Decentralised Emergence of Robust and Adaptive Linguistic Conventions in Populations of Autonomous Agents Grounded in Continuous Worlds

Jérôme Botoko Ekila[1], Jens Nevens[1], Lara Verheyen[1], Katrien Beuls[2,*], and Paul Van Eecke[1,*]

[1] *Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium*
[2] *Faculté d'informatique, Université de Namur, Belgium*

This paper introduces a methodology through which a population of autonomous agents can establish a linguistic convention that enables them to refer to arbitrary entities that they observe in their environment. The linguistic convention emerges in a decentralised manner through local communicative interactions between pairs of agents drawn from the population. The convention consists of symbolic labels (word forms) associated to concept representations (word meanings) that are grounded in a continuous feature space. The concept representations of each agent are individually constructed yet compatible on a communicative level. Through a range of experiments, we show (i) that the methodology enables a population to converge on a communicatively effective, coherent and human-interpretable linguistic convention, (ii) that it is naturally robust against sensor defects in individual agents, (iii) that it can effectively deal with noisy observations, uncalibrated sensors and heteromorphic populations, (iv) that the method is adequate for continual learning, and (v) that the convention self-adapts to changes in the environment and communicative needs of the agents.

Keywords: language emergence, multi-agent systems, autonomous agents, emergent communication, self-organisation, language evolution

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 still remain adaptive to changes in their environment and communicative needs (Beckner et al., 2009). Finally, the resulting languages effectively serve

* Joint last authors.

as an abstraction layer above the sensory-motor observations and internal mental representations of individual language users (Nevens et al., 2020). 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; Nevens et al., 2022; Doumen et al., 2023; Beuls and Van Eecke, 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.

The last decade has witnessed substantial progress when it comes to the application of these methodologies to a variety of tasks, including visual question answering (Das et al., 2017), solving puzzles (Foerster et al., 2016), negotiation (Cao et al., 2018), reference (Lazaridou et al., 2017), navigation (Sukhbaatar et al., 2016; Bogin et al., 2018; Mordatch and Abbeel, 2018) and coordination in self-driving cars (Resnick et al., 2017). At the same time, the application-oriented focus of the experiments has resulted in less attention to the modelling of the evolutionary mechanisms through which human languages have emerged and continue to evolve. For example, populations often consist of two agents only, populations are divided into agents that can speak and agents that can listen, learning is not decentralised, or agents have mind-reading capabilities. As a consequence, the languages that emerge do not exhibit the unique qualities of human languages that originally motivated the experimental paradigm (Van Eecke et al., 2022).

In this paper, we go back to the original motivation of computationally modelling the emergence and evolution of human-like languages and introduce a methodology through which a population of autonomous agents can establish, in a fully decentralised manner, a linguistic convention that enables them to refer to arbitrary entities that they observe in their environment. By taking part in situated and task-oriented communicative interactions, each agent gradually builds up its own inventory of associations between symbolic labels (word forms) and concept representations (word meanings). While all concept representations are individually constructed and grounded in an agent's own sensory-motor endowment and experiences, the linguistic systems of all individual agents are compatible on a communicative level. As opposed to prior work in this area (Wellens, 2012; Nevens et al., 2020), our methodology is applicable to any number or combination of continuous feature channels and is not limited to concepts or words that occur in existing natural languages. Apart from introducing the methodology, we also present a range of experiments that demonstrate the human-like qualities of the emergent languages. As such, we show that the methodology (i) enables a population to converge on a communicatively effective, coherent and human-interpretable linguistic convention, (ii) is naturally robust against sensor defects in individual agents, (iii) can effectively deal with noisy observations, uncalibrated sensors and heteromorphic populations, (iv) is adequate for continual learning, and (v) leads to languages that self-adapt to changes in the environment and communicative needs of the agents.

2 Methodology

The methodology that we introduce in this paper has its roots in the *language game paradigm*, a methodological framework that was originally conceived to computationally model the origins and evolution of language (Steels, 1995, 2003; Nevens et al., 2019; Van Eecke et al., 2022). Language game experiments simulate how populations of agents can learn to communicate with each other by taking part in pairwise, task-oriented and situated communicative interactions. Our methodology introduces an innovative way in which agents represent, invent, adopt and align concepts, and integrates these representations and processing mechanisms in a fairly standard language game set-up.

Experiment We define a language game experiment $E = (W, P, G)$ to be a coupling between a world W, a population P and a sequence $G = (g_j)_{j=1}^{i}$ of i communicative interactions, referred to as *games*.

Population The population $P = \{a_1, \ldots, a_k\}$ comprises a set of k autonomous agents. Each agent $a \in P$ is initialised with an empty linguistic inventory $I_a = \{\}$. Indeed, the agents do not know any concepts or words at the beginning of the experiment. Each agent is endowed with a set of l sensors $S_a = \{s_1, \ldots, s_l\}$ through which it can observe its environment. All sensors are required to map their output to values between 0 and 1. The number of sensors and their types are not necessarily the same for all agents in the population.

World The world $W = \{e_1, \ldots, e_m\}$ comprises a set of m entities. Each entity is represented through a feature vector X, with each dimension of this vector representing a particular sensor that agents can be endowed with. Depending on their individual endowment, agents can thus perceive an entity through a vector that comprises a subset of these dimensions. The values that are perceived on these dimensions might also differ from agent to agent, for example when noise or calibration differences are included in the experimental set-up. The feature vector X for an entity in W as perceived by agent $a \in P$ is notated as X_a.

Linguistic inventory The linguistic inventory I of an agent $a \in P$, notated 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 an entrenchment score s. s is bound between 0 and 1. F is an infinite set of word forms, typically enumerated through a regular expression.

Concept representation A concept representation $c = ((\omega_1, \mu_1, \sigma_1), ..., (\omega_l, \mu_l, \sigma_l))$ consists of a sequence of couplings between three numerical values ω, μ and σ. This sequence holds one such coupling for each sensor with which an agent is endowed. The weight value ω_i represents the importance of feature channel i for the concept, the mean value μ_i holds the prototypical value for the concept on this channel and the standard deviation value σ_i holds the standard deviation for the concept on this channel. Concepts are thus represented as a sequence of normal distributions, with one distribution being associated to each feature channel via a weight that indicates the importance of this feature channel for the concept.

Game Each game $g \in G$ proceeds as follows:

1. **Context selection** A context $C = \{e_1, \ldots, e_n\} \in W$ consisting of a subset of n entities is randomly selected from the world.

2. **Agent and role selection** Two agents $a_1, a_2 \in P$ are randomly selected from the population. a_1 is assigned the role of speaker $S = a_1$, while a_2 is assigned the role of listener $L = a_2$. Each agent perceives the world through its own sensors and has no access to the 'objective' feature vectors of C (see *World* above).

3. **Topic selection** A topic entity $T \in C$ is randomly selected from the context and is only disclosed to the speaker S. It is the task of S to draw the attention of the listener L to T using a word from the speaker's linguistic inventory $w \in I_S$.

4. **Conceptualisation and production** The speaker S computes the similarity $sim(c, X_S)$ between the concept representation $c = ((\omega_1, \mu_1, \sigma_1), \ldots, (\omega_l, \mu_l, \sigma_l))$ of each word in its linguistic inventory $w = (f, c, s) \in I_S$ and the perceived feature vector $X_S = (x_1, \ldots, x_l)$ for each entity in the context C. As specified in Equation 1, this is done by computing for each channel the z-score of the perceived value given the distributions stored in the concept representation. These z-scores are then mapped to values between 0 and 1 by first applying the exponential function to their absolute values and then computing the multiplicative inverse of the result. For each channel i in the concept representation, the resulting value is then weighted according to the weight value ω_i. ω_i is itself normalised by the sum of the weights on all channels. This normalisation step avoids an inherent bias towards concept representations with a higher number of relevant channels. The metric $sim(c, X_S)$ is then computed as the sum of the resulting values on all channels.

$$sim(c, X_a) = \sum_{i=1}^{l} \underbrace{\frac{\omega_i}{\sum_{k=1}^{l} \omega_k}}_{\text{normalised weight}} * \underbrace{\frac{1}{\exp(|\frac{x_i - \mu_i}{\sigma_i}|)}}_{\text{channel similarity}} \tag{1}$$

   All words in the speaker's linguistic inventory, i.e. $w \in I_S$, for which 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 as candidate words. As such, the set of candidate words corresponds to all words in the speaker'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 entrenchment score s and their discriminative power DP, which is itself computed as the similarity between c and T minus the similarity between c and the closest other entity in C. The word form f of the candidate word with the highest communicative adequacy is then uttered by S as the utterance U. U is shared between S and the listener L. If there are no candidate words in I_S:

   (4a) **Invention** A new word $w = (f, c, s)$ is added to the speaker's linguistic inventory 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. The concept representation $c = ((\omega_1, \mu_1, \sigma_1), \ldots, (\omega_l, \mu_l, \sigma_l))$ is initialised with $\mu_1 \ldots \mu_l$ being the values of the perceived feature vector X_S, $\sigma_1 \ldots \sigma_l$ being assigned a default initial value, and $\omega_1 \ldots \omega_l$ being assigned a default initial value as well. Then, f is uttered as U.

5. **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 computes the similarity between c and every entity in the context C using the similarity metric specified in Equation 1. L then points to the entity that

is most similar to c. If L does not know a word with the form U, no pointing happens and L signals that it could not understand.

6. **Feedback** If the listener L pointed to the topic entity T, the speaker S signals success. Otherwise, S signals failure and provides feedback by pointing to T.

7. **Alignment** If the game g was successful, the speaker S will increase the score s of the used word $w = (U, c, s) \in I_S$ by a fixed reward value. At the same time, the scores of the word's competitors, i.e. all other $w \in I_S$ that were earlier identified as candidate words (see *Conceptualisation and production* above), are decreased by a value that is proportional to how similar their concept representation is to the concept representation of the used word. This value is computed by multiplying the similarity between both concept representations by a fixed inhibition value. As specified in Equation 2, the similarity between two concept representations is computed as the sum over all channels of the Hellinger similarity between the corresponding distributions (Hellinger, 1909), multiplied by the similarity between their normalised weights and their average normalised weight.

$$
\begin{aligned}
sim(c_q, c_r) = \sum_{i=1}^{I} [& \overbrace{(1 - H(\mathcal{N}(\mu_{qi}, \sigma_{qi}^2), \mathcal{N}(\mu_{ri}, \sigma_{ri}^2)))}^{\text{Hellinger similarity}} \\
& * \underbrace{(1 - |\frac{\omega_{qi}}{\sum_{k=1}^{I} \omega_{qk}} - \frac{\omega_{ri}}{\sum_{k=1}^{I} \omega_{rk}}|)}_{\text{similarity of normalised weights}} \\
& * \underbrace{\frac{\frac{\omega_{qi}}{\sum_{k=1}^{I} \omega_{qk}} + \frac{\omega_{ri}}{\sum_{k=1}^{I} \omega_{rk}}}{2}}_{\text{average normalised weights}}]
\end{aligned}
\tag{2}
$$

The *Hellinger 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 *similarity of normalised weights* 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.

The listener L then collects all words in its linguistic inventory that can be 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). The scores of the word with form U and of the competing words are updated in the same way as is done for the speaker.

Both S and L will also update their concept representation associated to U based on the context C (as perceived by the individual agents). On 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). The weights on the channels are also updated for both S and L. In a first phase, the channels with a positive discriminative power (see *Conceptualisation and production* above) are identified, i.e. the channels that have a higher similarity to T than to any other entity in C according to Equation 1.

Table 1: Overview of parameters with standard default values.

Description	**Parameter**	**Default**
# agents in population	k	10
# entities in context	n	3...10
# sensors per agent	l	all (homomorphic)
initial entrenchment score	s_i	0.5
entrenchment reward	s_r	$+0.1$
entrenchment punishment	s_p	-0.1
entrenchment inhibition	s_{li}	$-0.02 * sim(c_q, c_r)$
initial standard deviation of c_i	σ_i	0.01
initial channel weight	ω_i	0.5
sigmoid function	f	$\frac{1}{1+e^{-1/2x}}$
channel weight reward	c_r	$+1$
channel weight punishment	c_p	-5

Then, all subsets of the powerset of all channels that at least contain the set of channels with positive discriminating power are considered, and the subset with the highest discriminative power for T with relation to C is selected. The weights on the channels in this subset are increased by a fixed step on a sigmoid function and the weights on the other channels are decreased by a fixed step on the same function. The weight values are thereby bounded between 0 and 1, with values becoming more stable as they approach 0 or 1.

If the game g was not successful, there are two distinct cases for alignment. If the failure was due to the L pointing to a different entity than T, S will decrease the score of $w = (U, c, s) \in I_S$ by a fixed value. L will also decrease the score of $w = (U, c, s) \in I_L$ by a fixed value and update c based on T with relation to C in the same way as if the game would have been successful. If the failure was due to L not knowing a word $w = (U, c, s)$, S will decrease the score of $w = (U, c, s) \in I_S$ by a fixed value and L will adopt the word as follows:

(7a) **Adoption** A new word $w = (U, c, s)$ is added to I_L, with s being assigned a default initial value. The concept representation $c = ((\omega_1, \mu_1, \sigma_1), ..., (\omega_l, \mu_l, \sigma_l))$ is initialised with $\mu_1 ... \mu_l$ being the values of the perceived feature vector X_L, $\sigma_1 ... \sigma_l$ being assigned a default initial value, and $\omega_1 ... \omega_l$ being assigned a default initial value as well.

The formal definition of the methodology specifies a number of parameters that need to be set when carrying out concrete experiments. A first set of parameters concerns the general experimental set-up. It specifies the number of agents in the population (k), the number of entities in the context of a single communicative interaction (n), and, per agent, a list of sensors with which it is endowed. The second set of parameters specifies how the scores of words are updated after each interaction. It specifies the initial score of words that are invented or adopted (s_i), along with update rules for words that were used successfully (s_r), for words that were used unsuccessfully (s_p), and for words that compete with words that were used successfully (s_{li}). The final set of parameters concerns the initialisation and updating of concept representations. It specifies the initial weight and standard deviation for feature channels in new concepts (ω_i and σ_i), the sigmoid function along which channel weights are increased or decreased (f) and the step on this function by which the weights are shifted in case of success (c_r) or failure (c_p). An overview of these parameters along with standard default

Table 2: Overview of hyperparameter search space.

Parameter	Tested values
s_r	$\{+0.01, +0.05, +0.1\}$
s_p	$\{-0.01, -0.05, -0.1\}$
s_{li}	$\{-0.05, -0.01, -0.02, -0.05, -0.1\} * sim(c_q, c_r)$
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	$\{-1, -5, -10\}$

values is provided in Table 1. Table 2 includes the space of hyperparameters explored for the baseline experiment. The best performing set corresponds to the standard default values specified in Table 1. Every subsequent experiment uses this same set of parameters.

3 Experimental validation

This section presents a range of experiments that were designed to serve as an initial validation of the methodology introduced in Section 2, as well as to demonstrate the robustness, flexibility and adaptivity of the emergent languages. Three datasets were chosen for this experimental validation, based on their public availability, the fact that they describe entities in terms of continuous features, and the diversity of domains that are covered. The first dataset (henceforth CLEVR) makes use of the images of the CLEVR dataset (Johnson et al., 2017), which were preprocessed according to the procedure described by Nevens et al. (2020). Concretely, the resulting dataset comprises 85,000 images, in which each depicted object is represented through a feature vector. The 20 dimensions of these feature vectors correspond to information obtained through computer vision techniques, including the number of corners of an object, its width-height ratio, its color channel values, and its position on the horizontal and vertical axes. The second dataset (henceforth WINE) concerns the Wine Quality dataset (Cortez et al., 2009), which holds information about 4898 wine samples along 11 dimensions that describe their physicochemical characteristics (e.g. acidity, residual sugar, alcohol and sulphates). The third dataset, called Credit Card Fraud Detection (Dal Pozzolo et al., 2014) (henceforth CREDIT), holds 284,807 entries of financial transactions described along 28 dimensions resulting from a principal component analysis. As such, the resulting datasets cover three very different types of data, ranging from visual scenes over physicochemical analyses to principal components extracted from financial transaction records.

Training and test splits for the three datasets were created in a two-stage process. The first stage consisted in the creation of scenes, i.e unique sets of entities that can serve as the context for a language game (see the *Context selection* step in Section 2). Each scene consists of 3 to 10 entities, with entities occurring in a training scene being excluded from being part of a test scene. For CLEVR, the distribution of scenes from the original training and test splits were kept, holding 70,000 and 15,000 scenes respectively. In the case of WINE, 90% of the wine samples were used to create 20,000 training scenes, and the remaining 10% were used to create 1,000 test scenes. For CREDIT, 90% of the financial transactions were used to create 40,000 training scenes and the remaining 10% were used to create 4,000 test scenes. In a second stage, the actual training and tests sets were constructed by randomly sampling from the training and test scenes. For each

dataset, training and test sets consisting of 1,000,000 and 100,000 scenes were compiled. The same scene can thus occur multiple times in the training or test set, but can never be part of both. The fact that the same scenes can occur multiple times does not entail that the same game is played multiple times. Indeed, many different games can be played in the same scene depending on the participating agents and the selected topic.

3.1 Emergence of a communicatively effective, coherent and interpretable convention

The first experiment validates the methodology on the three datasets, thereby adopting the default parameter settings listed in Table 1. The results are analysed in terms of three quantitative metrics: *degree of communicative success*, *degree of linguistic coherence* and *average linguistic inventory size*. The degree of communicative success reflects how successful a population of agents is at solving the game task. It is computed as the average outcome of the last 1,000 games, where success counts as 1 and failure as 0 (see the *Feedback* step of the game description is Section 2). The degree of linguistic coherence quantifies inhowfar 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 for each interaction as 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. This binary measure is then averaged over the last 1,000 interactions. Finally, the average linguistic inventory size reflects the number of distinct words that are in 'active use'. It is calculated as the average number of distinct words uttered by the agents during the last 1,000 games in which they took up the speaker role.

The experimental results obtained on the test portions of the three datasets are listed in Table 3. The table reports the average performance on the three metrics over 10 independent experimental runs[1], along with a value that indicates the spread of the results in terms of two standard deviations. The results show that the methodology enables a population of agents to converge on a communicatively effective and coherent linguistic convention in each of the task environments, with a degree of communicative success above 99.5% and a degree of linguistic coherence above 87.5%. The average linguistic inventory size of the agents revolves around 50 words in each of the task environments. Examples of words that have emerged during the different experiments are shown in Figure 1. Figure 1a visualises a word with the form "demoxu" that emerged in agent 1 in the CLEVR experiment and was fully entrenched after 1,000,000 games ($s = 1.0$). 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 1344 pixels (standard deviation of 76.8 pixels), a bounding box of 1574 pixels (standard deviation of 115 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 73% of all cases to refer to small spheres. Figure 1b visualises a word with the form "zapose" that emerged in agent 1 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

[1]All experiments were conducted on a 20-core INTEL Xeon Gold 6148 processor, paired with 32GB of RAM. One million sequential games (the amount of games in each experiment) were executed on this hardware in $\pm$ 8 hours.

Table 3: Results on the three test sets in terms of communicative success, linguistic coherence and linguistic inventory size. Mean and 2 standard deviations computed over 10 runs.

Dataset	**Communicative success** ↑	**Linguistic coherence** ↑	**Inventory size** ↓
CLEVR	99.65 ± 0.13	93.86 ± 1.09	46.72 ± 2.45
WINE	99.74 ± 0.15	88.67 ± 1.92	52.67 ± 2.93
CREDIT	99.67 ± 0.13	87.72 ± 2.50	51.43 ± 2.49

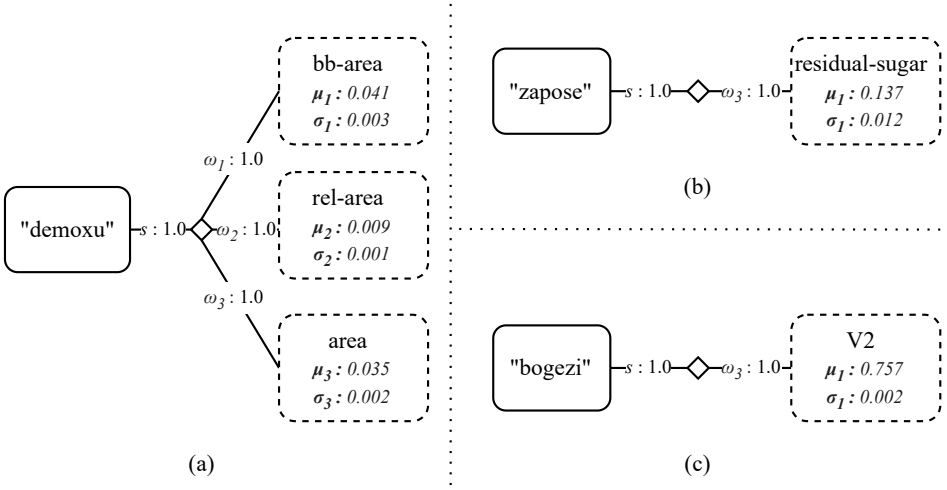

Figure 1: Examples of emerged concepts for the CLEVR (a), WINE (b) and CREDIT (c) datasets.

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 12,34 g/l (standard deviation of 1.39 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 "bogezi" that emerged in agent 1 in the CREDIT experiment and was fully entrenched after 1,000,000 interactions ($s = 1.0$). The concept representation of this word has specialised towards a low value range on a single relevant dimension ($\omega > 0.0$), namely the second PCA component. When it comes to interpretability, the three example words illustrate that the concept representations are interpretable up to the interpretability of the input dimensions. Indeed, if these dimensions are meaningful to humans, for example in the case of visual features, physico-chemical characteristics or other sensor measurements, the resulting concepts are equally human-interpretable. If these features correspond to dimensions that are more difficult to interpret by humans, such as PCA components, a communicatively effective and coherent linguistic convention with transparent concept representations still emerges, but the interpretation difficulty of the input dimensions percolates to the concept representations.

Figure 2 provides more insight into the evolutionary dynamics that take place during the training phase of the CLEVR experiment. The graph shows the degree of communicative success (solid line, left y-axis), the degree of linguistic coherence (dotted line, left y-axis) and the average linguistic inventory size (dashed line, right y-axis) as a function of the number of games that are played. The

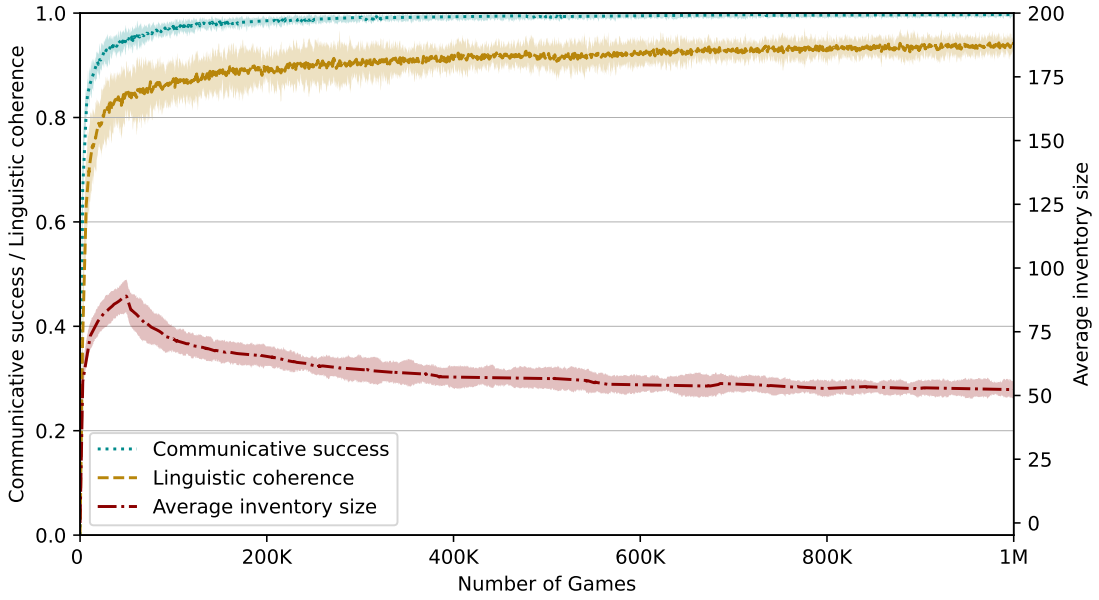

Figure 2: Evolutionary dynamics during the training phase of the CLEVR experiment: degree of communicative success, degree of linguistic coherence and average linguistic inventory size as a function of the number games that are played.

degree of communicative success starts at 0, as all agents start with an empty linguistic inventory. It rises to about 90% after 50,000 games, and continues to grow to over 99.5% over the course of the 1,000,000 games that are played. The degree of linguistic coherence roughly follows the same dynamics as the degree of communicative success, although the growth is much slower. After 1,000,000 games, the degree of linguistic coherence has reached about 90% as it continues to increase. 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 during the alignment phase of each game, the population converges towards a smaller set of words. The graph shows that the peak linguistic inventory size lies around 90 words, while an average of just over 50 words is reached after 1,000,000 games. Note that these numbers are not the same as those reported for the test set in Table 3. Indeed, words that are too specific to adequately describe previously unseen entities are never used by the agents at test time, reducing the number of words in 'active use'. The same dynamics as those shown in Figure 2 for CLEVR also materialise in the WINE and CREDIT experiments, but could not be included graphically due to space limitations.

3.2 Compositional generalisability of the emergent concepts

The second 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

Table 4: Results of the compositional generalisability experiments, showing a similar performance in both conditions.

Dataset	**Communicative success** ↑	**Linguistic coherence** ↑	**Inventory size** ↓
CoGenT A	99.63 ± 0.12	93.50 ± 1.96	47.63 ± 3.40
CoGenT B	99.62 ± 0.12	93.51 ± 1.83	47.58 ± 3.59

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 through the same two-stage process as the one that was used for creating the CLEVR, WINE and CREDIT datasets. The results of this experiment, which are provided in Table 4, show that the performance of the agents on test set A and test set B is very similar in terms of degree communicative success (99.63% vs. 99.62%), degree of linguistic coherence (93.50% vs. 93.51%) and average linguistic inventory size (47.63 words vs. 47.58 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 attribute combinations.

3.3 Applicability to heteromorphic populations

The third experiment assesses the applicability of the methodology to heteromorphic populations, in our case populations in which not all agents are equipped with the same combination of sensors. For this purpose, we set up a variation on the CLEVR experiment in which each individual agent has access to a randomly selected subset of the 20 dimensions that are provided by the dataset. This means in practice that almost all games are played by two agents that do not perceive the same entities through the same dimensions. Concretely, we run two instances of the experiment in which the agents are respectively endowed with combinations of 19 and 10 randomly selected sensors (HETERO-19 and HETERO-10). 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 (HOMO-19 and HOMO-10). In the homomorphic setting, the sensor combination is randomly selected for each agent at the beginning of each experimental run and remains constant throughout the experiment.

The test results of the experiment are listed in Table 5. When moving from the homomorphic to the heteromorphic setting, 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 coherence 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

Table 5: Results of the experiments that validate the applicability of the methodology to heteromorphic populations.

Condition	**Communicative success** ↑	**Linguistic coherence** ↑	**Inventory size** ↓
HOMO-19	99.66 ± 0.12	93.75 ± 1.26	46.34 ± 3.62
HETERO-19	98.47 ± 1.33	89.73 ± 3.62	48.25 ± 2.68
HOMO-10	99.60 ± 0.42	92.82 ± 2.93	47.33 ± 6.35
HETERO-10	85.55 ± 9.54	59.00 ± 14.54	52.68 ± 8.23

46.34 to 48.25 words and from 47.33 to 52.68 words. The experiment thereby confirms that a high degree of communicative success can still be reached even if agents are equipped with very 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 variation is reflected by the observed drop in degree of linguistic coherence and rise in average linguistic inventory size.

3.4 Robustness against sensor defects

The fourth experiment validates the robustness of the methodology against sensor defects that occur in individual agents. For this purpose, we run a version of the CLEVR experiment in which the agents suffer from a sudden malfunction after 500,000 games. 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 experiment are visualised in Figure 3 for experimental conditions in which the agents lose access to respectively 1 and 10 of their 20 sensors (DEFECT-1 and DEFECT-10). As is to be expected, the degrees of communicative success and linguistic coherence 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 linguistic coherence.

The results on the test set are provided for both conditions in Table 6 along with the results of the HETERO-19 and HETERO-10 experiments as a basis for comparison. The degree of communicative success amounts to 98.31% in the setting where one sensor malfunctions and to 89.95% in the setting with 10 malfunctioning sensors. The degree of linguistic coherence amounts to 90.15% and 66.61% respectively, while the average number of words in use amounts to 47.05 and 47.72 respectively. Note that the experimental conditions after the malfunction correspond in fact to those of the experiments with heteromorphic populations reported on in Section 3.3. 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 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.

3.5 Robustness against differences in perception

The fifth experiment assesses the robustness of the methodology against differences in the agents' perception 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

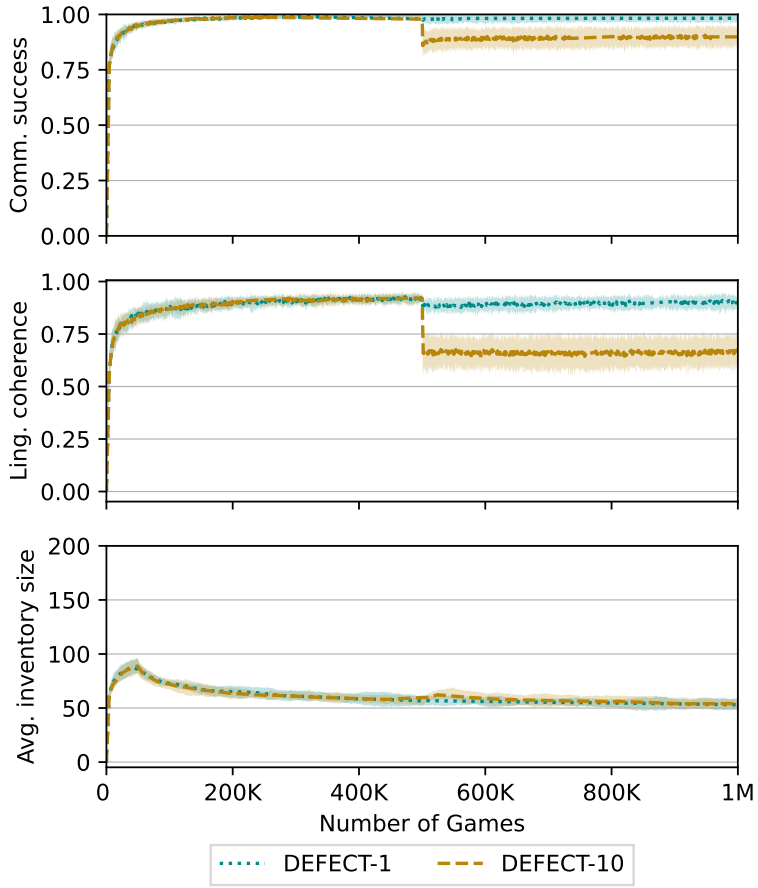

Figure 3: Evolutionary dynamics during the training phase of the CLEVR experiment in which each agent loses access to 1 or 10 sensors after 500,000 games.

Table 6: Results of the experiments that validate the robustness of the methodology against sensor defects in individual agents.

Condition	**Communicative success** ↑	**Linguistic coherence** ↑	**Inventory size** ↓
DEFECT-1	98.31 ± 1.67	90.15 ± 2.99	47.05 ± 3.82
HETERO-19	98.47 ± 1.33	89.73 ± 3.62	48.25 ± 2.68
DEFECT-10	89.95 ± 4.28	66.61 ± 7.55	47.72 ± 3.20
HETERO-10	85.55 ± 9.54	59.00 ± 14.54	52.68 ± 8.23

Table 7: Results of the experiments that assess the robustness of the methodology against differences in perception.

Condition	**Communicative success** ↑	**Linguistic coherence** ↑	**Inventory size** ↓
CLEVR	99.65 ± 0.13	93.86 ± 1.09	46.72 ± 2.45
SHIFT-0.1	99.62 ± 0.13	93.48 ± 2.02	46.82 ± 1.60
SHIFT-1	99.61 ± 0.14	93.73 ± 1.44	46.16 ± 4.01
NOISE-0.1	98.40 ± 0.63	82.65 ± 3.07	47.72 ± 2.28
NOISE-1	87.04 ± 2.44	46.58 ± 1.27	49.00 ± 6.38

Section 2). 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 shifting X_S and X_L with respect to X by a value that is individually set for each sensor of each agent at the beginning of each experimental run. 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 shifting X_S and X_L with respect to X by a value that is independently sampled for each sensor of each participating agent at the start of each game 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 experiment are provided in Table 7 in comparison to the original CLEVR experiment. We can see that a lack of calibration has no significant effect on the experimental results. The presence of sensor noise leads to a non-catastrophic decrease in degree of communicative success (from 99.65% to 98.40% and 87.04%). The decrease in degree of linguistic coherence is more substantial (from 93.86% to 82.65% and 46.58%) and is accompanied by a slight increase in the average linguistic inventory size (from 46.72 to 47.72 and 49.00). The experiment thereby shows that the methodology does not break down when faced with agents that perceive the world differently. It also confirms the trend observed in the previous experiments that more challenging experimental conditions lead to more variation in language use, while remarkable degrees of communicative success can still be achieved.

3.6 Adequacy for continual learning

The final experiment assesses the adequacy of the methodology for continual learning, focussing in particular on its robustness against catastrophic forgetting (McCloskey and Cohen, 1989; French, 1999). For the purposes of this experiment, we (i) train a model on the CLEVR training set, (ii) evaluate it on the CLEVR test set, (iii) continue to train it on the WINE training set, (iv) evaluate it on the WINE test set, and (v) evaluate it again on the CLEVR test set. The dynamics of the experiment are shown in Figure 4. Unsurprisingly, the first 1,000,000 games exhibit the same dynamics as those observed for the original CLEVR experiment shown in Figure 2. When moving to WINE after 1,000,000 games, the degrees of communicative success and linguistic coherence drop to 0 but rapidly increase again as if a new experiment would have started. The average linguistic inventory size quickly rises as many new words are invented to accommodate the WINE dataset. Then, this number gradually decreases as the words that are not adequate for describing wine samples cease to be used by the agents.

The results of the experiment are shown in Table 8. After training on CLEVR, a degree of

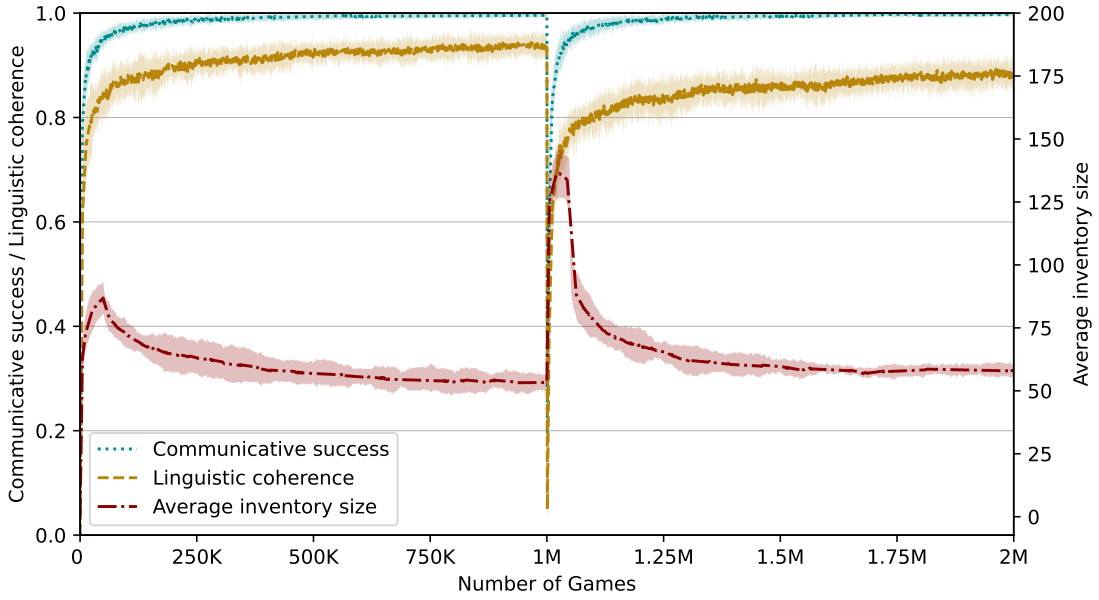

Figure 4: Evolutionary dynamics during the continual learning experiment, in which a model is first trained on CLEVR and then on WINE.

Table 8: Results of the continual learning experiment.

Condition	**Communicative success** ↑	**Linguistic coherence** ↑	**Inventory size** ↓
CLEVR	99.61 ± 0.11	93.75 ± 1.85	46.41 ± 2.67
CLEVR-WINE	99.72 ± 0.20	88.13 ± 2.34	52.40 ± 2.49
CLEVR-CONT	99.60 ± 0.14	93.72 ± 1.83	46.36 ± 2.34

communicative success of 99.61% is obtained on the CLEVR test set, along with a degree of linguistic coherence of 93.75% and an average linguistic inventory size of 46.41 (CLEVR). After continuing to train on WINE, the agents achieve a degree of communicative success of 99.72% on the WINE test set, along with a degree of linguistic coherence of 88.13% and an average linguistic inventory size of 52.40 (CLEVR-WINE). These numbers indeed match those recorded in the original CLEVR and WINE experiments. When evaluating the model on the CLEVR test set after training first on CLEVR and then on WINE, the results that are obtained do not deviate significantly from the results obtained before training on wine (CLEVR-CONT). The experimental results thereby confirm that the methodology is adequate for continual learning and is not susceptible to catastrophic forgetting.

4 Discussion and 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. The linguistic convention emerges in a decentralised manner through local, task-oriented and situated communica-

tive interactions that take place between pairs of agents drawn from the population. The linguistic convention takes the form of symbolic labels associated to concept representations that are grounded in a multi-dimensional, continuous feature space. These form-meaning associations are individually constructed by each agent and are shaped by their past successes and failures in communication. The methodology embodies the evolutionary dynamics of the language game paradigm (Steels, 1995, 2003; Nevens et al., 2019; Van Eecke et al., 2022) and integrates an innovative way in which agents represent, invent, adopt and align concept representations.

Along with a formal definition of the methodology, we have presented a range of experiments that serve as its initial validation and which demonstrate the desirable properties of the emergent *artificial natural languages*. As such, we have first applied the methodology to three datasets that contain very different types of data, ranging from visual scenes over physicochemical analyses to principal components extracted from financial transaction records. Yielding a communicatively effective, coherent and transparent linguistic convention in all three cases, the experiment shows that the effectiveness of the methodology is not limited to a particular domain or data type. Then, we reported on two experiments that confirm that the methodology is capable of compositional generalisation and that it remains effective when applied to heteromorphic populations. The fourth and fifth experiments demonstrate the robustness of the methodology against sensor defects and noisy observations, including those resulting from a lack of calibration. The final experiment validates the adequacy of the methodology for continual learning, focussing in particular on its resilience against catastrophic forgetting.

The research reported on in this paper constitutes a novel contribution to the state of the art as it lifts three consequential limitations that were never successfully overcome together in prior work. The first limitation concerns the emergent nature of the conceptual distinctions. Most prior approaches learn to ground a predefined set of concepts (Spranger and Beuls, 2016; Wang et al., 2016; Nevens et al., 2020; Verheyen et al., 2023). These concepts are symbolically annotated in training data and correspond to distinctions that occur in an existing natural language, typically English. As these concepts have emerged and evolved to fit the communicative needs and physical endowment of a community of human language users, they do not necessarily fit well the sensors and communicative tasks of artificial agents (Van Eecke et al., 2022). The second limitation concerns the circumstances under which the languages emerge and evolve. These are often too far removed from those under which human languages emerge and evolve to bring about evolutionary processes that yield emergent languages with the same desirable properties. In particular, populations sometimes consist of two agents only (Havrylov and Titov, 2017; Bouchacourt and Baroni, 2018; Noukhovitch et al., 2021), agents can either speak or listen, but not both, (Kottur et al., 2017; Mordatch and Abbeel, 2018; Chaabouni et al., 2021, 2022), or learning is not decentralised (Foerster et al., 2016; Kim and Oh, 2021). Finally, prior approaches that are not subject to the first two limitations are limited in their applicability, as they have not been generalised beyond the emergence of naming conventions (Steels and Loetzsch, 2012; Loetzsch, 2015; Steels et al., 2016), to continuous feature spaces (Wellens et al., 2008; Wellens, 2012), or to arbitrary combinations of feature channels (Vogt, 2005; Belpaeme and Bleys, 2005; Spranger, 2013; Steels, 2015; Vogt, 2015; Bleys, 2016; Spranger, 2016). By lifting these three limitations at the same time, the methodology introduced in this paper provides a model of how human-like linguistic conventions can emerge and evolve in populations of autonomous agents, which is, importantly, directly applicable to any dataset that situates entities in a continuous feature space.

Acknowledgements

The research reported on in this paper received funding from the EU's H2020 RIA programme under grant agreement no. 951846 (MUHAI), from the Research Foundation Flanders (FWO) through a postdoctoral grant awarded to PVE (grant no. 76929), from the Flemish Government under the 'Flanders AI Research Program' and from the Walloon Government under the *ARIAC by Digital Wallonia4.AI* program (project no. 2010235).

References

Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., and Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language learning*, 59:1–26.

Belpaeme, T. and Bleys, J. (2005). Explaining universal color categories through a constrained acquisition process. *Adaptive Behavior*, 13(4):293–310.

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

Beuls, K. and Van Eecke, P. (2024). Construction grammar and artificial intelligence. In Fried, M. and Nikiforidou, K., editors, *The Cambridge Handbook of Construction Grammar*. Cambridge University Press, Cambridge, United Kingdom. Forthcoming.

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

Bogin, B., Geva, M., and Berant, J. (2018). Emergence of communication in an interactive world with consistent speakers. In *Emergent Communication Workshop: NeurIPS 2018*.

Bouchacourt, D. and Baroni, M. (2018). How agents see things: On visual representations in an emergent language game. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 981–985. Association for Computational Linguistics.

Cao, K., Lazaridou, A., Lanctot, M., Leibo, J. Z., Tuyls, K., and Clark, S. (2018). Emergent communication through negotiation. In *6th International Conference on Learning Representations (ICLR 2018)*, pages 1–15.

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

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., and Piot, B. (2022). Emergent communication at scale. In *10th International Conference on Learning Representations (ICLR 2022)*, pages 1–30.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., and Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, 41(10):4915–4928.

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

Das, A., Kottur, S., Moura, J. M. F., Lee, S., and Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In Cucchiara, R., Matsushita, Y., Sebe, N., and Soatto, S., editors, *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2951–2960. IEEE Computer Society.

Doumen, J., Beuls, K., and Van Eecke, P. (2023). Modelling language acquisition through syntactico-semantic pattern finding. In Vlachos, A. and Augenstein, I., editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1317–1327. Association for Computational Linguistics.

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

Foerster, J., Assael, I. A., de Freitas, N., and Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In Lee, D., Sugiyama, M., Von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 2137–2145, Red Hook, NY, USA. Curran Associates Inc.

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135.

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

Havrylov, S. and Titov, I. (2017). Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 2146–2156, Red Hook, NY, USA. Curran Associates Inc.

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

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

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (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. IEEE Computer Society.

Kim, J. and Oh, A. (2021). Emergent communication under varying sizes and connectivities. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 17579–17591, Red Hook, NY, USA. Curran Associates Inc.

Kim, N. and Linzen, T. (2020). COGS: A compositional generalization challenge based on semantic interpretation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105. Association for Computational Linguistics.

Kottur, S., Moura, J., Lee, S., and Batra, D. (2017). Natural language does not emerge 'naturally' in multi-agent dialog. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967. Association for Computational Linguistics.

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

Loetzsch, M. (2015). *Lexicon formation in autonomous robots*. PhD thesis, Humboldt-Universität zu Berlin, Berlin, Germany.

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

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Bower, G. H., editor, *Psychology of Learning and Motivation 24*, pages 109–165. Academic Press.

Mordatch, I. and Abbeel, P. (2018). Emergence of grounded compositional language in multi-agent populations. In McIlraith, S. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 1495–1502. AAAI Press.

Nevens, J., Doumen, J., Van Eecke, P., and Beuls, K. (2022). Language acquisition through intention reading and pattern finding. In Calzolari, N. and Huang, C.-R., editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25. International Committee on Computational Linguistics.

Nevens, J., Van Eecke, P., and Beuls, K. (2019). A practical guide to studying emergent communication through grounded language games. In *AISB 2019 Symposium on Language Learning for Artificial Agents*, pages 1–8. AISB.

Nevens, J., Van Eecke, P., and Beuls, K. (2020). From continuous observations to symbolic concepts: A discrimination-based strategy for grounded concept learning. *Frontiers in Robotics and AI*, 7(84).

Noukhovitch, M., LaCroix, T., Lazaridou, A., and Courville, A. (2021). Emergent communication under competition. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 974–982.

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

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

Resnick, C., Kulikov, I., Cho, K., and Weston, J. (2017). Vehicle communication strategies for simulated highway driving. In *Emergent Communication Workshop: NeurIPS 2017*.

Schleicher, A. (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.

Spranger, M. (2013). Grounded lexicon acquisition - case studies in spatial language. In *Proceedings of the 2013 IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, pages 1–6. IEEE.

Spranger, M. (2016). *The evolution of grounded spatial language*. Language Science Press, Berlin, Germany.

Spranger, M. and Beuls, K. (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.

Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial Life*, 2(3):319–332.

Steels, L. (2003). The evolution of communication systems by adaptive agents. In Alonso, E., Kudenko, D., and Kazakov, D., editors, *Symposium on Adaptive Agents and Multi-Agent Systems*, pages 125–140.

Steels, L. (2015). *The Talking Heads experiment: Origins of words and meanings*. Language Science Press, Berlin, Germany.

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

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

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

Steels, L. and Szathmáry, E. (2018). The evolutionary dynamics of language. *Biosystems*, 164:128–137.

Sukhbaatar, S., Szlam, A., and Fergus, R. (2016). Learning multiagent communication with backpropagation. In Lee, D., Sugiyama, M., Von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 2244–2252, Red Hook, NY, USA. Curran Associates Inc.

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

Verheyen, L., Botoko Ekila, J., Nevens, J., Van Eecke, P., and Beuls, K. (2023). Neuro-symbolic procedural semantics for reasoning-intensive visual dialogue tasks. In Gal, K., Nowé, A., Nalepa, G. J., Fairstein, R., and Rădulescu, R., editors, *Proceedings of the 26th European Conference on Artificial Intelligence (ECAI 2023)*, pages 2419–2426, Amsterdam, Netherlands. IOS Press.

Vogt, P. (2005). The emergence of compositional structures in perceptually grounded language games. *Artificial intelligence*, 167(1–2):206–242.

Vogt, P. (2015). *How mobile robots can self-organise a vocabulary*. Language Science Press, Berlin, Germany.

Wang, S. I., Liang, P., and Manning, C. D. (2016). Learning language games through interaction. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2368–2378. Association for Computational Linguistics.

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

Wellens, P. (2012). *Adaptive Strategies in the Emergence of Lexical Systems*. PhD thesis, Vrije Universiteit Brussel, Brussels: VUB Press.

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