Why and how to study the impact of perception on language emergence in artificial agents

Xenia Ohmer, Michael Marino, Peter König*, Michael Franke*
{xenia.ohmer, michael.marino, peter.koenig, michael.franke}@uni-osnabrueck.de
Institute of Cognitive Science, University of Osnabrueck, Germany

*Authors contributed equally.

Abstract

The study of emergent languages in deep multi-agent simulations has become an important research field. While targeting different objectives, most studies focus on analyzing properties of the emergent language—often in relation to the agents' inputs—ignoring the influence of the agents' perceptual processes. In this work, we use communication games to investigate how differences in perception affect emergent language. Using a conventional setup, we train two deep reinforcement learning agents, a sender and a receiver, on a reference game. However, we systematically manipulate the agents' perception by enforcing similar representations for objects with specific shared features. We find that perceptual biases of both sender and receiver influence which object features the agents' messages are grounded in. When uniformly enforcing the similarity of all features that are relevant for the reference game, agents perform better and the emergent protocol better captures conceptual input properties.

Keywords: language emergence, deep learning, reinforcement learning, groundedness

Introduction

Sparked by the rapid advances in machine learning, there has been growing interest in studying language emergence in artificial agents that communicate to solve a common task. The underlying idea is that language derives its meaning from its use, and accordingly many of its aspects cannot be captured by supervised learning. Different research objectives come together in this new framework. On the one hand, it is used to analyze how artificial agents communicate and to improve this communication, for example in terms of learning speed, performance, or generalization ability (e.g., Das et al., 2019; Kharitonov & Baroni, 2020). On the other hand, it is used to investigate the pressures leading to the emergence of natural language properties, such as compositionality (e.g., Lazaridou, Hermann, Tuyls, & Clark, 2018; Harding Graesser, Cho, & Kiela, 2019; Rodríguez Luna, Ponti, Hupkes, & Bruni, 2020).

Fundamentally, a shift from supervised learning to grounding language in interactions with the environment requires an understanding of how linguistic expressions relate to that environment as experienced through perception (Harnad, 1990). This is definitely important when using deep multi-agent simulations for drawing conclusions about the emergence of natural language properties. In natural language, the formation of linguistic expressions is strongly influenced by perception, not only for concrete concepts like colors (Regier, Kay, & Khetarpal, 2007) but also abstract ones (Lakoff & Johnson, 1980). While neural networks commonly used as visual modules in artificial agents exhibit parallels to how the brain processes information, there are also important differences (Lake, Ullman, Tenenbaum, & Gershman, 2017). E.g., Peterson, Abbott, and Griffiths (2018) show that object representations in humans differ from those in neural networks in terms of similarity judgements, which in turn leads to different semantic relationships. Even if the objective is to improve communication between artificial agents, awareness of how to best process sensory input is crucial. For example, susceptibility to adversarial attacks or inherent biases of neural networks may pose challenges to the formation of efficient protocols. Taking these points into consideration, modern language emergence research should account for the entanglement of perception and the formation of language.

Although the field has experimented with various setups, the role of perceptual processes has been largely ignored. Many designs skip any form of perception by using symbolic input (e.g., Bouchacourt & Baroni, 2019; Kharitonov & Baroni, 2020). Others use pixel input, which is more realistic, and can capture natural differences in object appearance (e.g., Havrylov & Titov, 2017; Rodríguez Luna et al., 2020). Notably, Bouchacourt and Baroni (2018) examine the alignment between agents' internal representations and conceptual input properties to determine whether emergent language captures such properties or relies on low-level pixel information. However, all these setups extract object representations from pretrained convolutional neural networks (CNNs), and do not manipulate the perceptual process systematically. While emergent language and agents' representations can be related to differences in the input, the impact of differences in perception remains undetermined.

When using pretrained classifiers as visual modules, one cannot control the resulting similarities between object representations. However, developing a system of similarity relationships along relevant perceptual dimensions (e.g., color, shape, magnitude, texture) is an integral part of human concept formation and judgments of similarity are central for many cognitive processes (Gärdenfors, 2004). The transition from defining object similarities based on global perceptual resemblance, to having a system of dimension-sensitive similarities, is therefore an important step in child development (Smith, 1989). With this in mind, our goal is to use emergent language games as an experimental method for evaluating the impact of enforcing particular object relationships, based on perceptual differences along basic quality dimensions, in

order to explore the relationship between neural representations, similarity relationships, and emergent communication strategies.

Ideally, perceptual representations are manipulated directly, irrespective of input stimulus. We show how this can be achieved via *relational label smoothing* (Marino, Nieters, Heidemann, & Hertzberg, 2021). We use a conventional language emergence setup with two agents, a sender and a receiver, playing a reference game. In line with the studies above, we focus on visual perception and process the pixel inputs with CNNs. However, during CNN training, we manipulate the class labels such that for different conditions the resulting representational similarities between object classes vary. We use this setup to explore two different directions. First, we test whether perceptual biases are carried over into the emergent languages. More precisely, we ask if agents that perceive object similarities more strongly with respect to some features than others tend to ground their language in these specific features. We also evaluate the influence of sender versus receiver bias on such changes. Second, after showing that CNN representations do not preserve object similarities accurately, we test whether enforcing the preservation of similarity relationships for conceptually relevant features improves language emergence in terms of training process and emergent protocol.

General Methods

Code, results, and analyses are publicly available.[1]

Data set

We use the *3dshapes* data set (Burgess & Kim, 2018). The data set contains images of 3D shapes in an abstract room, with the following aspects being varied: floor color (10 values), wall color (10 values), object color (10 values), object size (8 values), object shape (4 values), and object orientation (15 values). We use a subset of four different object colors (red, turquoise, purple, yellow), and four different object sizes (equally spaced from smallest to largest); amounting to 96000 different images. For our purpose, we define objects by size, shape and color of the geometric shape, such that there are $4^3 = 64$ different objects. So, the term 'object' refers to an object class, such as 'tiny red cube', with each image representing an instance (or example) of such an object.

Communication game

Two agents, sender S and receiver R, play a reference game where one round of the game proceeds as follows.

1. A random object is selected as target, t.
2. S sees an image of t and produces a message. Messages have length L and consist of a sequence of words $(w_1, ..., w_L)$ from vocabulary V.
3. R sees a possibly different image of t and additionally k random distractor images, showing objects other than t (so differing in at least one of the three concept-defining features). Based on the message from S, it tries to select the target object.

If the receiver selects the target object, the agents receive a positive reward, $r = 1$, else $r = 0$. We use vocabulary size $|V| = 4$, message length $L = 3$, and $k = 2$ distractors in all simulations. In principle, this allows agents to use a distinct word for each object and thereby to achieve maximal reward. As the number of distractors is low, however, the agents may achieve relatively high rewards with suboptimal strategies. It is in the variation of such local solutions that we hope to identify linguistic differences that reflect perceptual biases.

Reinforcement learning

The sender maps the input object, o, to a probability distribution over messages, $\pi_S(m \mid o)$, by sequentially generating a probability distribution across words conditioned on the words produced so far. The receiver maps the input message onto a probability distribution over objects, $\pi_R(o \mid m)$. These distributions define the agents' policies. The agents minimize the negative expected reward, $-\mathbb{E}[r]$, and their trainable weights are updated using REINFORCE (Williams, 1992). During training, actions are sampled from the policies; during testing, the arguments of the maxima are used.

Agent model

Sender and receiver have very similar architectures. The model components and their interactions in the communication game are shown in Figure 1.

Vision module. Each agent uses a CNN as vision module, $v(\cdot)$. The CNN is pretrained on an object classification task, and the agents use the output of the penultimate fully connected layer as object representation. The weights of the vision module remain fixed during the communication game.

Language module. Each agent has a language module, $l(\cdot)$, consisting of an embedding layer and a gated recurrent unit (GRU) layer. The sender has an additional fully connected layer, which we call hidden-to-word, $htow(\cdot)$. At each time step t it projects the GRU hidden state, h_t, to a log probability distribution across words, forming the sender's policy $\pi_S(m = (w_1, ..., w_L) \mid o) = \prod_{t=1}^{L} \pi_S(w_t \mid w_{s<t}, o)$, with $\pi_S(w_t \mid w_{s<t}, o) \propto \exp\big(htow(h_t)\big)$.

Vision-to-hidden layer. Each agent has a fully connected layer, $vtoh(\cdot)$, mapping from vision to language module. For the sender, this layer is used to initialize the GRU hidden state. For the receiver, the dot product between vision-to-hidden output and final GRU hidden state determine its selection policy: $\pi_R(o \mid m) \propto \exp\big(vtoh_R(v_R(o)) \cdot l_R(m)\big)$.

Training

We use a train/validation split of 0.75/0.25.

CNN pretraining. The architecture consists of two convolutional layers with 32 channels, followed by two fully con-

[1] https://osf.io/83z5x/

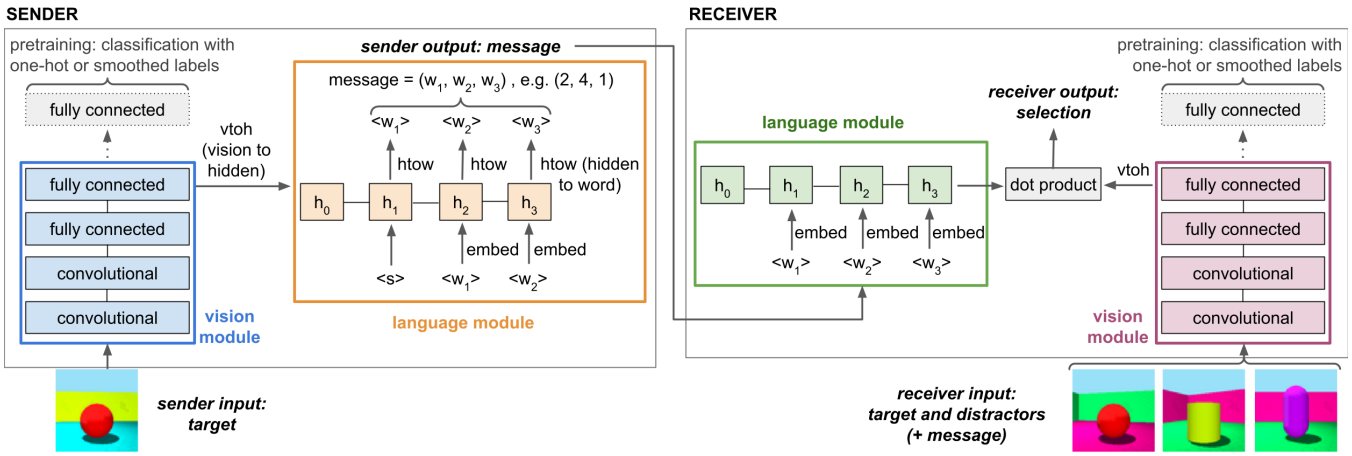

Figure 1: Schematic visualization of sender and receiver architecture and their interaction in one round of reference game. The initial input to the sender's language module, $\langle S \rangle$, is a zero vector of the same dimensionality as the embedding layer.

nected layers with 16 nodes, and a final softmax layer. The first convolutional layer is followed by a 2×2 max pooling layer. We use stochastic gradient descent (SGD) with learning rate 0.001 and batch size 128, and train for 200 epochs.

Communication game. We train agents for 150 epochs using Adam with learning rate 0.0005 and batch size 128. Embedding and GRU layer each have a dimensionality of 128. We add an entropy regularization term (Mnih et al., 2016) of 0.02 to sender and receiver loss to encourage exploration.

Relational label smoothing. In order to enforce perceptual biases in the CNN models, we use a form of relational label smoothing based on work by Marino et al. (2021), which calculates the target at training time as a sum of the usual one-hot target, $\mathbf{y}_0$, and a relational component, $\mathbf{y}_r$, according to

$$\mathbf{y} = \sigma \mathbf{y}_r + (1-\sigma)\mathbf{y}_0\,, \tag{1}$$

where $\sigma \in \mathbb{R}$ is the smoothing factor, controlling the strength with which the relationship(s) should be enforced.

To enforce object similarities along one specific object feature (or dimension), f, we make use of the single-level hierarchical version of relational label smoothing. If i is the true object class, we define superclass C_i as the set of object classes having the same feature value as i for f. Then $\mathbf{y}_r$ is given by

$$y_{r_{ij}} = \begin{cases} (n-1)^{-1} & j \in C_i \text{ and } i \neq j \\ 0 & \text{else} \end{cases}\,, \tag{2}$$

where n is the number of object classes in C_i. In order to enforce multiple feature relationships in a single model, we generalize the previous definition to include $\mathbf{y}_r$ to be a sum over relational components,

$$\mathbf{y}_r = \frac{1}{N}\sum_{f=1}^{N} \mathbf{y}_{r_f}\,, \tag{3}$$

where N is the number of feature relationships, and $\mathbf{y}_{r_f}$ represents the relational component from feature f.

Evaluation: Perception analysis metrics

Let $F = \{color, shape, size\}$ be the set of object features and A_f all values that feature $f \in F$ can take on, e.g. $A_{size} = \{tiny, small, big, huge\}$. All feature values together define the set of attributes, $A = \bigcup_{f \in F} A_f$.

Representational similarities. For every class (object type), $c_i \in C$, we extract the agent's visual representations for $N = 100$ images, using $v(\cdot)$, and calculate the average cosine similarity between all pairwise combinations of classes:

$$sim_{i,j} = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N} \cos\mathrm{sim}(v(o_k^i), v(o_l^j))\,,$$

where o^i denotes an instance of class c_i. As the vision module output lies in positive space $sim_{i,j} \in [0,1]$.

Feature-wise perceptual bias. Let $a_f^i \subset A_f$ be the value of feature f in class c_i. To analyze whether an agent is biased towards a specific feature f, we calculate the average similarity between object classes having the same value for that feature, as well as the average similarity between object classes having different values for that feature. The perceptual bias, $B(f)$, is calculated by subtracting the two values:

$$B(f) = Avg\left\{sim_{i,j}\mathbb{1}_{[a_f^i = a_f^j]}\right\} - Avg\left\{sim_{i,j}\mathbb{1}_{[a_f^i \neq a_f^j]}\right\},$$

for $i \neq j$. In our case $B(f) \in [-1,1]$, where 1 means maximal bias, 0 no bias, and -1 maximal anti-bias.

Evaluation: Language analysis metrics

Feature-wise effectiveness. We aim to measure how much information about individual features is contained in the messages. We extract 10000 random images from the training set

and use the trained sender to generate corresponding messages. Given a specific feature, f, we extract the object's feature value, $o_f \in A_f$, for each image. The conditional entropy of the objects' feature values, O_f, given the generated messages, M,

$$H(O_f \mid M) = -\sum_{m \in M,\, o_f \in O_f} p(m, o_f) \log \frac{p(m, o_f)}{p(m)},$$

quantifies how much information about feature f is not communicated in the messages. We can therefore define the feature-wise effectiveness score as

$$E(O_f, M) = 1 - \frac{H(O_f \mid M)}{H(O_f)},$$

where $H(O_f)$ is the marginal entropy of the feature values. The average effectiveness across all features,

$$\overline{E(O_f, M)} = \frac{1}{|F|} \sum_{f \in F} E(O_f, M),$$

can be used to measure how well all conceptually relevant features are communicated.

Zero-shot generalization. Zero-shot generalization measures how well the agents generalize to unfamiliar data. We retrain agents on a subset of the training data, leaving out specific objects for testing. We leave out four different objects with distinct values for each class-defining feature, such that all $3 \cdot 4 = 12$ feature values are covered.

Residual entropy. The residual entropy can be used to measure a strong form of compositionality, where each feature is encoded by the words at specific positions of the message (Resnick, Gupta, Foerster, Dai, & Cho, 2020). As the message length, $L = 3$, corresponds to the number of features, we consider all permutations of word-positions, $\pi = \{1,2,3\} \rightarrow \{1,2,3\}$. For each permutation, $p \in \pi$, we calculate the residual entropy given that the features, $f_1 = color$, $f_2 = size$, $f_3 = shape$, are encoded at the message positions given by the permutation,

$$RE(p) = \frac{1}{|F|} \sum_{i=1}^{|F|} \frac{H(O_{f_i} \mid M[p_i])}{H(O_{f_i})},$$

with messages, M, and feature values, O_f, as above. To measure compositionality, the permutation creating the smallest entropy is used, $RE = \arg\min_{p \in \pi} RE(p)$. The measure ranges from 0 (compositional) to 1 (not compositional).

Representational similarity analysis (RSA). RSA compares the similarity structure of two sets of representations (Kriegeskorte, Mur, & Bandettini, 2008). Like Bouchacourt and Baroni (2018) we use it for pairwise comparisons of sender space, receiver space, and input space—sender and receiver space being the respective RNN hidden states. We calculate the RSA score as the Spearman correlation between all pairwise cosine distances of representations in one space and all pairwise cosine distances of the corresponding representations in the other space. We represent objects by a k-hot encoding of their class defining features and use a subset of 50 random examples of each class.

Introducing perceptual biases

Our goal is to systematically manipulate the agents' perception. We aim to have four conditions, next to the unmanipulated *default*. Specific biases for either of the object-defining features—color, size, and shape—make up three of these conditions. E.g., in the *color* condition, color similarities are amplified. In addition, we experiment with an *all* condition, where we amplify similarities for all three features simultaneously. To introduce these biases, we apply relational label smoothing to the CNN training. For the individual feature conditions, we use hierarchical label smoothing, as in Equations (1) and (2), defining the superclasses by the respective feature values. E.g., in the *color* condition, if the training sample is a red object, the relational component, $\mathbf{y}_r$, is a uniform distribution of $\sigma/(16-1)$ across the class indices of the other 15 red objects. To calculate the relational component for the *all* condition, we average all relational components from the individual feature conditions, as in Equation (3).

Figure 2 illustrates the effects of label smoothing for the *default*, *color*, *size*, and *shape* conditions. Shown are pairwise similarities between object classes in the penultimate fully connected layer of the trained CNNs. Object features are structured periodically in the data set. For object class c, color is determined by $(c-1) \mod 16$, shape by $c-1 \mod 4$, and size by $\big((c-1) \mod 16\big)//4$, where $\mod$ is the modulo operator, and $//$ division without remainder. These periodic patterns are reflected in the similarity matrices. However, the patterns are not perfect as similarities in the penultimate layer are still influenced by the input topology and not entirely determined by the label distribution.

As the agents' vision modules use object representations from the penultimate layer, we quantify the CNN biases for that layer using the perceptual bias metric. The results in Table 1 show that targeting specific object features with the smoothing factor has the intended effect of inducing a perceptual bias for these features. The *color*, *shape*, and *size* networks are biased towards their specific feature, and only to that one. The *all* network is biased towards all three features although individual biases are weaker. Even if corrected for mutual attenuation they only increase to 0.117 for color, 0.099 for size, and 0.112 for shape. In addition, we find a color bias in the *default* condition. This inherent color bias is probably due to the networks' direct access to color information via the RGB channel input (Hill, Clark, Hermann, & Blunsom, 2020). In conclusion, per default object representations extracted from CNNs are biased towards color information, but relational label smoothing can shift this bias to (even multiple) other features.

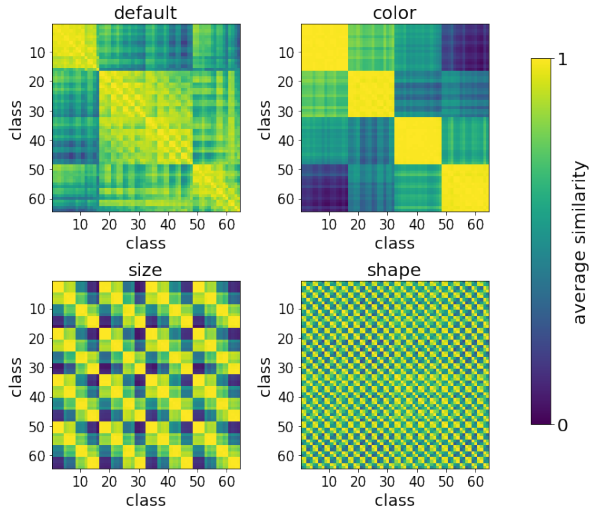

Figure 2: Pairwise class similarities of the penultimate CNN layer for *default*, *color*, *size*, and *shape* conditions.

Table 1: Feature-wise perceptual biases, $B(f)$, in the penultimate fully connected CNN layer for each condition.

	default	color	size	shape	all
$B(color)$	0.234	0.507	-0.020	-0.015	0.081
$B(size)$	0.023	-0.024	0.424	-0.016	0.056
$B(shape)$	0.004	-0.024	-0.021	0.371	0.074

The influence of differences in perception on emergent language

We now look at the influence of different perceptual biases on the emergent language, focusing on two different aspects. First, we test whether perceptual biases influence what the agents preferably talk about, i.e. what features they ground their messages in. Second, we examine whether amplifying similarities across all class-defining features (*all* condition) improves the training process or the emergent language.

Methods

For all CNNs (*default*, *color*, *shape*, *size*, *all*) we train a sender-receiver pair where both agents use the same network as vision module, and thus have the same bias. In addition, to evaluate the impact of sender versus receiver bias we run additional experiments combining a *default* sender with each type of receiver, and combining a *default* receiver with each type of sender. We report mean and standard deviation across ten runs for each agent combination.

Results

General performance. All agent pairs learn to communicate successfully, with average validation rewards ranging from 0.921 to 0.968, with chance being 0.333.

Effect of perceptual biases on language grounding. We begin by analyzing the effect of perceptual biases on emergent language when both agents have the same bias. We use the feature-wise effectiveness score to measure how much information the messages contain about specific features. The results for each type of bias and each feature are shown in Figure 3 (A). The five blocks on the x-axis show the perceptual bias conditions while the three colors encode the three features color, size and shape. In the *default* condition (left) the messages are strongly grounded in color features. This can be attributed to the inherent color bias of the default CNN. Each agent pair with color, size, and shape bias (central three blocks), grounds its messages to a large extent in the feature towards which it is biased. If similarities for all three features are amplified (right), the messages contain a relatively high amount of information about each feature.

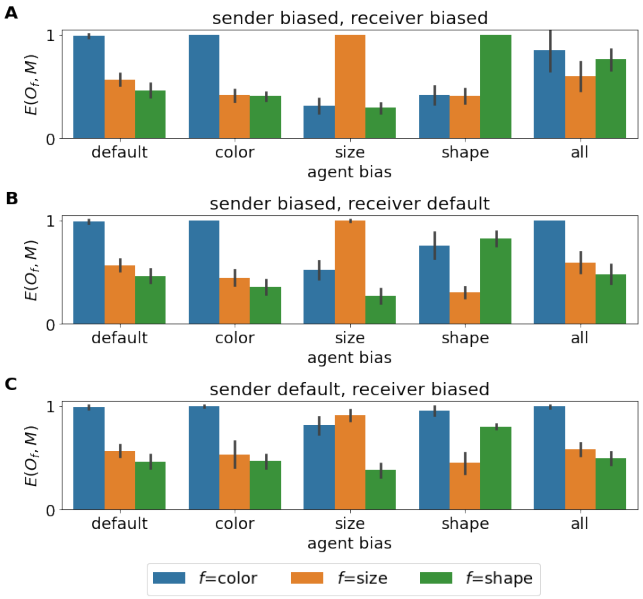

Figure 3: Feature-wise effectiveness for different pairings of senders and receivers: (A) biased sender and biased receiver, (B) biased sender and default receiver, (C) default sender and biased receiver. Perceptual biases are given on the x-axis, features (f) for calculating the effectiveness scores are color coded. We report means and standard deviations across ten runs.

Influence of sender versus receiver bias. Feature-wise effectiveness scores for varying sender bias in combination with a default receiver are shown in Figure 3 (B), and for varying receiver bias in combination with a default sender in Figure 3 (C). The results for *default* from part (A) are repeated as reference. Comparing part (B) to part (A) of the figure, and singling out the effects of color, shape and size biases, biasing only the sender has similar effects as biasing both agents. For each of these biases the language is

grounded largely in the corresponding feature. Still, the color bias of the *default* receiver leads to an increase in color effectiveness for the *size* and *shape* conditions. Comparing (C) to (B), also a receiver bias is carried over into the emergent language, even though its influence is weaker, which can be seen from the dominating color bias of the *default* sender. Looking at the *all* condition, an interesting pattern emerges. If both agents have an *all* CNN as in (A), the message information is more evenly distributed across all features than in the default condition. However, if either of the agents uses a default CNN, as in (B) or (C), this effect is reversed and the messages are mostly grounded in color, which is likely because the 'flexible' *all* agent adapts to the inherent color bias of the *default* agent. In sum, perceptual biases of both sender and receiver are reflected in the emergent language, but due to the asymmetry of communication the sender bias is more influential. Also, agents that rely strongly on all conceptually relevant object dimensions for perceptual categorization can flexibly adapt their language to suit communication partners with more narrow perceptual discrimination abilities.

Language improvement. We are interested in whether sharpening the agents' perception with respect to class-defining object features improves language learning or language properties. Table 2 compares a pair of *all* agents, having such enhanced perception, to a pair of *default* agents. Statistical significance in a two-tailed t-test with $\alpha = 0.01$ is indicated by an asterisk. In the upper half of the table the training process is evaluated. While *all* agents do not learn significantly faster than *default* agents, they do achieve significantly higher training and validation rewards.

In the lower half of the table the emergent languages are compared. Differences in zero-shot generalization are not significant. Generalization ability is largely driven by the size of the input space (Chaabouni, Kharitonov, Bouchacourt, Dupoux, & Baroni, 2020), and enforcing conceptually relevant similarities does not seem to yield an additional advantage. There is also no difference in residual entropy, and overall there is only little compositional structure in the languages. This is maybe not surprising given that even symbolic input—with fully structured object similarities—increases compositionality in comparison to pixel input but does not yield high absolute values (Lazaridou et al., 2018). However, looking at the average effectiveness score across all features, *all* agents communicate more conceptually relevant information than *default* agents. Together with the feature-wise effectiveness scores above, it seems that enforcing conceptually relevant similarity structures helps the agents overcome categorization biases, such that they can better communicate all relevant features instead of forming semantic categories based on individual features. Similarly, the RSA values show that given the right perceptual similarity structures, sender and receiver space each stay closer to the input space. While the RSA score between sender and receiver is typically high if communication works, RSA scores with respect to the input space indicate how much the emergent language captures conceptual rather than low-level input properties (Bouchacourt & Baroni, 2018). The higher values for *all* show that perceiving conceptual differences more clearly increases their use for linguistic reference.

Table 2: Evaluation of training process (top) and emergent language (bottom) for a sender-receiver pair with *default* vision modules, and one with *all* vision modules. Acquisition speed is given by the number of epochs until training reward $r \geq \theta$ is reached. Displayed are mean and standard deviation across ten runs, with better values highlighted. $*$ indicates statistical significance in a two-tailed t-test with $\alpha = 0.01$.

		default	all
train reward		0.956 ± 0.006	$\mathbf{0.968}^* \pm 0.006$
validation reward		0.959 ± 0.006	$\mathbf{0.968}^* \pm 0.006$
speed	$\theta = 0.87$	2.9 ± 1.1	$\mathbf{2.1} \pm 0.3$
	$\theta = 0.90$	5.0 ± 2.8	$\mathbf{2.8} \pm 0.7$
	$\theta = 0.93$	13.9 ± 9.5	$\mathbf{8.5} \pm 10.4$
zero-shot reward		$\mathbf{0.887} \pm 0.026$	0.860 ± 0.06
RE		0.773 ± 0.023	$\mathbf{0.752} \pm 0.028$
$\overline{E(O_f, M)}$		0.674 ± 0.028	$\mathbf{0.736}^* \pm 0.037$
RSA	sender-input	0.289 ± 0.034	$\mathbf{0.359}^* \pm 0.055$
	receiver-input	0.378 ± 0.021	$\mathbf{0.509}^* \pm 0.023$
	sender-receiver	$\mathbf{0.561} \pm 0.057$	0.546 ± 0.070

Discussion

Deep multi-agent language emergence simulations have become popular over the last years as a means to study language emergence in artificial agents themselves but also to draw inferences about the formation of natural language properties. Our experiments show that in a typical language emergence setup, agents' perceptual biases shape their linguistic biases. Importantly, such perceptual biases arise in default architectures and training conditions. For example, the organization of pixel inputs into dedicated color channels makes color information more easily accessible than other object information, and thereby induces a color bias. Given that this is just one of many ways in which neural networks process visual input differently from humans, future research should take into account the effects of such differences.

Besides, we investigate whether enforcing the 'right' similarity relations in the CNN representations improves emergent communication. Indeed, agents that perceive objects as similar along category-defining features achieve higher performance in the communication game and can more flexibly adapt to different communication partners than agents with a default CNN. In addition, the emergent language suffers less from semantic categorization biases and better captures all conceptually relevant features.

The simple task structure with few distractors allows the agents to resort to local strategies, which we exploit to identify linguistic biases. The simple data set, where object fea-

tures vary along specific dimensions, allows us to quantify these biases in a precise way. In future work, it is important to extend our findings to more complex setups, working with natural images, and more demanding communication games. In particular, the effect of enforcing contextually relevant features, should be reassessed for tasks that more strongly pressure the agents to develop optimal communication strategies, e.g. by increasing the number of distractors. Aside from that, we would like to extend our setup to study the reverse effect of how language can influence perception, and test whether the task-based formation of semantic categories can alleviate perceptual biases and improve visual processing.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - GRK 2340.

References

Bouchacourt, D., & Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP)* (pp. 981–985).

Bouchacourt, D., & Baroni, M. (2019). Miss Tools and Mr Fruit: Emergent communication in agents learning about object affordances. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)* (pp. 3909–3918).

Burgess, C., & Kim, H. (2018). *3D shapes dataset.* https://github.com/deepmind/3d-shapes.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020). Compositionality and generalization in emergent languages. *ArXiv*, *arXiv:2004.09124*.

Das, A., Gervet, T., Romoff, J., Batra, D., Parikh, D., Rabbat, M., & Pineau, J. (2019). TarMAC: Targeted multi-agent communication. In *Proceedings of the 36th international conference on machine learning (ICML)* (pp. 1538–1546).

Gärdenfors, P. (2004). *Conceptual spaces: The geometry of thought*. MIT press.

Harding Graesser, L., Cho, K., & Kiela, D. (2019). Emergent linguistic phenomena in multi-agent communication games. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3700–3710).

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1), 335–346.

Havrylov, S., & Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In *Advances in neural information processing systems (NeurIPS)* (Vol. 30, pp. 2149–2159).

Hill, F., Clark, S., Hermann, K. M., & Blunsom, P. (2020). Simulating early word learning in situated connectionist agents. In *Proceedings of the 42nd annual meeting of the cognitive science society (CogSci)* (pp. 875–881).

Kharitonov, E., & Baroni, M. (2020). Emergent language generalization and acquisition speed are not tied to compositionality. In *Proceedings of the third BlackboxNLP workshop on analyzing and interpreting neural networks for NLP* (pp. 11–15). Association for Computational Linguistics.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. In *International conference on learning representations (ICLR).*

Marino, M., Nieters, P., Heidemann, G., & Hertzberg, J. (2021). Manipulating class relationships via relational label smoothing. *Manuscript submitted for publication.*

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Harley, T., Lillicrap, T. P., ... Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd international conference on machine learning (ICML)* (p. 1928-–1937).

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, *104*(4), 1436–1441.

Resnick, C., Gupta, A., Foerster, J., Dai, A. M., & Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. In *Proceedings of the 19th international conference on autonomous agents and multiagent systems (AAMAS)* (p. 1125—1133).

Rodríguez Luna, D., Ponti, E. M., Hupkes, D., & Bruni, E. (2020). Internal and external pressures on language emergence: least effort, object constancy and frequency. In *Findings of the association for computational linguistics: EMNLP 2020* (pp. 4428–4437).

Smith, L. B. (1989). From global similarities to kinds of similarities: The construction of dimensions in development. In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 146–178). Cambridge University Press.

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