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

Jérôme Botoko Ekila[†] **Lara Verheyen**[†] **Jens Nevens**[†]
Katrien Beuls[◇*] **Paul Van Eecke**[†*]

[†] Artificial Intelligence Laboratory, Vrije Universiteit Brussel, Belgium
{jerome, lara.verheyen, jens, paul}@ai.vub.ac.be
[◇] Faculté d'informatique, Université de Namur, Belgium
katrien.beuls@unamur.be

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 37 publicly available datasets. Through a range of experiments, we 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).

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 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, 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

[*] Joint last authors.
This paper was conceived and written without the use of generative AI tools.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 2234–2257, 2026. https://doi.org/10.1162/TACL.a.799
Action Editor: Daniel Fried. Submission batch: 4/2026; Revision batch: 5/2026; Published 8/2026.

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 converge 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 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 method and experiments presented in this paper primarily distinguish themselves from earlier work through their focus on how local communicative interactions in fully decentralised populations can lead to emergent global conventions that are not only effective at solving their communicative task, but, crucially, are also robust against perceptual variation, morphological differences between agents, and changing environmental conditions. The evaluation of the method on a diverse selection of 37 publicly available datasets, ranging from physico-chemical analyses to real-world images, confirms at scale the generality of the method across environments. Both the learning process and the emergent languages are to a large extent transparent and human-interpretable, allowing the human experimenter to trace the gradual formation and alignment of conceptual systems, including the formation of niches that structure the conceptual space.[1]

2 Background and Related Work

Usage-based linguistics and constructionist approaches to language. From a usage-based perspective, linguistic knowledge is viewed as emerging from situated communicative interactions and grounded in the learner's experience (Tomasello, 2003; Bybee, 2010; Beuls and Van Eecke, 2024). In constructionist theories, such knowledge is formalised in terms of constructions, i.e. pairings of form and meaning, shaped by usage (Fillmore et al., 1988; Goldberg, 1995, 2006; Croft, 2001; Langacker, 2000; Traugott and Trousdale, 2013). Our experiments reflect this perspective in that agents start without a predefined lexicon or ontology and progressively develop grounded form-meaning pairings through interaction.

Models of language evolution. Computational models of language evolution have been used to study how communication systems can emerge and evolve. Early work has analysed signalling systems as equilibria in coordination games (Lewis, 1969; Skyrms, 2010). Later models studied the emergence of conventions in populations of embodied agents, both in simulation and in robotic systems (Batali, 1998; Cangelosi and Parisi, 1998; Steels, 1996; Kirby, 2001; Baronchelli and Diaz-Guilera, 2012). This line of research treats language as a culturally evolving system shaped by learning and interaction. Within this line of research, *iterated learning* models examine the *vertical transmission* across generations of language learners (Kirby, 2002; Smith et al., 2003; Ren et al., 2020), whereas *language game* models focus on *horizontal transmission*, where linguistic conventions emerge through repeated interactions among agents within a population (Steels, 1996; de Boer, 2001; Oudeyer, 2006; Steels, 2012).

Language game models. Our experiments and methodology are rooted in the experimental framework of language games (Steels, 1996; Nevens et al., 2019). 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) and in domain-specific continuous environments (Steels and Belpaeme, 2005; Bleys, 2015; Spranger and Beuls, 2016). Further experiments have also modelled the grounding of predefined concepts in perceptual data (Spranger and Beuls, 2016; Wang et al., 2016; Nevens et al., 2020). A key limitation of prior language game experiments concerns the reliance on strict assumptions about perceptual inputs, which has constrained the scale and complexity of the experimental set-ups and has largely confined this line of work to relatively simple, perceptual domains, such as colour or spatial categories.

[1]The source code for the experiments is publicly available at https://gitlab.ai.vub.ac.be/ehai/self-org-at-scale-tacl.

Neural emergent communication. Models of emergent communication with agents modelled as neural networks trained with reinforcement learning have addressed some of the limitations in scale and complexity of earlier language evolution experiments (Foerster et al., 2016; Havrylov and Titov, 2017; Kottur et al., 2017; Mordatch and Abbeel, 2018; Kharitonov et al., 2019; Noukhovitch et al., 2021; Chaabouni et al., 2022). These models have studied a variety of tasks, including reference games, in more realistic high-dimensional perceptual domains such as real-world images (Lazaridou et al., 2017).

A central question concerns the conditions under which robust and generalisable communication systems emerge. One such condition is the presence of populations rather than isolated agent pairs. Several experiments have examined how population size and interaction structure affect emergent languages (Sukhbaatar et al., 2016; Das et al., 2019; Raviv et al., 2019; Kim and Oh, 2021; Chaabouni et al., 2022; Michel et al., 2023; Rita et al., 2022; Lian et al., 2024; Lee, 2024), though the experimental findings remain mixed and dependent on task design and agent assumptions.

Another key factor is whether agents are restricted to fixed communicative roles. Most work assumes a division between speakers and listeners, which does not only depart from the interchangeability of speaker and listener roles in human language (Hockett, 1960), but also from the view of communication as a fundamentally interactive, bidirectional process of joint action (Clark, 1996). While some experiments have equipped agents with both production and interpretation capabilities (Choi et al., 2018; Portelance et al., 2021; Nikolaus, 2024; Wolff et al., 2024), these are typically restricted to two-agent set-ups. Two notable exceptions have explored bidirectional communication in populations (Cao et al., 2018; Graesser et al., 2019), though they differ in focus: the former focuses on a competitive setting (negotiation) rather than a cooperative one (reference), while the latter involves reference but agents were exposed to a supervised signal.

Finally, much of the neural literature has investigated a particular set-up of the reference game, in which a speaker describes a target entity and a listener selects it from a set of 'distractors' based on the utterance (e.g. Havrylov and Titov, 2017; Portelance et al., 2021; Chaabouni et al., 2022). Although this setting has proven effective for studying the emergence of language, it abstracts away from the broader situational context in which communication occurs. Recent work has shown that incorporating contextual information improves communicative efficiency (Gualdoni et al., 2024; Główka et al., 2024; Zhang et al., 2025) as it allows speakers to choose concepts with greater granularity (Ohmer et al., 2022; Kobrock et al., 2024). However, these experiments have so far only examined two-agent set-ups.

3 Problem Definition

We address a decentralised, multi-agent emergent communication problem in which agents must bootstrap a linguistic convention that allows them to draw each other's attention to arbitrary entities in their environment. Importantly, communicative interactions take place locally between pairs of agents, agents can both act as speakers and listeners, the environment does not come pre-categorised, and the emergent convention needs to be suitable for communication about previously unseen entities. The problem definition thereby brings together a number circumstances under which human language emerge and evolve, which have often been studied in isolation in prior work. More formally, the problem can be defined as follows:

Population. There exists a population $P = \{a_1, \ldots, a_n\}$ that consists of N 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 $W = \{e_1, \ldots, e_K\}$ that consists in a set of K entities. An observation of an entity by an agent a takes the form of a feature vector X_a of d 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. Importantly, every agent perceives the environment through its own sensors, so it cannot be assumed that all agents perceive a given entity identically or even represent it as a vector of the same dimensionality.

Interactions. Agents take part in a sequence of task-oriented communicative interactions. At the beginning of each interaction, a scene $C = \{e_1, \ldots, e_k\} \subset W$ of k entities from the world is

randomly created. Two agents are randomly selected from P. One is assigned the role of speaker (S) and the other the role of listener (L). 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. 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 is deliberately generic and can be straightforwardly instantiated in a variety of scenarios. For example, in a robotic scenario, agents may be equipped with sensors, in which case the perceived feature vector corresponds to sensor readings for a given entity. In simulated settings, scenarios could 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. In visual domains, agents may communicate about high-dimensional perceptual inputs derived from real-world images, where each image is treated as a single entity.

4 Methodology

The methodology that we present is grounded in usage-based and constructionist theories of language, in which linguistic structure emerges through situated communicative interactions and is shaped by usage (Tomasello, 2003; Beuls and Van Eecke, 2024). From this perspective, referential communication is a collaborative process: interlocutors coordinate on an intended referent in context (Clark and Wilkes-Gibbs, 1986; Clark, 1996; Pickering and Garrod, 2004). This coordination relies on the gradual accumulation of knowledge in the form of constructions, that is, form-meaning pairings, with each individual building up their own inventory over time. When existing constructions do not suffice, speakers may introduce new ones to resolve communicative impasses, which through repeated use become progressively more conventionalised and entrenched (Bybee, 1998, 2010; Langacker, 2000). Our methodology computationally operationalises these general principles with regard to the problem definition, building on prior language game research (Steels, 1995; Wellens et al., 2008; Nevens et al., 2020).

Linguistic inventory. Each agent $a \in P$ has its own linguistic inventory I_a, which is initially empty. Every word $w \in I_a$ is a triple $w = (f, c, s)$ consisting of a word form $f \in F$, a concept representation c and an entrenchment score $s \in [0, 1]$. F is an unbounded set of possible atomic word forms. In the experiments, forms are generated as atomic three-syllable consonant-vowel strings (e.g. "*demoxu*"). This generation procedure is arbitrary and chosen for ease of inspection. Word forms on their own thus carry no semantic information and play no role in the learning dynamics.

Concept representations. A concept representation $c = ((\omega_1, \theta_1) \dots (\omega_d, \theta_d))$ consists of a pair of a weight ω and a distribution θ for each dimension of the feature vector X perceived by the agent. For continuous dimensions, θ_i is a normal distribution parametrised by (μ_i, σ_i). For categorical dimensions, $\theta_i = \mathrm{f}_i$ is a count vector over the k possible categories, where $\mathrm{f}_{i,j}$ denotes the absolute frequency of category j. The weight ω_i represents the importance of dimension i for the concept. To evaluate how well a concept applies to a perceived entity, the similarity between a concept c and $X = (x_1, \dots, x_d)$ is computed dimension-wise and aggregated with normalised weights:

$$\mathrm{sim}_{\mathrm{c-e}}(c, X) = \sum_{i=1}^{d} \frac{\omega_i}{\sum_{k=1}^{d} \omega_k} \mathrm{sim}_{\mathrm{d}}(\theta_i, x_i) \quad (1)$$

$$\mathrm{sim}_{\mathrm{d}}(\theta_i, x_i) = \begin{cases} \exp\left(-\left|\frac{x_i - \mu_i}{\sigma_i}\right|\right) & \text{if continuous dim.} \\ \frac{\mathrm{f}_{i,x_i}}{\sum_{j=1}^{k} \mathrm{f}_{i,j}} & \text{if categorical dim.} \end{cases} \quad (2)$$

Normalising the weights avoids a bias towards concepts that distribute relevance across more dimensions.

Concepts therefore define prototypical categories grounded in the agent's own perceptual space (Rosch, 1973) and are iteratively refined as agents take part in communicative interactions. These representations serve as the basis for producing and interpreting utterances.

Conceptualisation and production. Given a scene C, a topic $T \in C$ and speaker S, S first retains only the words whose concepts discriminate T from all other entities in C:

$$K = \{w_i \in I_S \mid \text{sim}_{\text{c–e}}(c_i, T) > \max_{e \in C \setminus T} \text{sim}_{\text{c–e}}(c_i, e)\} \tag{3}$$

If $K \neq \varnothing$, the speaker S selects the candidate w^* with the highest *communicative adequacy*, defined as the product of its current score s and its *discriminative power*, i.e. its ability to discriminate T from the other entities in C:

$$w^* = \underset{w_i \in K}{\text{argmax}}\, s_i \cdot \left[\text{sim}_{\text{c–e}}(c_i, T) - \max_{e \in C \setminus T} \text{sim}_{\text{c–e}}(c_i, e)\right] \tag{4}$$

The word form of w^* is uttered by S as U to the listener L.

Invention. If $K = \varnothing$, the speaker S faces a communicative impasse and invents a new word $w = (f, c, s)$. The new form f is sampled from F, the concept c is initialised from the speaker's percept X_S of the topic, and the word is added to I_S. The form f is uttered as U.

Whenever an agent a creates a new word $w = (f, c, s)$ in its inventory I_a from a perceptual vector X_a, the concept c is initialised as follows. For continuous dimensions, μ_i is set to X_i and σ_i to the default value σ_{ini}. For categorical dimensions, the observed category in X_i is assigned a frequency of 1. All weights and the word score are initialised to the fixed values ω_{ini} and s_{ini}.

Comprehension and interpretation. Upon observing U, the listener checks whether it knows a word with that form. If so, let $w = (U, c, s) \in I_L$, L identifies the entity in the scene $e \in C$ that is most similar to c as the hypothesised topic T^*:

$$T^* = \underset{e_i \in C}{\text{argmax}}\, \text{sim}_{\text{c–e}}(c, e_i) \tag{5}$$

If $T^* = T$, the interaction is considered successful. Otherwise it fails, and T is disclosed to the listener 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. Successful use reinforces the selected word and inhibits competing alternatives. Let $w_U = (U, c_U, s)$ be the selected word. This word receives a fixed reward s_r, while competing candidates, i.e. the words in the candidate set K excluding w_U, are penalised in proportion to their similarity to c_U:

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

Thus, words that are more similar are considered stronger competitors and are punished harder. The similarity between two concept representations ($\text{sim}_{\text{c–c}}$), where D_f is the f-divergence between the two parametrised distributions (Hellinger, 1909), is defined as follows:

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

The listener L applies the same score update in its own inventory I_L, by collecting all words in I_L that it would consider candidates and constructs its own set K based on its own perceptual view of the scene.

After a successful interaction, both agents also update the concept c associated with U. For continuous dimensions, μ_i and σ_i are updated online using Welford's algorithm (Welford, 1962). For categorical dimensions, the observed category count is incremented. Dimension weights are updated according to discriminative power in the current scene: dimensions with positive discriminative power are rewarded by a fixed step ω_r on a sigmoid function, while the remaining dimensions are decreased by a fixed step ω_p on the same function. This bounds weights between 0 and 1, with weights 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 U, 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 relative to C in the same way as if the interaction had 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 c initialised based on the listener's percept X_L of the disclosed topic using the word initialisation procedure (see Invention).

Multi-word utterances. The methodology can be generalised to multi-word utterances. Instead of requiring a single word to uniquely discriminate the topic T in scene C, the speaker maintains an entity set E, initially equal to C, and iteratively selects words that reduce E while preserving T. At each step, candidate words are evaluated with respect to the current set E. For a candidate word w_i, selecting it updates $E \leftarrow \{e \in E \mid \mathrm{sim}_{\text{c–e}}(c_i, e) \geq \mathrm{sim}_{\text{c–e}}(c_i, T)\}$. Candidate words are ranked first by the reduction they achieve on E and, in case of ties, by communicative adequacy as defined in Equation 4. The procedure terminates when E contains only T or when the maximum utterance length ℓ is reached. The resulting utterance is the sequence of selected word forms. When $\ell = 1$, this reduces to the single-word utterance case. When $\ell > 1$, the speaker can combine several partially discriminative words into a compositional utterance. On the listener side, comprehension mirrors production by incrementally reducing the candidate set as words are processed sequentially. Alignment updates are applied stepwise: for each selected word, competitors are the words that achieve the same reduction of the current set E, and score updates are computed exactly as in the single-word case with respect to that step-specific competitor set.

5 Experimental setup

We instantiate the problem definition in 37 scenarios and highlight seven prototypical scenarios differing in feature type: continuous (CLEVR, WINE, MSCOCO, CELEBA, BIRDS), categorical (MUSHROOMS) and a combination of both (EXOPLANETS). The CLEVR scenario uses visual features extracted from images in the CLEVR dataset (Johnson et al., 2017). WINE relies on physico-chemical measurements of wine samples (Cortez et al., 2009). EXOPLANETS combines numerical and categorical features of exoplanets (Mishra, 2023). MUSHROOMS consists of categorical features describing mushroom attributes (Schlimmer, 1981). The three real-world images scenarios are MSCOCO (Lin et al., 2014), CELEBA (Liu et al., 2015) and BIRDS (Wah et al., 2011). Following common practice in the literature (Lazaridou et al., 2017; Chaabouni et al., 2022), each image is processed by a pre-trained vision encoder. In line with Kouwenhoven et al. (2024), we use DINOv2 (Oquab et al., 2024), yielding a dense vector representation that serves as the agent's perceptual input. Full processing details and dataset specifications for all 37 scenarios are provided in Appendix B.

Param.	Value	Description
n	10	# agents in population
k	10	# entities in scene
s_{ini}	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_{ini}	0.01	initial standard deviation
ω_{ini}	0.5	initial dimension weight
ω_r	$+1$	dimension weight reward
ω_p	-5	dimension weight punishment

Table 1: Overview of parameter settings.

In each scenario, the world W is defined as the set of entries from the underlying dataset. We hold out 25% of the entities in W 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 (unseen) test entities. The exception to this rule is CLEVR, where the original dataset splits are used, holding scenes of 3 to 10 entities.

In each experiment, unless specified otherwise, we train a population of 10 agents for 1M pairwise interactions on the training scenes and evaluate on 100K interactions on the testing scenes. Results are averaged over 10 independent runs. The chosen hyperparameters are shown in Table 1. These values have only been tuned on the CLEVR dataset and are used on all 37 datasets with no further fine-tuning.

In line with common practice in the field (Steels, 1999; Loetzsch, 2015), 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. It serves as the feedback signal through which agents reinforce or adjust their linguistic structures.

Dataset	# ent.	# cont.	# cat.	succ. (%)	conv. (%)	inv. size	inv. size (95%)
CLEVR	468K	20	0	$99.76_{\pm 0.08}$	$94.41_{\pm 1.46}$	$52.63_{\pm 1.73}$	$31.14_{\pm 1.11}$
WINE	5K	12	0	$99.76_{\pm 0.17}$	$88.34_{\pm 1.31}$	$77.12_{\pm 1.21}$	$60.25_{\pm 1.37}$
EXOPLANETS	5K	8	4	$99.67_{\pm 0.10}$	$92.30_{\pm 0.86}$	$79.35_{\pm 1.18}$	$54.47_{\pm 1.67}$
MUSHROOMS	8K	0	23	$98.10_{\pm 0.31}$	$86.55_{\pm 1.82}$	$265.40_{\pm 6.77}$	$76.32_{\pm 2.05}$
MSCOCO	159K	384	0	$95.77_{\pm 1.02}$	$95.81_{\pm 0.74}$	$17.63_{\pm 0.15}$	$14.67_{\pm 0.58}$
CELEBA	203K	384	0	$96.66_{\pm 1.88}$	$95.60_{\pm 1.58}$	$19.93_{\pm 0.92}$	$16.20_{\pm 1.04}$
BIRDS	12K	384	0	$95.88_{\pm 3.11}$	$96.11_{\pm 1.29}$	$19.50_{\pm 2.71}$	$15.00_{\pm 2.00}$

Table 2: Experimental results on the test sets of the seven featured scenarios after 1M training interactions. Mean and ± 2 standard deviations computed over 10 independent runs. Columns describe the dataset, number of entities, number of continuous and categorical dimensions, communicative success, conventionality and linguistic inventory size. Results for the remaining 30 datasets are included in Appendix C.

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 the listener had been the speaker. In prior work, this metric has been referred to as *lexical coherence* (Steels, 2015). Other formulations for this metric are *message agreement* (Kim and Oh, 2021) and *speaker synchronisation* (Rita et al., 2022).

Linguistic inventory size. The average linguistic inventory size (*inv. size*) is calculated as the average number of distinct words uttered by the agents. To estimate the most frequently used words, we quantify the number of distinct word forms required to account for 95% of all produced utterances (*inv. size (95%)*).

6 Results

The experiments address two central questions: whether the proposed methodology scales across diverse perceptual domains, and whether the resulting conventions remain effective when perception differs across agents. We first establish performance in homogeneous populations, then analyse the dynamics of the emergent conventions, before turning to experiments involving perceptual heterogeneity and multi-word utterances.

6.1 Reference in homogeneous populations

We first evaluate the methodology in a base setting with homogeneous populations, assessing its performance across the 37 datasets. Consistent with prior work, we assess zero-shot generalisation to unseen (in-distribution) samples by evaluating the resulting convention on entities not seen during training (Kottur et al., 2017; Lazaridou et al., 2018). Table 2 reports test set performance on the seven featured scenarios in terms of communicative success, conventionality and linguistic inventory size (including the metric to track the most frequently used words). In each scenario, the population reaches a degree of communicative success of over 98%, with a degree of conventionality above 86%. The average linguistic inventory size ranges from 17 to 265 words. A core of 14 to 76 words covers 95% of usage, indicating a heavy-tailed distribution. The evaluation results on the remaining 30 datasets are included as supplementary materials to this paper, and are in line with those obtained in the seven featured scenarios. 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.

Generalisability. Next, we assess 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 *generalisability* (Boldt and Mortensen, 2024; Lazaridou et al., 2018; Lee, 2024). This experiment can be interpreted as a controlled test of whether the convention remains robust under a shift in the environment, namely in the distribution of attribute combinations. We apply the methodology to the

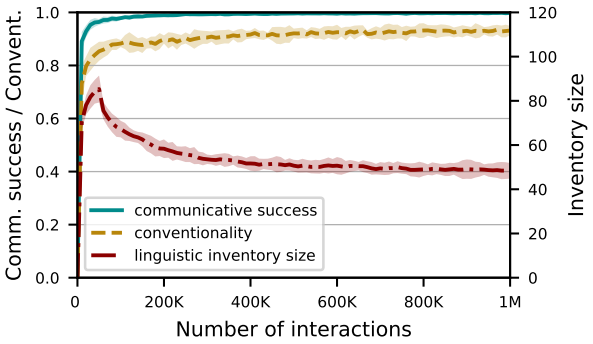


Figure 1: Evolutionary dynamics during the training phase of the CLEVR experiment with 10 agents. Mean shown as a line, shaded region indicates ±2 standard deviations over 10 runs.

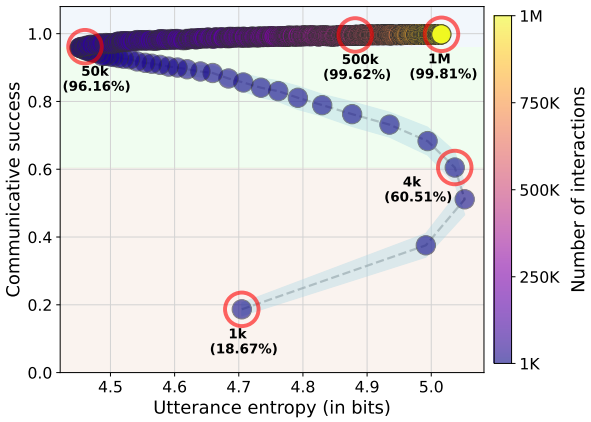


Figure 2: Communicative success plotted against utterance entropy $H(U)$ over training. Every circle aggregates 1000 interactions and colour encodes training progress.

CLEVR CoGenT dataset (Johnson et al., 2017), which was especially designed to test the robustness of intelligent systems against correlations that occur at test time but not during training. As such, a number of biases are built into the scenes. For example, in the training scenes, cubes are always grey, blue, brown or yellow. Test set A contains scenes that are subject to the same correlations, whereas test set B consists of scenes that are subject to different correlations. Test set B can be used to assess whether the learnt model generalises beyond the correlations that characterise the training set. The results show that the performance of the agents on test set B closely matches that on test set A, with a communicative success of 99.75% and 99.78% respectively. The 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.

Evolutionary dynamics. The evolutionary dynamics that take place during the training phase of the 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. Communicative success rises to about 96% after 50K interactions, and continues to grow to 99.81% 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). 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.

Evolution of an emergent inventory. We examine how the word usage frequency distribution evolves over time. Figure 2 visualises the evolution of communicative success against utterance entropy in the CLEVR scenario. Utterance entropy $H(U)$ is computed as the Shannon entropy of the empirical distribution over the agents' utterances. We identify three phases. In an early invention phase, entropy grows fast as agents invent and test out novel words. During an alignment phase, success rises sharply and entropy drops as agents converge on broadly effective but still coarse concept representations. Finally, a slow refinement phase follows, in which entropy rises again. In this final phase, few new words emerge, rather existing ones are gradually tuned or repurposed to fill remaining conceptual gaps. This phase progressively slows down as further gains in communicative success become marginal.

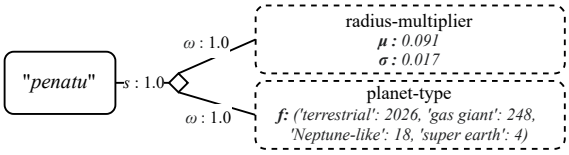


Figure 3: Example of a word emerged in the EXOPLANETS scenario. It has specialised towards two dimensions, one continuous ('radius-multiplier') and the other categorical ('planet-type').

Effect of the population size on convergence. We test how the approach scales with varying population sizes. Following Chaabouni et al. (2022), we test populations of up to 100 agents trained for 1 to 10M interactions on the CLEVR scenario (see detailed results in Appendix D). We observe similar dynamics as in Figure 1, but due to the number of interactions per agent decreasing as population size increases, it takes more interactions for larger population sizes to converge on an effective convention. However, the size of this convention (as measured by the linguistic inventory size) is smaller. We hypothesise that larger populations introduce more variants, allowing selection to favour the most effective variants. Similar findings were reported by Lian et al. (2024), who found that smaller populations tend to settle on effective yet sub-optimal, redundant conventions.

Example of an emerged word. Figure 3 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* (e.g. 'terrestrial'). 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.

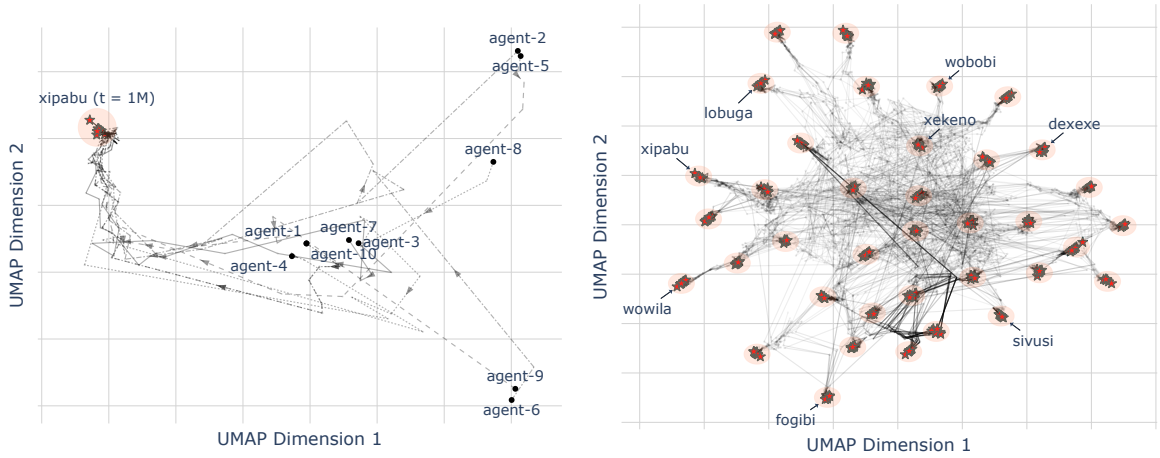
Trajectory of an emergent word. Figure 4 visualises the trajectories that words follow as they are shaped during training. Each word is represented at each time step by the concatenation of its parameters (score, weights and distributional parameters). These representations are projected in two dimensions using the Aligned-UMAP technique for temporal data (McInnes et al., 2018). Subfigure 4a shows the trajectory of the concept representation associated with the word "*xipabu*" in an experiment with 10 agents on the CLEVR scenario. Initially, the concept representations across the 10 agents are very different, as each was learnt locally from a specific interaction. Over time, the representations of the different agents align as a result of the evolutionary dynamics that take place. Subfigure 4b shows the trajectories of all words in the final linguistic inventories of the 10 agents. Figure 4 not only shows the alignment of concept representations but also the formation of niches that structure the conceptual space. These niches arise as competing words progressively differentiate in meaning as they are used, very much in the spirit of Bréal (1897)'s laws of *spécialité* and *répartition*.

6.2 Robustness to perceptual heterogeneity

So far, the experiments have considered homogeneous populations of agents where agents perceive entities through identical input feature spaces. These settings however abstract away from an important property of grounded communication, namely that meanings are tied to each agent's own embodiment. We therefore now turn to conditions under which agents do not perceive entities identically, testing whether successful conventions can still emerge when perception varies across agents or changes over time.

Uncalibrated sensors and noisy environments. The first two experiments assess the robustness of the methodology against differences in the agents' perception of the world (i.e. $X_S \neq X_L$). As shown in Chaabouni et al. (2022), even mild input noise, typically simulated by injecting Gaussian noise into agents' inputs, strongly reduces communicative success. In our setup, we simulate two distinct sources of perceptual variation. In a first scenario, a lack of calibration (Ca) is simulated by shifting X_a by a value that is individually sampled at the beginning of each experiment for each sensor of each agent from a normal distribution with a mean of 0 and a standard deviation of 0.001 or 0.01 (Ca1 and Ca2 conditions). In a second scenario, noise (No) is added to the sensor values by shifting X_a by a value that is independently sampled at the beginning of each interaction for each sensor of each agent from a normal distribution with a mean of 0 and a standard deviation of 0.001 or 0.01 (No1 and No2 conditions).

As seen in Table 3, a lack of calibration minimally impacts the emergent conventions. Adding random sensor noise does lead to less conventional



(a) Single word trajectory, 10 agents

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

Figure 4: Aligned-UMAP visualisations of the trajectories of concept representations over time.

languages, but the communication does not break down even under substantial amounts of noise.

Test	succ. (%)	conv. (%)	inv. size
Baseline			
-	$99.76_{\pm 0.08}$	$94.41_{\pm 1.46}$	$52.63_{\pm 1.73}$
Uncalibrated sensors and noisy environments			
Ca1	$99.77_{\pm 0.11}$	$93.05_{\pm 2.13}$	$54.45_{\pm 2.33}$
Ca2	$99.74_{\pm 0.10}$	$92.41_{\pm 1.28}$	$54.95_{\pm 1.50}$
No1	$98.81_{\pm 0.52}$	$81.08_{\pm 3.69}$	$54.69_{\pm 1.21}$
No2	$87.67_{\pm 0.91}$	$42.55_{\pm 2.67}$	$57.29_{\pm 2.20}$
Heteromorphic populations			
Ho1	$99.80_{\pm 0.07}$	$93.41_{\pm 1.08}$	$52.59_{\pm 2.09}$
He1	$98.12_{\pm 1.60}$	$87.42_{\pm 2.40}$	$53.73_{\pm 2.24}$
Ho2	$99.72_{\pm 0.16}$	$92.39_{\pm 2.72}$	$53.42_{\pm 3.06}$
He2	$83.47_{\pm 11.07}$	$57.03_{\pm 17.61}$	$65.49_{\pm 3.02}$
Robustness against sensor defects			
De1	$99.22_{\pm 1.05}$	$90.57_{\pm 2.60}$	$56.53_{\pm 1.80}$
De2	$93.82_{\pm 12.39}$	$77.43_{\pm 30.32}$	$57.90_{\pm 1.67}$

Table 3: Results of the robustness experiments in the CLEVR scenario. The results on three additional featured scenarios are included in Appendix E.

Heteromorphic populations. The next experiment assesses the applicability of the methodology to heteromorphic populations. Previous work has explored heterogeneity through differences in agents' internal design, such as architectural or learning asymmetries (Rita et al., 2022; Mahaut et al., 2025). Here, we focus on perceptual heterogeneity as in de Greeff and Belpaeme (2011) and Kim and Oh (2021), testing whether a convention can still emerge when agents differ in what they can sense. For this purpose, we set up variations on the first four featured scenarios in which each individual agent has access to a randomly selected subset of the d dimensions. Concretely, for every dataset, we run two instances of the experiment in which the agents are respectively endowed with combinations of $d - 1$ and $\frac{d}{2}$ randomly selected sensors (He1 and He2 conditions). 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 (Ho1 and Ho2).

As seen in Table 3, when moving from the homomorphic to the heteromorphic setting, we observe a similar trend as the one observed with the perceptual deviation experiments. While the conventionality of the language decreases, as agents will tend to use words that optimally fit their own sensory apparatus, they are still able to achieve a high degree of communicative success, even when equipped with a substantially different set of sensors.

Robustness against sensor defects. The last experiment validates the robustness of the methodology against sensor defects that occur in individual agents. We run a version of the experiment in which the agents suffer from a sudden malfunction after 500,000 interactions. Concretely, all agents suddenly lose access to one or half of their sensors (De1 and De2 conditions). The exact sensors that break down are randomly selected for each individual agent in each experimental run.

As seen in Table 3, the results show that the emergent languages become less conventional as the agents lose access to more sensors, but they still achieve high levels of communicative success. Using the He1 and He2 experiments as a basis for comparison, the experiment also demonstrates that the emergence of an effective linguistic convention prior to the malfunction remains beneficial even in the long term.

6.3 Comparison with baseline and prior work

To contextualise the proposed method, we compare it against a diagnostic baseline and the closest related approach from the literature. The baseline isolates the role of adaptive, interaction-driven concept formation.

Clustering baseline. As a diagnostic baseline, we combine k-means clustering (Lloyd, 1982) with a standard naming game (Steels, 1995). This baseline replaces adaptive, interaction-driven concept formation with a fixed discretisation of the perceptual space: each agent partitions the space into a set of categories learned from the training data, after which agents negotiate labels over these categories through a naming game. We evaluate several values of $k \in \{10, 50, 100, 250, 500\}$. Performance improves as k increases, with gains becoming marginal beyond $k = 100$. We report results with $k = 500$ as it yields the best performance. Full results are provided in Appendix F.

In homogeneous populations, this baseline provides a useful reference point, as clustering yields a nearly shared discretisation of the perceptual space and the task reduces to negotiating labels over shared categories. In this setting, the coordination problem is therefore relatively straightforward and achieves 97% success on the unseen test set. In contrast, when agents differ in their perceptions (i.e. conditions Ca1, Ca2, No1, No2, He1, He2, De1, De2 from Section 6.2), the learned clusters no longer align across agents. As a result, the labels cease to be as effective and communicative success collapses on the held-out test sets, ranging from 0.13% to 43%.

Comparison with neural approaches. To situate our methodology with respect to recent approaches to emergent communication on large-scale and high-dimensional inputs, we compare it with the closest comparable setup from the literature. As discussed in Section 2, most emergent communication experiments adopt a different game formulation, where the speaker observes only the topic while the listener observes the full scene. In some works, this setup has been referred to as the "*discrimination game*" (see e.g. Dessì et al., 2021; Chaabouni et al., 2021, 2022; Ben Zion et al., 2024). In contrast, we consider a situated setting in which both agents share the same context, as defined in Section 3. Only a small number of recent works study reference under such conditions, and these are typically restricted to two-agent setups and small synthetic datasets (Ohmer et al., 2022; Główka et al., 2024; Kobrock et al., 2024).

We therefore compare our methodology against Gualdoni et al. (2024), which ground their experiments in large-scale real-world image datasets and thus provide the closest available point of comparison for our approach. Their experiments use the ManyNames dataset (Silberer et al., 2020) which contains 25k real-world images annotated with free naming responses from native English speakers, spanning 442 distinct object names. Crucially, the annotations are never used for training, but serve only as a point of reference for analysing the emerged linguistic conventions. Gualdoni et al. (2024) construct scenes by pairing two entities detected within the same image, yielding pairs that are visually and semantically related. We follow the original processing pipeline of their paper.

Gualdoni et al. (2024) adopts the VQ-VIB framework of Tucker et al. (2022) which formulate the emergence of lexical categories as a multi-objective optimisation problem that trades off utility, informativeness and complexity. In their setting, a purely utility-driven objective leads to large linguistic inventories of ≈ 959 words, with smaller inventories emerging only when the three objectives are carefully balanced. In contrast, our approach is purely utility-driven and fully decentralised. In particular, we do not introduce any term that directly compares agents' internal representations, which is how informativeness is enforced in their framework. Despite this, when evaluated in the same setting, our method achieves higher communicative success on the held-out test set (98.77% vs. 95%) while producing substantially smaller inventories (≈ 408 words). These results suggest that, in a fully situated reference game with shared context, compact and effective conventions can emerge from interaction-level reward alone, without explicit regularisation terms. Finally, we evaluate the VQ-VIB approach in our experimental settings. As this ap-

proach is limited to two agents, we report results on populations of two agents and experiment with contexts of two and ten entities (see detailed results in Appendix F). In the *two agent, two entity* setting, both approaches perform well, achieving 99.97% (ours) and 95.29% (VQ-VIB). However, when the difficulty of the task is increased to contexts with ten entities, the performance of VQ-VIB substantially degrades ($\leq 42\%$), whereas our method maintains at least 82% across all conditions.

6.4 Multi-word utterances

The results on homogeneous and heteromorphic populations in Sections 6.1 and 6.2 demonstrate that the problem can be solved with single-word utterances ($\ell = 1$). When we allow longer utterances by setting $\ell > 1$, a trade-off emerges. Convergence becomes substantially slower: reaching the same communicative success achieved after 1M interactions with $\ell = 1$ requires more than 10M interactions when $\ell = 5$. At the same time, the resulting conventions are more conventional and substantially more compact. Across the seven featured scenarios, for $\ell = 1$, Table 2 shows inventories ranging from 17 to 265 words. With $\ell = 5$, by contrast, the inventories fall to between 17 and 48 words (see Appendix G). Agents converge on substantially smaller inventories because words can be reused across utterances rather than requiring many specialised single-word forms. Despite allowing multi-word utterances, agents still rely overwhelmingly on single-word utterances. During the early stages of training, roughly 80% of interactions involve a single word. After 1M training interactions, the evaluation on the held-out test set shows that 95.39% of utterances consist of a single word, 4.60% contain two words, and fewer than 0.01% contain more than two words.

The strong performance observed with $\ell = 1$ shows that, in the reference task, successful generalisation can be achieved without compositional (multi-word) utterances. This finding is consistent with previous work in the literature where, for example, Chaabouni et al. (2020) demonstrate that in neural agents there is no correlation between the degree of compositionality and the ability to generalise. Conklin and Smith (2023) suggest that this lack of correlation arises because, once a language is sufficiently structured to support generalisation, additional regularity does not necessarily improve performance. In our approach, however, the ability to generalise without multi-word utterances follows from the open-ended nature of the inventories. To resolve communicative impasses, agents reshape existing words or invent new ones, progressively carving up the conceptual space. When restricted to single-word utterances, this pressure to distinguish meanings cannot be realised through combinations of smaller units and must instead be encoded in individual words. As a result, some words come to express combinations of features that would otherwise be distributed across multiple words. Over time, these words stabilise and occupy distinct niches, allowing populations to establish successful conventions without compositional utterances, as different words progressively fill the remaining gaps.

7 Conclusion

This paper has introduced a fully decentralised methodology through which populations of autonomous agents converge on grounded linguistic conventions via local communicative interactions. An extensive evaluation on a diverse selection of 37 datasets, ranging from physico-chemical analyses to real-world images, confirmed its generality across environments. The proposed approach distinguishes itself from prior work by showing that effective global conventions can emerge from purely local interactions in population settings, without relying on shared representations, centralised training signals, perceptual discretisations or optimisation objectives beyond communicative success. It thereby brings together conditions that have typically been studied in isolation, in particular decentralisation, population-level dynamics, continuous perceptual grounding and perceptual heterogeneity. Not only are the resulting conventions highly effective at solving the reference task, they are also robust against perceptual variation, morphological differences between agents, and changing environmental conditions. At the same time, both the learning dynamics and the emergent languages remain transparent and human-interpretable, allowing the progressive structuring and alignment of conceptual spaces to be directly observed. Overall, our findings support the view that key properties of human language, most notably conventionality, robustness and adaptivity, can arise from usage-based mechanisms at the level of the local interaction alone, without the need for shared representations, central control or explicit regularisation objectives.

Acknowledgements

We are grateful to Fabio De Ponte, Liesbet De Vos, Jonas Gillain, Tom Lenaerts, Arno Temmerman, Maxime Toquebiau, Remi van Trijp and Jamie Wright for their insightful comments on an earlier version of this manuscript, as well as to the three anonymous TACL reviewers and the TACL action editor Daniel Fried for their invaluable work and constructive suggestions that have led to a better paper. This research was supported by the F.R.S.-FNRS-FWO WEAVE project HERMES I under grant numbers T002724F (F.R.S.-FNRS) and G0AGU24N (FWO), the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme and the AI Flagship project ARIAC by DigitalWallonia4.ai.

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.

Andrea Baronchelli and Albert Diaz-Guilera. 2012. Consensus in networks of mobile communicating agents. *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, 85(1):016113.

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

John Batali. 1998. Computational simulations of the emergence of grammar. In James R. Hurford, Michael Studdert-Kennedy, and Chris Knight, editors, *Approaches to the evolution of language: social and cognitive bases*, pages 405–426. Cambridge University Press, Cambridge, United Kingdom.

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.

Rotem Ben Zion, Boaz Carmeli, Orr Paradise, and Yonatan Belinkov. 2024. Semantics and spatiality of emergent communication. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 110156–110196, Red Hook, NY, USA. Curran Associates, Inc.

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. 2015. *Language Strategies for the Domain of Colour*. Language Science Press, Berlin, Germany.

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

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

Brendon Boldt and David Mortensen. 2024. A review of the applications of deep learning-based emergent communication. *Transactions on Machine Learning Research*.

Michel Bréal. 1897. *Essai de Sémantique: Science des Significations*. Hachette, Paris, France.

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

Joan Bybee. 1998. The emergent lexicon. In *CLS 34: The Panels*, pages 421–435, Chicago, IL, USA. Chicago Linguistics Society.

Joan Bybee. 2010. *Language, Usage and Cognition*. Cambridge University Press, Cambridge, United Kingdom.

Angelo Cangelosi and Domenico Parisi. 1998. The emergence of a "language" in an evolving population of neural networks. *Connection Science*, 10(2):83–97.

Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z. Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *6th International Conference on Learning Representations (ICLR 2018)*.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442. Association for Computational Linguistics.

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)*.

Edward Choi, Angeliki Lazaridou, and Nando de Freitas. 2018. Compositional obverter communication learning from raw visual input. In *6th International Conference on Learning Representations (ICLR 2018)*.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, United Kingdom.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Henry Conklin and Kenny Smith. 2023. Compositionality with variation reliably emerges in neural networks. In *11th International Conference on Learning Representations (ICLR 2023)*.

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.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford, United Kingdom.

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.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. TarMAC: Targeted multi-agent communication. In *36th International Conference on Machine Learning (ICML 2019)*, pages 1538–1546. PMLR.

Bart de Boer. 2001. *The Origins of Vowel Systems*. Oxford University Press, Oxford, United Kingdom.

Joachim de Greeff and Tony Belpaeme. 2011. The development of shared meaning within different embodiments. In *Proceedings of the 2011 IEEE international conference on development and learning (ICDL)*, volume 2, pages 1–6, New York, NY, USA. IEEE.

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.

Roberto Dessì, Eugene Kharitonov, and Marco Baroni. 2021. Interpretable agent communication from scratch (with a generic visual processor emerging on the side). In *Advances in Neural Information Processing Systems*, pages 26937–26949, Red Hook, NY, USA. Curran Associates Inc.

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. Association for Computational Linguistics.

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.

Charles J. Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64(3):501–538.

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

Jakob Foerster, Yannis M. 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.

Adele E. Goldberg. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press, Chicago, IL, USA.

Adele E. Goldberg. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press, Oxford, United Kingdom.

Laura Harding Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3700–3710. Association for Computational Linguistics.

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

Eleonora Gualdoni, Mycal Tucker, Roger P. Levy, and Noga Zaslavsky. 2024. Bridging semantics and pragmatics in information-theoretic emergent communication. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 21059–21078, Red Hook, NY, USA. Curran Associates, Inc.

Krzysztof Główka, Julian Zubek, and Joanna Rączaszek-Leonardi. 2024. Context-dependent communication under environmental constraints. *Cognitive Systems Research*, 88:101293.

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 D. Kiel, editor, *Knowledge management, organizational intelligence and learning, and complexity. The encyclopedia of life support systems*, pages 253–280. EOLSS Publishers, Oxford, United Kingdom.

Charles F. Hockett. 1960. The origin of speech. *Scientific American*, 203(3):88–97.

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.

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

Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. 2019. EGG: A toolkit for research on emergence of language in games. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 55–60. Association for Computational Linguistics.

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.

Simon Kirby. 2001. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, 5(2):102–110.

Simon Kirby. 2002. Learning, bottlenecks and the evolution of recursive syntax. In Ted Briscoe, editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, pages 173–203. Cambridge University Press, Cambridge, United Kingdom.

Kristina Kobrock, Xenia Isabel Ohmer, Elia Bruni, and Nicole Gotzner. 2024. Context shapes emergent communication about concepts at different levels of abstraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3831–3848. ELRA.

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.

Tom Kouwenhoven, Max Peeperkorn, Bram Van Dijk, and Tessa Verhoef. 2024. The curious case of representational alignment: Unravelling visio-linguistic tasks in emergent communication. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 57–71. Association for Computational Linguistics.

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

Ronald W. Langacker. 2000. A dynamic usage-based model. In Michael Barlow and Suzanne Kemmer, editors, *Usage-Based Models of Language*, pages 1–63. Chicago University Press, Chicago, IL, USA.

Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. 2018. Emergence of linguistic communication from referential games with symbolic and pixel input. In *6th International Conference on Learning Representations (ICLR 2018)*.

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)*.

Heeyoung Lee. 2024. One-to-many communication and compositionality in emergent communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20794–20811. Association for Computational Linguistics.

David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA, USA.

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. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV 2014)*, pages 740–755, Cham, Switzerland. Springer.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the

wild. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE Computer Society.

Stuart Lloyd. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137.

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.

Matéo Mahaut, Francesca Franzon, Robert Dessi, and Marco Baroni. 2025. Referential communication in heterogeneous communities of pre-trained visual deep networks. *Transactions on Machine Learning Research*.

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.

Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, and Angeliki Lazaridou. 2023. Revisiting populations in multi-agent communication. In *11th International Conference on Learning Representations (ICLR 2023)*.

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, DC, 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. International Committee on Computational Linguistics.

Jens Nevens, Paul Van Eecke, and Katrien Beuls. 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.

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).

Mitja Nikolaus. 2024. Emergent communication with conversational repair. In *12th International Conference on Learning Representations (ICLR 2024)*.

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, Richland, SC, USA. IFAAMAS.

Xenia Ohmer, Marko Duda, and Elia Bruni. 2022. Emergence of hierarchical reference systems in multi-agent communication. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5689–5706. International Committee on Computational Linguistics.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. DINOv2: Learning robust visual

features without supervision. *Transactions on Machine Learning Research*.

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.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Eva Portelance, Michael C. Frank, Dan Jurafsky, Alessandro Sordoni, and Romain Laroche. 2021. The emergence of the shape bias results from communicative efficiency. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 607–623. Association for Computational Linguistics.

Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2019. Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, 286(1907):20191262.

Yi Ren, Shangmin Guo, Matthieu Labeau, Shay Cohen, and Simon Kirby. 2020. Compositional languages emerge in a neural iterated learning model. In *8th International Conference on Learning Representations (ICLR 2020)*.

Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. 2022. On the role of population heterogeneity in emergent communication. In *10th International Conference on Learning Representations (ICLR 2022)*.

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

Eleanor H. Rosch. 1973. On the internal structure of perceptual and semantic categories. In Timothy E. Moore, editor, *Cognitive Development and Acquisition of Language*, pages 111–144. Academic Press, San Diego, CA, USA.

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.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905. International Committee on Computational Linguistics.

Brian Skyrms. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press, Oxford, United Kingdom.

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, New York, NY, USA. IEEE.

Kenny Smith, Simon Kirby, and Henry Brighton. 2003. Iterated learning: A framework for the emergence of language. *Artificial Life*, 9(4):371–386.

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

Luc Steels. 1996. Perceptually grounded meaning creation. In *Proceedings of the Second International Conference on Multi-Agent Systems*, pages 338–344, Washington, DC, 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. 2015. *The Talking Heads Experiment: Origins of Words and Meanings*. Language Science Press, Berlin, Germany.

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.

Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, pages 2244–2252, Red Hook, NY, USA. Curran Associates Inc.

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Harvard, MA, USA.

Elizabeth Closs Traugott and Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford University Press, Oxford, United Kingdom.

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

Mycal Tucker, Roger P. Levy, Julie Shah, and Noga Zaslavsky. 2022. Trading off utility, informativeness, and complexity in emergent communication. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 22214–22228, Red Hook, NY, USA. Curran Associates, Inc.

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.

Catherine Wah, Steve Branson, Pete Welinder, Pietro Perona, and Serge Belongie. 2011. Caltech UCSD Birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology. 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, 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.

Cornelius Wolff, Julius Mayer, Elia Bruni, and Xenia Ohmer. 2024. Bidirectional emergent language in situated environments. *arXiv preprint arXiv:2408.14649v2*.

Yuqing Zhang, Ecesu Ürker, Tessa Verhoef, Gemma Boleda, and Arianna Bisazza. 2025. NeLLCom-Lex: A neural-agent framework to study the interplay between lexical systems and language use. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10929–10945. Association for Computational Linguistics.

Appendix

A Hardware, training regime, tuned hyperparameters

All experiments were conducted on 20-core INTEL Xeon Gold 6148 processors, paired with 32GB of RAM. One million sequential interactions (the amount of communicative interactions in each experiment) were executed on this hardware in $\pm$ 6 to 8 hours. Table 4 includes the space of hyperparameters explored for the baseline CLEVR experiment. The best performing set of hyperparameters (in terms of communicative success and linguistic conventionalisation) are reported in the main text (see Table 1 in the main text). Every subsequent experiment uses this same set of parameters.

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

Table 4: Overview of hyperparameter search

B Data processing pipelines

Tabular datasets This paper uses 33 publicly available tabular datasets to validate the methodology. Each tabular dataset stores information in rows and columns, where rows represent entities and columns represent continuous or categorical features. The datasets can be broadly classified into one of three categories: (i) 7 contain only continuous features, (ii) 24 mix continuous and categorical features, (iii) and 2 contain only categorical features. The processing pipeline begins by removing columns containing all missing values and rows with any missing values. Duplicate rows are removed, keeping only the first occurrence. As some datasets represent discrete categorical information as integers, these 'continuous' features are converted to categorical features. Next, all continuous features are normalised. Finally, the datasets are divided into training and test sets using a 75%/25% split.

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

Real-world images We use four computer vision datasets: MSCOCO (v. 2017) (Lin et al., 2014), CELEBA (Liu et al., 2015), Caltech-UCSD BIRDS (Wah et al., 2011) and MANYNAMES (Silberer et al., 2020). MSCOCO (2017) contains over 159k naturalistic images of everyday scenes, CELEBA contains over 202k images of celebrity faces, Caltech-UCSD BIRDS contains approximately 12k images covering 200 bird species and the MANYNAMES dataset consists of 25k real-world images of everyday scenes. For MSCOCO we use the original dataset splits, while for CELEBA and BIRDS we use a 75%/25% train-test split. In contrast to the CLEVR setting, where agents communicate about objects within images, here each image is treated as a single entity and agents play language games about images as wholes. Each image is processed by a pre-trained DinoV2 model (Oquab et al., 2024), yielding a 384-dimensional embedding that serves as the agent's perceptual input. The MANYNAMES dataset is used for comparison with Gualdoni et al. (2024). We therefore follow, in this case, the exact processing pipeline of their paper.

C Results on the other 30 datasets

In Section 6.1 of the main text, the results on seven featured scenarios are presented. In Table 5, we present the full evaluation on the other 30 scenarios.

D Effect of population size on convergence

In Table 7, we test how the approach scales with population size by varying the number of k agents while keeping all other settings fixed (see Table 1 in the main text) in the CLEVR scenario. We test populations of up to 100 agents ($k \in \{2, 10, 25, 50, 100\}$). The population is trained for 1M (but also up to 10M) interactions and evaluated on 100K interactions.

E Experimental results demonstrating robustness of methodology

In Section 6.2 of the main text, four experiments are evaluated on CLEVR to demonstrate the robustness of our methodology (see paragraphs *Uncalibrated sensors and noisy environments*, *Heteromorphic populations*, *Robustness against sensor defects*). In Table 6, we present the results on the other three featured scenarios: WINERY, EXOPLANETS, and MUSHROOMS.

F Multi-word utterances

In Table 8, we test how the results vary when we vary the utterance length to maximally 5 words (i.e. $\ell = 5$). The population is trained for 10M interactions and evaluated for 100K interactions.

G Baseline and comparison with prior work

In Table 9, we evaluate a clustering-based baseline and a recent neural emergent communication approach (VQ-VIB) (Gualdoni et al., 2024) in the WINE scenario. For our approach and the clustering baseline population is trained for 1M interactions and evaluated one 100K interactions. For VQ-VIB, we run the released code of Gualdoni et al. (2024) on the WINE dataset using their best-performing configuration, which optimises solely for utility. We evaluate VQ-VIB with population of two agents, as their framework only supports this setting, and consider scenes containing two and ten entities.

Dataset	# ent.	# cont.	# cat.	succ. (%)	conv. (%)	inv. size	inv. size (95%)
Jadikar (2019)	37K	11	0	$99.66_{\pm 0.12}$	$89.59_{\pm 1.42}$	$76.57_{\pm 1.29}$	$56.18_{\pm 1.87}$
De Vito et al. (2008)	7K	15	0	$99.66_{\pm 0.15}$	$91.00_{\pm 1.07}$	$77.77_{\pm 1.20}$	$58.06_{\pm 2.04}$
Ma (2019)	10K	39	0	$99.53_{\pm 0.12}$	$92.21_{\pm 0.69}$	$86.67_{\pm 2.14}$	$51.40_{\pm 1.67}$
Vijaya et al. (2018)	303	16	0	$99.50_{\pm 0.27}$	$87.74_{\pm 2.66}$	$88.63_{\pm 1.40}$	$56.50_{\pm 1.50}$
Tuameh (2023)	1K	99	0	$99.46_{\pm 0.10}$	$93.46_{\pm 0.96}$	$81.22_{\pm 2.48}$	$41.69_{\pm 1.07}$
Brooks and Pope (1989)	2K	6	0	$99.14_{\pm 0.21}$	$90.11_{\pm 2.01}$	$116.89_{\pm 3.19}$	$59.38_{\pm 1.43}$
Dal Pozzolo et al. (2014)	284K	30	1	$99.77_{\pm 0.07}$	$88.35_{\pm 1.13}$	$73.24_{\pm 1.17}$	$59.02_{\pm 1.47}$
Boksha (2024)	8K	7	1	$99.71_{\pm 0.14}$	$90.35_{\pm 1.14}$	$65.49_{\pm 1.20}$	$49.26_{\pm 1.16}$
Kadiwal (2021)	2K	9	1	$99.69_{\pm 0.15}$	$89.51_{\pm 1.36}$	$66.60_{\pm 1.09}$	$49.99_{\pm 1.70}$
Smith et al. (1988)	768	8	1	$99.65_{\pm 0.11}$	$91.10_{\pm 1.01}$	$82.90_{\pm 1.49}$	$55.93_{\pm 1.35}$
Olteanu (2020)	10K	58	1	$99.64_{\pm 0.12}$	$88.83_{\pm 1.67}$	$74.89_{\pm 1.56}$	$55.93_{\pm 1.76}$
Koklu and Ozkan (2020)	14K	16	1	$99.57_{\pm 0.25}$	$92.58_{\pm 1.51}$	$78.54_{\pm 1.63}$	$47.89_{\pm 1.34}$
Agrawal (2017)	54K	7	3	$99.56_{\pm 0.11}$	$93.47_{\pm 0.88}$	$95.50_{\pm 2.12}$	$57.70_{\pm 1.83}$
Wolberg et al. (1993)	569	31	1	$99.53_{\pm 0.19}$	$91.34_{\pm 1.27}$	$75.42_{\pm 2.66}$	$47.96_{\pm 1.48}$
USDA (2023)	5K	66	1	$99.52_{\pm 0.21}$	$89.19_{\pm 1.59}$	$82.23_{\pm 2.18}$	$56.41_{\pm 1.56}$
Kottarathil (2022)	611K	7	1	$99.50_{\pm 0.19}$	$92.65_{\pm 0.65}$	$91.90_{\pm 1.03}$	$52.48_{\pm 1.72}$
Sejnowski and Gorman (1988)	208	60	1	$99.44_{\pm 0.48}$	$87.03_{\pm 2.45}$	$65.51_{\pm 3.19}$	$41.93_{\pm 1.63}$
Lo (2024)	4K	3	1	$99.34_{\pm 0.10}$	$91.35_{\pm 1.63}$	$124.89_{\pm 2.91}$	$66.00_{\pm 1.60}$
Er (2024)	51K	126	2	$99.34_{\pm 0.59}$	$89.14_{\pm 2.53}$	$82.30_{\pm 1.45}$	$56.22_{\pm 1.37}$
Jikadara (2024)	1K	10	10	$98.94_{\pm 0.23}$	$89.64_{\pm 2.20}$	$146.78_{\pm 2.02}$	$64.18_{\pm 1.78}$
Lainguyn123 (2024)	6K	7	13	$98.87_{\pm 0.22}$	$88.98_{\pm 1.85}$	$140.76_{\pm 3.05}$	$67.86_{\pm 1.49}$
Romero-Hernandez (2022)	2K	18	10	$98.83_{\pm 0.32}$	$92.34_{\pm 1.25}$	$89.67_{\pm 1.59}$	$43.68_{\pm 1.49}$
Mujtaba (2024)	765	3	1	$98.78_{\pm 0.34}$	$91.68_{\pm 1.52}$	$87.15_{\pm 1.52}$	$41.95_{\pm 1.82}$
Bart (2015)	3K	9	8	$98.62_{\pm 0.62}$	$92.58_{\pm 2.79}$	$107.90_{\pm 2.23}$	$46.10_{\pm 1.79}$
François (2024)	1K	4	3	$98.50_{\pm 0.41}$	$93.65_{\pm 2.29}$	$118.48_{\pm 2.87}$	$48.56_{\pm 1.32}$
Khorasani (2024)	1K	10	7	$98.40_{\pm 0.31}$	$93.87_{\pm 0.57}$	$84.78_{\pm 2.69}$	$33.18_{\pm 1.43}$
Ms (2024)	1K	7	2	$97.42_{\pm 1.43}$	$90.67_{\pm 1.63}$	$93.27_{\pm 1.03}$	$42.75_{\pm 2.32}$
Fisher (1936)	147	4	1	$97.27_{\pm 0.59}$	$80.52_{\pm 2.12}$	$91.68_{\pm 2.56}$	$48.61_{\pm 0.95}$
Banik (2018)	339	33	5	$95.60_{\pm 1.69}$	$86.96_{\pm 3.02}$	$84.62_{\pm 2.18}$	$40.70_{\pm 2.44}$
Bohanec and Rajkovič (1998)	2K	0	7	$99.08_{\pm 0.15}$	$92.02_{\pm 1.04}$	$139.61_{\pm 3.89}$	$42.08_{\pm 1.27}$

Table 5: Experimental results on the held-out test sets of the 30 remaining scenarios. 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.

Dataset	Condition	succ. (%)	conv. (%)	inv. size	inv. size (95%)
Uncalibrated sensors and noisy environments					
WINE	Baseline	$99.76_{\pm 0.17}$	$88.34_{\pm 1.31}$	$77.12_{\pm 1.21}$	$60.25_{\pm 1.37}$
WINE	Ca1	$99.71_{\pm 0.15}$	$88.83_{\pm 1.89}$	$77.95_{\pm 1.34}$	$58.96_{\pm 1.34}$
WINE	Ca2	$99.58_{\pm 0.21}$	$88.04_{\pm 1.66}$	$77.94_{\pm 2.14}$	$57.75_{\pm 2.14}$
WINE	No1	$97.61_{\pm 0.46}$	$72.73_{\pm 2.51}$	$68.89_{\pm 0.82}$	$53.92_{\pm 1.69}$
WINE	No2	$76.89_{\pm 2.46}$	$38.17_{\pm 2.22}$	$78.74_{\pm 3.64}$	$37.47_{\pm 2.13}$
EXOPLANETS	Baseline	$99.67_{\pm 0.10}$	$92.30_{\pm 0.86}$	$79.35_{\pm 1.18}$	$54.47_{\pm 1.67}$
EXOPLANETS	Ca1	$99.46_{\pm 0.29}$	$90.98_{\pm 1.27}$	$79.79_{\pm 1.77}$	$54.71_{\pm 1.33}$
EXOPLANETS	Ca2	$97.93_{\pm 1.22}$	$87.95_{\pm 3.06}$	$85.98_{\pm 1.21}$	$51.45_{\pm 1.42}$
EXOPLANETS	No1	$94.23_{\pm 0.88}$	$69.54_{\pm 2.38}$	$74.25_{\pm 1.66}$	$46.38_{\pm 1.49}$
EXOPLANETS	No2	$68.89_{\pm 1.86}$	$44.54_{\pm 1.94}$	$134.95_{\pm 6.81}$	$30.25_{\pm 1.00}$
Heteromorphic populations					
WINE	Ho1	$99.73_{\pm 0.08}$	$88.71_{\pm 1.01}$	$77.62_{\pm 1.39}$	$58.79_{\pm 1.41}$
WINE	He1	$93.47_{\pm 2.35}$	$77.46_{\pm 4.22}$	$77.65_{\pm 1.39}$	$56.58_{\pm 1.86}$
WINE	Ho2	$99.62_{\pm 0.23}$	$89.81_{\pm 2.04}$	$86.18_{\pm 4.92}$	$56.28_{\pm 2.63}$
WINE	He2	$63.51_{\pm 4.36}$	$41.10_{\pm 4.97}$	$110.52_{\pm 6.53}$	$36.38_{\pm 1.34}$
EXOPLANETS	Ho1	$99.59_{\pm 0.33}$	$92.89_{\pm 1.43}$	$82.46_{\pm 5.30}$	$50.90_{\pm 1.35}$
EXOPLANETS	He1	$93.23_{\pm 4.83}$	$78.32_{\pm 9.00}$	$84.16_{\pm 2.46}$	$50.43_{\pm 1.63}$
EXOPLANETS	Ho2	$96.31_{\pm 6.10}$	$91.55_{\pm 12.46}$	$151.20_{\pm 10.37}$	$40.87_{\pm 5.83}$
EXOPLANETS	He2	$18.29_{\pm 30.45}$	$8.15_{\pm 23.67}$	$211.97_{\pm 30.45}$	$84.04_{\pm 27.02}$
MUSHROOMS	Ho1	$97.78_{\pm 1.17}$	$86.07_{\pm 2.90}$	$270.08_{\pm 16.32}$	$74.92_{\pm 3.07}$
MUSHROOMS	He1	$95.04_{\pm 1.91}$	$80.66_{\pm 3.17}$	$285.29_{\pm 9.13}$	$76.57_{\pm 2.74}$
MUSHROOMS	Ho2	$88.97_{\pm 7.13}$	$81.70_{\pm 5.56}$	$180.72_{\pm 34.70}$	$60.26_{\pm 3.68}$
MUSHROOMS	He2	$53.37_{\pm 4.93}$	$30.56_{\pm 5.59}$	$176.84_{\pm 5.53}$	$73.16_{\pm 2.33}$
Robustness against sensor defects					
WINE	De1	$97.72_{\pm 3.86}$	$84.18_{\pm 5.48}$	$78.44_{\pm 1.39}$	$56.12_{\pm 2.52}$
WINE	De2	$88.46_{\pm 23.29}$	$69.27_{\pm 36.39}$	$84.17_{\pm 5.82}$	$53.70_{\pm 1.35}$
EXOPLANETS	De1	$97.69_{\pm 4.67}$	$87.18_{\pm 8.53}$	$80.80_{\pm 1.94}$	$51.12_{\pm 2.38}$
EXOPLANETS	De2	$73.37_{\pm 55.87}$	$60.92_{\pm 62.47}$	$102.84_{\pm 23.96}$	$51.22_{\pm 2.53}$
MUSHROOMS	De1	$96.75_{\pm 1.70}$	$83.95_{\pm 5.73}$	$257.68_{\pm 12.71}$	$75.38_{\pm 3.02}$
MUSHROOMS	De2	$83.53_{\pm 28.77}$	$66.04_{\pm 41.85}$	$225.97_{\pm 25.49}$	$72.96_{\pm 1.96}$

Table 6: Results of the robustness experiments on the held-out test set in Section 5.3 of the main paper on the CLEVR, WINE and EXOPLANETS and MUSHROOMS datasets. Mean and 2 standard deviations computed over 10 independent runs.

# agents	succ. (%)	conv. (%)	inv. size	inv. size (95%)
1M interactions				
$k = 2$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$	$60.40_{\pm 2.14}$	$40.40_{\pm 1.65}$
$k = 10$	$99.76_{\pm 0.08}$	$94.41_{\pm 1.46}$	$52.63_{\pm 1.73}$	$31.14_{\pm 1.11}$
$k = 25$	$99.01_{\pm 0.45}$	$91.75_{\pm 1.95}$	$49.09_{\pm 1.21}$	$25.33_{\pm 0.48}$
$k = 50$	$97.79_{\pm 1.19}$	$89.28_{\pm 3.52}$	$45.71_{\pm 2.29}$	$21.34_{\pm 1.26}$
$k = 100$	$92.69_{\pm 1.96}$	$81.14_{\pm 2.94}$	$45.61_{\pm 1.74}$	$17.99_{\pm 0.71}$
10M interactions				
$k = 10$	$99.97_{\pm 0.02}$	$97.28_{\pm 0.76}$	$68.36_{\pm 1.79}$	$40.49_{\pm 1.65}$
$k = 100$	$99.16_{\pm 0.42}$	$93.39_{\pm 1.45}$	$68.66_{\pm 5.10}$	$26.23_{\pm 1.02}$

Table 7: Effect of population size in the CLEVR scenario. All runs are trained for 1M interactions except the two final rows, which are trained for 10M interactions. Mean and 2 standard deviations computed over 10 independent runs.

Dataset	# ent.	# cont.	# cat.	succ. (%)	conv. (%)	inv. size	inv. size (95%)
CLEVR	468K	20	0	$99.66_{\pm 0.16}$	$98.73_{\pm 0.13}$	$17.10_{\pm 1.15}$	$15.50_{\pm 0.56}$
WINE	5K	12	0	$99.59_{\pm 0.35}$	$98.05_{\pm 1.29}$	$30.67_{\pm 1.76}$	$26.00_{\pm 1.00}$
EXOPLANETS	5K	8	4	$99.14_{\pm 0.56}$	$96.53_{\pm 3.21}$	$31.53_{\pm 1.12}$	$25.57_{\pm 1.69}$
MUSHROOMS	8K	0	23	$99.05_{\pm 0.98}$	$92.72_{\pm 0.96}$	$48.17_{\pm 3.70}$	$31.23_{\pm 1.40}$
MSCOCO	159K	384	0	$98.03_{\pm 0.62}$	$98.56_{\pm 0.60}$	$21.83_{\pm 1.58}$	$17.75_{\pm 0.72}$
CELEBA	203K	384	0	$98.41_{\pm 0.67}$	$98.12_{\pm 2.05}$	$26.01_{\pm 2.72}$	$19.24_{\pm 1.36}$
BIRDS	12K	384	0	$97.97_{\pm 0.68}$	$98.43_{\pm 0.56}$	$21.14_{\pm 0.89}$	$17.48_{\pm 0.69}$

Table 8: Experimental results on the held-out test set of the seven featured scenarios after training agents for 10M interactions, when agents are allowed to produce multi-word utterances ($\ell = 5$). Mean and 2 standard deviations computed over 10 independent runs. The columns describe the dataset, number of entities, number of continuous and categorical dimensions, communicative success, conventionality and linguistic inventory size.

	This work			Clustering			VQ-VIB	
Setting	**2-2**	**2-10**	**10-10**	**2-2**	**2-10**	**10-10**	**2-2**	**2-10**
Baseline	$99.97_{\pm 0.02}$	$100.00_{\pm 0.00}$	$99.52_{\pm 0.19}$	$99.74_{\pm 0.05}$	$97.60_{\pm 0.16}$	$97.60_{\pm 0.18}$	$95.29_{\pm 1.86}$	$47.33_{\pm 23.41}$
Ca1	$99.77_{\pm 0.09}$	$99.69_{\pm 0.22}$	$98.81_{\pm 0.39}$	$39.11_{\pm 3.86}$	$38.33_{\pm 3.60}$	$28.58_{\pm 1.14}$	$94.57_{\pm 1.73}$	$32.28_{\pm 27.48}$
Ca2	$96.69_{\pm 1.99}$	$89.46_{\pm 4.80}$	$80.02_{\pm 4.01}$	$37.52_{\pm 2.36}$	$36.99_{\pm 3.49}$	$27.57_{\pm 1.27}$	$93.94_{\pm 2.14}$	$29.81_{\pm 24.34}$
No1	$99.72_{\pm 0.13}$	$99.46_{\pm 0.13}$	$98.99_{\pm 0.25}$	$39.92_{\pm 3.05}$	$38.36_{\pm 2.87}$	$28.46_{\pm 1.06}$	$92.79_{\pm 3.00}$	$42.89_{\pm 26.56}$
No2	$96.09_{\pm 1.08}$	$82.25_{\pm 1.33}$	$83.62_{\pm 1.06}$	$37.06_{\pm 3.56}$	$36.72_{\pm 2.96}$	$27.42_{\pm 0.98}$	$87.21_{\pm 6.03}$	$38.88_{\pm 26.12}$
Ho1	$99.97_{\pm 0.02}$	$100.00_{\pm 0.00}$	$99.51_{\pm 0.11}$	$39.87_{\pm 4.39}$	$39.33_{\pm 4.34}$	$28.83_{\pm 2.37}$	$93.45_{\pm 3.27}$	$47.13_{\pm 26.64}$
Ho2	$99.99_{\pm 0.01}$	$99.99_{\pm 0.01}$	$99.29_{\pm 0.30}$	$42.36_{\pm 6.27}$	$41.68_{\pm 5.23}$	$32.04_{\pm 6.11}$	$93.84_{\pm 1.27}$	$45.69_{\pm 20.71}$
He1	$97.18_{\pm 2.87}$	$99.92_{\pm 0.18}$	$93.66_{\pm 2.91}$	$29.17_{\pm 12.69}$	$28.98_{\pm 11.83}$	$18.76_{\pm 3.93}$	$93.87_{\pm 2.66}$	$39.92_{\pm 18.11}$
He2	$84.35_{\pm 12.75}$	$98.27_{\pm 2.71}$	$64.10_{\pm 4.42}$	$8.52_{\pm 6.22}$	$7.92_{\pm 5.09}$	$3.97_{\pm 1.07}$	$91.05_{\pm 2.38}$	$34.42_{\pm 26.83}$
De1	$98.57_{\pm 1.94}$	$99.89_{\pm 0.32}$	$96.29_{\pm 2.20}$	$37.09_{\pm 5.13}$	$33.26_{\pm 3.79}$	$25.35_{\pm 4.00}$	$93.55_{\pm 3.26}$	$43.76_{\pm 27.05}$
De2	$88.64_{\pm 4.29}$	$99.43_{\pm 0.44}$	$80.79_{\pm 3.36}$	$14.80_{\pm 7.94}$	$13.03_{\pm 6.29}$	$12.54_{\pm 1.60}$	$86.38_{\pm 5.24}$	$38.54_{\pm 29.31}$

Table 9: Comparison against the clustering baseline k-means (Lloyd, 1982) and the neural VQ-VIB approach (Gualdoni et al., 2024). Results on held-out test sets in the WINE scenario. Mean and 2 standard deviations computed over ten independent runs. In the column labels, x-y denotes a setting with x agents and scenes containing y entities, where one entity is the topic and the others act as distractors.