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

Anonymous ACL submission

Abstract

This paper introduces a general methodology through which a population of autonomous agents can converge on a linguistic convention that enables them to refer to arbitrary entities in their environment. The linguistic convention emerges in a decentralised manner through local communicative interactions between pairs of agents drawn from the population. The emergent convention consists of associations between symbolic labels (word forms) and subsymbolic concept representations (word meanings) that are grounded in a continuous feature space. We confirm the generality and scalability of the method through its evaluation on a wide and diverse selection of 34 publicly available datasets. We also demonstrate the robustness of the method against perceptual variation, including in heteromorphic populations, as well as the ability of the emergent conventions to self-adapt to changes in the environment.

1 Introduction

Human languages are evolutionary systems, which emerge and evolve through local communicative interactions between members of a linguistic community. Processes of variation and selection are at play during each and every communicative interaction, at the level of concepts, words and grammatical structures (Schleicher, 1869; Darwin, 1871; Maynard Smith and Szathmáry, 1999; Oudeyer and Kaplan, 2007; Steels and Szathmáry, 2018). Variants are introduced as creative solutions to communicative impasses and are selected for based on their linguistic, cognitive and physical fitness (Grice, 1967; Echterhoff, 2013; Van Eecke et al., 2022). The evolutionary and self-organising nature of human languages gives rise to a number of unique qualities. First of all, such decentralised, self-organising systems are known to be robust and to be able to self-repair substantial perturbations (Heylighen, 2001; Pfeifer et al., 2007). Second, populations of language users converge on shared conventions that remain adaptive to changes in their environment and communicative needs (Beckner et al., 2009). Finally, the resulting languages serve as an abstraction layer above the sensory observations and internal mental representations of individual language users (Nevens et al., 2020; Beuls and Van Eecke, 2024; Garside et al., 2025). Indeed, while linguistic forms can be observed and shared, their meanings remain tied to each language user's individual physical and cognitive embodiment.

This agent-based and evolutionary perspective on the human ability to communicate through language has served as a starting point for the development of a range of computational methodologies that model how artificial agents can co-construct emergent languages that satisfy their communicative needs (see e.g. Steels and Belpaeme, 2005; Beuls and Steels, 2013; Foerster et al., 2016; Lazaridou et al., 2017; Mordatch and Abbeel, 2018; Chaabouni et al., 2021, 2022; Nevens et al., 2022; Doumen et al., 2023; Lian et al., 2024). Rather than modelling the learning of an existing natural language, which has emerged and evolved to fit the communicative needs of a population of human language users, these methodologies allow for *artificial natural languages* to emerge and evolve to optimally support the embodiment, environment and communicative needs of populations of artificial agents. These languages are artificial in the sense that they do not exist outside the experimental set-up, yet natural in the sense that they emerge and evolve through the same evolutionary principles as human languages do.

In this paper, we focus on the emergence of linguistic conventions that associate symbolic labels (referred to as *word forms*) to subsymbolic concept representations (referred to as *word meanings*). We introduce a methodology through which a population of autonomous agents tasked with verbally referring to entities in their environment can con-

verge on a conceptually grounded vocabulary that is adequate for solving their reference task. The linguistic convention emerges in a decentralised manner through local, task-oriented and situated communicative interactions that take place between pairs of agents drawn from the population. Importantly, the entities in the environment of the agents do not come pre-categorised, but are perceived by the agents as points in a multi-dimensional, continuous feature space. As they take part in situated communicative interactions, the agents gradually converge on a vocabulary that associates shared word forms with internal concept representations that are personal yet compatible on a communicative level.

The main contribution of the paper with respect to the state of the art lies in the generality and scalability of the method. We demonstrate its direct applicability in wide variety of scenarios through evaluation on a diverse selection of 34 publicly available datasets. We also demonstrate that the emergent convention indeed exhibits a number of qualities typically associated with human linguistic communication. In particular, we show that the methodology is naturally robust against perceptual deviation, which leads to languages that self-adapt to changes in the environment of the agents.

2 Problem Definition

We address a de-centralised, multi-agent emergent communication problem. More specifically, the problem concerns the bootstrapping of a linguistic convention that agents can use for drawing each other's attention to arbitrary entities in their environment. Importantly, communicative interactions always take place locally between two agents from the population, agents need to be able to act both as speakers and as listeners, the environment does not come pre-categorised, and the emergent convention needs to be suitable for communication about previously unseen entities. More formally, the problem can be defined as follows:

Population There exists a population $P = \{a_1, \ldots, a_k\}$ that consists in a set of k autonomous agents. Agents have no access to each other's internal state nor to any centralised knowledge base, and start out as 'blank slates' without any words, concepts or knowledge about the world.

World There exists a world $W = \{e_1, \ldots, e_m\}$ that consists in a set of m entities. An observation of an entity by an agent a takes the form of a feature vector X_a of l dimensions, for example resulting from the agent's sensor read-outs. The dimensions of such a vector can be continuously-valued, categorically-valued or a combination of both. Values on continuous dimensions can be assumed to be in the range $[0, 1]$, but it cannot be assumed that all agents perceive a given entity identically or even as a vector of the same dimensionality.

Interactions Agents take part in a sequence $G = (g_j)_{j=1}^{i}$ of i task-oriented communicative interactions. At the beginning of each interaction $g \in G$, a scene $C = \{e_1, \ldots, e_n\} \subset W$ of n entities from the world is randomly created. Two agents $a_p, a_q \in P$ are randomly selected from the population, where a_p is assigned the role of speaker ($S = a_p$), while a_q is assigned the role of listener ($L = a_q$). A topic entity $T \in C$ is randomly selected from the scene and is only disclosed to S. S is tasked with drawing the attention of L to T by producing an utterance U that is passed on to L. L should then identify $T \in C$. Success occurs if L correctly identifies T. In case of failure, T is disclosed to L. After the interaction, both agents are informed about whether the interaction succeeded or failed. Identification or disclosure of entities always happens in terms of the agents' own perceived feature vectors, i.e. X_S for S and X_L for L.

The formal definition of the problem was designed to be generic and is straightforwardly instantiable in a variety of scenarios. An intuitive scenario would involve a population of robotic agents that are each equipped with a set of sensors. The values recorded by an agent's sensors for a given entity would then yield the perceived feature vector for that agent for that entity. Other scenarios involve populations of simulated agents communicating about entities that are stored as entries in tabular datasets. In such cases, agents 'perceive' a given entry as the vector composed of that entry's (normalised) column values. The problem definition will be instantiated in 34 different scenarios below. For illustrative purposes, we will focus in particular on four prototypical scenarios, with environments that are perceived in continuous dimensions (CLEVR and WINE), in categorical dimensions (MUSHROOMS) or in a combination of both (EXOPLANETS). The CLEVR scenario makes use of the images from the CLEVR dataset (Johnson et al., 2017), which were preprocessed according

to the procedure described by Nevens et al. (2020). The dataset comprises 85K images, in which 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). The WINE scenario is based on the Wine Quality dataset (Cortez et al., 2009), which holds information about 6497 wine samples along 12 dimensions that are all continuously-valued and describe their physicochemical characteristics (e.g. acidity, residual sugar, sulphates). The EXOPLANETS scenario features 4575 exoplanets, described along a combination of 8 continuously-valued dimensions (e.g. planet radius, orbital period) and 4 categorically-valued dimensions (e.g. planet type, detection method) (Mishra, 2023). Finally, the MUSHROOMS scenario features 8124 mushrooms described along 23 categorical dimensions (e.g. poisonous, habitat) (Schlimmer, 1981).

In each scenario, the world W is defined as the set of entries from the underlying dataset. For methodological reasons, 25% of the entities in W are held out for testing purposes. At the beginning of each interaction, a new scene is created by randomly selecting 10 entities from W, with the constraint that training scenes can only hold training entities and that test scenes can only hold test entities. The exception to this rule is CLEVR, where the original dataset already consists of train and test splits holding scenes of (3 to 10) entities, which we adopt in our experiments.

In line with common practice in the field (Steels, 1999; Loetzsch, 2015; Van Eecke et al., 2022), the results are analysed in terms of three quantitative metrics both during training and at test time:

Degree of communicative success The degree of communicative success reflects how successful a population of agents is at solving the task. It is computed as the average outcome of all interactions, where success counts as 1 and failure as 0.

Degree of conventionality The degree of conventionality quantifies to what extent the different agents in the population would produce the same utterance under the same circumstances, thereby measuring convergence towards a predictable linguistic convention. It is computed by averaging over all interactions a binary measure that indicates whether the listener agent would have used the same utterance as the one produced by the speaker agent to describe the topic entity, if this agent would have been the speaker.

Linguistic inventory size The average linguistic inventory size is calculated as the average number of distinct words uttered by the agents.

3 Methodology

In order to solve the problem defined above, agents need to be able to represent concepts and words, and update them based on the communicative interactions they take part in. The resulting learning dynamics should ensure that a communicatively adequate and conventional language emerges in the population. Let us first generically define how agents will represent concepts and words:

Linguistic inventory The linguistic inventory I of an agent $a \in P$, denoted as I_a, is a potentially empty set of words, with each word $w \in I$ being a coupling $w = (f, c, s)$ between a word form $f \in F$, a concept representation c and a score $0 \leq s \leq 1$. F is an infinite set of word forms (enumerated through a regular expression). Each agent is initialised with an empty linguistic inventory.

Concept representations A concept representation $c = ((\omega_1, \theta_1) \dots (\omega_l, \theta_l))$ is a sequence of couplings between a weight value ω and a distribution θ. This sequence holds one such coupling for each dimension in the feature vectors that an agent perceives. Depending on whether a dimension is continuously-valued or categorically-valued, θ will be a normal distribution parametrised by a mean μ and standard deviation σ, i.e. $\theta = (\mu, \sigma)$, or an empirical distribution $\theta = f$, where f corresponds to absolute frequencies of categories. The weight value ω represents the importance of a particular dimension for a concept. Concepts are thus represented as a sequence of distributions, with one distribution being associated to each observed dimension via a weight that indicates the importance of this dimension for the concept.

Concrete representations for concepts and words are learnt as agents take part in communicative interactions. These interactions follow the protocol defined in Section 2: a scene $C = \{e_1 \dots e_{10}\}$ of 10 entities is created, a topic $T \in C$ is selected, and two agents are assigned the roles of speaker S and listener L. Then, S needs to produce an utterance U to draw the attention of L to T:

Conceptualisation and production The speaker S computes the concept similarity $\mathrm{sim_c}(c, X_S)$ between the concept representation $c = ((\omega_1, \theta_1) \dots (\omega_l, \theta_l))$ of each word in its linguistic inventory $w = (f, c, s) \in I_S$ and the perceived feature vector $X_S = (x_1, \dots, x_l)$ for each entity in the context C. As formalised in Equations 1 and 2, the concept similarity $\mathrm{sim_c}$ between a concept c and a feature vector X is defined as the sum over all dimensions of the dimension similarity $\mathrm{sim_d}$ between the distribution θ for a given dimension in the concept and the value x for the same dimension in the feature vector, weighted by the weight value ω for that dimension in the concept. Weight values are normalised to sum to 1 across dimensions to avoid an inherent bias towards concept representations with a higher number of relevant channels. The dimension similarity $\mathrm{sim_d}$ is defined for continuous dimensions as the z-score of the value for this dimension in the feature vector given the distribution for this dimension in the concept representation (mapped between 0 and 1), and for categorical dimensions as the relative frequency of the category for this dimension in the feature vector with respect to the frequencies of categories in the concept representation.

$$\mathrm{sim_c}(c, X) = \sum_{i=1}^{l} \underbrace{\frac{\omega_i}{\sum_{k=1}^{l} \omega_k}}_{\text{normalised weight}} \underbrace{\mathrm{sim_d}(\theta_i, x_i)}_{\text{dimension similarity}} \quad (1)$$

$$\mathrm{sim_d}(\theta, x) = \begin{cases} \exp\left(-\left|\frac{x-\mu}{\sigma}\right|\right) & \text{if continuous dim.} \\ \frac{f_x}{\sum_{i=1}^{k} f_i} & \text{if categorical dim.} \end{cases} \quad (2)$$

All words $w \in I_S$ in the speaker's linguistic inventory for which it holds that the similarity between their concept representation c and the perceived feature vector for the topic entity T is larger than the similarity between c and any other entity in C are collected into a set of candidate words K (see Eq. 3). K thus groups all words in S' inventory that distinguish the topic entity from the other entities in the context. Then, the candidate words are ranked according to their communicative adequacy, computed as the product of their score s and their discriminative power, which is itself computed as the similarity between c and T minus the similarity between c and the closest other entity in C (see Eq. 4). The word form f of the candidate word with the highest communicative adequacy w^* is then uttered by S as the utterance U. U is shared with the listener L.

$$K = \{w_i \in I_S \mid \mathrm{sim}(c_i, T) > \max_{e \in C \setminus T} \mathrm{sim_c}(c_i, e)\} \quad (3)$$

$$w^* = \underset{w_i \in K}{\operatorname{argmax}}\, s_i * \underbrace{\left[\mathrm{sim_c}(c_i, T) - \max_{e \in C \setminus T} \mathrm{sim_c}(c_i, e)\right]}_{\text{discriminative power}} \quad (4)$$

Invention If there were no candidate words in I_S (i.e. $K = \varnothing$), S adds a new word $w = (f, c, s)$ to I_S, with f being randomly selected from the infinite set of forms F (see *Linguistic inventory* above) and s being assigned a default initial value s_i. $c = ((\omega_1, \theta_1) \dots (\omega_l, \theta_l))$ is initialised based on the perceived feature vector X_S. For continuous features, where $\theta = (\mu, \sigma)$, $\mu_1 ... \mu_l$ are initialised with the values of X_S and $\sigma_1 ... \sigma_l$ are assigned a default initial value σ_i. For categorical features, where $\theta = f$, the frequency of the observed category is set to 1. Finally, the weight values $\omega_1 ... \omega_l$ are assigned a default initial value ω_i. Then, f is uttered as U.

Comprehension and interpretation The listener L observes the utterance U. If L knows a word with the form U, i.e. $w = (U, c, s) \in I_L$, L identifies the entity in the context $e \in C$ that is most similar to c as the hypothesised topic T^*:

$$T^* = \underset{e_i \in C}{\operatorname{argmax}}\, \mathrm{sim_c}(c, e_i) \quad (5)$$

If L correctly identifies the topic entity, i.e. $T^* = T$, the interaction is considered successful. Otherwise, the interaction is considered a failure and T is disclosed to L as X_L. After each communicative interaction, both S and L will update the words and concept representations in their respective linguistic inventories I_S and I_L. We distinguish between successful interactions and failed interactions:

Successful interaction update After a successful interaction, S will increase the score s of the used word $w_U = (U, c, s) \in I_S$ by a fixed reward value s_r. At the same time, S will decrease the scores of the word's *competitors*, i.e. all other $w \in I_S$ that were earlier identified as belonging to the set of candidate words K, by a value that is proportional to how similar their concept representation is to the concept representation of the used word. This is done by multiplying the similarity

between both concept representations with a fixed inhibition value s_{li}. Thus, words that are more similar are considered stronger competitors and are punished harder (see Eq. 6).

$$s \leftarrow s + \begin{cases} s_r & \text{if } w = w_U \\ s_{li} \cdot \text{sim}_{\text{cc}}(c, c_U) & \text{if } w \in K \setminus w_U \end{cases} \quad (6)$$

The similarity between two concept representations is computed as the sum over all channels of the complement of the Hellinger distance (a type of f-divergence, Hellinger, 1909) between the corresponding distributions, multiplied by the similarity between their normalised weights and their average normalised weight (see Eq. 7). The *distribution similarity* component is included to reflect the relative importance of the similarity between the distributions for corresponding channels, where closer distributions lead to a higher similarity. The *normalised weight similarity* component is included to reflect the relative importance of the similarity between the weights on corresponding channels, where a smaller difference between the weights indicates a higher similarity between the channels. Finally, the *average normalised weights* component is included to reflect that channel similarities are more meaningful if channel weights are higher, with channels holding a higher average weight contributing more to the overall similarity score.

$$\begin{aligned} \text{sim}_{\text{cc}}(c_q, c_r) = \sum_{i=1}^{l} & \underbrace{(1 - D_f(\theta_{q,i} \parallel \theta_{r,i}))}_{\text{distribution similarity}} \\ & * \underbrace{\left(1 - \left|\frac{\omega_{q,i}}{\sum_{k=1}^{l} \omega_{q,k}} - \frac{\omega_{r,i}}{\sum_{k=1}^{l} \omega_{r,k}}\right|\right)}_{\text{normalised weight similarity}} \\ & * \underbrace{\frac{\frac{\omega_{q,i}}{\sum_{k=1}^{l} \omega_{q,k}} + \frac{\omega_{r,i}}{\sum_{k=1}^{l} \omega_{r,k}}}{2}}_{\text{average normalised weights}} \end{aligned} \quad (7)$$

Similarly, L will collect all words in its linguistic inventory that are considered candidate words according to the procedure described in *Conceptualisation and production* above (now based on I_L and X_L instead of I_S and X_S), and update their scores as defined in Equation 6.

Both S and L will also update their concept representation associated to U based on the context C. For continuous features, in each channel i, they update μ_i and σ_i to include their perceived feature vector X_S or X_L using Welford's online algorithm (Welford, 1962). For categorical features, the frequency of the observed category f_i is incremented. To update the weights $\omega_1 \dots \omega_l$, the dimensions with a positive discriminative power (see *Conceptualisation and production* above) are identified and increased by a fixed step c_r on a sigmoid function $\sigma(x)$. The weights from the other dimensions are decreased by a fixed step c_p in the same function. The weight values are thereby bounded between 0 and 1, with values becoming more stable as they approach 0 or 1.

Failed interaction update After a failed interaction, S will decrease the score of $w_U = (U, c, s) \in I_S$ by a fixed value s_p. If L knew a word with the observed form, L will decrease the score of $w_U = (U, c, s) \in I_L$ by a fixed value s_p and update its c based on T with relation to C in the same way as if the interaction would have been successful. If L did not know a word $w = (U, c, s)$, L will adopt the word as follows:

Adoption A new word $w = (f, c, s)$ is added to I_L, with f being the observed utterance U and s being assigned a default initial value s_i. $c = ((\omega_1, \theta_1) \dots (\omega_l, \theta_l))$ is initialised based on the perceived feature vector X_L. For continuous features, where $\theta = (\mu, \sigma)$, $\mu_1 ... \mu_l$ are initialised with the values of X_L and $\sigma_1 ... \sigma_l$ are assigned a default initial value σ_i. For categorical features, where $\theta = f$, the frequency of the observed category is set to 1. Finally, the weight values $\omega_1 ... \omega_l$ are assigned a default initial value ω_i.

4 Results

We evaluate the methodology on a wide and diverse collection of 34 publicly available datasets. We use the experimental set-up described in Section 2, train for 1M interactions and evaluate on 100K interactions that only feature entities not seen during training, averaging over 10 independent experimental runs. The methodology was defined in Section 3 in terms of a number of parameters that need to be set when running a concrete experiment. We have optimised these parameters (except # agents and # entities) on the training portion of CLEVR (see Supplementary Materials) and obtained the values reported in Table 1. We have then evaluated the methodology on all 34 datasets using the same parameter settings, with no dataset-specific fine-tuning.

The results of evaluation on the test sets in terms

Parameter	Value	Description
k	10	# agents in population
n	10	# entities in context
s_i	0.5	initial word score
s_r	$+0.1$	word score reward
s_p	-0.1	word score punishment
s_{li}	-0.02	competitor score punishment
σ_i	0.01	initial standard deviation
ω_i	0.5	initial dimension weight
c_r	$+1$	dimension weight reward
c_p	-5	dimension weight punishment
$\sigma(x)$	$\frac{1}{1+e^{-1/2x}}$	sigmoid function

Table 1: Overview of parameter settings.

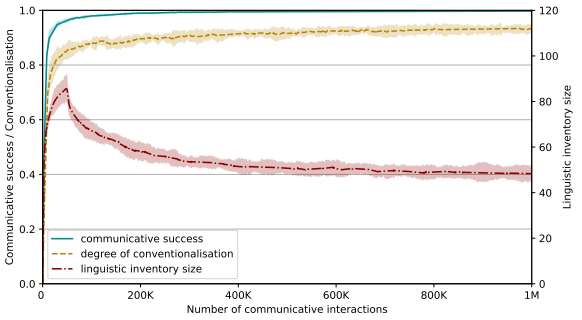


Figure 1: Evolutionary dynamics during the training phase of the CLEVR experiment: degree of communicative success, degree of conventionality and average linguistic inventory size as a function of the number of communicative interactions.

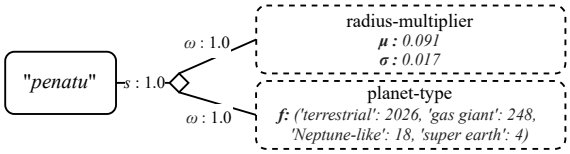


Figure 2: Example of a word emerged in the EXOPLANETS scenario, which has specialised towards two dimensions, one continuous ('radius-multiplier') and the other categorical ('planet-type').

of degree of communicative success, degree of conventionality and linguistic inventory size, as defined in Section 2, are shown in Table 2. In all 34 scenarios, the population reaches a degree of communicative success of over 95%, with a degree of conventionality above 80%. The average linguistic inventory size ranges from 55 to 267 words. These results confirm that the populations consistently converge on communicatively effective and conventional languages with a limited number of words as compared to the number of entities in the training data.

5 Analysis

The evolutionary dynamics that take place during the training phase of the prototypical CLEVR experiments are visualised in Figure 1. The graph shows the degree of communicative success (solid line, left y-axis), the degree of conventionality (dashed line, left y-axis) and the average linguistic inventory size (dashdotted line, right y-axis) as a function of the number of communicative interactions that took place and averaged over a sliding window of 5K interactions. The degree of communicative success starts at 0, as all agents start with an empty linguistic inventory. It rises to about 90% after 50K interactions, and continues to grow to 99.72% over the course of the 1M interactions that take place. The degree of conventionality roughly follows the same dynamics as the degree of communicative success, although the growth is much slower. After 1M interactions, the degree of conventionality has reached 93.30%. The average linguistic inventory size shows the typical 'overshoot pattern' that is found in many language emergence experiments (Van Eecke et al., 2022). Indeed, many words emerge during the initial phase of the experiment, as the individual agents are constantly faced with the need to invent. Then, as a result of the rewarding and punishing of words, the population converges on a smaller inventory size. The graph shows that the peak linguistic inventory size lies around 90 words, while an average of 48.42 words is reached after 1M interactions.

Figure 2 shows a word with the form "*penatu*" that emerged in agent 1 in the EXOPLANETS scenario and was fully entrenched after 1M interactions ($s = 1.0$). Two dimensions are important in the concept representation of this word ($\omega > 0.0$): the continuously-valued dimension *radius-multiplier* (expressed in earth radii) and the categorically-valued dimension *planet-type* (either 'terrestrial', 'gas giant', 'Neptune-like' or 'super earth'). De-normalising the *radius-multiplier* value reveals that "*penatu*" prototypically refers to terrestrial-type exoplanets with a radius around 81% of the earth's radius.

Figure 3 visualises the trajectories that concepts follow as they are shaped during training, projected in two dimensions using the Aligned-UMAP technique for temporal data (McInnes et al., 2018). Subfigure 3a shows the trajectory of the concept representation associated to the word "*xipabu*" in each of the 10 agents during one of the CLEVR runs. Ini-

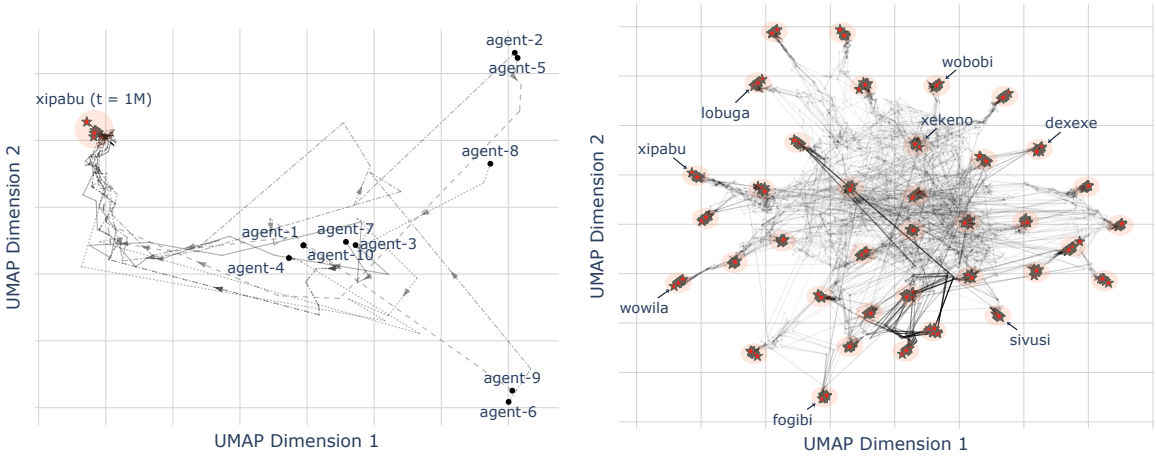
Dataset	# ent.	# cont.	# cat.	comm. success (%)	convent. (%)	inventory size
Johnson et al. (2017) (CLEVR)	468K	20	0	99.74 (~0.09)	93.27 (~1.46)	55.56 (~3.43)
Cortez et al. (2009) (WINE)	5K	12	0	99.64 (~0.20)	87.40 (~1.57)	78.50 (~5.08)
De Vito et al. (2008)	7K	15	0	99.62 (~0.18)	90.52 (~1.37)	78.40 (~3.20)
Jadikar (2019)	37K	11	0	99.48 (~0.65)	88.50 (~2.08)	76.00 (~3.40)
Ma (2019)	10K	39	0	99.41 (~0.09)	91.99 (~0.85)	92.50 (~5.97)
Vijaya et al. (2018)	303	16	0	99.28 (~0.77)	84.86 (~3.68)	86.00 (~3.50)
Tuameh (2023)	1K	99	0	99.29 (~0.22)	92.99 (~2.16)	86.10 (~5.65)
Brooks and Pope (1989)	2K	6	0	98.86 (~0.13)	91.17 (~3.29)	126.30 (~4.57)
Boksha (2024)	8K	7	1	99.70 (~0.06)	90.27 (~0.99)	65.90 (~3.78)
Mishra (2023) (EXOPLANETS)	5K	8	4	99.67 (~0.10)	92.30 (~0.86)	80.50 (~4.74)
Kadiwal (2021)	2K	9	1	99.65 (~0.17)	88.45 (~1.32)	65.90 (~3.57)
Smith et al. (1988)	768	8	1	99.63 (~0.26)	91.06 (~1.95)	82.90 (~5.26)
Agrawal (2017)	54K	7	3	99.55 (~0.14)	93.28 (~1.08)	96.20 (~6.56)
Dal Pozzolo et al. (2014)	284K	30	1	99.55 (~0.19)	86.37 (~1.99)	75.60 (~2.67)
Koklu and Ozkan (2020)	14K	16	1	99.50 (~0.24)	92.23 (~1.20)	77.40 (~5.38)
Kottarathil (2022)	611K	7	1	99.41 (~0.25)	92.80 (~1.46)	91.50 (~4.55)
Wolberg et al. (1993)	569	31	1	99.36 (~0.45)	90.38 (~3.00)	77.20 (~4.29)
Olteanu (2020)	10K	58	1	99.29 (~0.92)	86.50 (~2.35)	77.00 (~3.13)
Sejnowski and Gorman (1988)	208	60	1	99.27 (~0.64)	87.46 (~2.74)	68.50 (~4.65)
Er (2024)	51K	126	2	99.16 (~0.42)	87.90 (~1.92)	83.20 (~5.65)
USDA (2023)	5K	66	1	99.15 (~0.77)	87.04 (~2.64)	85.10 (~4.58)
Lo (2024)	4K	3	1	98.98 (~0.30)	89.23 (~1.52)	128.70 (~5.23)
Jikadara (2024)	1K	10	10	98.94 (~0.11)	89.77 (~3.18)	148.20 (~4.78)
Lainguyn123 (2024)	6K	7	13	98.92 (~0.15)	88.70 (~1.62)	142.70 (~8.78)
Mujtaba (2024)	765	3	1	98.57 (~0.64)	91.68 (~2.59)	89.40 (~2.99)
Romero-Hernandez (2022)	2K	18	10	98.38 (~0.74)	90.08 (~2.46)	88.00 (~4.42)
François (2024)	1K	4	3	98.14 (~0.57)	91.50 (~2.31)	119.30 (~5.19)
Khorasani (2024)	1K	10	7	98.06 (~0.75)	92.70 (~1.41)	84.30 (~4.27)
Bart (2015)	3K	9	8	97.93 (~0.62)	90.20 (~2.79)	107.78 (~4.47)
Ms (2024)	1K	7	2	97.88 (~1.04)	92.05 (~1.42)	93.90 (~6.05)
Fisher (1936)	147	4	1	96.41 (~1.03)	83.57 (~4.27)	92.60 (~7.44)
Banik (2018)	339	33	5	95.11 (~2.23)	86.91 (~3.68)	86.90 (~8.05)
Bohanec and Rajkovič (1998)	2K	0	7	99.19 (~0.22)	92.31 (~0.83)	120.20 (~9.32)
Schlimmer (1981) (MUSHROOMS)	8K	0	23	98.22 (~0.38)	86.79 (~2.07)	267.60 (~14.65)

Table 2: Experimental results on the 34 test sets. Mean and 2 standard deviations computed over 10 independent experimental runs. The columns describe the dataset, number of entities, number of continuous dimensions, number of categorical dimensions, communicative success, conventionality and linguistic inventory size.

tially, the concept representations of the 10 agents are very different, as each was learnt locally from a specific interaction. Over time, the concept representations of the different agents align as a result of the evolutionary dynamics that take place. Subfigure 3b shows the trajectories of all words in the final linguistic inventories of the ten agents. Not only does the figure show the alignment of concept representations, but also the formation of niches that structure the conceptual space.

6 Further experiments

Through a range of additional experiments, we also showcase that the emergent languages indeed exhibit a number of qualities typically associated with human linguistic communication. For reasons of conciseness, we summarise the conclusions here and attach a full specification of the experimental set-ups and results as supplementary material to this paper. A first experiment confirms the compositional generalisability of the emergent convention, i.e. its adequacy to refer to entities that exhibit previously unseen attribute combinations. Two experiments demonstrate the robustness of the methodology against perceptual differences, where the values of the perceived feature vectors for speaker and listener are substantially different. A fourth experiment confirms the applicability of the methodology to heteromorphic populations, where different agents perceive the environment in different dimensions, in this case because they are equipped with a different set of sensors. A final experiment confirms that the emergent conventions self-adapt to changes in the environment, in this case due to sudden sensor defects. Overall, these additional experiments show that the methodology does not



(a) Single word trajectory, 10 agents

(b) Trajectories of all words with $s > 0$, 10 agents[1]

Figure 3: Aligned-UMAP visualisations of the trajectories of concept representations over time.

break down under these challenging conditions.

7 Related Work

Our problem definition and general experimental framework build on a long tradition of population-based models of language emergence, collectively called *language game experiments* (Steels, 1996; de Boer, 2001; Oudeyer, 2006; Steels, 2012). The basic mechanisms were established through experiments on the emergence of grounded naming conventions (Steels and Loetzsch, 2012; Loetzsch, 2015; Steels et al., 2016), later moving to grounded concept learning in categorical environments (Wellens et al., 2008; Wellens, 2012) and in domain-specific continuous environments (Steels and Belpaeme, 2005; Bleys, 2016; Spranger and Beuls, 2016). The grounding of predefined concepts in perceptual data has also been modelled within this paradigm (Spranger and Beuls, 2016; Wang et al., 2016; Nevens et al., 2020). The main limitation of prior language game experiments resides in their limited applicability resulting from strict assumptions about input data.

Other related work on emergent communication has been less concerned with modelling the conditions under which human languages emerge (see Section 2) (Foerster et al., 2016; Havrylov and Titov, 2017; Kottur et al., 2017; Bouchacourt and Baroni, 2018; Mordatch and Abbeel, 2018; Noukhovitch et al., 2021; Chaabouni et al., 2021; Kim and Oh, 2021; Chaabouni et al., 2022). The problems that are addressed and the methodologies that are presented are thereby rather distantly related to ours, while definitely valuable and interesting in their own right.

8 Conclusion

This paper has introduced a methodology through which a communicatively effective, robust and adaptive linguistic convention can emerge in a population of autonomous agents. Along with a formal definition of the methodology, we have presented an extensive evaluation on 34 publicly available datasets and reported on a number of experiments that demonstrate the desirable properties of the emergent *artificial natural languages*. The research reported on in this paper constitutes a substantial contribution to the state of the art as it lifts three consequential limitations that were never jointly overcome in prior work. First, the conceptual systems are truly emergent and grounded in the perceptions of the agents. They do not need to correspond to any predefined ontology or set of concepts occurring in an existing natural language. Second, the circumstances under which the conventions emerge reflect key properties of those under which human language emerge: populations consist of more than two agents, agents can both speak and listen, and learning is fully decentralised. Finally, the method is general and thereby directly applicable, even without fine-tuning, to any dataset that describes any kind of entity in terms of any combination of continuously-valued and categorically-valued dimensions.

[1] An interactive version of this figure is accessible at https://anonymous-git-links.github.io/grounded-vocabularies/trajectory.html.

References

Shivam Agrawal. 2017. Diamonds dataset. Kaggle. Retrieved on 2025-01-13.

Rounak Banik. 2018. The complete pokemon dataset. Kaggle. Retrieved on 2025-01-20.

Austin Cory Bart. 2015. CORGIS datasets project. Kaggle. Retrieved on 2025-01-13.

Clay Beckner, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. 2009. Language is a complex adaptive system: Position paper. *Language learning*, 59:1–26.

Katrien Beuls and Luc Steels. 2013. Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLOS ONE*, 8(3):e58960.

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. What about machines? *Computational Linguistics*, 50(4):1277–1311.

Joris Bleys. 2016. *Language strategies for the domain of colour*. Language Science Press, Berlin, Germany.

Marko Bohanec and Vladislav Rajkovič. 1998. Car evaluation dataset. Retrieved on 2025-01-20.

Victor Boksha. 2024. Banana quality dataset. Kaggle. Retrieved on 2025-01-20.

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

Thomas F. Brooks and Dennis S. Pope. 1989. Airfoil self-noise dataset. UCI Machine Learning Repository. Retrieved on 2025-01-20.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2021. Communicating artificial neural networks develop efficient color-naming systems. *Proceedings of the National Academy of Sciences*, 118(12):e2016569118.

Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. Emergent communication at scale. In *10th International Conference on Learning Representations (ICLR 2022)*, pages 1–30.

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.

Andrea Dal Pozzolo, Olivier Caelen, Yann-Aël Le Borgne, Serge Waterschoot, and Gianluca Bontempi. 2014. Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10):4915–4928.

Charles R. Darwin. 1871. *The descent of man, and selection in relation to sex*, 1st edition, volume 1. John Murray, London, United Kingdom.

Bart de Boer. 2001. *The origins of vowel systems*. Oxford University Press, Oxford, United Kingdom.

Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. 2008. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical*, 129(2):750–757.

Jonas Doumen, Katrien Beuls, and Paul Van Eecke. 2023. Modelling language acquisition through syntactico-semantic pattern finding. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1317–1327.

Gerald Echterhoff. 2013. The role of action in verbal communication and shared reality. *Behavioral and Brain Sciences*, 36(4):354–355.

Aakash Er. 2024. Indian sign language hand landmarks dataset. Kaggle. Retrieved on 2025-01-20.

Ronald Aylmer Fisher. 1936. Iris dataset. UCI Machine Learning Repository. Retrieved on 2025-01-20.

Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 2137–2145, Red Hook, NY, USA. Curran Associates Inc.

Thomas François. 2024. World's best restaurants dataset. Kaggle. Retrieved on 2025-01-20.

Daniel J. Garside, Audrey L. Y. Chang, Hannah M. Selwyn, and Bevil R. Conway. 2025. The origin of color categories. *Proceedings of the National Academy of Sciences*, 122(1):e2400273121.

Paul Grice. 1967. Logic and conversation. In Paul Grice, editor, *Studies in the Way of Words*, pages 41–58. Harvard University Press, Cambridge, MA, USA.

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

Ernst Hellinger. 1909. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271.

Francis Heylighen. 2001. The science of self-organization and adaptivity. In Lowell Douglas Kiel, editor, *Knowledge management, organizational intelligence and learning, and complexity. The encyclopedia of life support systems*, pages 253–280. EOLSS Publishers, Oxford, United Kingdom.

Munirdin Jadikar. 2019. Gas turbine CO and NOx emission dataset. UCI Machine Learning Repository. Retrieved on 2025-01-20.

Bhavik Jikadara. 2024. Brand laptops dataset. Kaggle. Retrieved on 2025-01-20.

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.

Aditya Kadiwal. 2021. Water quality dataset. Kaggle. Retrieved on 2025-01-20.

Vala Khorasani. 2024. Electric vehicle charging patterns dataset. Kaggle. Retrieved on 2025-01-20.

Jooyeon Kim and Alice Oh. 2021. Emergent communication under varying sizes and connectivities. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 17579–17591, Red Hook, NY, USA. Curran Associates Inc.

Murat Koklu and Ilker Ali Ozkan. 2020. Multiclass classification of dry beans using computer vision and machine learning techniques. *Computers and Electronics in Agriculture*, 174:105507.

Prasoon Kottarathil. 2022. Bitcoin historical dataset. Kaggle. Retrieved on 2025-01-13.

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. Association for Computational Linguistics.

Lainguyn123. 2024. Student performance factors dataset. Kaggle. Retrieved on 2025-01-20.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *5th International Conference on Learning Representations (ICLR 2017)*, pages 1–11.

Yuchen Lian, Tessa Verhoef, and Arianna Bisazza. 2024. NeLLCom-X: A comprehensive neural-agent framework to simulate language learning and group communication. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 243–258, Miami, FL, USA. Association for Computational Linguistics.

Ta-Wei Lo. 2024. Fish species sampling dataset. Kaggle. Retrieved on 2025-01-20.

Martin Loetzsch. 2015. *Lexicon formation in autonomous robots*. Ph.D. thesis, Humboldt-Universität zu Berlin, Berlin, Germany.

Yi Lan Ma. 2019. League of legends diamond ranked games dataset. Kaggle. Retrieved on 2025-01-20.

John Maynard Smith and Eörs Szathmáry. 1999. *The origins of life: From the birth of life to the origin of language*. Oxford University Press, Oxford, United Kingdom.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Aditya Mishra. 2023. Nasa exoplanets. Kaggle. Retrieved on 2025-03-07.

Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1495–1502, Washington, D.C., USA. AAAI Press.

Christoffer Ms. 2024. Pokemon with stats and image dataset. Kaggle. Retrieved on 2025-01-20.

Ahmad Mujtaba. 2024. Color detection dataset. Kaggle. Retrieved on 2025-01-20.

Jens Nevens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25, Gyeongju.

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).

Michael Noukhovitch, Travis LaCroix, Angeliki Lazaridou, and Aaron Courville. 2021. Emergent communication under competition. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 974–982.

Andrada Olteanu. 2020. GTZAN dataset - music genre classification. Kaggle. Retrieved on 2025-01-20.

Pierre-Yves Oudeyer. 2006. *Self-organization in the evolution of speech*. Oxford University Press, Oxford, United Kingdom.

Pierre-Yves Oudeyer and Frédéric Kaplan. 2007. Language evolution as a darwinian process: Computational studies. *Cognitive Processing*, 8(1):21–35.

Rolf Pfeifer, Max Lungarella, and Fumiya Iida. 2007. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318(5853):1088–1093.

Omar Romero-Hernandez. 2022. Customer personality analysis dataset. Kaggle. Retrieved on 2025-01-20.

August Schleicher. 1869. *Darwinism tested by the science of language. English translation of Schleicher 1863, translated by Alex V. W. Bikkers*. John Camden Hotten, London, United Kingdom.

Jeff Schlimmer. 1981. Mushroom dataset. UCI Machine Learning Repository. Retrieved on 2025-01-20.

Terrence J. Sejnowski and Paul R. Gorman. 1988. Connectionist bench (sonar, mines vs. rocks) dataset. UCI Machine Learning Repository. Retrieved on 2025-01-20.

Jack W. Smith, James E. Everhart, Will C. Dickson, William C. Knowler, and Richard S. Johannes. 1988. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 261–265, Los Angeles, CA, USA. IEEE Computer Society Press.

Michael Spranger and Katrien Beuls. 2016. Referential uncertainty and word learning in high-dimensional, continuous meaning spaces. In *Proceedings of the 2016 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 95–100. IEEE.

Luc Steels. 1996. Perceptually grounded meaning creation. In *Proceedings of the Second International Conference on Multi-Agent Systems*, pages 338–344, Washington, D.C., USA. AAAI Press.

Luc Steels. 1999. *The Talking Heads experiment: Volume I. Words and Meanings*. Best of Publishing, Brussels, Belgium.

Luc Steels. 2012. Self-organization and selection in cultural language evolution. In Luc Steels, editor, *Experiments in Cultural Language Evolution*, pages 1–37. John Benjamins, Amsterdam, Netherlands.

Luc Steels and Tony Belpaeme. 2005. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489.

Luc Steels and Martin Loetzsch. 2012. The grounded naming game. In Luc Steels, editor, *Experiments in Cultural Language Evolution*, volume 3, pages 41–59. John Benjamins, Amsterdam, Netherlands.

Luc Steels, Martin Loetzsch, and Michael Spranger. 2016. A boy named Sue: The semiotic dynamics of naming and identity. *Belgian Journal of Linguistics*, 30(1):147–169.

Luc Steels and Eörs Szathmáry. 2018. The evolutionary dynamics of language. *Biosystems*, 164:128–137.

Muhanned Tuameh. 2023. Physical exercise recognition dataset. Kaggle. Retrieved on 2025-01-20.

USDA. 2023. FNDDS nutrient values database. U.S. Department of Agriculture: Agricultural Research Service. Retrieved on 2025-01-20.

Paul Van Eecke, Katrien Beuls, Jérôme Botoko Ekila, and Roxana Rădulescu. 2022. Language games meet multi-agent reinforcement learning: A case study for the naming game. *Journal of Language Evolution*, 7(2):213–223.

L. Farras Vijaya, Melike Dilekci, and Bala Baskar. 2018. Steel strength dataset. Kaggle. Retrieved on 2025-01-20.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378. Association for Computational Linguistics.

Barry Payne Welford. 1962. Note on a method for calculating corrected sums of squares and products. *Technometrics*, 4(3):419–420.

Pieter Wellens. 2012. *Adaptive Strategies in the Emergence of Lexical Systems*. Ph.D. thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

Pieter Wellens, Martin Loetzsch, and Luc Steels. 2008. Flexible word meaning in embodied agents. *Connection Science*, 20(2–3):173–191.

William Wolberg, Olvi Mangasarian, and W. Nick Street. 1993. Breast cancer wisconsin diagnostic dataset. UCI Machine Learning Repository. Retrieved on 2025-01-13.