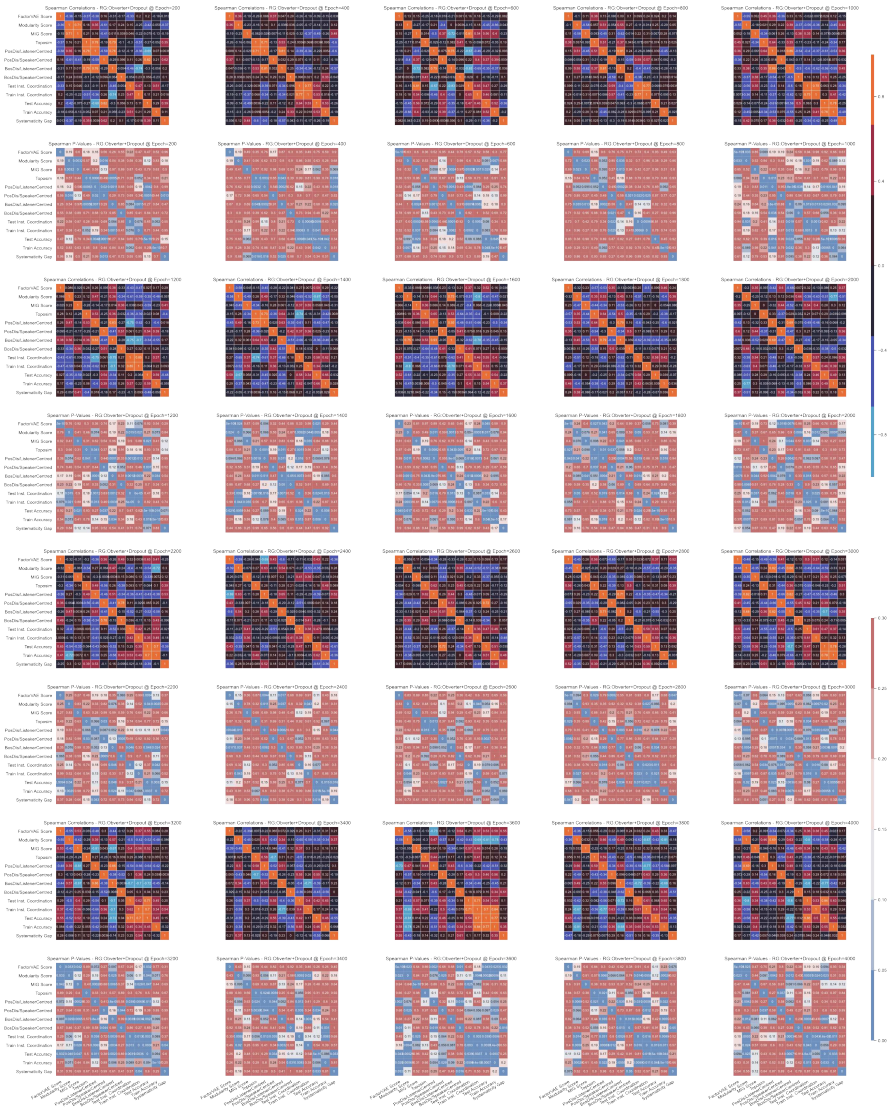


Figure 6: Spearman ρ correlation coefficients (odd rows) and p-values (even rows) matrices between the different metrics, from epoch 200 to 4000, with an interval of 200 epochs between measures, in the context of **RG:Obverter+Dropout**.