Curiosity-Driven Partner Selection Accelerates Convention Emergence in Language Games

Chin-wing Leung
University of Warwick
Coventry, United Kingdom
chin-wing.leung@warwick.ac.uk

Paolo Turrini
University of Warwick
Coventry, United Kingdom
p.turrini@warwick.ac.uk

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@ai.vub.ac.be

ABSTRACT

In language games a speaker and a listener attempt to coordinate on a shared mapping between words and concepts. The usual approach in the literature is to study convention emergence in well-mixed populations, where pairs of agents are randomly matched to play the role of speaker and listener, respectively. This way of pairing agents can be shown to promote the emergence of a unifying common language in the long run. Despite the theoretical guarantee, convention emergence can be very slow and practically unfeasible, especially in large populations with many words and concepts. Here, we propose an alternative approach, where we allow agents to selectively partner with other agents based on their past experience. To this aim, we study Boltzmann Q-learning agents that are curiosity-driven, i.e., more likely to choose partners they misunderstood in the past. We show that this selection method significantly accelerates convention emergence when compared against a random-matching baseline and is even more pronounced in graph generation models restricting agents' communication channels. By inspecting the evolution of the agents' interaction frequency we see that partner selection induces low treewidth and high degree variance at the early stages of learning, to then converge to a regular graph, which allows for settling misunderstandings in the population at a faster rate than the traditional approaches.

KEYWORDS

Language Games; Partner Selection; Reinforcement Learning

ACM Reference Format:
Chin-wing Leung, Paolo Turrini, and Ann Nowé. 2025. Curiosity-Driven Partner Selection Accelerates Convention Emergence in Language Games. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

The emergence of language is considered one of the distinctive features of intelligent social behaviour [21, 22] and it has been strictly linked with the development of a theory of mind [8], as well as the capacity of individuals to establish social ties and adapt to one another [10]. Partner selection plays a significant role in establishing social conventions and norms [20] and it is no surprise that both developmental and evolutionary psychology have emphasised the importance of social ties in language development [27, 31].

From the computational point of view, the problem of language coordination is more complex than a traditional coordination game [7]. Typical language games are characterised by large action spaces mapping words to concepts, as well as an asymmetric relation between a speaker and a listener. These two attempt to come up with a shared mapping between words and concepts by repeated interactions, what is commonly known as a Naming Game (NG) [29]. The features of the NG make the complexity of equilibrium computation nontrivial and, as we shall see, naive learning approaches hardly scale up. Indeed, although feasible in theory [6, 33], learning in well-mixed populations is not guaranteed to be fast. With large action spaces, reaching uniform convergence in the induced uniformly random graphs is often unfeasible in practice. However, the tools of RL can come to our rescue. Partner selection can itself be modelled as a decision and, albeit often constrained by the environment, agents can be equipped with stochastic policies dictating who they would like to interact with. This approach was successfully attempted to solve social dilemmas [1], extending Q-learning agents with the possibility of keeping Q-values for each potential partner and choosing accordingly. Language games present challenges of their own, as the occurrence of positive payoffs (e.g., successful communication) is generally sparse, especially in the initial learning phase, where as a result of randomisation each agent speaks a different language with high probability. One might argue that partner selection can only worsen the emergence of unifying conventions, as agents would be taking local decisions on who to interact with only based on their own utility. Indeed, this is what would happen if agents chose to be paired with those they understood the most, quickly leading to closed linguistic communities. As it turns out, the opposite direction, incentivising agents to connect to unlike-minded partners, induces dynamic interaction structures that quickly promote unifying conventions.

Contribution. In this paper, we study language games where agents autonomously learn a mapping between words and concepts and, for the first time, equip them with the capacity to choose the partners whom to interact with and learn from. We model our agents as Boltzmann Q-learners that are curiosity-driven, i.e., more likely to choose partners they misunderstood in the past. Agents keep track of the Q-values for their potential partners, estimating the expected reward associated with the interaction. This is a measurement of how well they expect to understand others and the basis for choosing them as partners. We show that this learning method significantly accelerates convention emergence when compared against a random-matching baseline, using established metrics such as Average Communicative Efficacy, Percentage of

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

Agents Converged to a Convention and Dominant Lexicon Specificity. By inspecting the evolution of the interaction frequency, i.e., how often agents are chosen by others, we see that curiosity-driven partner selection induces low treewidth and high degree variance at the early stage of learning, to then approximate a regular graph later on. In other words, some agents are chosen significantly more often than others early on, to then revert to the mean at the later stages. Intuitively, the initial tree-like structure is beneficial for building fewer and stronger local conventions quickly, while the regular graph structure at the later stage merges them into one, which turns out to promote convention emergence at a faster rate than the traditional approach. We also considered language coordination in restricted communication channels, where agents can only talk to a subset of others, implementing them as regular, scale-free and small-world networks. The result shows that the performance gap between partner selection and random matching is even more pronounced. Finally, we look at the role of inverse temperature, which determines how curiosity-driven agents are, showing the effect of different values on convergence rates.

Related Research. The language coordination problem has drawn the attention of researchers in the computational and social sciences for decades, with various attempts to come up with consensus-reaching mechanisms.

Simple spreading approaches based on strategy copy-transfer [2] were shown not to be sufficient to achieve consensus and, leveraging the tools from evolutionary algorithms, they were extended to define complex agent behaviours on information transfer, lexicon selection, innovation and self-protection [26]. The enriched framework showed promising results in terms of convergence speed and reached largely shared conventions, under various graph generation models, e.g., scale-free and small-world networks. However, besides not guaranteeing consensus, this approach requires extensive additional built-in architecture. To resolve the issue, [12] introduce influencer agents as network seeds and simplify the agent's behaviours, while achieving comparable performance to the previous method. Others [14] propose a network-aware utility for lexicon selection and allow agents to rewire their links to others, which contributes to faster convergence speed as well as overall quality.

None of these approaches, though, model populations of utility-maximising agents that learn from their local interactions, and instead resort to spreading-based mechanisms, where agents disseminate and adopt language conventions according to some predetermined rules, or manipulate the network properties altogether, either forcing change by introducing seeds or granting access to other agents' connections for decision-making purposes.

In recent years, learning-based approaches have arisen as a distributed optimisation framework for the emergence of language among self-interested agents. Inspired by double-Q learning [35], two learning algorithms, multiple-Q and multiple-R, were proposed [36], which significantly improve the speed of convergence in multiagent scenarios. However, the proposed algorithms were shown to work on multi-stage pure coordination games, which are an oversimplified version of the classical NG [29]. The emergence of conventions was also studied on static networks from an RL point of view, showing the effect of various network structures on convergence [28].

Multiagent reinforcement learning (MARL) methods have proven effective in striking a good balance between the complexity of the language game formulation and the requirement for learning among autonomous decision-makers. The language coordination problem was formulated as a MARL problem using bidirectional dynamic Q-learning [34], demonstrating that full coordination can be achieved with distributed learning as well, matching the results obtained using the more constrained diffusion models. [24] show that agents are able to learn a more stable and structured language in the reconstruction games when they are trained at different learning speeds. Others [11, 15, 19] study the effect of population size and communication networks among agents on the learning outcome in the referential games.

While the MARL approach showed significant progress in terms of learning capabilities, the underlying well-mixed population dynamics were not tested for large-scale problems, which provides an important bottleneck when we want to achieve coordination in a complex state space. This paper is about driving the randomness of population dynamics in language games using a partner selection approach, where we optimise the agents' exploration-exploitation tradeoff to select who to interact with. A similar approach combining partner selection with Q-learning was successfully employed when studying the emergence of cooperation in social dilemmas [1], where agents use epsilon-greedy Q-learning to select partners that have given them higher rewards in the past. Even minimal forms of partner selection were shown to be key in leading to a cooperative society among Q-learning agents [16].

All in all, we know from evolutionary biology that cooperation does not emerge in societies of self-interested agents unless a specific mechanism is at play [20]. Coordination games do not share the same incentive structures of social dilemmas, and the emergence of consensus is still possible without partner selection. However, as it turns out, having a say on whom to interact accelerates learning across the board and makes practically unfeasible coordination problems achieve a successful solution.

Paper Structure. In Section 2 we provide the necessary technical background describing the language coordination problem as well as Boltzmann and bidirectional Q-learning. Section 3 describes the algorithm we use for partner selection, the communication restrictions and the experimental setup. Section 4 analyses the results. We conclude by discussing various other potential directions.

2 PRELIMINARIES

We first review the problem of language coordination, introducing the Naming Game (NG) as well as the performance metrics used in language coordination. We then recall Q-learning and bidirectional Q-learning for NG.

2.1 The Language Coordination Problem

Following [12, 14, 26] our framework features a population of N agents aiming to develop a common lexicon (word-to-concept mapping) through local interactions. We refer to C as the set of concepts and W as the set of words, assuming throughout that $|C| = |W|$ and thus ideally seeking a bijective function that is uniformly shared in the population. The language coordination is attempted through repeated plays of the Naming Game (NG), which is an interaction

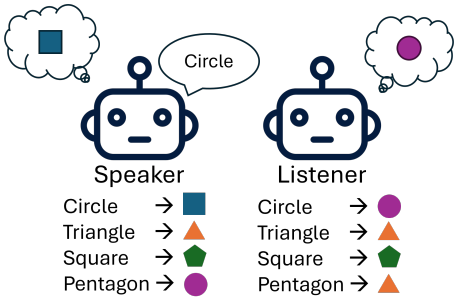


Figure 1: An example of an unsuccessful communication in the Naming Game (NG). The speaker's intended choice of word does not match the listener's understanding.

happening between two agents drawn from the population following some distribution. The agents will revise their lexicon based on the outcome of the game, and the goal is to achieve the highest communication efficacy with others.

2.1.1 The Naming Game. Two agents, drawn from a distribution, take the role of speaker and listener, respectively. Then, a concept $c \in C$ is randomly selected and revealed to the speaker. The speaker will utter the word $w \in W$ referring to that concept in their own language, i.e., will select the word w matching c according to their current policy. Upon hearing the word, the listener will interpret it into a concept $c' \in C$ and respond. The game is successful if the response matches the original concept $c = c'$, with both agents receiving the reward of 1; and receiving -1 each, otherwise. Figure 1 shows an example of an unsuccessful communication.

2.1.2 Performance Metrics. To evaluate the agents' communicative performance and the quality of the produced lexicon, we adopt four key metrics used in previous research [12, 26]:

Average Communicative Efficacy (ACE): It measures the percentage of communicative success in NG when the agents are matched uniformly at random, reflecting the global level of coordination of the current population.

Percentage of Agents Converged to a Convention (ACC): The mode lexicon is defined as the lexicon used by the most number of agents. The ACC measures the size of the mode convention in the population.

Dominant Lexicon Specificity (DLS): The quality of a lexicon is measured by the lexicon specificity, i.e., the proportion of words that can identify a single concept. Let W_c be the set of words that are mapped to the concept c in a given policy. The specificity of concept c in that given policy is then evaluated as $S_c = \frac{1}{|W_c|}$. If no word is mapped to a concept, then $S_c = 0$. The lexicon specificity S is defined as the average of the specificity of all concepts:

$$S = \frac{\sum_{c \in C} S_c}{|C|}$$

Therefore, the lexicons with maximum specificity ($S = 1$) are those with one-to-one mappings. In that case, the communication between agents is unambiguous. Note that a lexicon with 100% specificity would not be particularly useful if adopted by only a few agents, therefore the DLS measures the specificity of the mode lexicon, which is adopted by the highest number of agents.

Number of Distinct Lexicons (NDL) : This is the number of distinct lexicons, i.e., mappings, adopted by the agents. The maximum value equals N, where every agent has its own language. The minimum value equals 1, where everyone shares the same lexicon.

In our experiments, we will use these metrics to analyse the effectiveness of partner selection strategies and compare them to random matching, over various network structures and hyperparameters.

2.2 Q-Learning

Our agents use Q-learning, a widely established Reinforcement Learning algorithm [37] and act on a Markov decision process (MDP). Each agent maintains a Q-table for each state-action pair (s, a) to estimate the expected accumulated reward of using each action $a \in \mathcal{A}$ under each state $s \in \mathcal{S}$. Suppose the agent at state s has performed action a, let G be the accumulated reward, then the Q-value for the state-action pair (s, a) is updated as follows:

$$Q(s, a) \leftarrow Q(s, a) + \alpha(G - Q(s, a)) \tag{1}$$

where $\alpha \in (0, 1)$ is the learning rate. An exploration mechanism aims to strike a good balance between exploitation and exploration such that the performance of the agent is maximised during learning while ensuring the convergence guarantees are met. Boltzmann exploration is a commonly used mechanism, where the stochastic policy $\boldsymbol{\pi}(s) = (\pi(s, a_1), ..., \pi(s, a_N)) = (\pi_1, ..., \pi_N) \in \Delta$ is evaluated as

$$\pi_i = \frac{e^{\tau Q(s,a_i)}}{\sum_{j=1}^{d} e^{\tau Q(s,a_j)}} \tag{2}$$

where τ is a parameter known as the inverse temperature.

2.2.1 Bidirectional Q-learning for the Naming Game. Bidirectional Q-learning is a method proposed in the literature [34] for solving the problem of language coordination. By exploiting the speaker-listener dual relationship, each agent stores a single Q-table for decision-making purposes. Let $Q(c, w)$ be the Q-value of a concept-word pair. The action is then selected greedily, that is, the word is selected by $w = argmax_w Q(c, w)$ if the agent is in the speaker role, and the concept is selected by $c' = argmax_c Q(c, w)$ if the agent is in the listener role. In this manner, we are able to construct the lexicon (the word-to-concept mapping) and the reverse lexicon (the concept-to-word mapping) for each agent. The lexicon will be used to compile various statistics (ACC, DLS and NDL) to evaluate the status of convention emergence. Note that compiling the same statistics for the reverse lexicon would yield the same results, therefore we disregard the reverse lexicon in our discussion.

Upon receiving the reward r from the NG, the Q-value is updated accordingly.

$$Q(c, w) \leftarrow Q(c, w) + \alpha_{NG}(r - Q(c, w)) \tag{3}$$

for the speaker, and

$$Q(c', w) \leftarrow Q(c', w) + \alpha_{NG}(r - Q(c', w)) \tag{4}$$

for the listener, where α_{NG} is the learning rate for NG. To prevent the emergence of synonyms, the Q-values for other actions are

reduced when the agent encounters a success. That is

$$Q(c, w') \leftarrow Q(c, w') + \alpha_{NG}(-r - Q(c, w'))$$

for $w' \neq w$ for a successful speaker, and

$$Q(c'', w) \leftarrow Q(c'', w) + \alpha_{NG}(-r - Q(c'', w))$$

for $c'' \neq c'$ for a successful listener.

In [34] agents are paired at random, while concepts and words are chosen greedily. In our approach, we keep the greedy selection of concepts and words, but partners are instead selected through low temperature, e.g., curiosity-driven, Boltzmann exploration.

3 LANGUAGE WITH PARTNER SELECTION

Consider a population of agents learning to play the Naming Game (NG) from pairwise interactions. The goal is to come up with a choice of lexicon that maximises the reward from the NG through repeated gameplay. Our computational model is described in Algorithm 1. The agents are initialised with the learning rate for NG α_{NG}, the learning rate for partner selection (PS) α_{PS} and the inverse temperature for PS τ_{PS} (line 1). Depending on the existence of a constrained communication network, this is initialised with the graph-specific parameters (line 2, see Section 3.1 for the network models). At each iteration, each agent i selects an opponent from their neighbours to play the NG sequentially (line 5). Thus, the set of available actions is $A_{PS} = \{agent_1, ..., agent_{|N(i)|}\}$ for state $s = PS$, where $N(i)$ is the set of agents in the population except itself, or the neighbours for agent i when a constrained communication network is considered. Partner selection is carried out through Boltzmann exploration. NG is then initialised (lines 6-10, see Section 2.1.1 for the game description), where the indices S and L indicate the roles of speaker and listener, respectively. Upon receiving the reward r, both agents will update their Q-value for the NG accordingly (lines 11-12). The agent who has made the selection will then update the corresponding PS policy (line 13).

Algorithm 1 Language coordination with partner selection

Input: $Na, Nc, T, \alpha_{NG}, \alpha_{PS}, \tau_{PS}, Gp$

1: Initialize $Agents$ with $N, \alpha_{NG}, \alpha_{PS}, \tau_{PS}$
2: Initialize $Graph = randomNetwork(Gp)$
3: **for** $iteration = 1$ to T **do**
4: **for** $i = 1$ to N **do**
5: $j = Agents[i].choosePartner(Graph[i])$ with eq.(2)
6: $S, L = randomAssign([i, j])$
7: $c1 = randomChoice(Nc)$
8: $w1 = Agents[S].chooseWord(c1)$
9: $c2 = Agents[L].chooseConcept(w1)$
10: $r = 2 * (c1 == c2) - 1$
11: $Agents[S].trainLC(c1, w1, r)$ with eq.(3)
12: $Agents[L].trainLC(w1, c2, r)$ with eq.(4)
13: $Agents[i].trainPS(j, r)$ with eq.(1)
14: **end for**
15: **end for**

3.1 Constraining the Communication Network

As we shall see, partner selection induces specific networks on the interaction frequencies of the agents, when starting from a complete graph. On top of that, we can study the partner selection process on a priori constrained communication networks, where agents can only perform partner selection with a subset of others, following the hypothesis that constraining the interaction structure has an effect on convention emergence, as well as language [12, 14, 28]. Intuitively, if the network is extremely sparse, a unifying convention will be less likely to emerge fast. Compared to communication without restriction (the complete graph), introducing a network structure will increase the separation between agents, favouring the emergence of local conventions at the expense of a collective shared one. Especially when evaluating the resulting language at the population level using random matching, highly irregular graph structures will necessarily hinder successful communication. However, as we shall see, partner selection can still have pronounced positive effects then.

Parameters such as degree distribution, centrality and starting conditions of central highly influential nodes will be key in determining how a communication network will shape the resulting convention. We compare three main graph generation models:

Regular (RG) We generate them using the configuration model [30], setting the node degree to be 20.

Scale-Free (SF) We generate them using the Barabási–Albert model [3] setting the initial sampling parameter to be $m = 11$. The network is formed by sequentially adding a node with m edges preferentially attached to high-degree nodes until $N = 100$. The average degree is 19.58, the closest to an average degree of 20 we can get in a BA graph [1].

Small-World (SW) This is generated using the Watts–Strogatz model [38] setting parameters $k = 20, p = 0.3$, The network is formed by first creating a ring of $N = 100$ nodes joined to its k nearest neighbours, then each edge (u, v) is randomly replaced with a new edge (u, w) with probability p. The average degree is 20.

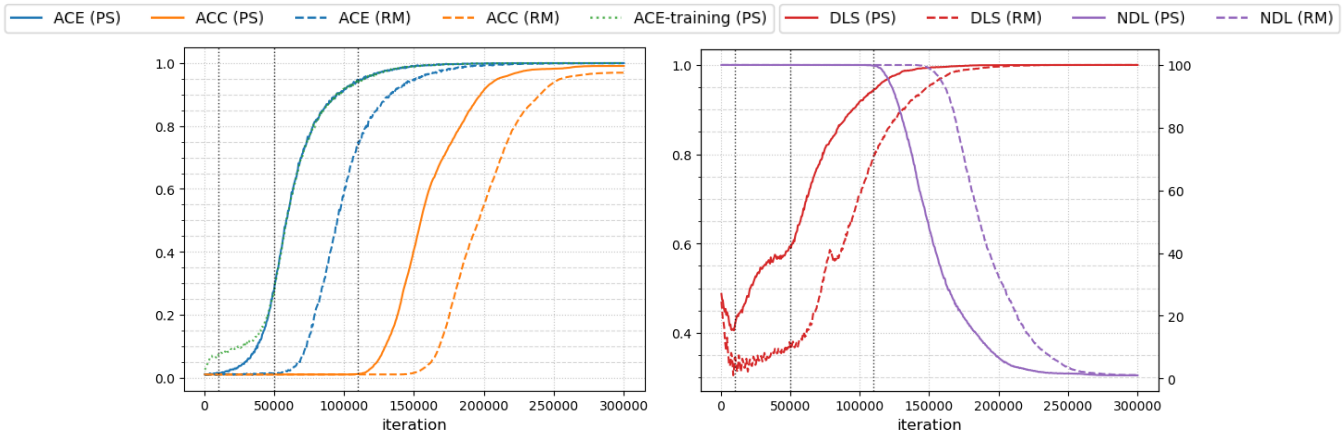
Each graph generation model will output the network that will be fixed throughout training and evaluation. All networks are generated with the NetworkX library [13].

3.2 Experimental Parameters

In our experiments, we work with a population of $N = 100$ agents, each attempting to map a set of $|W| = 100$ words, to $|C| = 100$ concepts. The population performance is evaluated every 100 iterations, and measured against the Average Communicative Efficacy (ACE) and the percentage of Agents Converged to a Convention (ACC). Partner selection will typically induce interaction frequencies that do not follow a regular graph. To avoid biasing the evaluation by correlating it to the induced network structure, we conduct it using random matching. In other words, we still judge language performance in a population by disregarding agents' connections.

We conducted the experiments up to 300, 000 iterations, and all the statistics are averaged over 100 simulations. The learning rate

[1] We have conducted the same experiments with $m = 12$, resulting in 21.12, the closer to 20 from above average degree achievable in a BA. Results and conclusions remain the same.



(a) Population Performance **(b) Mode Lexicon Quality**

Figure 2: Figure (a) shows the performance of agents in the language game across iterations averaged over 100 simulations, when trained under random matching (RM) and partner selection (PS), measured by the Average Communicative Efficacy (ACE) and the percentage of Agents Converged to a Convention (ACC). The Average Communicative Efficacy during training (ACE-training) is also presented for agents trained under PS. Figure (b) shows the statistics on the most popular lexicon, measured by the Dominant Lexicon Specificity (DLS), and the Number of Distinct Lexicons (NDL) in the population. The learning rate for the NG is $\alpha_{NG} = 0.7$, and the learning rate and the inverse temperature for partner selection are $\alpha_{PS} = 0.05$ and $\tau_{PS} = -15$, respectively. The results are averaged over 100 simulations. The dashed lines are marked to indicate the ending of each phase when agents are learning under PS. Agents trained under PS clearly outperform those trained under RM.

α_{NG} for NG is optimised over $[0, 1]$ with step size 0.05, to achieve the best ACC at $T = 300,000$ under random matching, in the case of the complete graph. This serves as the baseline for the best performance achieved on language coordination without partner selection. The learning rate α_{PS} and inverse temperature τ_{PS} for partner selection are optimised through a grid search over $[0, 1]$ with step size 0.05, and $[-25, 25]$ with step size 5, respectively, to achieve the best ACC at the end of training. The optimised parameters are $\alpha_{NG} = 0.7, \alpha_{PS} = 0.05$ and $\tau_{PS} = -15$.

4 EXPERIMENTAL ANALYSIS

In this section, we present our results, showing how partner selection accelerates convention emergence in language games. We then look at the evolution of the interaction frequency graph, tracking how partner selection shapes the graph structure during learning. We also look at the effectiveness of partner selection across restricted communication networks obtained by graph generation models. Finally, we zoom in on the key role of the inverse temperature τ_{PS} in convention emergence.

4.1 Partner Selection versus Random Matching

When agents are allowed to select their partner actively, convention emerges much quicker than it would under random matching. Figure 2a displays the population performance in language games for learning under random matching (RM) and partner selection (PS), measured against Average Communicative Efficacy (ACE) and Percentage of Agents Converged to a Convention (ACC). The Average Communicative Efficacy during training (ACE-training) is also presented for agents trained under PS. Note the ACE during training and evaluation are identical for agents trained under RM. Figure 2b displays the statistics on the most popular lexicon, measured by the Dominant Lexicon Specificity (DLS), and the Number of Distinct Lexicons (NDL) in the population. We can see how the introduction of partner selection has largely promoted the speed of convergence and the quality of the most popular lexicon. It takes more than 255,000 iterations to achieve a 95% convergence rate under RM, while it takes around 210,000 to achieve the same level of convergence with PS.

Looking at the results in Figure 2, we can observe four different phases of learning under partner selection.

- Phase 1 (iterations 0 to 10,000). The quality of the lexicon has decreased. The ACE-training starts to be higher than the ACE in evaluation.
- Phase 2 (iterations 10,000 to 50,000). The ACE-training is significantly higher than the ACE in evaluation. The DLS ends the decrease and rises to 0.6.
- Phase 3 (iterations 50,000 to 110,000). The ACE-training matches the ACE in evaluation. The ACE and DLS grow steadily, and the communication success achieves 95%.
- Phase 4 (iterations 110,000 to 300,000). The ACC increases and goes up to 99% at the end of training, and the NDL decreases correspondingly.

In the first phase, we observe no performance difference between agents trained with partner selection and random matching. Yet, the lexicon quality is higher when we train the agents allowing for partner selection. Perhaps counterintuitively curiosity-driven agents get a higher ACE-training and are thus able to form local conventions quickly. This is because, at the beginning of training,

the communication is almost always unsuccessful (99%), as agents are effectively guessing at random. With the ability to choose their partner, agents tend to experience a loss with high probability, and are therefore more willing to select that partner once more. This encourages agents to choose a stable opponent to play the NG for some time, before changing their selection. It turns out this will reduce the discrepancy between different languages quickly and improve communications locally, since coordinating with a smaller number of agents is easier than in a larger pool, thus increasing the ACE-training. In the second phase, agents start to come up with a more accurate lexicon. On the other hand, agents trained under random matching have struggled to improve communications and the DLS keeps at a low level. In the third phase, The ACE-training and the ACE in evaluation match. This reflects agents having learnt to select their partners approximately randomly. This is key since, at the later stage of learning, we need to avoid the formation of local dialects, and random matching is an efficient way to prevent this. In the fourth phase, the ACE has reached 95%, meaning that agents' policies are approximately 95% the same. The final stretch has agents coming up with a high-quality lexicon seeing communication effectiveness maximised.

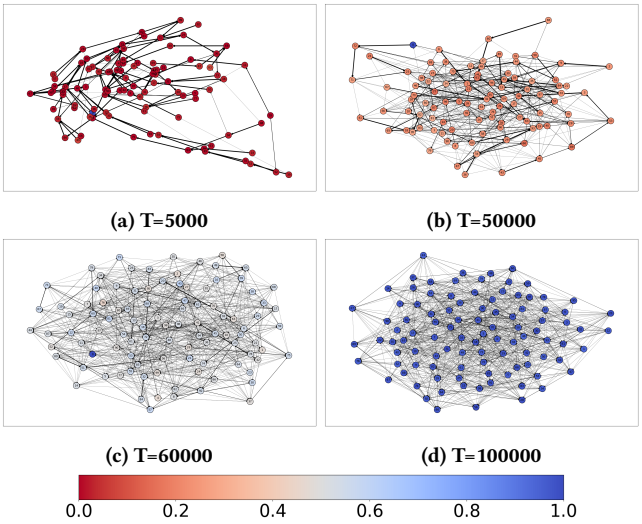


Figure 3: The partner selection frequency graph generated from a representative simulation, with nodes representing agents and edges representing partner choice. Lexicon similarity is calculated with respect to a reference agent to show how the population is converging towards the same convention. As lexicon similarity between agents grows, the selection graph turns from tree-like into a regular graph.

4.2 Interaction Frequency Graph

We present the partner selection frequency graph from a representative simulation in Figure 3. The nodes represent the agents in the population. The edges draw the directions from the partner selection over the last 10 iterations at specific times $T \in \{5000, 50000, 60000, 100000\}$. The thickness of the edge reflects the frequency of the same agent being selected - the thicker the edge the more often the agent is chosen by others. The colour of a node measures the lexicon similarity between that node and a reference agent 1, which is the percentage of word-to-concept matches between their lexicons. Blue represents high similarity, and red low similarity.

At the earlier stage of learning ($T = 5000, 35000$), the frequency graph exhibits a tree-like structure, where agents mainly choose the same partner over the last 10 iterations. As the similarity between agents rises, the graph becomes more like a regular graph. The tree structure is beneficial to maintaining the lexicon similarity at a sufficient level, as local conventions are passed over.

It is worth observing that, although we are only displaying a single simulation in Figure 3, the partner selection frequency graph exhibits a similar shape in all others. Figure 4 plots the treewidth of the partner selection graph presented above, as well as the variance of the in-degree, by summing all agents' partner selection policies. The statistics are averaged over 100 simulations. Treewidth is a standard measurement of how close a graph is to a tree [4]. A graph with a treewidth of 1 is exactly a tree, while a clique has a treewidth of $N - 1$, where N is the number of vertices. The treewidth is compiled using the minimum degree heuristic [5]. We also plot in-degree variance, as a dual measure [9]. We can see how the plots support our analysis of interaction frequency graphs.

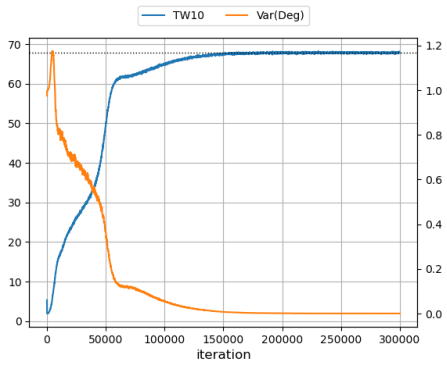


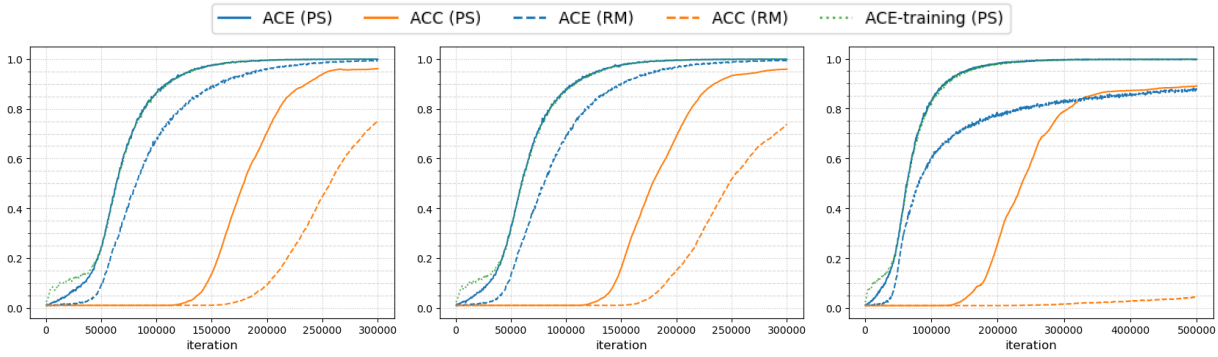
Figure 4: The blue line (with the left axis) presents the treewidth (TW10) statistics of the partner selection frequency graph across iterations, the corresponding treewidth statistics under random matching is 67.84 marked with the black dotted line. The orange line (with the right axis) presents the in-degree variance (Var(Deg)) from aggregating agents' partner selection policies across iterations, the corresponding in-degree variance under random matching is 0.

4.3 Constrained Communication Networks

Figures 5 and 6 show the evolution of the various performance metrics in constrained communication networks. We can see how agents trained under PS outperform those trained under RM. The latter ones have suffered from being separated into a variety of local conventions, which is reflected in the low value of ACC throughout the simulation. In a more challenging scenario of small-world networks, the ACC under RM is still at 1.52% after 300, 000 iterations, while, for PS, the value goes up to 78.88%. Compared to the case of the complete network, the performance gap on ACE between PS

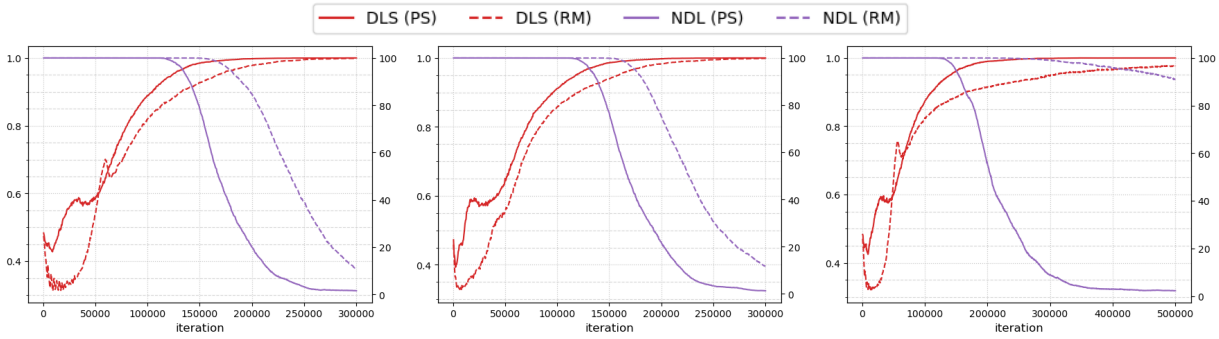
Table 1: Iterations needed to achieve specific ACCs under different interaction topologies

ACC (%)	*Complete*		*Regular*		*Scale-Free*		*Small-World*	
	RM	PS	RM	PS	RM	PS	RM	PS
80	222,200	183,400	-	211,100	-	213,400	-	306,300
85	229,300	189,800	-	219,100	-	221,300	-	336,100
90	240,100	197,700	-	235,500	-	234,300	-	-
95	255,300	209,800	-	255,000	-	277,700	-	-
ACC (%) at $T = 300,000$	96.94	99.09	75.15	96.11	73.69	95.89	1.52	78.88



(a) Regular (RG) **(b) Scale-Free (SF)** **(c) Small-World (SW)**

Figure 5: Performance across iterations over different communication networks, measured by the ACE, ACC, and ACE-training, for agents trained under PS. The average degree of the regular graph, the scale-free network, and the small-world network are 20, 19.58, **and** 20 **respectively. The convention emerges significantly faster when agents are trained using PS.**



(a) Regular (RG) **(b) Scale-Free (SF)** **(c) Small-World (SW)**

Figure 6: Mode lexicon statistics, measured by DLS and NDL in the population. The average degree for regular, scale-free and small-world networks are 20, 19.58, **and** 20, **respectively. DLS generally achieves the maximum faster when agents are trained under PS. The DLS under RM has raised above the DLS under PS for a short period for RG and SW.**

and RM has reduced when we have constrained communication networks. This is because the evaluation of communication success is based on matching agents with their neighbours, and a lower average degree (compared to the complete network) favours the emergence of local conventions. The ACC does not suffer from this fact and we therefore see a large performance gap when it is used as a measurement. When we look at the statistics for mode lexicon, the DLS achieve its peak faster when agents are trained under PS. Unlike for complete graphs, we can see the differences in DLS between PS and RM have reduced, and that the DLS under RM has raised above the DLS under PS for a short period under the cases of RG and SW. This is because, when a small number of agents coordinate locally, they will tend to improve lexicon quality, which in turn improves ACE, but also makes local conventions stronger and harder to merge. After all, a high-quality lexicon is not useful if the adoption rate is too low.

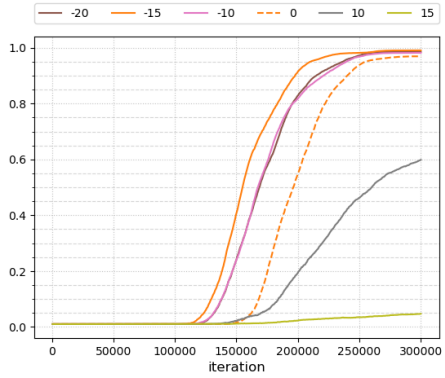


Figure 7: The ACCs across iterations for different inverse temperatures $\tau_{PS} \in \{-20, -15, -10, 0, 10, 15\}$ under PS, the learning rates are fixed at $\alpha_{NG} = 0.7, \alpha_{PS} = 0.05$. As τ_{PS} decreases, the rate of convergence increases, until τ_{PS} reaches -15 (the orange solid line). When $\tau_{PS} = 0$ (the orange dashed line), the partner selection policy is equivalent to random matching.

Table 1 displays the statistics for the iterations needed to achieve specific ACCs for constrained communication networks under RM and PS. We compared the benchmark convergence rates from 80% to 95%. In the complete graph, agents trained under PS always achieve the benchmarks about $40,000$ iterations faster than RM. In the case of regular graphs and scale-free networks, the agents trained under RM cannot even achieve an 80% convergence rate at the end of training at $T = 300,000$, while agents trained under PS achieve beyond 95% convergence at the end of training. For small-world networks, we have extended the training to $T = 500,000$ and the ACC for agents trained under PS attained 89.06% at the end of the training, while for RM still at the value of 4.34%. All in all, this illustrates the increased effectiveness of partner selection across the different communication networks.

4.4 On the Role of the Inverse Temperature

Finally, we examine the key role of inverse temperature τ_{PS} in partner selection, which dictates the mode of exploration the agents employ. The intensity of τ_{PS} affects the greediness of action selection, while the sign of τ_{PS} affects the preference towards larger or smaller Q-value. In the case of $\tau_{PS} = 0$, it reduces to the case of random matching. Figure 7 presents the ACCs across iterations for different inverse temperatures $\tau_{PS} \in \{-20, -15, -10, 0, 10, 15\}$.

We can see the performance of PS surpasses RM when τ_{PS} becomes negative. In this case, agents do not keep interacting with those other agents they have had success with but, rather, are more willing to connect to those with lower associated Q-values, in other words, they are curiosity-driven. Our study shows that “talking to strangers" is generally a good strategy, as the lower expected value of the partner choice allows for settling lexical differences faster. The result also explains the behaviour of agents during different phases of learning. In the earlier phases, the communication failures reduce the Q-values for certain neighbours, causing the agent to be more willing to communicate with them further. Therefore, we can see agents selecting a stable partner for a period of time before moving on to selecting others. In the later phases of learning, where some lexicon uniformity is established, the agents become focused on increasing the communication efficacy with every agent in a balanced manner, which results in partner selection converging towards random matching. If we considered the case of positive τ_{PS}, the above points would be reversed, and would then result in the emergence of a multitude of languages and therefore be worse in terms of global convention emergence.

When the value of τ_{PS} drops below -15, the performance on language coordination starts to reduce. This shows the greediness in partner selection should be at the "right" level, having a good balance between focusing on less familiar agents, as well as communicating with everyone effectively.

5 DISCUSSION

While the standard approach to convention emergence in language games hinges on the theoretical guarantee that well-mixed populations will eventually converge to a unique fully shared mapping between words and concepts, this is often unfeasible to achieve in practice. In this paper, we have shown that curiosity-driven partner selection leads to convention emergence by transforming the interaction structure in a distributed manner, settling misunderstandings among agents at a faster rate, without sacrificing the convergence guarantees. This is even more evident when constraining communication channels upfront, where random matching performs poorly in comparison to optimised Boltzmann Q-learners taking decisions on who to partner with.

Despite the progress, a number of research questions are still left unanswered. The interplay between language games and partner selection makes the theoretical analysis more challenging, as the main tool of evolutionary game theory for studying population dynamics, the replicator equation, abstracts away from spatio-temporal considerations [25]. This begs the question of whether a mean dynamics analysis can still provide meaningful insights on the policy change, perhaps under fixed partner selection strategies as identified in the simulations, or we would need to explore variance using stochastic dynamics [17, 32]. Moreover, while we know that a uniform convention will emerge quickly with partner selection, we do not know what this convention will look like, or whether some of its high-level properties could be predicted from the starting conditions. A similar point could be made regarding how agents select one another. Recent contributions [16, 18] have shown that partner selection rules co-evolve with in-game strategies to promote cooperation in social dilemmas. What the co-evolving partner selection rules are that sustain convention emergence in language games is unclear if agents can learn as well the hyperparameters. Finally, networks with a multitude of relationship types [23] and correlated starting lexicons are worth investigating.

ACKNOWLEDGMENTS

CL and PT acknowledge the support of the Leverhulme Trust for the Research Grant RPG-2023-050 and the TAILOR Connectivity Fund (Agreement 29). PT also acknowledges travel support from the European Union’s Horizon 2020 research and innovation programme under Grant Agreement No 951847. AN acknowledges the support of the Flemish Government (AI Research Program).

REFERENCES

[1] Nicolas Anastassacos, Stephen Hailes, and Mirco Musolesi. 2020. Partner Selection for the Emergence of Cooperation in Multi-Agent Systems Using Reinforcement Learning. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*. AAAI Press, 7047–7054. https://doi.org/10.1609/aaai.v34i05.6190

[2] Norman TJ Bailey. 1975. *The mathematical theory of infectious diseases and its applications*. Number 2nd ediition.

[3] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *science* 286, 5439 (1999), 509–512.

[4] Umberto Bertele and Francesco Brioschi. 1973. On non-serial dynamic programming. *J. Comb. Theory, Ser. A* 14, 2 (1973), 137–148.

[5] Hans L Bodlaender and Arie MCA Koster. 2010. Treewidth computations I. Upper bounds. *Information and Computation* 208, 3 (2010), 259–275.

[6] Tilman Börgers and Rajiv Sarin. 1997. Learning Through Reinforcement and Replicator Dynamics. *Journal of Economic Theory* 77, 1 (1997), 1–14. https://doi.org/10.1006/jeth.1997.2319

[7] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence* (Madison, Wisconsin, USA) *(AAAI '98/IAAI '98)*. American Association for Artificial Intelligence, USA, 746–752.

[8] Jill de Villiers. 2007. The interface of language and Theory of Mind. *Lingua* 117, 11 (2007), 1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006 Language Acquisition between Sentence and Discourse.

[9] Reinhard Diestel. 2012. *Graph Theory, 4th Edition*. Graduate texts in mathematics, Vol. 173. Springer.

[10] Robin Dunbar. 1998. *Grooming, Gossip and the Evolution of Language*. Harvard University Press.

[11] Nicole Fitzgerald. 2019. To populate is to regulate. *arXiv preprint arXiv:1911.04362* (2019).

[12] Henry Franks, Nathan Griffiths, and Arshad Jhumka. 2013. Manipulating convention emergence using influencer agents. *Autonomous Agents and Multi-Agent Systems* 26 (2013), 315–353.

[13] Aric Hagberg, Pieter J Swart, and Daniel A Schult. 2008. *Exploring network structure, dynamics, and function using NetworkX*. Technical Report. Los Alamos National Laboratory (LANL), Los Alamos, NM (United States).

[14] Mohammad Hasan, Anita Raja, and Ana Bazzan. 2015. Fast convention formation in dynamic networks using topological knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[15] Jooyeon Kim and Alice Oh. 2021. Emergent communication under varying sizes and connectivities. *Advances in Neural Information Processing Systems* 34 (2021), 17579–17591.

[16] Chin-wing Leung and Paolo Turrini. 2024. Learning Partner Selection Rules that Sustain Cooperation in Social Dilemmas with the Option of Opting Out. In *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems / ACM, 1110–1118.

[17] Chin-wing Leung, Shuyue Hu, and Ho-fung Leung. 2024. The Stochastic Evolutionary Dynamics of Softmax Policy Gradient in Games. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (Auckland, New Zealand) *(AAMAS '24)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1101–1109.

[18] Chin-wing Leung, Tom Lenaerts, and Paolo Turrini. 2024. To Promote Full Cooperation in Social Dilemmas, Agents Need to Unlearn Loyalty. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization, 111–119. https://doi.org/10.24963/ijcai.2024/13 Main Track.

[19] Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, and Angeliki Lazaridou. [n.d.]. Revisiting Populations in multi-agent Communication. In *The Eleventh International Conference on Learning Representations*.

[20] Martin A. Nowak. 2006. Five Rules for the Evolution of Cooperation. *Science* 314, 5805 (2006), 1560–1563. https://doi.org/10.1126/science.1133755 arXiv:https://www.science.org/doi/pdf/10.1126/science.1133755

[21] Martin A Nowak and Natalia L Komarova. 2001. Towards an evolutionary theory of language. *Trends in Cognitive Sciences* 5, 7 (2001), 288–295. https://doi.org/10.1016/S1364-6613(00)01683-1

[22] Martin A. Nowak and David C. Krakauer. 1999. The evolution of language. *Proceedings of the National Academy of Sciences* 96, 14 (1999), 8028–8033. https://doi.org/10.1073/pnas.96.14.8028 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.96.14.8028

[23] Davide Nunes and Luis Antunes. 2015. Modelling structured societies: A multi-relational approach to context permeability. *Artificial Intelligence* 229 (2015), 175–199.

[24] Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. 2022. On the role of population heterogeneity in emergent communication. *arXiv preprint arXiv:2204.12982* (2022).

[25] Carlos P. Roca, José A. Cuesta, and Angel Sánchez. 2009. Evolutionary game theory: Temporal and spatial effects beyond replicator dynamics. *Physics of Life Reviews* 6, 4 (2009), 208–249. https://doi.org/10.1016/j.plrev.2009.08.001

[26] Norman Salazar, Juan A Rodriguez-Aguilar, and Josep L Arcos. 2010. Robust coordination in large convention spaces. *Ai Communications* 23, 4 (2010), 357–372.

[27] Bambi Schieffelin and E. Ochs. 1984. *Language acquisition and socialization: Three developmental stories and their implications*. Cambridge University Press, United Kingdom, 276–320.

[28] Sven Van Segbroeck, Steven de Jong, Ann Nowé, Francisco C. Santos, and Tom Lenaerts. 2010. Learning to coordinate in complex networks. *Adapt. Behav.* 18, 5 (2010), 416–427. https://doi.org/10.1177/1059712310384282

[29] Luc Steels. 1995. A self-organizing spatial vocabulary. *Artificial life* 2, 3 (1995), 319–332.

[30] Angelika Steger and Nicholas C Worlmald. 2008. Generating random regular graphs quickly. *Combinatorics, Probability and Computing* 8, 4 (2008), 377–396.

[31] Michael Tomasello. 2010. *Origins of human communication*. MIT Press, Cambridge, Mass.; London.

[32] Arne Traulsen, Christoph Hauert, Hannelore De Silva, Martin A. Nowak, and Karl Sigmund. 2009. Exploration dynamics in evolutionary games. *Proceedings of the National Academy of Sciences* 106, 3 (2009), 709–712. https://doi.org/10.1073/pnas.0808450106 arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0808450106

[33] Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. 2003. A selection-mutation model for q-learning in multi-agent systems. In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems* (Melbourne, Australia) *(AAMAS '03)*. Association for Computing Machinery, New York, NY, USA, 693–700. https://doi.org/10.1145/860575.860687

[34] 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 (2022), 213–223.

[35] Hado van Hasselt. 2010. Double Q-learning. *Advances in neural information processing systems* 23 (2010).

[36] Yixi Wang, Wenhuan Lu, Jianye Hao, Jianguo Wei, and Ho-fung Leung. 2018. Efficient convention emergence through decoupled reinforcement social learning with teacher-student mechanism. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 795–803.

[37] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3 (1992), 279–292.

[38] Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature* 393, 6684 (1998), 440–442.