TEXREL: A GREEN FAMILY OF DATASETS FOR EMERGENT COMMUNICATION WITH RELATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a new dataset TEXREL as a playground for the study of emergent communications, in particular for relations. TEXREL provides rapid training and experimentation, whilst being sufficiently large to avoid overfitting. We use TEXREL to compare Sender models, compare with a related dataset, examine the effect of changing meaning space size, and perform a case-study for using TEXREL in place of symbolic input.

1 INTRODUCTION

Emergent communications is the study of the linguistic behavior of agents with no pre-training on natural human languages, when placed in a situation where inter-agent communications is needed in order to maximize performance. We can investigate characteristics of the resulting language, such as compositionality; and the extent to which the agents learn to communicate at all. Agents can learn to communicate pictures to each other, e.g. Lazaridou et al. (2018), or to negotiate with each other, e.g. Cao et al. (2018). In general, the resulting emergent language has limited compositionality. For example, Lazaridou et al. (2018) presented results showing that even when the agents have learned to solve a task with 98% accuracy, the topographic similarity - a measure of compositionality - might be only around 0.16-0.26. The resulting languages do not tend to clearly show certain key characteristics of human languages, such as the formation of atomic re-usable units of tokens, i.e. words.

We hypothesize that in order to increase the compositionality of the emergent languages, we need to increase the dimensionality of the underlying meaning space, such that the only feasible way for models to be able to store the language is to store it in factorized form. For example, if a language has 10 words for colors, and 10 words for shapes, then a model need memorize only these 20 words in order to describe all possible combinations of colors and shapes. However, if a model uses a unique non-compositional word for each combination of colors and shapes, then the model will need to memorize $10^2 = 100$ such words, which is a heavier burden.

Thus, a key step to increasing the compositionality of emergent languages is to train agents in an environment of sufficient complexity, that is with many meaning space dimensions. Our work presents a dataset, TEXREL, which provides an experimental playground for learning emergent languages in relatively high dimensional meaning spaces. In this work, we experiment with meaning spaces with up to 6 dimensions.

We find counter-intuitively that increasing the dimensionality of the meaning space does not increase traditional metrics of compositionality, such as topographic similarity (ρ, (Brighton and Kirby, 2006))

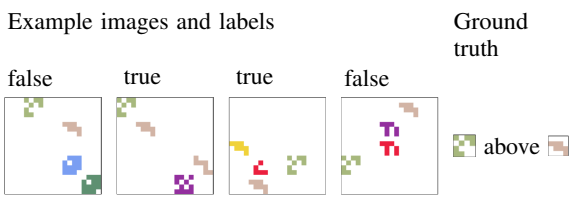

Figure 1: Example images, labels, and ground truth relation for TEXREL dataset Relations task (best viewed in color)

nor of more recent compositional metrics such as TRE (Andreas, 2019). Whilst this could show that increasing the meaning space does not increase underlying latent ground truth compositionality, we argue that our counter-intuitive result might instead be because existing compositionality metrics do not correlate perfectly with underlying latent ground-truth compositionality. Thus our results show that there could be an opportunity to develop new compositionality metrics, or to refine existing ones.

One way to create a high-dimensional dataset is to use symbolic inputs, e.g. Study 1 in Lazaridou et al. (2018). However, symbolic input is essentially a language in compositional form, where each token is a single word describing one attribute. Thus it is unclear whether any emergent compositional language is a reflection of a tendancy of the agents to learn compositional representations, or to simply reflect the compositional representation of the input.

An alternative approach, which we use in this work, is to represent meanings using images. Each image contains one or more objects, each having shape and color. We can use the relative positioning of two objects to express relations between objects, adding additional meaning dimensions. An existing dataset, Shapeworld, Andreas et al. (2018), provides such a dataset. However, as we shall see the training set is small, and models capable of learning on the training set quickly overfit, as alluded to in Andreas et al. (2018). TEXREL provides a much larger training set, 100k training examples, each with 256 images, compared to 9k training images, each with 6 images.

We seek an experimental playground for emergent communication which not only provides high dimensional meaning spaces, using non-symbolic input, but which should ideally be relatively fast to train on. We seek thus to provide images of relatively low resolution, which are friendly to convolutional networks. We note from e.g. Khurshudov (2015) that convolutional networks might pay more attention to the textures of objects than to their outline shape. We thus generate textures, rather than solid-textured shapes. Shapeworld by comparison provides objects with identical solid texture, differing in outline shape. We argue that our approach of using textures allows the use of lower dimensional images, which are easier for a convolutional network to learn. Thus we argue that our family of datasets is 'green', that is easy to use in relatively low computational resource environments.

Our contributions are as follows:

- propose a new dataset, TEXREL, which provides a playground for emergent communications
  - uses non-symbolic inputs, i.e. images
  - fast to train on
  - is much larger than comparable existing emergent communication datasets
  - provides a high dimensional underlying meaning space
- we provide extensive baselines and empirical studies using TEXREL:
  - compare potential sender agent architectures
  - compare TEXREL with Shapeworld
  - examine the effect of meaning space size and dimensionality on compositionality
  - provide a case-study of using TEXREL in place of symbolic inputs, for fast experimentation
- propose new metrics:
  - clustering precision and clustering recall as measures of language expressivity and consistency, respectively
  - PTRE: derivative of TRE which assigns low compositionality to languages having low expressivity

2 BACKGROUND

2.1 REFERENTIAL TASK

We target a referential task, e.g. Lazaridou et al. (2018), in the context of emergent communications. See Figure 2: a sender agent receives labeled sender images, $(X_{trn}, Y_{trn}) = \{(x_{trn,1}, y_{trn,1}), \ldots, (x_{trn,n}, y_{trn,n})\}$ for each example, generated from some underlying hypothesis h. The sender emits a linguistic utterance, $\hat{u}$. A receiver agent receives $\hat{u}$, along

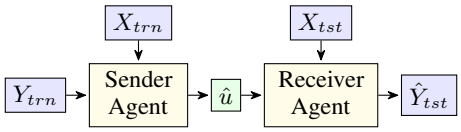

Figure 2: Emergent communications referential task

with unlabeled receiver images, $X_{tst} = \{x_{tst,1}, \ldots, x_{tst,n}\}$, and predicts the correct labels, $Y_{tst} = \{y_{tst,1}, \ldots, y_{tst,n}\}$. We hope that the Sender agent will learn to represent the underlying concept h in the generated linguistic utterance $\hat{u}$. Often, X_{trn} consists of a single example, and each $y_{tst,i}$ is always implicitly true. In our own work, we wish to represent relations, so we extend to the more general case of multiple labeled sender examples.

Each hypothesis can be represented as a list of attributes, potentially tree-structured, for example. (red), (square), (red, square), or left-of((red, square), (blue, circle)). In the general case, we denote the number of attributes in a hypothesis as n_{att}, and the number of possible values for each attribute as n_{val}. For example, if each hypothesis consists of three specific shapes that must exist in an image, then $n_{att} = 3$. If there are 8 possible shapes, then $n_{val} = 8$. We use 'meaning space', $\mathcal{M}$, to denote the space of all possible hypotheses. If n_{att} is 3, and n_{val} is 8, then $|\mathcal{M}| = 8^3$.

2.2 METRICS

2.2.1 METRICS OF COMPOSITIONALITY

Brighton and Kirby (2006) proposed topological similarity for measuring the compositionality of utterances, by comparison to a ground-truth description of the underlying concept. We denote topological similarity as ρ, in line with the notation in Lazaridou et al. (2018). Given pairs of hypotheses h_i, h_j and associated utterances, $\hat{u}_i, \hat{u}_j$ ρ is the Spearman's rank correlation between the distances between pairs of hypotheses $D_h(h_i, h_j)$, and the distances between the associated pairs of utterances $D_u(\hat{u}_i, \hat{u}_j)$. For example, we would like the representation $\hat{u}$ for the hypothesis (big,red, circle) to be more similar to (small, red, circle) than to (small, pink, square).

Holdout accuracy on unseen objects, acc_{new}, can be seen as a measure of compositionality. In training the agent could have seen red objects, square objects, but not red, square objects. If the agents are using compositional communication, we expect them to generalize well to these novel combinations. Note that whilst ρ depends only on the sender, acc_{new} depends also on the receiver.

A recent metric of compositionality is TRE, which measures the extent to which an evaluation model taking as input a ground truth concept can generate the emitted linguistic utterances, under certain compositional constraints. We experiment with TRE in a later experiment in this work.

2.3 CLUSTERING METRICS

In the domain of clustering (Driver and Kroeber, 1932), given gold standard clusters and predicted clusters, the quality of the clustering can be evaluated by calculating 'precision' and 'recall', e.g. Haponchyk et al. (2018). For each predicted cluster, we assign the most frequent gold standard cluster to it, and calculate precision as $prec = \frac{1}{N} \sum_{j=1}^{\hat{k}} \max_i |c_i \cap \hat{c}_j|$. Cluster recall is correspondingly $rec = \frac{1}{N} \sum_{j=1}^{k} \max_i |\hat{c}_i \cap c_j|$.

2.3.1 METRICS OF EXPRESSIVITY AND CONSISTENCY

The expressivity of a language is the extent to which a language can be used to represent different meanings. Lazaridou et al. (2018) used the lexicon size as a measure of the ambiguity of emerged languages. Higher lexicon size corresponds to higher expressivity. An issue with lexicon size is that it is not normalized, and is difficult to compare across different meaning spaces.

Consistency of a language is the extent to which the same utterance is used to represent the same underlying meaning. Dagan et al. (2020) used Jaccard Similarity, which is the ratio of the intersection of two sets divided by their union.

Table 1: Description of each task type in TEXREL, where $\mathcal{C}$ is space of available colors, $\mathcal{S}$ is space of available textures, and $\mathcal{P}$ is space of available prepositions.

Task type	Description	n_{att}
Col$n$	Includes n objects of colors $\{\mathcal{C}_{c_1}, \ldots, \mathcal{C}_{c_n}\}$, where $\{c_1, \ldots, c_n\}$ are sampled for each example	n
Tex$n$	Includes n objects of textures $\{\mathcal{S}_{s_1}, \ldots, \mathcal{S}_{s_n}\}$, where $\{s_1, \ldots, s_n\}$ are sampled for each example	n
TexCol$n$	Includes n objects having texture and color $\{(\mathcal{S}_{S_1}, \mathcal{C}_{c_1}), \ldots, (\mathcal{S}_{s_n}, \mathcal{C}_{c_n})\}$, where $\{(s_1, c_1), \ldots, (s_n, c_n)\}$ are sampled for each example	$2n$
Rel	Includes an object of texture and color $(\mathcal{S}_{s_1}, \mathcal{C}_{c_1})$ positioned $\mathcal{P}_{p_1}$ relative to an object of texture and color $(\mathcal{S}_{s_2}, \mathcal{C}_{c_2})$, where $(c_1, s_1, p_1, c_2, s_2)$ are sampled for each example	5

2.4 EMERGENT COMMUNICATION OF RELATIONS

Andreas et al. (2018) presented a relations task, Shapeworld. A model receives four positive example images of a hypothesis h, and must then predict whether a test image is an example of the hypothesis. Andreas showed that by partitioning the model into a Sender and Receiver model, with a linguistic bottleneck, the performance improved. Different from our work, the agents were not trained end-to-end, but supervised, using English language annotations for the utterances u. Andreas used a prototypical sender to embed the sender utterances, followed by an LSTM (Hochreiter and Schmidhuber, 1997) decoder to generate u. The receiver used a second LSTM to embed u, and formed the dot product of this embedding with an embedding of the test image.

Mu et al. (2020) showed that adding a linguistic bottleneck between two agents was un-necessary: it was sufficient to use decoding the English language description of the hypothesis as an auxiliary loss.

In our own work, we seek to emerge a language between two agents, with no English language supervision. This is a challenging exploration problem for the two agents.

3 OUR WORK

3.1 TEXREL DATASET FAMILY

We seek to construct a family of datasets to study the emergence of relations, that is sufficiently lightweight to allow fast experimentation, whilst being large enough that agents do not overfit.

Table 1 shows the task types provided with TEXREL. These task types allow experimentation with varying n_{att} of the underlying meaning space. For example, Col3 has $n_{att} = 3$, TexCol3 has $n_{att} = 6$, and Rel has a $n_{att} = 5$ (n_val is not constant across attributes for Rel). We carry out experiments in later sections on the effect of n_{att} on measured compositionality. $\mathcal{C}$ is the space of available colors, where the colors are $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_{N_{colors}}\}$. Similarly $\mathcal{S}$ is the space of textures, and $\mathcal{P}$ is the space of prepositions. Because of symmetry, we include only two prepositions 'above' and 'right-of'.

We wanted to avoid the agents taking short-cuts as much as possible. If negative examples are sampled uniformly from the space of all possible examples, then only communicating a single attribute might be sufficient to achieve reasonable accuracy. To mitigate this possibility, we keep all negative examples 'tight' to the manifold of positive examples by constructing negative examples from positive examples, in which we change just a single attribute.

Each image is associated with a ground truth label, True or False. In addition, for each example, we provide an English language description of the underlying ground truth concept, and a tree-structured representation, that could be used for example with TRE. Table 2 shows example annotations.

We added distractor objects to each image, to increase the size of the state space, and thus aim to discourage the sender from simply sending the entire state.

For each task, we carve a holdout set of objects which are not presented at training time. Task-specific details on holdout set construction can be found in the Appendix. Following the approach in Andreas

Table 2: Examples of English language annotations made available with each TEXREL example.

Task	English language annotation	Tree-structured annotation
Col2	has-colors color1 color5	('has-colors', ('color1', 'color5'))
Tex2	has-shapes shape2 shape4	('has-shapes', ('shape2', 'shape4'))
TexCol2	has-shapecolors color4 shape1 color5 shape7	('has-shapecolors', (('color4', 'shape1'), ('color5', 'shape7')))
Rel	color0 shape6 above color1 shape6	('above', (('color0', 'shape6'), ('color1', 'shape6')))

et al. (2018), we name the eval datasets using the training objects 'val_same' and 'test_same', and the eval datasets using the holdout objects 'val_new' and 'test_new'. Given the scarcity of objects in the holdout set for some task types, we draw distractors for 'val_new' and 'test_new' from the union of the training and holdout objects.

For each task, we create a training dataset of 100,000 examples. Each example comprises 128 labeled sender images, and 128 labeled receiver images. For each set of 128 labeled images, 64 are positive examples of the underlying concept. Each evaluation set has 1024 examples.

3.2 METRICS

3.2.1 CLUSTER PRECISION AND RECALL FOR MEASURING EXPRESSIVITY AND CONSISTENCY

We propose to use cluster precision to measure language expressivity, and cluster recall to measure language consistency. For each example, the cluster gold label is taken to be the underlying hypothesis h, and the cluster predicted label is the generated utterance $\hat{u}$. Then, if many hypotheses map to just a few utterances, cluster precision will be low, corresponding to low expressivity. Similarly, if a single hypothesis maps to multiple utterances, then cluster recall will be low, corresponding to low consistency. Compared to lexicon size, cluster precision is normalized, and can be compared across different meaning spaces.

3.2.2 PTRE

TRE (Andreas, 2019) measures compositionality by minimizing the discrepancy between a composition function of the hypotheses, and the generated utterances. When this discrepancy is near zero, the generated utterances are assumed to be compositional. However, languages with low expressivity, having many hypotheses mapping to the same utterance, can be easily fitted by any model, even in the absence of any underlying compositionality. Thus, TRE will report high compositionality for non-compositional, low expressivity languages.

We propose to correct for this issue by dividing TRE by cluster precision, and denote the resulting metric by PTRE.

4 RELATED WORK

Our work relates primarily to Shapeworld, Andreas et al. (2018), which is a dataset of images depicting relations between colored shapes. By comparison with Shapeworld, our work provides a significant larger dataset, and uses textures rather than solid filled shapes, which believe allows convolutional networks to learn faster on TEXREL, and thus allows faster experimentation.

Other datasets containing relations include CLEVR (Johnson et al., 2017) and CUB-200 (Welinder et al., 2010). CLEVR is a dataset of high-resolution 3-dimensional images, created using the Blender application (Blender Foundation, 2002); along with english language annotations. By comparison with our work, CLEVR images are more aesthetically pleasing, however might require significantly longer training time. Note that CLEVR does incorporate some notion of texture, since objects can be either shiny metal, or matte rubber. CLEVR comprises 100,000 images, along with around 1 million associated questions. TEXREL contains 100k examples, each having 256 images, for a total of 2.5 million images: significantly larger. CLEVR targets the visual question-answering ('VQA') setting, and does not clearly map to usage in emergent communications. We experimented with using

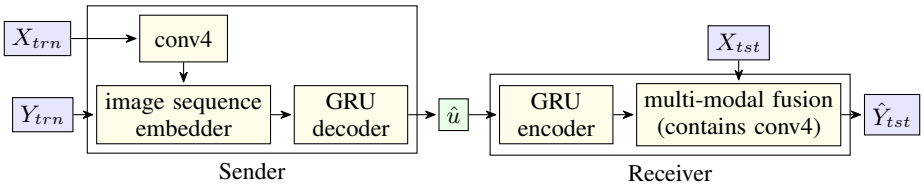

Figure 3: Detailed architecture for sender and receiver models

the CLEVR code to create a dataset for emergent communication, but found that the ray-tracing generation process is relatively slow; and the resulting images are large, and slow to train on.

CUB-200 (Welinder et al., 2010) is a dataset of photos of birds, comprising 40-60 images for each of 200 species of birds, along with English language annotations. CUB200 along with the annotations from Reed et al. (2016) was used for example in Mu et al. (2020), where it was used as a meta-learning task. The ground truth annotations are only available in free-form text, which limits the use of compositional metrics, such as ρ, in an emergent communication setting. The dataset is relatively small, and agents in an emergent communication setting might overfit to training examples. Lastly, the images are photos, so convolutional networks might need considerable training time in order to form effective representations.

Our work relates to the SHAPES dataset (Andreas et al., 2016). SHAPES dataset is similar to the Shapeworld dataset, in that it provides images of solid filled colored shapes. Like CLEVR, SHAPES targets the VQA setting. The SHAPES dataset comprises 64 training images, along with 244 unique questions. By comparison with TEXREL, the entire SHAPES training dataset contains four times fewer images than a single training example from TEXREL. Dagan et al. (2020) used a modified version of the SHAPES dataset, with 80k training examples. However, each example comprised a single colored shape, without distractor objects, so the task was relatively simpler than our own tasks.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

All results are the mean over 5 runs, unless otherwise indicated. Each run used an NVIDIA 2080Ti GPU. Where early stopping was used, we set the patience to 10, and evaluated acc_{val_same} set every 300 steps. Batch size was 32, unless otherwise indicated. For the emergent communication linguistic representations, we use a vocab size of 21, and an utterance length of 10. We use L1-distance when calculating ρ.

Code and data will be made available at [1].

5.1.1 ARCHITECTURE

Figure 3 shows the architecture of the sender and receiver agents. The sender agent uses a convolutional neural network, 'conv4' (Snell et al., 2017), to embed each of the images in X_{trn}. The 'image sequence embedder' takes the sequence of embedded images $\{e_1, \ldots, e_{m_{trn}}\}$, along with labels $\{y_{trn,1}, \ldots, y_{trn,m_{trn}}\}$ and embeds them into a single embedding e_s. A GRU decoder (Cho et al., 2014) decodes e_s into a linguistic utterance, $\hat{u}$. The receiver agent uses a second GRU to embed $\hat{u}$ into e_r, then uses multi-modal fusion to output predictions $\hat{Y}_{tst} = \{\hat{y}_{tst,1}, \ldots, \hat{y}_{tst,m_{tst}}\}$ for test images $X_{tst} = \{x_{tst,1}, \ldots, x_{tst,m_{tst}}\}$, based on e_r. We use a prototypical sender as the image seq embedder, $e_s = \sum_{i=1}^{m_{trn}} 1[y_{trn,i}] \cdot e_i$ (Andreas et al., 2018), where $1(\cdot)$ is an indicator function. Following Andreas et al. (2018), we use cosine similarity as the multi-modal fusion, $\hat{y}_{tst,i} = \sigma(e_r^T \cdot \text{conv4}(x_{tst,i}))$, where σ is the sigmoid function.

Table 3: Image sequence embedder architectures

Image sequence embedder	Description
RNNOverCNN	Encode the images using a convolutional neural network (CNN, LeCun et al. (1989), then pass through a residual neural network (RNN, Hopfield (1982))
ConvLSTM	LSTM built from convolutions instead of projections (Shi et al., 2015)
Stack	Stack the planes of all images together (each image comprises red/green/blue planes), then pass through a CNN
MaxCNN	Pass each image through a CNN, max pool
AvgCNN	Pass each image through a CNN, average pool
Prot	'Prototypical': Pass positive images through a CNN, take mean (Andreas et al., 2018)

Table 4: Comparison of effect of sender architectures on acc_{test_new} using Relations task of TEXREL dataset, with 2 distractors. All runs are for 5 minutes. Each result is the mean over 5 runs.

Task	RNNOverCNN	ConvLSTM	Stack	MaxCNN	AvgCNN	Prot
Rels	0.22	0.25	**0.84**	0.50	0.58	0.78

5.2 SEARCH FOR EFFECTIVE SENDER AND RECEIVER ARCHITECTURES

We used TEXREL to search for effective Sender and Receiver agent architectures for emergent communication. Due to space constraints, we present results only for the Sender agent. Results for the Receiver agent can be found in the Appendix. For the Sender agent, the image sequence embedders we tested are described in Table 3. When using the Prot sender model, only positively labeled images were used. For other Sender models, the labels were introduced by adding an additional feature plane to each image, either all 1s, or all 0s, according to y.

5.2.1 SUPERVISED TRAINING OF SENDER MODELS

We trained each model supervised, given ground truth English utterances $\{u_1, \ldots, u_{n_{trn}}\}$, along with X_{trn} and Y_{trn}. We trained each model for 5 minutes elapsed, because we were optimizing for models which train quickly. We used the Relations task from TEXREL, with 2 distractors.

Table 4 shows the results. Due to space constraints, we present only acc_{test_new}, and only for the Relations task. Interestingly, the simplest models, StackedInputs and PrototypicalSender, did better than more complex models, such as RNNOverCNN and ConvLSTM. This is perhaps because we were optimizing for training time. Similarly, Prot did better than AvgCNN, even though Prot discards images for which $y_{trn} = 0$.

5.2.2 END TO END TRAINING

We retained some of the best models from the supervised Sender training, and paired them with Receiver models, then trained end-to-end, using early stopping on acc_{val_same}. Due to space constraints, we present results only for the Cosine receiver model here. Additional results can be found in the Appendix. We sampled utterances $\hat{u}$ from the Gumbel distribution (Jang et al., 2017), parameterized by the output from the Sender model. We used the Relations task.

Table 5: Comparison on end-to-end architectures. Results are evaluated using test_new. Time is in minutes. Each result is mean over 5 runs.

Sender	Receiver	Steps	Time	acc	ρ	prec	rec
Stacked	Cosine	16k	36	**0.65**	**0.16**	**0.05**	0.69
Prot	Cosine	17k	40	0.63	0.15	**0.05**	**0.81**

[1] url provided at publication; zip file of code submitted as addendum

Table 5 shows the results. Stack performed slightly better than Prot, when paired with CosineReceiver. However, Prot performed almost on par. The results align with the choice of Andreas et al. (2018) to use a Prot sender and a Cosine Receiver.

5.3 COMPARISON WITH SHAPEWORLD DATASET

Table 6: Comparison between TEXREL and ShapeWorld datasets for an emergent communication scenario. 'LSL' code means the modified L^3 implementation provided with the LSL paper. 'Shapeworld+aug' dataset denotes Shapeworld dataset with data augmentation. Each result is mean over 5 runs.

		train	test_same				test$_{new}$			
Code	Dataset	acc	acc	ρ	prec	rec	acc	ρ	prec	rec
LSL	Shapeworld	0.62	0.50				0.50			
LSL	Shapeworld+aug	0.47	0.50				0.50			
ours	Shapeworld	0.57	0.51	0.01	0.01	0.7	0.52	0.01	0.01	0.6
ours	TEXREL	0.68	0.67	0.08	0.03	**0.99**	0.63	0.15	0.05	**0.81**

We compare TEXREL to the Shapeworld relations dataset. We evaluate Shapeworld using two codebases: a modified version of the implementation of 'Learning with Latent Language' ('L^3', Andreas et al. (2018)) provided with 'Learning with Shaped Language' ('LSL', Mu et al. (2020)); and our own codebase. In LSL and L^3, the linguistic utterance is used as a supervisory signal during the training, based on English language annotations. In emergent communication, no such supervision is provided: the agents emerge their own language. We thus modified the LSL code to enable end-to-end learning, without supervised pre-training. The LSL code-base also adds data augmentation, which negatively samples data from other concepts. We disabled this by default. The LSL codebase uses soft sampling at training time for the intermediate latent utterances. In our own codebase we used Gumbel sampling. Other training hyper-parameters remained the same. When using TEXREL dataset, we chose the Relations task, with 2 distractors.

Figure 6 shows the results. When training end-to-end, in an emergent communication scenario, the agents failed to learn to co-ordinate when using Shapeworld dataset: both acc$_{test_same}$ and acc$_{test_new}$ were at chance, 0.50. However, using TEXREL, whilst the task remained challenging, with acc$_{test_new}$ $= 0.63$, the agents did learn to co-ordinate, and ρ reached 0.15, showing the beginnings of compositional communication.

5.4 EFFECT OF SIZE OF MEANING SPACE ON METRICS OF COMPOSITIONALITY

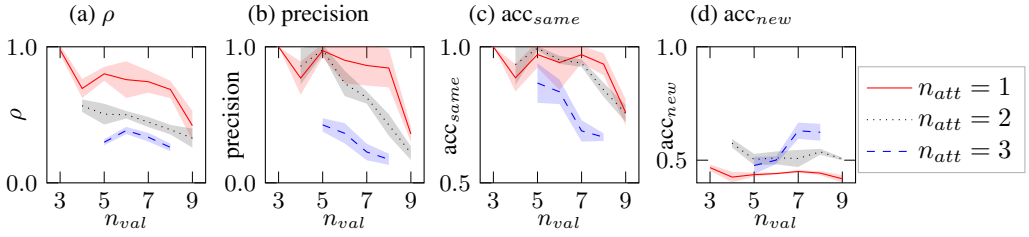

Figure 4: Experiments on varying size of meaning space $\mathcal{M}$, using TEXREL Texture datasets. Shaded areas are 95% confidence intervals, over 5 seeds.

We hypothesized that increasing the number of attributes, n_{att}, would increase compositionality, and thus increase ρ. To test this, we used the TEXREL Texture datasets, and varied both the number of textures in the hypothesis, n_{att}, and the number of available textures, n_{val}. Each training run was for 5k training steps.

Figure 4 shows the results. Surprisingly, we found that ρ actually decreased with n_{att} and n_{val}. We hypothesize that this is an artifact of the way that ρ is calculated. Take the case of a single entity. No compositionality is possible, since there is only a single concept to represent. Indeed acc$_{test_new}$ is consistently at chance, aligned with compositional generalization being impossible for

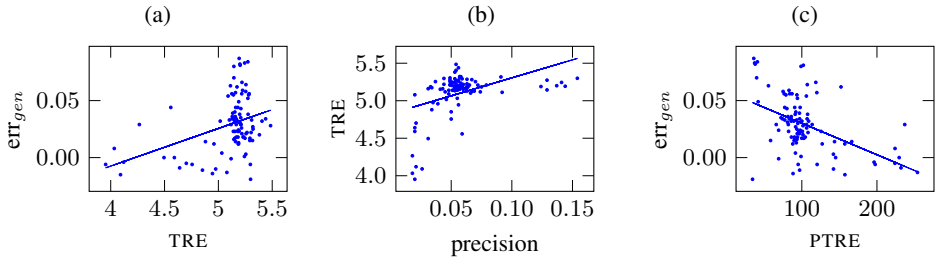

Figure 5: Experiments on TRE based on Andreas (2019) section 7, using TEXREL. err_{gen} means generalization error. Each run is for 5k training steps. Each point represents the result of a single run.

$n_{att} = 1$. ρ is the Spearman's rank correlation between the distances between pairs of utterances, and the distances between corresponding ground-truth concept labels. For a single entity, the distance between ground-truth concept labels $\in \{0, 1\}$. However, the distances between pairs of utterances will vary. Thus, as we increase n_{val}, the Spearman's rank correlation will fall, and ρ will decrease. That ρ changes for a non-compositional scenario suggests there could be an opportunity to introduce a new compositional metric, e.g. by an appropriate normalization of ρ.

Looking at acc_{val_new}, we can see that increasing n_{val} did in fact improve acc_{val_new}, in line with our hypothesis that increasing n_{val} will increase compositionality.

5.5 REPRODUCTION OF TRE SECTION 7 EXPERIMENTS USING NON-SYMBOLIC INPUT

As an example of using TEXREL to run experiments using symbolic data to use non-symbolic data, we reproduce experiments from section 7 of Andreas (2019). We target the experiments that investigate the relationship between compositionality and generalization. Figure 5 (a) shows the relationship between TRE and generalization error, for 100 runs using a Relations dataset from TEXREL. Note that low TRE is a measure of high compositionality. We can see that generalization error decreases with lower TRE, that is with increased compositionality, which goes against conventional wisdom, e.g. Kottur et al. (2017) or Lazaridou et al. (2018), but aligns with Andreas's results. These experiments ran in 17 hours, on a single 2080Ti GPU. Therefore the compact size of the images in TEXREL allows for rapid, low-cost experimentation on non-symbolic image inputs.

We were interested by the negative correlation between generalization error and compositionality. From Figure 5 (b), we note that TRE decreases with decreasing cluster precision, i.e. with less expressive languages. We hypothesize that expressivity is a confounding factor. We added a correction to TRE, by dividing by cluster precision, to form PTRE. Figure 5 (c) shows the relationship between PTRE and generation error. Generalization error in fact increases with decreasing PTRE. Thus, the finding that generalization increases with compositionality could be an artifact of a reduction in expressivity. We leave further investigation of the relationship between TRE, PTRE, and underlying ground-truth compositionality to future work.

6 CONCLUSION

We have presented TEXREL a dataset targeted at fast, green experimentation for emergent communication. TEXREL provides a number of challenging tasks, such as relations learning. We showed that TEXREL compares favorably with existing relations datasets, and experimented with using TEXREL on a number of tasks related to emergent communication. We hope that TEXREL can provide a helpful experimental playground that is an alternative to using symbolic data, and that allows for fast, green experimentation on emergent communications.

7 REPRODUCIBILITY

Full code is provided in the addendum, along with instructions in the README.md. Full code will be published to github following acceptance. The datasets can be created from scratch, using the

provided code. In addition, a link to download the data from Google Drive has been provided in the README.md.

Each experiment scenario was run multiple times, except where noted (eg experiments on TRE section 7); and the mean reported. 95% confidence interval (CI95) was provided where space allowed, e.g by adding shading to graphs.

8 ETHICS

This work does not involve human subjects. It contains no obviously harmful insights, methodologies or applications. There are no obvious conflicts of interest or sponsorship to note. There are no obvious discrimination/bias/fairness concerns to report. There are no obvious issues with privacy, security, or legal compliance. All data provided was artificially generated, and does not present privacy or other issues. We have done our due diligence to ensure the integrity and reproducibility of our research.

REFERENCES

Jacob Andreas. 2019. Measuring compositionality in representation learning. In *International Conference on Learning Representations*.

Jacob Andreas, Dan Klein, and Sergey Levine. 2018. Learning with latent language. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2166–2179, New Orleans, Louisiana. Association for Computational Linguistics.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Blender Foundation. 2002. Blender. `https://www.blender.org/`. Accessed: 2021-05-22.

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. *arXiv preprint arXiv:1804.03980*.

Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. 2018. Gated-attention architectures for task-oriented language grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. 2020. Co-evolution of language and agents in referential games. *arXiv preprint arXiv:2001.03361*.

Harold Edson Driver and Alfred Louis Kroeber. 1932. *Quantitative expression of cultural relationships*, volume 31. Berkeley: University of California Press.

Iryna Haponchyk, Antonio Uva, Seunghak Yu, Olga Uryupina, and Alessandro Moschitti. 2018. Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2310–2321.

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

John J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Artem Khurshudov. 2015. Suddenly a leopard print sofa appears. `https://archive.is/PhExN`. Accessed: 2021-05-22.

Satwik Kottur, José Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967, Copenhagen, Denmark. Association for Computational Linguistics.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *International Conference on Learning Representations*.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Dipendra Misra, John Langford, and Yoav Artzi. 2017. Mapping instructions and visual observations to actions with reinforcement learning. *arXiv preprint arXiv:1704.08795*.

Jesse Mu, Percy Liang, and Noah Goodman. 2020. Shaping visual representations with language for few-shot classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4823–4830, Online. Association for Computational Linguistics.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting.

Jake Snell, Kevin Swersky, and Richard S Zemel. 2017. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-ucsd birds 200.

Table 7: Holdout approach by task type.

Task type	Holdout approach
Col$n$	Set aside several colors
Tex$n$	Set aside several textures
TexCol$n$	Set aside several pairs of colors and textures
Rel	Set aside several pairs of colors and textures

APPENDICES

.1 HOLDOUT SETS

Table 7 describes the task-specific holdout set construction approaches.

.2 RECEIVER MODEL SEARCH

Table 8: Multi-modal fusion architectures

Multi-modal fusion	Description
Concat	Concatenate encoded utterance and encoded image, then project, e.g. Misra et al. (2017)
Cosine	Dot product of encoded utterance and encoded image, eg Lazaridou et al. (2018)
GatedAtt	Image is encoded using a convolutional neural network (CNN, e.g. LeCun et al. (1989). Encoded utterance is used as attention (Bahdanau et al., 2015) over the output planes of the CNN output (Chaplot et al., 2018)
AllPlaneAtt	Encoded utterance is used to give attention over feature planes of all layers of the CNN, not just the output of the final layer
Configurable Convolution Kernel ('CCK')	Encode utterance, use as weights in convolutional network (Chen et al., 2015)

Table 8 depicts the Receiver models we searched over. Note that in the general case, the convolutional network cannot be factorized out of the multi-modal fusion. For example, for CCK, the encoded utterance is used as the weights for the convolutional network, and this comprises the entire entanglement between the encoded utterance and the receiver images.

Table 9 shows the results for supervised training of the Receiver models in isolation, using English language utterances u. The task is Relations.

Table 10 shows the results for end to end training of pairs of Sender and Receiver agents. The task is Relations. We sample each predicted utterance $\hat{u}$ from the Gumbel distribution, parameterized by the output of the Sender model.

Table 9: Comparison of effect of receiver architectures on acc_{test_test} and acc_{test_new} using TEXREL dataset. 'col$n$' is Colors task, where n is the number of entities; 'shp$n$' is Shapes task, 'sc$n$' is shapes-colors task, and 'rels' is Relations task. In all cases, two distractor objects are added to each image. All runs are for 5 minutes. Each result is the mean over 5 runs.

(a) Effect of receiver architecture on acc_{test_same}.

Receiver architecture	col1	col2	col3	shp1	shp2	shp3	sc1	sc2	sc3	rels
Concat	0.61	0.50	0.51	0.60	0.53	0.51	0.50	0.50	0.50	0.50
Cosine	**1.00**	**1.00**	**1.00**	0.96	0.99	**1.00**	0.98	0.84	0.75	0.84
GatedAtt	0.99	**1.00**	**1.00**	**1.00**	**1.00**	0.98	0.98	0.84	0.73	0.78
AllPlaneAtt	**1.00**	**1.00**	0.99	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.83**	**1.00**
CCK	0.56	0.52	0.50	0.50	0.51	0.50	0.50	0.50	0.50	0.50

(b) acc_{test_same}

(c) Effect of receiver architecture on acc_{test_new}.

Receiver architecture	col1	col2	col3	shp1	shp2	shp3	sc1	sc2	sc3	rels
Concat	0.38	0.46	0.47	0.36	0.42	0.45	0.49	0.50	0.50	0.50
Cosine	0.38	**0.99**	0.97	0.41	0.83	0.96	0.80	0.74	**0.71**	0.74
GatedAtt	0.40	0.96	0.97	0.37	**0.87**	0.94	0.84	0.74	0.67	0.72
AllPlaneAtt	0.50	0.83	**0.98**	**0.50**	0.78	**0.99**	**0.95**	**0.92**	0.70	**0.98**
CCK	**0.58**	0.49	0.52	**0.50**	0.52	0.50	0.50	0.50	0.50	0.51

(d) acc_{test_new}

Table 10: Comparison on end-to-end architectures. Each result is mean over 5 runs. Utterances are sampled from Gumbel distributions. The underlying task is Relations. Early stopping on acc_{val_same}

Sender	Receiver	Steps	Time (mins)	test_new acc	ρ	prec	rec
StackedInputs	AllPlaneAtt	3k	11	0.53	0.04	0.04	0.63
StackedInputs	Cosine	16k	36	**0.65**	**0.16**	**0.05**	0.69
StackedInputs	FeatPlaneAtt	14k	30	0.60	0.10	0.04	0.67
Prototypical	AllPlaneAtt	11k	30	0.57	0.11	0.04	0.80
Prototypical	Cosine	17k	40	0.63	0.15	**0.05**	**0.81**
Prototypical	FeatPlaneAtt	17k	40	0.60	0.13	**0.05**	0.77