

Appendix

Emergent Communication in Continuous Worlds: Self-Organisation of Conceptually Grounded Vocabularies at Scale

Anonymous ACL submission

A Source Code and Datasets

The source code can be found within the CODE zip folder. It contains all the necessary resources to replicate and run the experiments described in the paper. Inside, you'll find two main components: DATA and EXP. The DATA folder houses the code to download, preprocess and format the datasets. The EXP folder houses the code to reproduce the experiments of the paper.

B Hardware, Training Regime, Tuned Hyperparameters

All experiments were conducted on a 20-core INTEL Xeon Gold 6148 processor, paired with 32GB of RAM. One million sequential interactions (the amount of communicative interactions in each experiment) were executed on this hardware in ± 6 to 8 hours. Table 1 includes the space of hyperparameters explored for the baseline CLEVR experiment. The best performing set of hyperparameters (in terms of communicative success and linguistic conventionalisation) are reported in the main text (see Figure 1 in the main text). Every subsequent experiment uses this same set of parameters.

Param.	Tested values
s_r	$\{+0.01, +0.05, +0.1\}$
s_p	$\{-0.01, -0.05, -0.1\}$
s_{li}	$\{-0.05, -0.01, -0.02, -0.05, -0.1\}$
σ_i	$\{0.001, 0.005, 0.01, 0.05, 0.1\}$
ω_i	$\{0.1, 0.2, 0.5, 0.75, 1.0\}$
c_r	$\{+1, +5, +10\}$
c_p	$\{-1, -5, -10\}$

Table 1: Overview of hyperparameter search

C Data Preprocessing Pipelines

C.1 Tabular Datasets

This paper uses 33 publicly available tabular datasets to validate the methodology. Each tabular

dataset stores information in rows and columns, where rows represent entities and columns represent continuous or categorical features. The datasets can be broadly classified into one of three categories: (i) 7 contain only continuous features, (ii) 24 mix continuous and categorical features, (iii) and 2 contain only categorical features. The pre-processing pipeline begins by removing columns containing all missing values and rows with any missing values. Duplicate rows are removed, keeping only the first occurrence. As some datasets represent discrete categorical information as integers, these 'continuous' features are converted to categorical features. Next, all continuous features are normalised. Finally, the datasets are divided into training and test sets using a 75%/25% split.

C.2 CLEVR

As described in the main text, the CLEVR scenario uses images from the CLEVR dataset (Johnson et al., 2017), preprocessed following the method outlined by Nevens et al. (2020). The dataset contains 85,000 images, each depicting 3 to 10 geometric objects. We retain the original data splits, with 70,000 images for training and 15,000 for testing. After processing, each depicted object is represented through a feature vector. The 20 dimensions of these feature vectors are continuously-valued and correspond to information obtained through computer vision techniques (e.g. width-height ratio, colour channel values, x-axis position, etc.).

D Examples of emerged concepts for the CLEVR, WINE and MUSHROOMS datasets

In Section 5 of the main text, an example of an emerged concept for the EXOPLANETS dataset was provided. In Figure 1, we provide three additional examples of emerged concepts for the CLEVR (Johnson et al., 2017), WINE (Cortez et al., 2009)

and MUSHROOMS (Schlimmer, 1981) datasets.

Figure 1a visualises a word with the form “xekeno” that emerged in an agent in the CLEVR experiment and was fully entrenched after 1,000,000 communicative interactions ($s = 1.0$). The concept representation of this word includes three relevant dimensions ($\omega > 0.0$): *area*, *bb-area* and *rel-area*. The values on these dimensions respectively represent, normalised on a scale between 0 and 1, the number of pixels within an entity’s boundaries, the number of pixels within an entity’s rectangular bounding box, and the ratio between an entity’s area and the number of pixels in the entire image. When mapping the *bb-area* and *rel-area* values back to raw pixel counts, we can interpret that the word prototypically refers to entities with an area of 1228.8 pixels (standard deviation of 76.8 pixels), a bounding box of 1420.8 pixels (standard deviation of 115.2 pixels), and covering just under 1% of the image. In human terms, these are objects with a small visible surface that fill a large part, yet not all, of their bounding box. When looking at agent 1’s use of this word throughout the experiment, it is indeed used in 63% of all cases to refer to small spheres.

Figure 1b visualises a word with the form “rix-esu” that emerged in an agent in the WINE experiment and was fully entrenched after 1,000,000 interactions ($s = 1.0$). The concept representation of this word has specialised towards a single relevant dimension ($\omega > 0.0$), namely the amount of residual sugar. When mapping the μ and σ values back to grams per liter, we can interpret that the concept representation prototypically refers to entities with a residual sugar content of 12,14 g/l (standard deviation of 1.51 g/l). In human terms, the concept can thus be used to refer to medium sweet wines.

Figure 1c visualises a word with the form “nivena” that emerged in an agent in the MUSHROOMS experiment and was fully entrenched after 1,000,000 interactions ($s = 1.0$). The MUSHROOMS dataset consists of 8125 entities, each described by 23 categorical features. The concept representation of this word has specialised towards four relevant dimensions. We can interpret this concept to prototypically refer to all entities (i.e. mushrooms) that have attached gills (*gill-attachment: attached*), the color of the stalk to be orange (*stalk-color-above-ring: orange*, *stalk-color-below-ring: orange*) and have either an brown or orange veil (*veil-color: (brown, orange)*). When mapping this

combination back to the dataset, we identify 192 mushrooms that are described by this combination of categorical features.

E Experimental results demonstrating robustness of methodology

E.1 Experiment testing compositional generalisability

The first experiment assesses the generality of the emergent concepts in terms of their adequacy to refer to entities that exhibit previously unseen attribute combinations, a challenge referred to as *compositional generalisability* (Johnson et al., 2017; Kim and Linzen, 2020). We therefore apply the methodology to a variation on CLEVR that is based on the CLEVR CoGenT dataset (Johnson et al., 2017). CLEVR CoGenT was especially designed to test the robustness of intelligent systems against correlations that occur at training time but not at test time. As such, a number of biases are included in the scenes by imposing restrictions on the composition of entities. In particular, in the training scenes, all cubes are either grey, blue, brown or yellow, while cylinders are always red, green, purple, or cyan. Test set A contains scenes that are subject to the same correlations. Test set B however consists of scenes that are subject to a different set of correlations, with cubes always being red, green, purple or cyan, and cylinders always being grey, blue, brown or yellow. There are no restrictions on the colour of spheres in either of the splits. Test set A can be used to assess how well a learnt model performs in a standard machine learning setting, in which the training and test sets are drawn from the same distribution. Test set B can be used to assess whether the learnt model generalises beyond the correlations that characterise the training set. For the purposes of this experiment, we built a training set and two test sets using the CLEVR CoGenT images using the same data preprocessing pipeline as CLEVR. The results of this experiment, which are provided in Table 2, show that the performance of the agents on test set A and test set B is very similar in terms of degree communicative success (99.78% vs. 99.75%), degree of linguistic conventionalisation (93.04% vs. 93.01%) and average linguistic inventory size (54.40 words vs. 55.50 words). The compositional generalisability experiment thereby confirms that the emerged linguistic convention does not break down when faced with the need to refer to entities that instantiate previously unseen

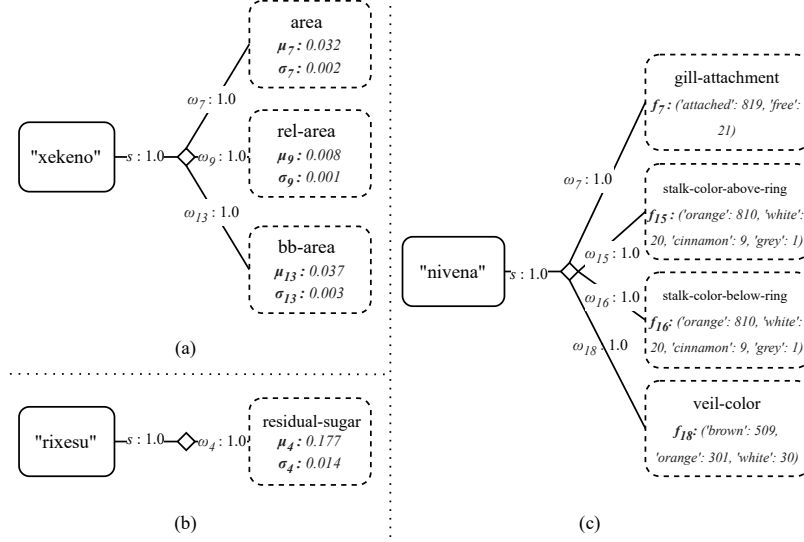


Figure 1: Examples of emerged concepts for the CLEVR (a), WINE (b) and MUSHROOMS (c) datasets. Blue dashed lines denote learned gaussian distributions over continuous features, while red dashed lines denote learned categorical distributions.

attribute combinations.

E.2 Experiments with uncalibrated sensors and noisy environments

The second and third experiment assesses the robustness of the methodology against differences in the agents' perception of the continuous domains of the world, which corresponds in our experiments to the way in which the perceived feature vectors X_S and X_L are computed from an entity's 'objective' feature vector X (see *World* in Section 2 in the main text). Concretely, we simulate two different scenarios. In the first scenario, the agents record different sensor values because of a lack of calibration. This is simulated by, at the beginning of each experimental run, shifting X_S and X_L with respect to X by a value that is individually set for each sensor of each agent. These values are sampled from a normal distribution with a mean of 0 and a standard deviation of either 0.1 (SHIFT-0.1), simulating slight calibration differences, or 1.0 (SHIFT-1), simulating substantial calibration differences. In the second scenario, the sensor values recorded by the agents are subject to noise. This is simulated by, at the start of each communicative interaction, shifting X_S and X_L with respect to X by a value that is independently sampled for each sensor of each participating agent from normal distributions with a mean of 0 and a standard deviation of 0.1 (NOISE-0.1) or 1.0 (NOISE-1).

The results of the perceptual difference exper-

iment are provided in Table 3 in comparison to the original CLEVR, WINE and EXOPLANETS experiments. Note that the MUSHROOMS dataset is not included here, as it only contains categorical data. We observe that a lack of calibration (SHIFT-0.1 and SHIFT-1) has no significant effect on the experimental results. Interestingly, we observe scenario specific responses to the different levels of noise. The presence of sensor noise in the CLEVR scenario leads to a non-catastrophic decrease in degree of communicative success (from 99.74% to 98.81% and 87.67%) accompanied by a substantial decrease in degree of conventionalisation (from 93.27% to 81.08% and 42.55%) and a slight increase in the average linguistic inventory size (from 55.96 to 55.90 and 57.90). In the WINE and EXOPLANETS scenarios these effects are more pronounced. For example, the degrees communicative success for EXOPLANETS goes from 99.67% to 94.23% and 68.89% which is paired with a significant increase in the size of the linguistic inventory. We observe that more challenging experimental conditions lead to greater variation in language use, yet remarkable levels of communicative success can still be achieved, even when agents perceive the world differently.

E.3 Experiments with heteromorphic populations

The fourth experiment assesses the applicability of the methodology to heteromorphic populations,

Dataset	comm. success (%)	convnt. (%)	inventory size
CoGenT A	99.78 (~0.10)	93.04 (~1.05)	54.40 (~3.20)
CoGenT B	99.75 (~0.10)	93.01 (~1.00)	55.50 (~4.72)

Table 2: Results of the compositional generalisability experiments, showing a similar performance in both conditions.

Dataset	Condition	comm. success (%)	convnt. (%)	inventory size
CLEVR	BASELINE	99.74 (~0.09)	93.27 (~1.46)	55.56 (~3.43)
CLEVR	NOISE-0.1	98.81 (~0.52)	81.08 (~3.69)	55.90 (~4.41)
CLEVR	NOISE-1	87.67 (~0.91)	42.55 (~2.67)	57.90 (~2.13)
CLEVR	SHIFT-0.1	99.77 (~0.11)	93.05 (~2.13)	52.90 (~3.45)
CLEVR	SHIFT-1	99.74 (~0.10)	92.41 (~1.28)	56.40 (~5.87)
WINE	BASELINE	99.64 (~0.20)	87.40 (~1.57)	78.50 (~5.08)
WINE	NOISE-0.1	97.61 (~0.46)	72.73 (~2.51)	68.40 (~1.78)
WINE	NOISE-1	76.89 (~2.46)	38.17 (~2.22)	80.60 (~7.31)
WINE	SHIFT-0.1	99.71 (~0.15)	88.83 (~1.89)	77.60 (~2.91)
WINE	SHIFT-1	99.58 (~0.21)	88.04 (~1.66)	77.70 (~5.12)
EXOPLANETS	BASELINE	99.67 (~0.10)	92.30 (~0.86)	80.50 (~4.74)
EXOPLANETS	NOISE-0.1	94.23 (~0.88)	69.54 (~2.38)	72.10 (~3.45)
EXOPLANETS	NOISE-1	68.89 (~1.86)	44.54 (~1.94)	135.80 (~14.04)
EXOPLANETS	SHIFT-0.1	99.46 (~0.29)	90.98 (~1.27)	80.00 (~3.65)
EXOPLANETS	SHIFT-1	97.93 (~1.22)	87.95 (~3.06)	86.50 (~5.99)

Table 3: Results of the experiments on the CLEVR, WINE and EXOPLANETS datasets that assess the robustness of the methodology against differences in perception.

in our case populations in which not all agents are equipped with the same combination of sensors. For this purpose, we set up variations on the four prototypical scenarios (CLEVR, WINE, MUSHROOMS and EXOPLANETS) in which each individual agent has access to a randomly selected subset of the l dimensions. This means in practice that most interactions consist of two agents that perceive entities with different sensors. Concretely, for every dataset, we run two instances of the experiment in which the agents are respectively endowed with combinations of $l-1$ and $l/2$ randomly selected sensors (HET-ONE and HET-HALF). In order to establish a meaningful basis for comparison, we also run a version of the experiment with homomorphic populations in which the agents are endowed with the same number of sensors (HOM-ONE and HOM-HALF). In the homomorphic setting, a single random subset of sensors is selected for the entire population at the beginning of each experimental run.

The test results of the experiment are listed in Table 4. When moving from the homomorphic to the heteromorphic setting, for CLEVR ($l = 20$), the degree of communicative success decreases from 99.66% to 98.47% with 19 out of 20 sensors available and from 99.60% to 85.55% with only 10 out of 20 sensors available. The degree of linguistic

conventionalisation drops to a larger extent, from 93.75% to 89.73% and from 92.82% to 59.00%. At the same time, the average linguistic inventory size increases from 46.34 to 48.25 words and from 47.33 to 52.68 words. Across all scenarios with the HET-ONE instances, the decrease in performance in terms of communicative success remains relatively limited. In the HET-HALF instances, we do observe a significant drop-off in terms of success, which can be partly attributed to the amount of dimensions l of each dataset and the nature of the data. For instance, in the EXOPLANETS scenario, each agent is equipped with a random subset of 6 sensors out of a total of 12 ($l = 12$). In this setting, the degree of communicative success averages 18.2%, with a standard deviation of 30.45%. This high variability across runs reflects the influence of the specific sensor combinations sampled. As agents perceive the environment through divergent subsets of sensors the experimental condition becomes increasingly challenging, especially when l is small. The experiment confirms that communicative success can still be reached even if agents are equipped with different combinations of sensors. Unsurprisingly, there is more variation in the words that are used by the agents in the heteromorphic setting, as agents will tend to use words that optimally fit their own sensory apparatus. This increased vari-

ation is reflected by the observed drop in degree of linguistic conventionalisation and rise in average linguistic inventory size.

E.4 Robustness against sensor defects

The fifth experiment validates the robustness of the methodology against sensor defects that occur in individual agents. For this purpose, we run a version of the four prototypical scenarios (CLEVR, WINE, MUSHROOMS and EXOPLANETS) in which the agents suffer from a sudden malfunction after 500,000 interactions. To simulate this malfunction, all agents lose access to a predefined number of sensors, which are randomly selected for each individual agent. The dynamics of the CLEVR experiments are visualised in Figure 2 for experimental conditions in which the agents lose access to respectively 1 and half of their l sensors (DEF-ONE and DEF-HALF). As is to be expected, the degrees of communicative success and conventionalisation drop at the moment of the malfunction. As the linguistic convention adapts to the new circumstances, we observe a temporary rise in the average linguistic inventory size and a partial recovery of the degrees of communicative success and conventionalisation.

The results on the test set for DEF-ONE and DEF-HALF are provided in Table 5 along with the results of the HET-ONE and HET-HALF experiments as a basis for comparison. Concretely, we are comparing the effect of agents having different sensors since the beginning of the experiment (HET-ONE and HET-HALF) to a sudden breakdown of different sensors halfway the experiment (DEF-ONE and DEF-HALF). On CLEVR the degree of communicative success amounts to 99.22% in the setting where one sensor malfunctions and to 93.82% in the setting with 10 malfunctioning sensors. The degree of linguistic conventionalisation amounts to 90.57% and 77.43% respectively, while the average number of words in use amounts to 56.10 and 55.80 respectively. We observe the same dynamics in the other three scenario's. Note that the experimental conditions after the malfunction correspond in fact to those of the experiments with heteromorphic populations reported on in Section E.3. When comparing both results, we can see that the performance after the malfunction is still better in terms of all three metrics than the performance achieved in the experiments where the agents never had access to all sensors. The experiment thereby demonstrates on the one hand that

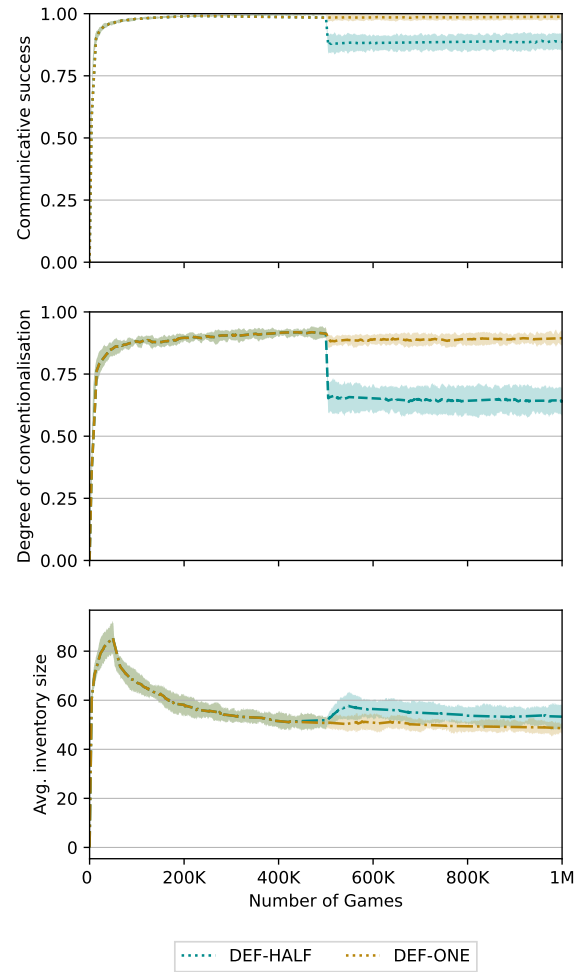


Figure 2: Evolutionary dynamics during the training phase of the CLEVR experiment in which each agent loses access to 1 or 10 sensors ($l = 20$) after 500,000 communicative interactions.

the methodology is robust against extensive sensor defects in individual agents, and on the other hand that the emergence of an effective linguistic convention before a malfunction can remain beneficial even in the long term.

References

- Paulo Cortez, Antonio Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. [Modeling wine preferences by data mining from physicochemical properties](#). *Decision Support Systems*, 47(4):547–553.
- 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2901–2910, Honolulu.

Dataset	Condition	comm. success (%)	convent. (%)	inventory size
CLEVR	HOM-ONE	99.80 (~0.07)	93.41 (~1.08)	52.70 (~3.37)
CLEVR	HET-ONE	98.12 (~1.60)	87.42 (~2.40)	54.50 (~3.72)
CLEVR	HOM-HALF	99.72 (~0.16)	92.39 (~2.72)	53.30 (~4.32)
CLEVR	HET-HALF	83.47 (~11.07)	57.03 (~17.61)	62.60 (~6.33)
WINE	HOM-ONE	99.73 (~0.08)	88.71 (~1.01)	77.40 (~2.91)
WINE	HET-ONE	93.47 (~2.35)	77.46 (~4.22)	76.60 (~2.63)
WINE	HOM-HALF	99.62 (~0.23)	89.81 (~2.04)	86.80 (~4.59)
WINE	HET-HALF	63.51 (~4.36)	41.10 (~4.97)	113.50 (~12.77)
EXOPLANETS	HOM-ONE	99.59 (~0.33)	92.89 (~1.43)	81.60 (~6.54)
EXOPLANETS	HET-ONE	93.23 (~4.83)	78.32 (~9.00)	86.80 (~5.43)
EXOPLANETS	HOM-HALF	96.31 (~6.10)	91.55 (~12.46)	155.11 (~10.74)
EXOPLANETS	HET-HALF	18.29 (~30.45)	8.15 (~23.67)	199.70 (~107.66)
MUSHROOMS	HOM-ONE	97.78 (~1.17)	86.07 (~2.90)	269.80 (~17.31)
MUSHROOMS	HET-ONE	95.04 (~1.91)	80.66 (~3.17)	288.60 (~14.99)
MUSHROOMS	HOM-HALF	88.97 (~7.13)	81.70 (~5.56)	179.22 (~37.22)
MUSHROOMS	HET-HALF	53.37 (~4.93)	30.56 (~5.59)	176.00 (~11.66)

Table 4: Results of the experiments on the CLEVR, WINE, EXOPLANETS and MUSHROOMS datasets that validate the applicability of the methodology to heteromorphic populations.

Dataset	Condition	comm. success (%)	convent. (%)	inventory size
CLEVR	DEF-ONE	99.22 (~1.05)	90.57 (~2.60)	56.10 (~3.41)
CLEVR	HET-ONE	98.12 (~1.60)	87.42 (~2.40)	54.50 (~3.72)
CLEVR	DEF-HALF	93.82 (~12.39)	77.43 (~30.32)	55.80 (~3.33)
CLEVR	HET-HALF	83.47 (~11.07)	57.03 (~17.61)	62.60 (~6.33)
WINE	DEF-ONE	97.72 (~3.86)	84.18 (~5.48)	79.30 (~3.68)
WINE	HET-ONE	93.47 (~2.35)	77.46 (~4.22)	76.60 (~2.63)
WINE	DEF-HALF	88.46 (~23.29)	69.27 (~36.39)	81.10 (~4.18)
WINE	HET-HALF	63.51 (~4.36)	41.10 (~4.97)	113.50 (~12.77)
EXOPLANETS	DEF-ONE	97.69 (~4.67)	87.18 (~8.53)	79.80 (~4.61)
EXOPLANETS	HET-ONE	93.23 (~4.83)	78.32 (~9.00)	86.80 (~5.43)
EXOPLANETS	DEF-HALF	73.37 (~55.87)	60.92 (~62.47)	98.60 (~19.93)
EXOPLANETS	HET-HALF	18.29 (~30.45)	8.15 (~23.67)	199.70 (~107.66)
MUSHROOMS	DEF-ONE	96.75 (~1.70)	83.95 (~5.73)	269.80 (~20.44)
MUSHROOMS	HET-ONE	95.04 (~1.91)	80.66 (~3.17)	288.60 (~14.99)
MUSHROOMS	DEF-HALF	83.53 (~28.77)	66.04 (~41.85)	236.30 (~23.33)
MUSHROOMS	HET-HALF	53.37 (~4.93)	30.56 (~5.59)	176.00 (~11.66)

Table 5: Results of the experiments on the CLEVR, WINE, MUSHROOMS and EXOPLANETS datasets that validate the robustness of the methodology against sensor defects in individual agents.

Najoung Kim and Tal Linzen. 2020. [COGS: A compositional generalization challenge based on semantic interpretation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105. Association for Computational Linguistics.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 2020. [From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning](#). *Frontiers in Robotics and AI*, 7(84).

Jeff Schlimmer. 1981. [Mushroom dataset](#). UCI Machine Learning Repository. Retrieved on 2025-01-20.