
This is the **published version** of the bachelor thesis:

Fiorillo Polo, Marco. *Language Emergence in Artificial Agents*. Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Intel·ligència Artificial)

This version is available at <https://ddd.uab.cat/record/336310>

under the terms of the  license.

Language Emergence in Artificial Agents

Marco Fiorillo

Bachelor's Degree in Artificial Intelligence

Universitat Autònoma de Barcelona (UAB)

Tutor: Marco Schorlemmer (IIIA-CSIC / UAB)

Academic year 2025–2026

Abstract—This project studies how artificial agents develop communication protocols when playing referential games: tasks where a sender must describe a target image so that a receiver can identify it among distractors. Two agents, each built around a visual encoder and an LSTM message generator, are trained on a synthetic dataset of 10,000 three-dimensional scenes defined by shape, colour, and size. We extend the setup to a population of five agent pairs and run ablation studies over vocabulary size and distractor count. Results reveal three distinct training phases linked to the activation of the length penalty, including a sharp compression event followed by a recovery to higher accuracy than before compression. Evaluation covers compositionality, grounding, and generalisation using standard emergent-language metrics. The goal is to understand how environmental structure, internal constraints (in this case a least-effort pressure on message length together with the size of the symbol inventory $|\mathcal{V}|$), and social dynamics (training a whole population of agents rather than a single pair) collectively shape the language that emerges.

Index Terms—emergent communication, referential games, multi-agent learning, compositionality, language emergence

I. INTRODUCTION

Emergence is a phenomenon where complex patterns arise from simple local interactions. This is a concept rooted in philosophy of science and later formalised in complex systems theory [1]. In this concrete experiment, we will go into one particular branch: *language emergence*.

Despite significant progress in emergent communication research, systematic investigations of how environmental structure (the statistics of the scenes themselves; which attributes appear, how often, and how distinguishable they are), social dynamics (training a whole population of interacting agents rather than a single fixed pair) and internal constraints jointly shape language properties are limited. Most studies vary one factor at a time: either internal constraints or population dynamics. This project addresses that gap by studying both in a new experimental framework using synthetic 3D scenes and *referential games*. Referential games can be briefly explained as follows:

- 1) A collection of images $\mathcal{I} = \{i_n\}_{n=1}^N$ is given, from which a target image t is drawn, along with K distractor images $\{d_k\}_{k=1}^K$.
- 2) The system consists of two agents: a sender S_ϕ and a receiver R_θ . Where θ and ϕ are the weights of the neural networks that implement the perception and message-generation networks of these agents.
- 3) When observing the target image t , the sender S_ϕ produces a message z_{msg} — a sequence of discrete symbols

drawn from a vocabulary $\mathcal{V}$, where $|\mathcal{V}|$ denotes the vocabulary size and L denotes the maximum sequence length.

- 4) The receiver R_θ is then presented with the message z_{msg} alongside the full candidate set (the distractors) K , and must correctly identify the target image t from among the distractors.

II. OBJECTIVES

The general objective of this project is to understand how internal constraints and social dynamics shape the structure, consistency, and generalisation of the languages that emerge between artificial agents in referential games.

We break this down into eight specific objectives (O1–O8). The first four build the ground we need to study the phenomenon at all: the environment, the agents, a working baseline, and the way we measure everything. The remaining four are the scientific core, and turn our two questions (whether internal pressure reshapes the protocol, and whether a population converges on something more structured) into things that we can actually test. Each objective has a single, explicit criterion that tells us whether we have met it, and we report how each one turned out in Section VI (Table I).

- O1 Environment.** Build a synthetic 3D-scene dataset, and the generator behind it, in which the visual attributes (shape, colour and size) can be varied independently and combined with controllable augmentations. We treat this as met when the generator produces the full 10,000-scene dataset over the defined attribute space.
- O2 Agents and perceptual validity.** Implement a sender–receiver pair whose visual encoders are trained from scratch, with no pretrained priors, and ensure the perception actually works before we ask the agents to communicate at all. We treat this as met when our ResNet-18 encoder reaches a validation recall of at least 0.80 on the shape–colour binding proxy task.
- O3 Baseline Protocol.** Train a single sender–receiver pair that genuinely solves the referential game, giving us a pivot point for later comparison. We treat this as met when validation accuracy sits well above the chance level of $1/K = 0.20$, taking ≥ 0.70 as the target.
- O4 Evaluation framework.** Put together a selection of standard metrics in similar literature, covering morphology and semantics, and apply exactly the same selection of metrics to every condition so that the numbers are

comparable. We treat this as met when all conditions are scored with the same metric set on a single common validation set, ready for analysis.

- O5 Internal pressure as an abrupt change.** Test whether the least-effort pressure reorganises the communication protocol abruptly, at an identifiable moment in training. We treat this as met when we can localise a non-gradual drop in mean message length to a training window that coincides with the activation of the length penalty λ .
- O6 Isolating the driver.** Use ablations to pin down if compression is really caused by the length penalty, and not by the vocabulary size or the difficulty of the task. We treat this as met when compression appears in every run with $\lambda > 0$ and is absent when $\lambda = 0$, regardless of $|\mathcal{V}|$.
- O7 Population structure.** Test whether training a population, rather than an isolated pair, pushes the agents towards a language that is both more shared and more structured. We treat this as met when the population shows cross-sender agreement above a random-message baseline and, at the same time, a higher topographic similarity than any single pair.
- O8 Generalisation.** Check whether the protocols that emerge actually generalise beyond the exact conditions they were trained on. We treat this as met when zero-shot accuracy on a harder, unseen setting — a larger candidate set K — stays above the matching chance level $1/K$.

A. Terminology

It is important to distinguish carefully among terms, since they are used in various ways across the diverse literature. Although these terms originate from linguistics and semiotics, they have been adapted in the literature to describe artificial agent communication systems. Thus, highlighting the most important terms is needed in this section:

- **Symbol:** A symbol is an atomic discrete unit drawn from a fixed vocabulary which, by itself, carries no inherent meaning but can be combined with other symbols to form larger structures. In semiotics, symbols are *signifiers* whose meaning arises through convention [2] rather than intrinsic properties [3]. The ideas of *language as use* is commonplace in the field of emergent language. In artificial communication systems, symbols are typically discrete tokens generated by agents to encode information [4].
- **Message:** A message is a sequence of one or more symbols produced by a sender and transmitted to a receiver; it represents the minimal communicative act in the system [5].
- **Communication protocol:** A communication protocol refers to the set of rules, mappings, and agent architectures that define how symbols and messages are generated, interpreted, and used to achieve a task.
- **Language:** Finally, we reserve the term language to describe the emergent system of structured communication

that arises when messages display systematic or compositional patterns across contexts, potentially supporting abstraction and generalisation [6].

III. RELATED WORK

At the root of this investigation is the Lewis signaling game, which can be found in David Lewis’ *Convention* [7]. It is the foundation of modern referential games in multi-agent learning. This game has been adapted to multi-agent communication with neural networks as seen in Lazaridou et al. (2017) [5], one of the first references on emergent communication in referential games. Early deep learning experiments can resemble classifiers or look-up tables rather than complex communicative systems, since they do not introduce compositionality or generalisation. However, they introduce concepts such as grounding (the property by which symbols are anchored to real-world concepts) [4] and purity (how well the symbols agents use align with human-recognisable categories), which are essential to assess the performance of the communication protocol. This and the next paper mentioned will function as the technical baseline for experimentation.

Luna et al. (2020) [8] demonstrated that internal constraints (least-effort) and external environmental variations (object constancy, frequency, brightness in the scenes) promote abstract, compositional, and generalisable communication. This is at the core of the project, since we are also interested in designing constraints that will enrich the quality of the language in question and measure how its attributes evolve.

In Chaabouni et al. (2023) [9], the paper addresses population training and shows that partitioned population training do not encourage compositional structure and convergence across populations more effectively. This is a key concept of this investigation as well as we will see further.

Because the field is still young and experimental setups vary widely, systematic investigations of environmental factors influencing language emergence remain limited. Although emergent language research has explored many architectures and learning strategies, this part remains underexplored. Prior work typically isolates a single force: Rodríguez-Luna et al. [8] study internal least-effort pressure in a single pair, while Chaabouni et al. [9] study population dynamics without a controlled internal-pressure ablation. To our knowledge, no work varies both within one framework. This project fills the gap by measuring how internal message constraints and population training jointly shape the emergent language under a single common metric battery.

IV. METHODOLOGY

For the entire experiment, we developed a specific program to generate the dataset. Following a successful validation of the synthetic data generation pipeline, we released the code publicly as an open-source repository¹. Made in python, it mainly uses Tkinter for a simple UI, numpy for numerical arrays, and trimesh for 3D image generation and rendering.

¹<https://github.com/polarised/sceneshapemaker>

TABLE I
ACHIEVEMENT OF THE SPECIFIC OBJECTIVES AGAINST THEIR SUCCESS CRITERIA. STATUS IS REPORTED ON THE DATA AVAILABLE AT SUBMISSION; O7 IS MARKED *partial* AS IT RESTS ON A SINGLE POPULATION RUN (SEE SECTION VI FOR THE SEED-VARIANCE LIMITATION).

Obj.	Success criterion	Status	Evidence
O1	Dataset generator over the defined attribute space, released as a reproducible tool	Met	sceneshapemaker (open source)
O2	Encoder recall ≥ 0.80 on shape-colour binding	Met	recall = 0.882
O3	Baseline accuracy ≥ 0.70 ($\gg 1/K = 0.20$)	Met	0.86
O4	Same metric battery on a common held-out set for all conditions	Met	Table III
O5	Non-gradual length drop localised to the λ -activation window	Met	$7.4 \rightarrow 2.3$ tokens, steps 2500–3500
O6	Compression in all $\lambda > 0$ runs, absent at $\lambda = 0$, independent of $ \mathcal{V} $	Met	Table II
O7	Population agreement above random <i>and</i> TopSim above any single pair	Partial	agreement 0.51; TopSim 0.125 vs ≤ 0.05 (single seed)
O8	Zero-shot accuracy above chance at larger K	Partial	test for robustness that matches validation

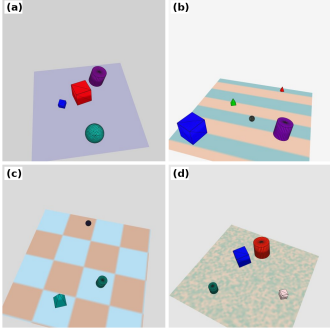


Fig. 1. Four Scenes extracted from the Training Dataset

A. Synthetic Data Creation

Synthetic images must promote generalisation rather than memorisation, while offering a rich but controlled environment where meaning can be systematically derived from multiple visual attributes. The characteristics of scenes are as follows:

- shapes: [cube, sphere, cylinder, pyramid]
- colour: [red, blue, green, yellow, cyan, magenta, orange, purple, white, teal]
- size: [small, large]

This synthetic image generation software we created allows to select the amount of scenes, resolution, objects per scene, layout and camera settings, camera elevation, and other augmentation techniques such as floor colour changes and patterns, colour jitter, lighting (ambient or diffuse), noise patterns, 3D rotations and camera zoom jitter. The chosen settings for training were 0-2 objects of each shape with a 512×512 resolution as well as all the augmentation techniques aforementioned. All rendered scenes use sharp object contours, as softer lighting and colour variations produced samples that were difficult to distinguish even for human observers. This type of synthetic image generation is commonplace in emergent language experiments as seen in [10], [11], [12]. The size of the dataset for this project is of 10.000 scenes. An example of four images can be found at Fig.1

B. Agent Architecture

An architectural decision for the Agent Architecture is the Vocabulary Size ($|\mathcal{V}|$) and the Message Length (L). For this case, the best option is to use $|\mathcal{V}| = 10$ (a total of 10 symbols) and $L=8$ (8 symbols at max for a message).

$|\mathcal{V}| = 10$ is chosen to sit close to the scale of the attribute space the agents must distinguish (4 shapes and 10 colours). A symbol to each dedicated feature would lead to 14 symbols total, but we deliberately propose fewer. This shortfall is the point: the sender cannot give every attribute a unique label and is therefore pushed to reuse and combine symbols across positions rather than memorise one symbol per concept. For this investigation an underlying idea is to not take human semantics as a starting point, so keeping the vocabulary slightly shorter is a good idea to try to avoid replicating human semantics in the process. Reserving index 0 for EOS leaves exactly 9 content symbols. A larger vocabulary would allow the sender to memorise one symbol per scene rather than compose several symbols together; a smaller one would theoretically be unable to refer to all attributes individually, and would have to cluster many values in a symbol.

$L = 8$ reflects the maximum number of symbols a human would need to communicate a scene perfectly: up to 2 objects of each of 4 shapes must be described, and the most compact unambiguous encoding requires one symbol for shape and one for colour per object — giving a worst-case of $4 \times 2 = 8$ symbols. The cap is therefore not arbitrary but tight to the human reference length for the hardest possible scene in the dataset, since we will apply internal constraints on message length anyway, also being aware of not replicating human semantics.

1) *Sender Agent Architecture:: Convolutional Neural Network (ResNet-18)*: As the "eyes" of the sender in the communication, we train a CNN from scratch. The architecture used is ResNet-18 [13], which is the most appropriate since it is a standard architecture that is light and solves vanishing gradients [14](an important concern when training from scratch). Building a custom convolutional architecture from scratch falls outside the scope and time constraints of this investigation, yet using a pretrained model would silently

undermine the experiment’s core claim. Pretrained weights encode visual representations already organised around human conceptual categories: handing the agents a CNN that already “knows” what a sphere is — because it was trained to classify one — means the emergent protocol would simply rediscover and relabel pre-existing human categories rather than construct its own. The language that emerges would then reflect the inductive biases of ImageNet supervision, not the structure of the referential task itself. ResNet-18 is therefore used as a fixed architectural choice, initialised from random weights, so that the visual representations and the communication protocol co-emerge entirely from game signal. The final output of this module is a 512-dimensional scene embedding.

a) LSTM message generation:: A linear layer projects this embedding with a tanh activation to produce the LSTM’s initial hidden state. We use tanh since it compresses the representation in $[-1, 1]$, which is the same range as LSTM’s internal gates operate in. The LSTM helps us transform the “vision” vector into a concatenation of symbols. However, language emergence does not imply that these symbols must refer to the amount of shapes, the type of shape or its colour, but the language will form itself in a way where it can express whatever structure best solves the referential task. The mapping from visual features to symbols is entirely unconstrained: the protocol that emerges may be compositional, encoding attributes systematically across positions, or it may be holistic, treating whole scenes as unanalysed units.

The previous token for the LSTM as initial input is the Start of Sequence [SOS] tag. For each of the L steps:

- It takes previous token embedding and current memory state, then it updates both. Hidden state now encodes everything the sender has said so far.
- A linear layer turns the hidden state into scores. One per vocabulary symbol (for visualisation, “abcdefghi” + EOS).
- We add small random noise to the scores, then apply a softmax, specifically Gumbel-Softmax sampling [15]. The result of this Gumbel-Softmax is a soft one-hot vector: a vector of 10 values that sum to 1, with nearly all the weight on one symbol. The temperature τ controls how peaked this distribution is. This is necessary since we cannot backpropagate with one-hot vectors, but cannot discretise to symbols by passing a numerical vector.
- Finally, we track a variable s_t (*still speaking*), which starts at $s_0 = 1.0$ and represents the probability that no EOS has been emitted up to step t . At each step it is updated as:

$$s_t = s_{t-1} \cdot (1 - P(\text{EOS}_t)) \quad (1)$$

If EOS was sampled with high probability, s_t collapses towards 0 and stays there; if EOS was not sampled, it remains close to 1. Summing s_t across all L steps gives the expected message length $\bar{\ell} = \sum_{t=1}^L s_t$, which is added to the training loss as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \cdot \bar{\ell} \quad (2)$$

where $\lambda = 0.1$ and CE stands for Cross Entropy. This creates pressure on the sender to terminate as soon as the message is enough, rather than filling all L symbols systematically.

This sender architecture is based on prior experiments such as [16], [17], [10].

2) Visual Encoder Validation: To ensure that ResNet-18 was the appropriate visual encoder for this experiment, we conducted a standalone validation experiment prior to any referential training. The purpose was to verify that the architecture was capable of solving a meaningful perception task on the synthetic dataset. Without this sanity check, any failure in the referential game could not be attributed to the communication protocol with confidence, since it might simply reflect a blind encoder.

The proxy task was shape–colour binding: given a scene image, the model was required to predict which (shape, colour) pairs were present. With 4 shapes and 10 colours, this yielded 40 possible binary labels in total, one per combination. Since each scene contained at most 2–3 objects, no more than 3 of the 40 labels were ever active simultaneously, producing severe class imbalance. To address this, we applied `pos_weight` during training, upweighting the positive class proportionally to its under-representation, which is standard practice for multi-label tasks with sparse ground truth.

Training ran for 60 epochs with Adam at a learning rate of 10^{-3} — the conventional starting point for Adam on randomly initialised networks of this scale. The dataset for this proxy task was exactly the real experiment dataset length (5000 scenes for the proxy task). By epoch 60, the encoder reached a validation recall of 0.882 and a validation accuracy of 0.753. Recall was the chosen criterion rather than accuracy precisely because, under heavy class imbalance, a model that never fires on any label achieves near-perfect accuracy while detecting nothing. We therefore set a recall above 0.8 as the threshold: it constituted genuine evidence that the encoder had learned to consistently identify which shape–colour combinations were present in a scene, which is exactly the perceptual competence the referential game demands. Having cleared that bar, we adopted ResNet-18 as the visual module for both the sender and the receiver.

3) Receiver Agent Architecture:: The receiver mirrors the sender’s architecture in its broad structure, though no components are shared between the two agents. The receiver maintains its own visual encoder so that the discrete message remains the only bridge between them — had the encoders been shared, the agents would have entered the game already speaking a common visual language, which would have trivialised the communicative challenge. For the same reason, the receiver holds its own token embedding matrix: the question of how to emit a symbol and the question of how to interpret one are fundamentally distinct, and tying those matrices would impose an artificial symmetry with no linguistic justification.

Where the receiver diverges from the sender is in its output. Rather than producing a message, it reconstructs the incoming signal into a 512-dimensional “imagination” vector [18] —

a point in the same embedding space its visual encoder produces. This design choice makes it possible, during evaluation, to inspect what the receiver inferred the target looked like given only the message, independently of whether that belief was correct.

a) Message interpretation (LSTM): The incoming message arrives as a sequence of soft one-hot vectors of shape $(L, |\mathcal{V}|)$ (which, as said previously, having soft one-hot vectors is essential so that we can backpropagate later on the whole sender and receiver architecture). We multiply by the receiver’s token embedding matrix to turn each symbol into a 64-dimensional vector (thus, translating into the receiver representation of symbols).

The LSTM produces a hidden state h_t at each step. We collapse the sequence into a single vector by taking a mask-weighted mean. The decision of doing a mask-weighted mean is to not just divide by L all the time (since not all of L is used for messages) but rather weight only the used symbols.

$$\bar{h} = \frac{\sum_{t=1}^L s_t \cdot h_t}{\sum_{t=1}^L s_t} \quad (3)$$

where $\bar{h}$ is the final pooled message representation passed to the imagination head, and s_t is the *still speaking* variable. This is multiplied to the weighted average to make sure to not just take the last LSTM state, but rather take into account all of them with their own weight of how sure we are that the message had all the content of the message.

Note that the length mask s_t is applied twice: once on the embedded input tokens before the LSTM, and once on the hidden states in the weighted mean. This is necessary because even when a zero vector is fed into an LSTM, the hidden state at that step still carries forward information from previous steps through the LSTM’s internal memory. Masking the output makes these h to not contribute to the final pooled representation.

b) Imagination and candidate matching: A linear layer projects $\bar{h}$ into a 512-dimensional “imagination” vector $\hat{z}_{\text{msg}}$, living in the same space the visual encoder produces. In parallel, the K candidate images are passed through the receiver’s CNN to produce K embeddings $z_{c,k}$. Now candidates and message live in the same space.

The receiver then scores each candidate by cosine similarity against its imagination, scaled by a learned temperature, similar to the mechanism seen in CLIP[19]:

$$\text{logits}_k = e^{\sigma} \cdot \frac{\hat{z}_{\text{msg}} \cdot z_{c,k}}{\|\hat{z}_{\text{msg}}\| \|z_{c,k}\|} \quad (4)$$

Cosine bounds the similarity range, and the learned temperature $\exp(\sigma)$ sharpens the distribution so the softmax distribution does not collapse to a uniform distribution. Cross-entropy against the true target index then trains the whole pipeline end-to-end. This receiver design follows [16], [10]. For a simple visualisation on Sender-Receiver Architecture (see Figure 6).

C. First Experiment: Individual Referential Communication

The objective of this experiment is to establish a baseline emergent communication protocol between a single sender–receiver pair, and to measure how it responds to controlled environmental and internal pressures.

1) Training Loop: At each training step, we sample a batch of $B = 64$ target images from the training split. The sender S_ϕ processes each target and produces a discrete message z_{msg} via the Gumbel-Softmax LSTM loop described in Section IV-B1. The receiver R_θ then scores the $K = 5$ candidate images against its imagined embedding $\hat{z}_{\text{msg}}$ and returns logits over the lineup. The joint loss is backpropagated through both agents in a single pass — no separate pre-training, no frozen components (total of about 16 components to which loss flows):

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(R_\theta(z_{\text{msg}}, \mathcal{K}), y) + \lambda \cdot \bar{\ell} \quad (5)$$

Where $\mathcal{K} = \{k_1, \dots, k_n\}$ is the set of K candidate images presented to the receiver, and $y \in \{1, \dots, K\}$ is the ground-truth index of the target within that lineup. This means that a single game outcome simultaneously shapes how both agents see the world, how the sender encodes it into symbols, and how the receiver interprets those symbols back into a visual expectation. Both ϕ and θ are updated jointly by a single Adam optimiser at learning rate 10^{-4} , with gradient clipping at norm 5.0 (to prevent instability in the early phase when both visual encoders start from random weights).

Training runs for 30,000 steps. The Gumbel-Softmax temperature τ is annealed linearly from 1.0 to 0.5 over the course of training, progressively sharpening the symbol distributions towards discrete one-hots without breaking the gradient path.

In the first few epochs of training, the accuracy and loss failed dramatically. This is called silent collapse: In the early training phase both agents are random. The receiver can’t use the message at all because it doesn’t understand it yet, so the game loss gradient is pure noise from the sender’s perspective. The only clean, consistent gradient signal is the length penalty, which always points toward shorter messages. So the sender rationally learns to say nothing before the receiver has had any chance to learn to listen. In order to prevent silent collapse of the sender, we let the communication channel establish before adding the internal pressures of minimum message length. This is done with the warmup hyperparameter, which controls how many steps this restriction applies (i.e, 2000 steps before activating the internal pressure of message length). Finally, validation is performed every 500 steps using hard sampling ($\tau = 0$) on a fixed held-out set with deterministic distractor lineups (so that accuracy is comparable across checkpoints).

2) Ablation Studies: To isolate the contribution of individual design decisions, we run four controlled ablations against the baseline, changing only one hyperparameter at a time (not accumulative):

- $\lambda = 0.0$: removes the least-effort penalty. It poses the question: Can the referential task alone drive message compression? Does explicit length pressure improve

compositionality? In the baseline experiment, before the length pressure is activated (in the first 2000 steps), message length actually converges towards 7.4 symbols per message, which means that agents already establish an effective message length without restrictions.

- $|\mathcal{V}| = 20$: doubles the vocabulary size, this ablation asks whether a larger vocabulary leads to less compositional pressure and more symbol-concept memorisation.
- $|\mathcal{V}| = 5$: reduces the vocabulary to 1 EOS + 4 content symbols, less than half the baseline. With fewer symbols available, sender is forced to reuse and combine them to convey the same information, creating the strongest possible compositionality pressure. This tests the lower bound of vocabulary size.
- $K = 10$: doubles the candidates per round, making the task harder and tests whether it can entail richer communication.

Each ablation is evaluated with the same metrics as the baseline and compared directly against it.

D. Second Experiment: Population Referential Communication

The main objective of this experiment is to investigate how emergent languages evolve in a population of agents, emphasising convergence, standardisation, and compositionality.

a) Technical Setup::

- Agents: Multiple Senders and Receivers (e.g., 5–10 each).
- Pairing strategy: Randomly assign Sender–Receiver pairs per round; each agent interacts with several partners over training (partitioned population)[9].
- Task and Environment: Same referential counting task and synthetic images as in Experiment 1.
- Communication: Discrete-symbol sequences, same as Experiment 1.

For each sender–receiver pair, the sender observes a target image and produces a discrete-symbol message using its CNN–LSTM encoder, while the receiver processes the message alongside a set of candidate images to identify the correct target (exactly as in experiment 1). The pairwise loss is typically cross-entropy over the receiver’s selection probabilities or mean squared error if predicting continuous attributes, and gradients are backpropagated through both sender and receiver networks. To encourage compositional and efficient communication, an optional least-effort term can be added to penalise longer or more complex messages [8]. Population-level generalisation emerges because each agent interacts with multiple partners across training rounds: the loss is effectively averaged over many pairwise interactions, forcing agents to converge on a communication protocol that is interpretable and effective across the population, rather than being specific to a single partner. Importantly, this means that we do not need to compute a gradient over all possible sender–receiver combinations at once; each agent naturally experiences a gradient signal representative of the broader population through repeated interactions with different partners.

Environmental pressures can be applied as in the prior experiment, and we can also check for *language convergence*; whether multiple agents adopt shared symbols and compositional structures. We can also do Zero-shot evaluation on unseen objects, scenes, or new agent pairings not seen during training, and compare all metrics explained in the next section (purity, compositionality, topographic sim. etc.) between the two main experiments and the parameters tweaked in each experiment.

V. RESULTS

We first look at the raw signals we log at every evaluation checkpoint (validation accuracy, cross-entropy loss, mean message length, and vocabulary usage), which give a first picture of how the protocol forms and where the interesting events happen during training. The evaluation standards for emergent-language experiments are extracted from a 2024 survey [20] gaining knowledge of the most popular and standardised methods (see Figure 7 for the full map, based on references). We will mainly focus in evaluating semantics and morphology, since things like phonetics or syntax (although syntax is slightly mentioned) are either not related to the scope of the investigation or need of the aforementioned to be able to be measured.

Throughout this section we read each metric against the success criteria set out in Section II: the training signals and the ablations are in dialogue with the internal-pressure objectives (O5, O6). Purity and the token–shape decomposition to grounding and structure. Sender agreement and topographic similarity to the population objective (O7), and the zero-shot tests to generalisation (O8). Table I collects the findings.

A. A First View of the Training Signals

1) *Baseline Training Dynamics*: The baseline run ($|\mathcal{V}| = 10$, $K = 5$, $\lambda = 0.1$) has completed the full 30,000 steps, and looking at the accuracy and the mean length curves together, there are three main phases.

In the warmup phase (steps 0–2000), before the length penalty λ is activated, accuracy rises from 0.202 to 0.758 while mean message length stabilises naturally around 7.4 tokens; very short of the maximum $L = 8$. This gives a lot of information as how the agents already start to communicate before any explicit compression pressure exists, and the length they settle on is below the maximum on its own. This suggests the referential task itself is sufficient to drive some degree of message efficiency, a finding we ought to investigate further with the $\lambda = 0.0$ ablation.

At step 2500, when the effects of λ are already noticeable, a sharp compression occurs: mean length drops from 7.47 to 3.32 in a single evaluation window and accuracy dips to 0.709. This is consistent with the sender over-compressing faster than the receiver can adapt: there is a co-adaptation delay between both agents. A second dip occurs at step 3500 (0.640), as the sender continues compressing toward ≈ 2.3 tokens, again making a small delay in the receiver’s adaptation.

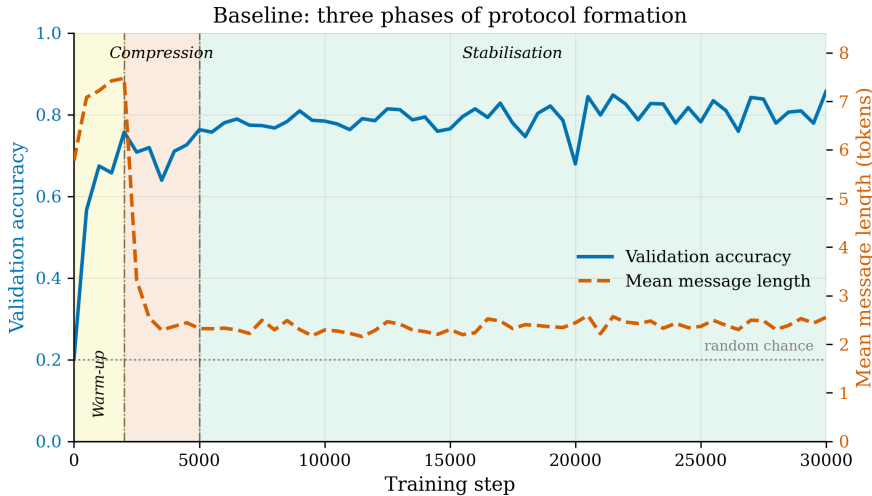


Fig. 2. Baseline run ($|\mathcal{V}| = 10$, $K = 5$, $\lambda = 0.1$): validation accuracy (left axis) and mean message length (right axis) over training. The three shaded regions mark the warm-up phase (before λ activates at step 2000), the compression event, and the stabilisation plateau. Accuracy recovers above its pre-compression high once the receiver adapts to the shortened messages.

In the following plateau/stabilisation phase (steps 5000 onward), the receiver is adapting to the compressed message. Mean length plateaus at 2.3–2.5 tokens approx. and accuracy recovers to 0.858, going over the pre-compressed high of 0.758, while cross-entropy falls from ≈ 0.45 around the compression event to ≈ 0.29 at the end. This suggests that shorter, more efficient messages benefit communication if the receiver adapts accordingly. The use of vocabulary also goes down over this window, from all 9 content symbols to 5 at the end, which suggests the agents settle on a smaller working set of symbols once the protocol is stable.

2) *Ablation and Population Signals*: Table II collects the final-checkpoint signals for the four ablations and the population run next to the baseline for comparison. A few patterns stand out. The $\lambda = 0.0$ run never compresses (6.91 tokens, all 9 symbols kept active) and reaches the highest peak accuracy of any condition (0.927) although its late phase is noisier, which confirms that the task can succeed without the length penalty but produces a structurally different, long protocol. The $|\mathcal{V}| = 20$ run compresses like the baseline yet uses 9 of its 19 symbols at its best checkpoint, consistent with a larger vocabulary giving less reuse pressure, while $|\mathcal{V}| = 5$ matches the baseline almost exactly in both length and accuracy despite having less than half the symbols, which is exactly the reuse pressure we were looking for in the investigation. The $K = 10$ run is the only ablation that we left behind for the memory errors arising with it and the time limitation the investigation has. Finally, the population run (5×5 , evaluated over the full $N \times N$ pairings) reaches a best mean accuracy of 0.650 (min 0.587, max 0.687) — below the single pair, as expected since each sender must now be understood by several receivers — but keeps the full vocabulary active and compresses slightly harder, to 2.21 tokens. Together, these first-view signals confirm that the length penalty is the dominant force shaping

TABLE II
TRAINING SIGNALS ACROSS ALL CONDITIONS. MEAN LENGTH AND VOCABULARY USAGE ARE MEASURED AT THE BEST-ACCURACY CHECKPOINT. POPULATION ACCURACY IS THE MEAN OVER THE 25 PAIRS.

Condition	Best acc.	Final acc.	Mean len.	Vocab used
Baseline	0.858	0.858	2.56	5 / 9
$\lambda = 0.0$	0.927	0.837	6.91	9 / 9
$ \mathcal{V} = 20$	0.864	0.761	2.44	9 / 19
$ \mathcal{V} = 5$	0.832	0.827	2.48	4 / 4
$K = 10$	0.090	0.090	5.78	9 / 9
Population 5×5	0.650	0.642	2.21	9 / 9

the protocol, a claim the morphological metrics in the next subsection will examine in detail.

B. Morphological Results

Morphological metrics aim to evaluate the structure and formation of words within a language. All metrics in this section were computed on the best checkpoint of each run in terms of accuracy. The validation set is of 1000 scenes and uses hard sampling (no Gumbel-Softmax in this case is required, since we work only forwards). Table III collects all the values together for cross comparison.

a) *Referential Accuracy*: Accuracy is the simplest measure. The single-agent ablations achieve very good accuracy for the task at hand, where the worst performing is the population model. This is expected, as population requires different receivers to process correctly instead of one; the task becomes inherently harder. Accuracy is only a measure of

³Sender agreement reaches a peak of 0.513 at step 19,000 and a final value of 0.434 at step 30,000. The single reported figure in Table III should be read as the converged regime rather than the exact final checkpoint.

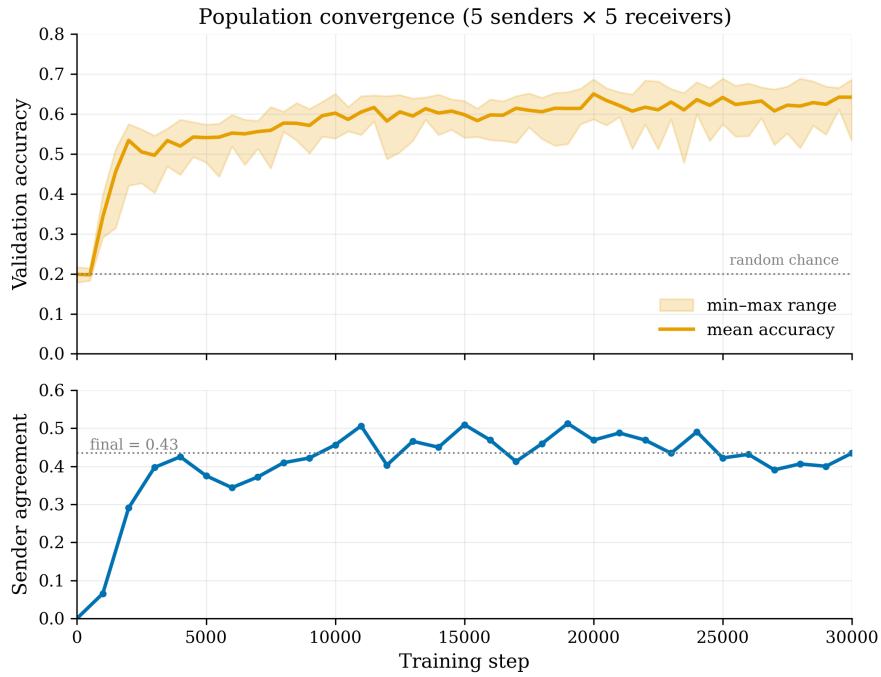


Fig. 3. Population run (5×5). Top: mean validation accuracy with the min-max range across the 25 pairings shaded. Bottom: sender agreement, the fraction of scenes for which two randomly chosen senders emit an identical message.³

model health, but not a metric that will help us conclude about this investigation.

b) Average Message Length.: Average Message Length is a simple and good measure to measure information density. As Table III shows, every condition with the length penalty active converges to a comparable short length (baseline 2.56, $|\mathcal{V}| = 20$ at 2.44, $|\mathcal{V}| = 5$ at 2.48) regardless of vocabulary size. The clear reading is that it is the length penalty λ , and not the vocabulary size or the task difficulty, that drives the compression. Population compresses the hardest of all with a 2.21. Short messages built from the most common and compositional symbols are a safer bet across agents. Compression, convergence and compositionality reinforce between each other: messages will be much smaller to be able to be understood better in the tasks.

c) Distinct Appearances (DA).: Distinct Appearances in this particular case measures how much agents reuse their symbols. A high value towards 1 means symbols are recycled across many messages. The pattern matches the ideas stated; $|\mathcal{V}| = 5$, with only 4 content symbols, has the highest reuse among the working runs (0.388): with so few symbols it is forced to recycle them, which reinforces one of our hypotheses; reducing vocabulary size increases vocabulary reuse (in which ablation $|\mathcal{V}| = 5$ now hints for compositionality). $|\mathcal{V}| = 20$ has the lowest (0.169): with a large vocabulary, as expected. The $K = 10$ run shows a high value (0.509). As we know, this run did not work as intended but it is a good control group to evaluate that from such poor performance, it sends the same long, repetitive message, pushing DA upwards but obviously not meaning anything of matter for the investigation.

d) Singular Value Decomposition (SVD).: SVD checks whether individual symbols specialise for particular shapes, or whether they all carry the same vague shape information [5]. For each condition we build a table whose rows are the content symbols and whose columns are the four shapes, where each entry counts how often a symbol appears in scenes of a given dominant shape. We normalise each row into a distribution — “when this symbol is used, which shape is the scene usually about?” — and factorise the table with SVD. The result tells us how concentrated the symbols’ behaviour is: if one direction explains almost all of it, every symbol points to roughly the same shape distribution and no symbol is individually shape-selective; if the variance is spread out, different symbols specialise for different shapes.

The collapsed $K = 10$ run sits at the degenerate extreme (99.7% on the first component), its rows essentially identical, as expected from a protocol that says the same thing regardless of the scene. $|\mathcal{V}| = 20$ is the most spread (82.3%; the rest on the second and third components), so with a larger inventory some individual symbols do specialise. The baseline (92.5%), $\lambda = 0.0$ (88.0%) and the population (91.7%) fall in between. $|\mathcal{V}| = 5$ is the most concentrated of the working runs (98.3%), but this is the limitation showing through, not a clean negative result: with only four content symbols the protocol expresses itself through combinations (‘da’, ‘ac’, ‘aaa’), and a per-symbol decomposition cannot see that. The high first-component share tells us its symbols are not shape-selective in isolation; it says nothing about whether the combinations are.

C. Semantic Results

Semantic results measure whether the symbols agents use actually carry meaning: whether they refer consistently to concepts, combine to express new ones, and generalise beyond training.

a) Grounding — Purity: Purity asks how cleanly messages line up with human-recognisable categories: we group the validation scenes by the message they receive and, for each group, measure how concentrated it is on a single shape or colour [5]. On its own the number means little: even a protocol that says nothing scores well simply by matching the most common class, so we always compare it against the relevant base rate. Shapes are not evenly distributed across scenes⁴ — cube is the dominant shape in about 45% of them, the others less — while the ten colours are roughly even (about 10% each). Read against those base rates (Figure 4), shape purity (about 0.69) sits clearly above its 0.45 base rate, roughly 1.5 times chance, so the messages do carry real shape information. Colour purity (about 0.19) sits above its 0.107 base rate as well, roughly 1.8 times chance. Both attributes are therefore partially grounded, with colour the stronger of the two. The two runs with the highest raw purity, $\lambda = 0.0$ and the collapsed $K = 10$, only look pure because they spread their scenes over hundreds of different messages (423 and 807 out of 1000); that inflates purity automatically and is not grounding. This indicates the compressed protocol grounds in both human-recognisable attributes to a meaningful degree, rather than relying solely on features we did not label.

b) Compositionality — Topographic Similarity: Topographic similarity is the main test for whether language is compositional, and often used to compare neural network layers or brain activity patterns [10]. The intuition is to see whether similar scenes produce similar messages, and different scenes produce different ones. A compositional language establishes this link, and the form of the message mirrors the meaning. A high value equals a more structured language, and a value near zero means that the words are arbitrary labels without internal logic. For the single pair experiments, the values are all small (between -0.01 and 0.05). This is due to the high compression it reached. Our protocols compressed down to one to two symbols, and this is very little structure for topographic similarity to evaluate. This makes this metric, for the single pair ablation, weaker for conclusions: it can either mean the message is *holistic* (one message = one scene), or that it is compositional but topographic similarity has not enough space to evaluate so. On the other side, population experiments have a higher topographic similarity value than single pair ablations, reaching 0.125. This value is still not high in absolute terms, but it is informative that proportionally, communication protocols developed in population settings can develop *more* compositional structure than an isolated pair. Notably, the population reaches this higher value despite

having the shortest messages of all conditions, which suggests the additional structure is genuine and not an artifact of message length.

c) Compositionality — Positional Disentanglement: Positional disentanglement [21] asks whether each position in the message has a fixed job, for example “the first symbol always says the shape and the second always says the colour”. It is a way of discussing semantics along with syntax, commonplace in emergent communication benchmarks [22], [23]. Across every condition, values are tiny (between 0.001 and 0.021). As with topographic similarity this is partly a floor effect, since messages of one or two symbols leave almost no positions to carry information; beyond that, it points to holistic single-agent behaviour where order has no semantic information.

d) Consistency (Variance of the Laplacian): Consistency asks a simple question: when the agents see the same kind of scene, do they describe it with the same symbols? To make “the same kind of scene” precise we group scenes by their concept, which in our experiment is the count of each shape in the image, that is the vector [cube, sphere, cylinder, pyramid]. Two scenes with one cube and two spheres share a concept regardless of colour or placement. For each concept we collect every symbol the sender used to describe its scenes and build a histogram of how often each symbol of the vocabulary appears. If the language is consistent, this histogram is sharp [24], with most of the weight on one or two symbols; if it is inconsistent, the histogram is flat and spread across many symbols. We summarise the sharpness of that histogram with the Variance of the Laplacian (the variance of its second difference), where a higher value means a more peaked, more consistent usage. Because the score is normalised by the number of samples, the absolute numbers are small and are meaningful only relative to each other, not on their own.

The short-message runs are the most concentrated (baseline 0.017, $|\mathcal{V}| = 5$ 0.014, population 0.021), while the long-message runs are near zero ($\lambda = 0.0$ 0.001, $K = 10$ 0.0003). This fits the rest of the picture: when messages are compressed, each concept ends up tied to a small set of symbols, so usage is sharp; when messages are long and varied, usage spreads out. We note, however, that this depends on our linguistic definition of the concepts: the agents could well just measure colour or size; it is important to not impose our inherent concepts to the evaluation of the language. This is a way to see, on proportion among models, the consistency of the communication protocol.

e) Generalisation — Zero-Shot Evaluation: The most standard metric for Generalisation is doing Zero Shot Evaluation [8], [21], [22]. A very interesting event are “metonymic” links as seen in Lazaridou et al. (2017) [5], in which the agents are able to encapsulate a concept like “sea” inside a “dolphin” label (the nearest one conceptually). We evaluate generalisation on a held out set of 1.000 scenes. This set is the same attribute set as training (four shapes, ten colours) but renders scenes with unseen object positions, camera angles and higher object counts (up to three objects for a single shape instead of two for training). This zero-shot generalisation

⁴Per-shape counts are sampled evenly; the skew in the dominant shape arises from taking the most frequent shape per scene, where tied counts are resolved in a fixed shape order.

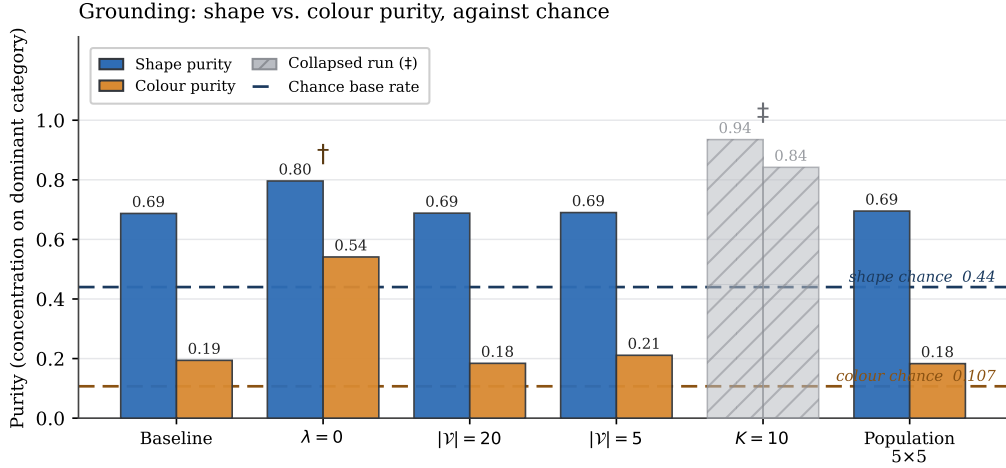


Fig. 4. Dominant-shape purity (blue) and colour purity (ochre) per condition, against chance base rates. [†] $\lambda=0$: valid run, purity inflated by many unique messages. [‡] $K=10$: collapsed run (greyed), purity is an artefact.

misses the actual premise it takes on other studies as the aforementioned [5], since it only tests robustness, and not new unseen features encoding. This is a natural next step for this project, but it could not be met in this current investigation due to a time limitation. Across every working condition, accuracy on this set stays within a few points of validation. The protocols are thus robust to new viewpoints, augmentations, and harder line-ups. It is not mentioned on Figure 5 since we do not believe it is relevant to evaluate actual communication protocol metrics but rather to check the sanity of our protocol.

f) Population — Sender Agreement.: Sender agreement only applies to the population. It asks whether the different senders speak the same language as each other: it counts how often two randomly chosen senders, shown the same scene, produce exactly the same message. A value of 1 would mean they are perfectly in sync, and 0 would mean every sender invented its own private language.

Our population reaches a sender agreement of 0.513. It is not a full convergence to a single shared protocol, but it is well above zero and shows that the senders genuinely share a large part of their language. Read together with the higher topographic similarity, this is the core evidence that social dynamics push the population towards a more shared and more structured protocol than a single isolated pair produces.

Table III brings all the metrics together. For the population, every value is taken from the initial pair (sender 0, receiver 0).

VI. CONCLUSIONS

This project set out to study whether internal constraints and social dynamics shape the language that emerges between artificial agents, and at what point those forces produce qualitatively different communication. The morphological and

grounding metrics (message length, symbol reuse, purity, and the token–shape singular value decomposition) have now been computed on every condition; the one piece still outstanding is a proper zero-shot generalisation test, which we come back to at the end. We read the two hypotheses against the specific objectives of Section II (Table I).

Our first hypothesis was that internal pressure reorganises the protocol not gradually but at identifiable points during training, and the baseline run supports it clearly (O5). The protocol does not drift smoothly towards its final form: it goes through three distinct phases, and the transition is sharp. Once the length penalty engages, mean message length collapses (from around 7.4 to roughly 2.3 tokens in a single evaluation window), accuracy dips while the receiver lags behind, and only afterwards does the system stabilise at an accuracy higher than before compression (0.858, over the pre-compression high of 0.758). The ablations then tell us what drives it (O6): removing the penalty ($\lambda = 0$) is the only condition that never compresses, while every penalty-active run converges to a comparable short length regardless of vocabulary size, so the compression is a consequence of the least-effort pressure (and not of $|\mathcal{V}|$). The $|\mathcal{V}| = 5$ ablation reinforces this from the other side: with only four content symbols the sender is forced to recycle them, giving the highest symbol reuse of any working run.

Grounding, read against the right base rates, is more positive than accuracy alone would suggest. Shape purity (0.69) sits above its 0.45 base rate (cube is the dominant shape in about 45% of scenes), roughly 1.5 $\times$ chance, and colour purity (0.19) sits at roughly 1.8 $\times$ its 0.107 chance level. Both attributes are partially grounded, with colour somewhat stronger, consistent with colour being the attribute that best separates the candidates as seen in Figure 4. The compressed protocol therefore grounds partially in both shape and colour, more cleanly in colour, while likely also leaning on features we did not label (position, gradients, proportions). A way to see it is that

⁵The $K = 10$ run scores highly on Distinct Appearances only because it repeats a single long, uninformative message; this is an artefact of its collapse, not evidence of genuine symbol reuse (see Section V-B).

TABLE III
CROSS COMPARISON OF ALL CONDITIONS ON THE HELD-OUT VALIDATION SET (BEST CHECKPOINT).

Condition	Acc.	Len.	DA	Perp.	TopSim	PosDis	VoL	Pur _{sh}	Pur _{col}	ZS _K
Baseline	0.858	2.56	0.206	2.98	0.048	0.021	0.0169	0.687	0.194	0.706
$\lambda = 0.0$	0.927	6.91	0.314	4.89	0.045	0.009	0.0010	0.796	0.541	0.852
$ \mathcal{V} = 20$	0.864	2.44	0.169	5.22	0.000	0.009	0.0039	0.688	0.184	0.736
$ \mathcal{V} = 5$	0.832	2.48	0.388	3.14	0.030	0.018	0.0139	0.690	0.211	0.690
$K = 10$	0.090	5.78	0.509	8.47	-0.010	0.001	0.0003	0.935	0.842	0.051
Population 5×5	0.650	2.21	0.184	2.74	0.125	0.020	0.0209	0.695	0.183	0.434

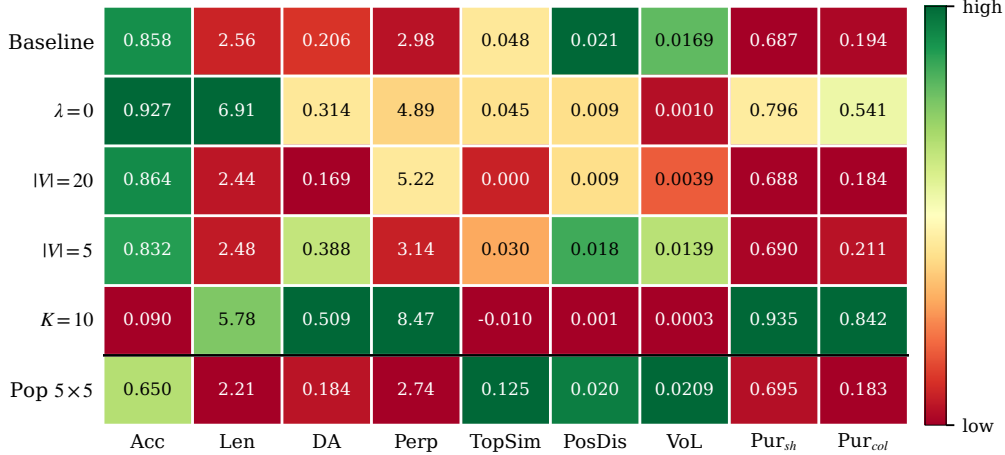


Fig. 5. Metric comparison across conditions. Lower is better for average length and perplexity.⁵

agents did not gravitate toward what an object conceptually or visibly is; they gravitated toward whatever most cheaply told one scene from another. In this case, grounding is less the discovery of the scene’s “true” categories than the negotiation of whichever distinctions happen to do the communicative work.

Our second hypothesis was that population training pushes communication towards a more shared and more structured protocol than an isolated pair, and the population results point in this direction (even if not overwhelmingly, O7). As expected, mean accuracy is lower than for a single pair (0.65, since each sender now has to be understood by several receivers). But the population reaches a higher topographic similarity (0.125) than any single pair (≤ 0.05) despite having the shortest messages of all conditions, and its senders agree on a large fraction of scenes (agreement peaks at 0.513 around step 19,000 and settles near 0.43, well above the zero of a private-language regime). Read together, these two facts are the core evidence that social dynamics add genuine structure: the agents are not just succeeding at the task, they are converging on a common way of describing scenes that an isolated pair does not develop.

It is worth being explicit about what is not fully settled. O7 is met only partially: it rests on a single population seed, and the population’s purity and SVD figures were taken from one

sender–receiver pair (sender 0, receiver 0) rather than averaged over the full $N \times N$ pairings, so they should be read as indicative. The harder-task axis is also incomplete: the $K = 10$ ablation was left behind (memory errors plus the time the investigation had), and its near-chance, single-message protocol survives only as a control. And the compositionality metrics (topographic similarity and positional disentanglement) hit a floor under one-to-two-symbol compression, while the per-symbol SVD cannot see combinations, so for $|\mathcal{V}| = 5$ we still cannot cleanly separate a holistic protocol from a compressed compositional one (a limitation of the metrics under heavy compression, not of the protocol itself).

The generalisation objective O8 is met in a robustness sense. On the held-out set, every working condition stays within a few points of its validation accuracy. The protocols therefore generalise to novel renderings and denser, harder scenes, but generalisation over genuinely novel attributes was not tested.

That opens some objective points to the most natural continuation of this work. The single agent and population experiments should be repeated across several seeds and scored over the full set of pairings (rather than a single extracted pair), the $K = 10$ ablation re-run once the memory issue is fixed, and the $|\mathcal{V}| = 5$ protocol probed with combination-aware metrics that a per-symbol decomposition simply cannot see. Calculating zero-shot over actually unseen features (new

shapes or colours) is also a very coherent next step to evaluate further the compositionality of the protocol.

ACKNOWLEDGEMENTS

Many thanks to Marco Schorlemmer (IIIA-CSIC / UAB) for his guidance, feedback, and continued support throughout the development of this project. This work was carried out as part of the Bachelor’s Thesis programme of the Grau en Intel·ligència Artificial at the Universitat Autònoma de Barcelona (UAB), academic year 2025–2026.

REFERENCES

- [1] P. W. Anderson, “More is different,” *Science*, vol. 177, no. 4047, pp. 393–396, 1972.
- [2] L. Wittgenstein, *Philosophical Investigations*. Oxford: Blackwell, 1953.
- [3] F. d. Saussure, *Course in General Linguistics*. New York: McGraw-Hill, 1916, originally published 1916; English translation by Wade Baskin.
- [4] S. Harnad, “The symbol grounding problem,” *Physica D*, vol. 42, no. 1–3, pp. 335–346, 1990.
- [5] A. Lazaridou, A. Peysakhovich, and M. Baroni, “Multi-agent cooperation and the emergence of (natural) language,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [6] L. Steels, “The synthetic modeling of language origins,” *Evolution of Communication*, vol. 1, no. 1, pp. 1–34, 1997.
- [7] D. Lewis, *Convention: A Philosophical Study*. Harvard University Press, 1969.
- [8] D. Rodríguez Luna, E. M. Ponti, D. Hupkes, and E. Bruni, “Internal and external pressures on language emergence: least effort, object constancy and frequency,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 4428–4437. [Online]. Available: <https://aclanthology.org/2020.findings-emnlp.397/>
- [9] R. Chaabouni, F. Strub, F. Altché, E. Tarassov, C. Tallec, E. Davoodi, K. W. Mathewson, O. Tieleman, A. Lazaridou, and B. Piot, “Emergent communication at scale,” in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: <https://openreview.net/forum?id=AUGBfDIV9rL>
- [10] A. Lazaridou and D. Rodríguez Luna, “Emergence of linguistic communication from referential games with symbolic and pixel input,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [11] R. Lowe, J. Pouget-Abadie, A. Tamar, J. Harb *et al.*, “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)*. Hong Kong, China: Association for Computational Linguistics, November 3–7 2019, pp. 3700–3710.
- [12] P. P. Liang, J. Chen, R. Salakhutdinov, L.-P. Morency, and S. Kottur, “On emergent communication in competitive multi-agent teams,” in *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, 2020.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *arXiv preprint arXiv:1512.03385*, 2015. [Online]. Available: <https://arxiv.org/abs/1512.03385>
- [14] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.
- [15] E. Jang, S. Gu, and B. Poole, “Categorical reparameterization with Gumbel-Softmax,” *arXiv preprint arXiv:1611.01144*, 2016.
- [16] S. Havrylov and I. Titov, “Emergence of language with multi-agent games: Learning to communicate with sequences of symbols,” in *arXiv preprint arXiv:1705.11151*, 2017.
- [17] E. Choi, A. Lazaridou, and N. De Freitas, “Compositional obverter communication learning from raw visual input,” in *6th International Conference on Learning Representations (ICLR) Conference Track Proceedings*, 2018. [Online]. Available: <https://openreview.net/forum?id=rknt2Be0->
- [18] R. Ben Zion, B. Carmeli, O. Paradise, and Y. Belinkov, “Semantics and spatiality of emergent communication,” in *38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.
- [19] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” *arXiv preprint arXiv:2103.00020*, 2021.
- [20] J. Peters, C. W. de Puiseau, H. Tercan, A. Gopikrishnan, G. A. L. D. Carvalho, C. Bitter, and T. Meisen, “Emergent language: A survey and taxonomy,” *arXiv preprint*, 2024.
- [21] R. Chaabouni, E. Kharitonov, D. Bouchacourt, E. Dupoux, and M. Baroni, “Compositionality and generalization in emergent languages,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2020, pp. 4427–4442. [Online]. Available: <https://aclanthology.org/2020.acl-main.407/>
- [22] S. Kottur, J. Moura, S. Lee, and D. Batra, “Natural language does not emerge ‘naturally’ in multi-agent dialog,” in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [23] D. Akkerman, P. Le, and R. G. Alhama, “The emergence of compositional languages in multi-entity referential games: from image to graph representations,” in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024, pp. 18 713–18 723. [Online]. Available: <https://aclanthology.org/2024.emnlp-main.1042/>
- [24] S. Verma and J. Dhar, “Emergence of writing systems through multi-agent cooperation (student abstract),” in *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 13 941–13 942, student abstract. [Online]. Available: <https://arxiv.org/abs/1910.00741>

APPENDIX

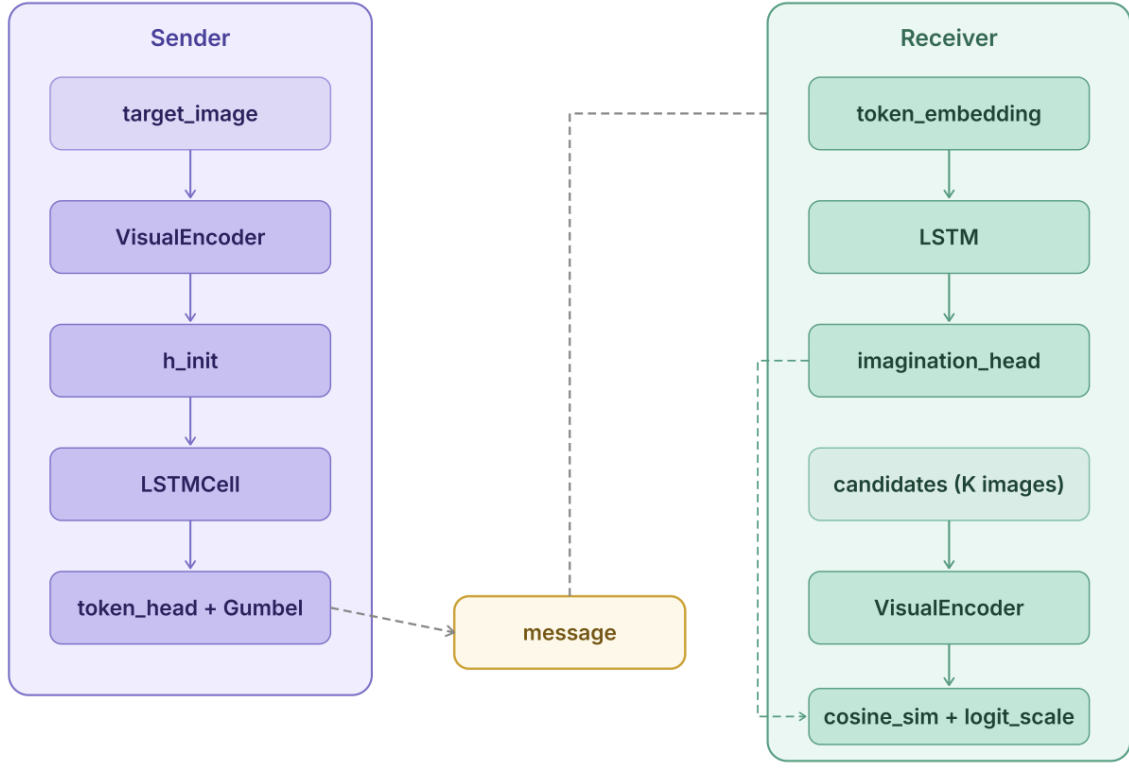


Fig. 6. Sender-Receiver architecture.

TABLE IV
ARCHITECTURAL HYPERPARAMETERS FOR THE FIRST EXPERIMENT.

Symbol	Value	Where	What it controls
$ \mathcal{V} $	10	Sender, Receiver	Vocabulary size: 1 EOS + 9 content symbols
L	8	Sender, Receiver	Maximum message length (hard cap)
K	5	Data, training	Candidates per round (1 target + 4 distractors)
d_{emb}	512	Sender CNN, Receiver CNN	Visual / imagination embedding size
d_{hid}	256	Sender LSTM, Receiver LSTM	LSTM hidden state size
d_{tok}	64	Sender, Receiver	Per-symbol embedding fed into the LSTM

TABLE V
TRAINING HYPERPARAMETERS FOR THE FIRST EXPERIMENT.

Symbol	Value	Where	What it controls
τ	$1.0 \rightarrow 0.5$	Sender	Gumbel-Softmax temperature (annealed)
T_0	0.07	Receiver	CLIP-style cosine logit temperature; learned
λ	0.1	Training loss	Weight on $\bar{\ell}$ for least-words pressure
T_{warm}	2,000	Training loss	Steps before λ is active
$H \times W$	228×228	Data	Image input resolution to the CNNs
Optimiser	Adam	Training	One optimiser over $\phi \cup \theta$
Learning rate	10^{-4}	Training	—
Batch size	64	Training	—
Train / val split	80/20	Data	Scene-level split, seeded for reproducibility

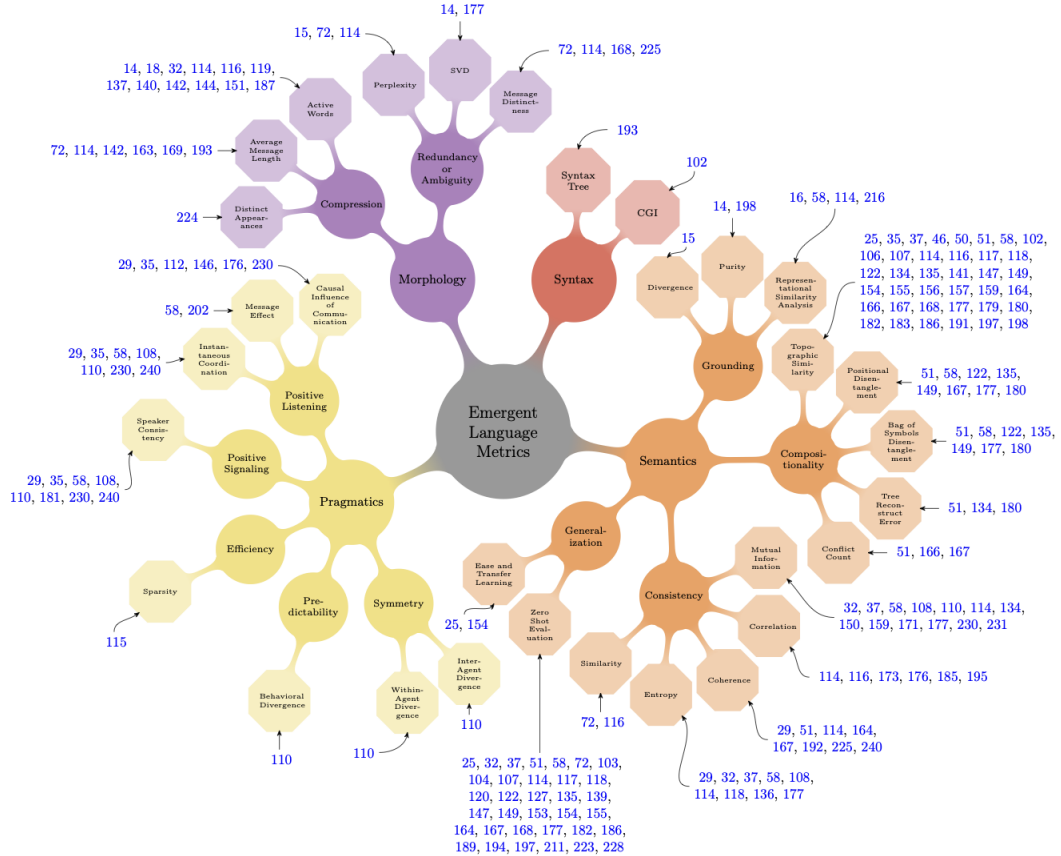


Fig. 7. Evaluation metrics map for emergent language. Numbers are citations from [20] and bear no relation to the bibliography of this investigation.