Universiteit Leiden
The Netherlands

Data Science & Artificial Intelligence

Emergent Communication in Aging Populations

Jan Dziewoński

Supervisors:
Tessa Verhoef & Flor Miriam Plaza del Arco

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/12/2025

Abstract

When neural network agents learn to communicate through interaction, population turnover poses a challenge: as agents are replaced, the shared language can drift, degrading communication between generations. This thesis investigates whether age-dependent plasticity, where younger agents learn quickly and older agents maintain stable representations, can reduce language drift in populations with turnover. The hypothesis draws on evidence from human language acquisition, where children's high plasticity and adults' stability may together preserve cross-generational intelligibility.

Three experiments test this hypothesis using populations of symmetric agents playing referential games. Experiment 1 establishes that static populations of 2 to 20 agents reliably develop shared languages. Experiment 2 demonstrates that turnover causes cumulative drift: cross-generational communication degrades regardless of turnover rate or population size, though smaller populations and slower turnover preserve intelligibility longer. Experiment 3 compares uniform plasticity against age-based plasticity under turnover. Populations with uniformly low plasticity fail because agents cannot adapt quickly enough to integrate newcomers. Populations with uniformly high plasticity fail because the language changes faster than stable conventions can form. The age-based condition, combining both properties, substantially outperforms all alternatives on cross-generational metrics, with stable older agents anchoring the language while plastic younger agents efficiently acquire it.

These findings provide computational evidence that age-based plasticity substantially reduces language drift under turnover. The combination of high and low plasticity within a population proves more effective than either extreme alone, supporting the hypothesis that heterogeneous plasticity contributes to language stability in communicating populations.

Contents

1 Introduction

When autonomous agents learn to communicate, they must develop a shared language that all participants can use effectively. Research in emergent communication has shown that deep neural network agents, trained through reinforcement learning, can develop such languages from scratch to solve cooperative tasks [LPB17, FAdFW16, LB20]. A common scenario is the referential game, where the sender describes a target object and the receiver identifies it among distractors [LB20]. Early work in this area focused on single sender-receiver pairs, but the languages emerging from pairwise training tend to be highly specialized and lack properties found in human languages [CKDB19, BB18]. In response, researchers have scaled emergent communication to populations of multiple agents [CSA+22], introduced symmetric architectures where sender and receiver modules are tied together [LVB24], and explored introducing population turnover [CLL+20]. These modifications bring the simulations closer to the conditions under which human languages evolve, but they also introduce new dynamics, particularly the challenge of maintaining a stable, transmittable language. Population turnover creates pressure that static populations do not face. Newcomers must learn the language from existing agents, and agents who entered at different times must remain mutually intelligible. As each group learns from its predecessors, the language naturally drifts: symbols shift in usage, and agents from distant generations may struggle to communicate even if each generation performs well internally. Prior work has examined how turnover affects compositionality and learnability [CLL+20, LB19], but less attention has been paid to mechanisms that might regulate drift and preserve cross-generational communication.

Human languages operate under similar dynamics, and one hypothesis from cognitive science proposes a mechanism that balances adaptability with stability [Gop20]. Children acquire language with high plasticity, readily adapting to the conventions around them, while adults maintain more stable representations [HTP18]. This asymmetry may allow languages to both evolve (thanks to flexible young learners) and maintain cross-generational intelligibility, with older speakers providing stable learning targets that younger speakers can efficiently acquire. Computational work supports this hypothesis: Verhoef and de Boer [VdB11] found that simulated populations with age-dependent learning rates preserved communication systems better than populations with uniform rates. This thesis investigates whether an analogous mechanism affects neural network agents playing referential games - a domain where the interaction between plasticity heterogeneity and language stability remains unexplored. Specifically, it examines whether introducing age-dependent plasticity into populations with turnover can improve language stability and cross-generational accuracy.

1.1 Background and Motivation

Emergent communication research has both practical and scientific motivations. On the practical side, learned communication protocols could enable flexible coordination in multi-agent systems. These applications have been explored in domains including autonomous navigation, robotic coordination, and distributed network management [BM24]. On the scientific side, emergent communication simulations provide a controlled environment for testing theories of language evolution: unlike human populations, where developmental, social, and historical factors blend with each other, simulations allow researchers to manipulate variables such as learning dynamics and population composition independently [BM24]. Realizing either goal, however, requires emergent languages that remain stable as populations change. When agents enter and leave a population over time, the

shared communication protocol can undergo drift: gradual, cumulative changes in the language. If drift proceeds too quickly, mutual intelligibility between older and newer agents degrades, and the language can lose its communicative function, become harder to analyze by researchers or deviate too much from the desired structure[LPT20]. Understanding how to regulate drift is therefore an important challenge for the field.

Several lines of prior work have examined how turnover and population structure shape emergent languages. Research on iterated learning demonstrates that imperfect transmission across generations creates pressure toward compositionality, i.e. a structure where parts of a message correspond to parts of the meaning (e.g., separate symbols encoding an object's shape and color) [LHTC18]. Ren et al. [RGL+20] showed that periodically resetting neural agents and requiring them to relearn from others amplifies compositional structure. Li and Bowling [LB19] introduced the concept of ease of teaching: when agents are periodically replaced, implicit selection favors languages that newcomers can learn quickly. Cogswell et al. [CLL+20] confirmed that such cultural transmission improves compositional generalization compared to static populations. Work on population heterogeneity complements these findings. Rita et al. [RSG+22] found that asymmetric speaker-listener dynamics produce higher compositionality and stability, indicating that variation in learning speed across agents shapes language properties. A gap remains, however. Studies of heterogeneity vary learning parameters but do not combine them with turnover dynamics. Turnover studies, in turn, typically employ homogeneous agents without modelling how plasticity might change over an agent's lifetime. Furthermore, much of this work uses asymmetric sender-receiver architectures rather than symmetric agents, and most studies measure compositionality or task success rather than cross-generational communication stability.

The hypothesis motivating this thesis - introducing age-based plasticity to populations with turnover - is inspired by evidence from human language acquisition. Hartshorne et al. [HTP18] analyzed data from nearly 670,000 English speakers and found that grammar-learning ability remains high until approximately age 17, after which it starts declining. Gopnik [Gop20] proposed that this developmental pattern reflects an exploration-exploitation trade-off: children's high plasticity enables adaptation to local linguistic conventions, whereas adults' more stable representations anchor the language against rapid drift. Computational work provides support for this claim. Verhoef and de Boer [VdB11] simulated vowel system emergence in populations where younger agents learned faster than older ones. They compared age-structured populations against control populations with consistently high or consistently low learning rates. The age-structured populations performed better than either control condition: mutual intelligibility between initial and final generations was higher, and the vowel systems maintained greater complexity. The authors attributed this to older agents providing stable learning targets for newcomers. This thesis extends their approach to neural network agents playing a variation of the referential game with explicit population turnover, testing whether age-dependent plasticity reduces language drift and improves cross-generational communication accuracy.

1.2 Research Questions and Objectives

The primary research question explored by this thesis is:

Does age-dependent plasticity - where younger agents learn faster and older agents maintain more stable representations - improve language stability and cross-generational communication in populations with agent turnover?

To address this question, the thesis pursues three objectives corresponding to a sequence of experiments:

1. Establish baseline behaviors by documenting how population size affects language emergence and task accuracy in static, homogeneous populations without turnover or aging mechanisms.

2. Characterize turnover effects by measuring how agent replacement impacts communication success, language drift and other language metrics across agents from different generations.

3. Evaluate age-dependent plasticity by comparing three configurations under turnover: homogeneous high-plasticity populations (all agents learn quickly), homogeneous low-plasticity populations (all agents learn slowly), and heterogeneous populations where plasticity decreases with agent age, comparing them to each other and the previous turnover-only scenarios.

The hypothesis motivating this research is that age-dependent plasticity may act as a stabilizing force against language drift. If younger agents adapt quickly to existing conventions while older agents anchor those conventions through lower plasticity, the population could balance linguistic adaptability with stability - similar to patterns observed in human language communities. Furthermore, I hypothesize that this stabilizing effect depends on heterogeneity: populations with mixed plasticity levels should outperform both uniformly high-plasticity populations (where no agent provides a stable foundation) and uniformly low-plasticity populations (where agents adapt too slowly to converge efficiently before they are replaced). I also expect that populations combining turnover with age-dependent plasticity will exhibit reduced drift and improved cross-generational communication compared to populations with turnover alone. However, this outcome is not guaranteed. If the hypothesis is not supported - if plasticity produces no measurable effect or even accelerates drift - it would suggest that the mechanism functions differently in neural network agents than in the human systems that inspired it, or that other factors dominate language stability in this setting. Either outcome would contribute to understanding the conditions under which age-based plasticity influences emergent communication.

1.3 Thesis Overview

The remainder of this thesis is structured as follows:

1. **Theoretical Background**
   This section covers the core concepts needed to understand the research. It explains the field of emergent communication, the mechanics of the referential game, the metrics used to track language drift, and the cognitive science behind aging and plasticity.

2. **Related Work**
   This section reviews the relevant literature. It discusses previous studies on multi-agent populations, the effects of turnover on language structure, different agent architectures, and the role of heterogeneity in populations.

3. **Methodology**
   This section describes the experimental framework. It details the extension of the EGG framework [KCBB19], the design of the game and agents, the logic behind the turnover and plasticity mechanisms, and the implementation of the evaluation metrics.

4. **Experiments and Results**
   This section presents the three experiments. Experiment 1 establishes baselines for static populations. Experiment 2 analyzes how agent turnover drives language drift. Experiment 3 tests the main hypothesis by comparing age-based plasticity against homogeneous control groups.

5. **Discussion**
   This section interprets the findings. It connects the results to theories of language evolution, discusses the limitations of the current model, and suggests directions for future research.

6. **Conclusion**
   This section summarizes the main contributions of the thesis and the key findings regarding age-dependent plasticity.

2 Theoretical Background

2.1 Computational Linguistics and Emergent Communication

The field of computational linguistics has traditionally focused on analyzing, modeling, and processing existing human languages. In these scenarios, models are typically trained in a supervised manner on large static datasets, learning to capture statistical regularities [LB20]. Emergent communication represents a noticeable departure from this approach. Instead of modeling a pre-existing language, this research simulates scenarios in which a communication system arises from scratch between autonomous agents. In these simulations, agents are not provided with a lexicon or grammar; rather, they must develop a shared protocol to coordinate their actions and solve cooperative tasks. Consequently, language is treated not as a static dataset to be learned, but as a dynamic functional tool that evolves through interaction [LB20].

The primary scientific motivation for simulating emergent communication lies in the difficulty of studying language evolution empirically. Unlike biological evolution, which leaves physical evidence, the origins of human language have left no fossil record, making it difficult to test hypotheses about how linguistic structure evolved from pre-linguistic signaling [WRUW03]. Computational simulations provide a solution to this problem by allowing researchers to conduct controlled experiments with artificial agents. By creating environments where the parameters of interaction, perception, and population dynamics can be manipulated, researchers can observe how these constraints influence the properties of the resulting communication system [BM24]. This approach allows for the isolation of specific variables, such as memory bottlenecks [KMLB17], channel noise, or critical period effects[VdB11], to determine their specific contributions to language stability and change. This thesis specifically investigates one such potentially stabilizing mechanism: age-dependent plasticity, examining how varying temperature and learning rate across an agent's lifespan affects the system's evolution.

While early work in this field utilized symbolic agents, the widespread adoption of deep reinforcement learning has fundamentally expanded the scope of emergent communication research [LB20]. In modern deep reinforcement learning setups, agents are typically modeled as neural networks that perceive raw high-dimensional inputs rather than pre-processed symbolic representations. This shift is significant because it forces the agents to solve the problem of grounding: they must learn

to map continuous perceptual inputs to discrete communicative symbols [LB20]. The agents in this thesis face this precise challenge, required to map feature vectors representing target objects to sequences of discrete symbols without any prior linguistic knowledge. Because the agents have no agreement on what symbols mean at the start of the training, the semantics of the emergent language are defined entirely by their utility in solving the task at hand. If a specific sequence of symbols allows the population to maximize their shared reward, that sequence acquires meaning within that specific community [[FAdFW16], discussed in [LB20]].
This functional perspective frames communication as a coordination problem. The agents function as a distributed system where the language is the interface allowing them to transfer information about their internal states or observations [LB20]. However, because the communication protocol is learned through trial and error, it is subject to instability. Just as human languages change over time, emergent protocols can drift, particularly when the composition of the agent population changes [BM24, Section 4.4]. To study these dynamics, the field uses frameworks that strip communication down to its most essential components. Perhaps the most common of these involves a sender describing a target object to help a receiver identify it among distractors, known as the referential game.

2.2 Lewis Signaling Game and Referential Communication

The referential game is derived from the Lewis signaling game, a framework introduced by David Lewis to explain how linguistic meaning can arise from pure convention [Lew08]. In Lewis's original design, a sender observes some state of the world and transmits a signal to a receiver, who must then take an appropriate action. What is important, is that initially the signals carry no particular meaning; rather, the meaning emerges through repeated interaction as both agents converge on conventions that allow them to coordinate successfully. When sender and receiver consistently associate the same signals with the same world states, communication succeeds and both agents benefit. This framework demonstrates that shared meaning does not have to be provided in advance but can arise naturally from the pressure to coordinate alone. This makes the game a suitable model for studying language emergence in artificial agents.
The referential game implements Lewis's framework for neural network agents [LPB17]. In each round, a sender observes a target feature vector x and produces a message m, which is a sequence of discrete tokens drawn from a finite vocabulary V. A receiver observes the message along with a candidate set $C = \{x, d_1, \ldots, d_{k-1}\}$ containing the target and $k-1$ distractor feature vectors, and must identify which candidate is the target. If the receiver selects the correct target, both agents receive a reward of 1; otherwise, they receive 0. Because the vocabulary tokens are arbitrary symbols with no predetermined semantics, any meaning they acquire throughout the training will emerge entirely from their utility in solving the task.
Training proceeds by optimizing both agents to maximize communication success. Accuracy is computed as the proportion of games in which the receiver correctly identifies the target:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\hat{x}_i = x_i\} \tag{1}$$

where N is the number of games, x_i is the target in game i, $\hat{x}_i$ is the receiver's prediction, and $\mathbf{1}\{\cdot\}$ is an indicator function returning 1 when its argument is true and 0 otherwise. The receiver is

trained by minimizing cross-entropy loss:

$$\mathcal{L}_R = -\log P(x \mid m, C) \tag{2}$$

where $P(x \mid m, C)$ is the probability the receiver assigns to the correct target x given message m and candidate set C.

While the two-agent referential game provides a simple framework for studying emergent communication, researchers have developed various extensions that capture a richer environment. These include multi-agent populations [CSA+22, GCK19, RSG+22], symmetric architectures where individual agents possess both sending and receiving capabilities [LVB24], and mechanisms for population turnover that introduce new agents over time while retiring others [CLL+20]. Analyzing the resulting languages requires metrics beyond communication accuracy, e.g. measures of compositionality, language similarity, drift, and other properties that capture what kind of communication system has emerged.

2.3 Language Properties and Metrics

Communication accuracy measures whether agents successfully complete the referential task, but it reveals little about the language itself. Two populations can achieve similar accuracy while developing languages that are different [CKB+20]. Both succeed at the task, but the languages may differ in important ways. To understand what kind of communication system has emerged, researchers measure additional properties such as language similarity across agents, compositionality of the message structure, and stability of the language over time. Each property can be quantified through metrics that aim to capture it accurately. The choice of metrics often might depend on the experimental setup. Measures of language drift, for instance, might require a point of reference to compare against. This section introduces the language properties most relevant to this thesis along with common approaches to measuring them. The specific implementations of these metrics for the thesis appear in Section 4.

2.3.1 Language Similarity

When agents learn to successfully communicate, we are tempted to automatically assume that they are speaking the same language. In reality, each agent could develop its own idiosyncratic protocol. Graesser et al.[GCK19] show that pairs of agents could successfully communicate with each other but failed completely when tested against themselves. Each agent had developed its own language to which the other had adapted. This finding highlights that communication accuracy between specific partners does not guarantee a population-wide shared language. This phenomenon is one of the motivations for tracking language similarity, which aims to answer the following question: To what extent do different agents in a population produce consistent messages when describing the same input?

Researchers typically measure this by collecting messages from multiple senders and computing pairwise distances between them. The most common distance measure is edit distance, which counts the minimum number of symbol insertions, deletions, and substitutions needed to transform one message into another [RSG+22, MRM+22]. Similarity is then expressed as one minus this distance, so identical messages yield a similarity of one and completely different messages yield values closer to zero. When averaged across all inputs and all agent pairs, this produces a single score indicating

how synchronized the population's language is. Another name for this metric is synchronization, as described in Rita et al.[RSG+22].

2.3.2 Language Drift

Language drift broadly refers to gradual changes in a language over time. The term implies movement away from some reference point: an original state of the language, or a desired standard. Drift becomes problematic when it decreases mutual intelligibility or causes the language to lose properties that made it useful in the first place. The specific reference point and the way researchers measure deviation from it depend on the experimental context.

Most research on language drift has focused on agents that are first pretrained on human language and then finetuned through interaction to maximize task rewards [LB20, LCK19, LSS+20]. In this setting, the reference point is natural language itself, and drift is a divergence from human-like communication. Lazaridou et al.[LB20] identify three types of drift in this context. Structural drift involves loss of grammatical properties, such as fluency degradation or repetitive patterns. Semantic drift occurs when the meanings of words shift, so that symbols come to denote different concepts than they originally did. Pragmatic drift arises when agents develop private conventions that differ from how humans would interpret the same messages. These categories all aim to describe different aspects of the same phenomenon.

When agents develop a language from scratch rather than starting from human language, the dynamics of language change take a different form and have received less attention in the literature [BM24]. Most research on emergent communication with generational transmission has focused on beneficial changes: studies using iterated learning show that languages tend to become more compositional and structured as they pass through successive generations of learners [RGL+20, KGS14, CLL+20]. Graesser et al. have studied related events such as language contact between previously isolated populations [GCK19].

However, the possibility that an emergent language might change in ways that reduce mutual intelligibility across generations has been less explored. When old agents leave and new agents join a population, small changes can accumulate over time. Eventually, agents from distant generations may struggle to communicate even though each generation performed well with itself and its immediate neighbors. To capture this phenomenon, this thesis introduces two metrics that measure how far the language has moved from its initial state. **Cross-generational accuracy** pairs agents from different time periods and measures whether they can still successfully complete the referential task together. If a sender from an early generation produces a message and a receiver from a later generation must identify the target, the metric records whether communication succeeds. **Cross-generational similarity** compares the messages that agents from different time periods produce for the same input, using normalized edit distance in the same manner as the language similarity metric described above. Together, these metrics provide a way to quantify language stability or change in emergent communication over time. For the exact implementation see Section 4. These metrics compare each generation to the founding population, measuring cumulative drift. This thesis also uses previous-generation variants that compare consecutive generations, capturing the rate of change at each transition rather than total divergence from the original language.

2.3.3 Compositionality

Compositionality is one of the most studied properties in emergent communication research [BM24]. It refers to the principle that the meaning of a complex expression derives from the meanings of its constituent parts and the rules used to combine them. For instance, in English, the phrase "red triangle" functions because "red" contributes color information and "triangle" contributes shape information independent of each other. A compositional emergent language should exhibit similar structure: distinct parts of the message (symbols) should correspond systematically to distinct parts of the input attributes (e.g., color or shape) [CKB+20].

Researchers care about compositionality for several reasons. A compositional language should generalize better to new inputs, since agents can recombine known symbols rather than memorizing separate messages for every possible input [LHTC18]. The paralell here is that once we know what words "red" and "triangle" mean, we can easily grasp the concept of a "red triangle". Compositional languages might also be easier for new agents to learn, since the systematic structure provides regularities that learners can exploit [CKB+20]. Finally, compositional languages are more interpretable to human researchers, who can identify what each symbol means and predict how agents will describe new objects.

The most widely used measure of compositionality is topographic similarity [BK06, LHTC18, BM24]. If a language is compositional, similar meanings should (in general) map to similar messages. Topographic similarity quantifies this by computing the correlation between pairwise distances in the input space and pairwise distances in the message space. A high correlation indicates that the language preserves structural relationships, e.g. inputs that differ by one attribute tend to receive messages that differ by roughly one symbol.

However, topographic similarity has faced substantial criticism. Chaabouni et al. [CKB+20] demonstrate that languages can achieve high topographic similarity through different mechanisms, some of which align with intuitive notions of compositionality better than others. More importantly, Kharitonov and Baroni [KB20] and Chaabouni et al. [CKB+20] show that compositionality and generalization do not correlate as strongly as researchers initially expected. Agents can successfully generalize to novel input combinations without developing transparently compositional languages, apparently finding other strategies to solve the task. Chaabouni et al. [CSA+22] further confirm this finding in large-scale experiments with complex image inputs. This thesis uses simpler discrete feature vectors rather than images, but these critiques suggest caution when interpreting topographic similarity scores.

In response to these limitations, Chaabouni et al. [CKB+20] introduced two additional measures. Positional disentanglement (posdis) checks whether symbols at specific positions in the message consistently encode specific attributes of the input. For instance, in English adjective-noun phrases the first position might always encode color while the second always encodes shape ("red square", "blue circle" etc.). Bag-of-symbols (bosdis) disentanglement captures whether the presence of specific symbols, regardless of their position, consistently indicates specific attributes. This corresponds to how conjunctions work: "dogs and cats" has the same meaning as "cats and dogs." These measures are more specific than topographic similarity but they also make stronger assumptions about the form that composition should take.

Compositionality is not the primary focus of this thesis. Nevertheless, this thesis reports topographic similarity, positional disentanglement, and bag-of-symbols disentanglement for two reasons. First, these metrics are widely used in the field, and reporting them allows comparison with prior work and

contributes additional data points to ongoing debates about their validity. Second, compositionality may interact with the phenomena under investigation. Prior work suggests that compositional languages are easier for new agents to learn [CKB+20], which could affect how quickly newcomers acquire the population's language and how much the language changes as agents are replaced. The experiments in this thesis span different population configurations, turnover rates, and plasticity conditions, which may reveal how these factors influence compositional structure.

2.3.4 Other Metrics

Researchers often also track simpler metrics that provide additional context about the emergent language. This thesis measures two such metrics. Mean message length indicates how many symbols agents typically use to describe inputs. If agents consistently produce messages near the maximum allowed length, the channel capacity may be a bottleneck. Vocabulary usage measures what proportion of available symbols agents actually use; a language that uses only a few symbols might indicate a solution that does not exploit the full expressive capacity of the channel. These additional metrics can help identify potential bottlenecks or inefficiencies and provide a fuller picture of what kind of communication system has emerged.

2.4 Aging and Plasticity in Language Acquisition

One of the most robust findings in cognitive science is that children acquire language more easily than adults [HTP18]. Individuals who learn a second language in childhood are often indistinguishable from native speakers, whereas those who begin in adulthood typically retain accents and make grammatical errors throughout their lives. This observation led to the critical period hypothesis, which states that language acquisition occurs most easily during a certain developmental window where the brain exhibits the most plasticity. After this window closes, language learning becomes substantially more difficult. The causes are not entirely clear to this day and the proposed explanations include neural maturation, interference from a well-learned first language, reduced memory, and social or environmental changes that happen during the transition to adulthood [HTP18].

Hartshorne et al. [HTP18] conducted the largest empirical study of this trajectory, analyzing 669,498 English speakers. They found that grammar-learning ability is preserved until approximately 17-18 years of age - later than previous estimates which placed the decline around puberty [Len67]. While precise parameters remain debated [vdSSBvH22], the core finding holds: learning ability is high during childhood and declines with age. For this thesis, the exact timing matters less than the established observation that age-dependent changes in learning ability are real and substantial: children and adults differ in how they acquire language.

Gopnik [Gop20] proposes a theoretical framework that reframes the critical period as a solution to a well-known computational problem: the exploration-exploitation trade-off. In reinforcement learning and optimization, agents must balance exploring unknown possibilities against exploiting known rewards.

In complex environments, there is no truly optimal way to balance exploration and exploitation, but one effective strategy is to begin with broad, high-temperature search that samples widely across the space of possibilities, then gradually cool to narrower, low-temperature search that converges on stable solutions [KGJV83, Gop20]. Gopnik argues that human development demonstrates

precisely this strategy. Children exhibit high plasticity, broad hypothesis search, and characteristics suited to exploration. Adults, by contrast, display more stable behavior, focused attention, and characteristics suited to exploitation. This division of labor allows a human to explore the space of possible languages, behaviors, and structures during a protected period of immaturity, then exploit that knowledge efficiently in adulthood [Gop20]. In the context of this thesis, this framework suggests that populations containing both high-plasticity children and low-plasticity adults might behave differently than homogeneous populations: older agents could provide stable learning targets that anchor the system against drift while younger agents could have the liberty to explore and learn the language quicker.

Computational work by Verhoef and de Boer [VdB11] provides evidence for this hypothesis. They simulated vowel system emergence in populations of agents playing imitation games, comparing populations with age-dependent learning rates against homogeneous control populations. In age-structured populations, younger agents learned with larger step sizes than older agents. The study found that age-structured populations preserved their vowel systems significantly better than either homogeneous high-plasticity (all children) or homogeneous low-plasticity (all adult) populations. Mutual intelligibility between initial and final generations was higher, and the complexity of the vowel systems was better maintained. These findings align with observations from sociolinguistics: Labov [Lab07] found that unbroken transmission from adults to children preserves linguistic structure better than adult-to-adult contact alone. However, their agents used pre-made vowel representations as opposed to learning from scratch. Section 5.4 tests whether age-dependent plasticity produces similar stabilizing effects when neural network agents must develop a shared language through referential games.

3 Related Work

3.1 Emergent Communication in Populations

Experiments with human participants show that larger communities develop more systematic and structured languages [RMLA19]. This finding suggests that population dynamics play an important role in shaping language structure and motivates extending emergent communication simulations beyond single pairs of agents. Having only two agents learn to communicate together can also cause some problems. Agents trained in pairs can develop highly specialized protocols that lack properties found in human languages [BB18, CKDB19]. Additionally, Graesser et al. [GCK19] showed that agents trained in pairs develop idiolects instead of a single shared language and can completely fail on self-communication. These limitations caused researchers to scale emergent communication to multi-agent populations [MRM+22].

Graesser et al. [GCK19] were one of the first to develop a multi-agent framework. They varied connections between agents within a population and studied different populations coming into contact with each other. They found that when two isolated populations start communicating with each other, they either converge to the majority protocol (the bigger group) or, when the populations are similar in size, they develop a "creole" language with lower complexity. They also found that when multiple populations are connected in a "chain", there will form a linguistic continuum where neighboring languages are more mutually intelligible than distant ones. In summary, this research has confirmed that certain complex features of language evolution (like merging of two

populations) can be simulated with a relatively simple setup which can utilize similar agents as two-agent scenarios. An important contribution to the research of populations were community-based autoencoders introduced by Tieleman et al. [TLM+19], where encoders (senders) and decoders (receivers) are randomly paired with different partners at every training step. Because the sender does not know the identity of the receiver they cannot rely on partner-specific shortcuts ("inside jokes") to communicate. Instead, this uncertainty forces the system to converge to a single language that can be understood by any member of the community. In another study, Kim and Oh [KO21] investigated varying connectivity patterns in more detail, within larger populations (e.g. over 100 agents). They found that a phenomenon similar to the lingusitic continuum emerges within a single community. If the population were not fully connected (when speakers can only communicate with nearby partners), the agents formed local dialects, whereas fully-connected populations converged to a single language.

Despite these (and other) promising results, scaling to larger populations has not consistently reproduced the benefits observed in human experiments. Chaabouni et al. [CSA+22] conducted large-scale experiments with populations up to 100 agents on realistic image datasets and found no improvement in language generalization or robustness from population size alone. They did however propose additional mechanisms possible only in populations (imitation of best performing speakers and voting among listeners) that were slightly beneficial. Rita et al. [RSG+22] reproduced this null effect of size and attributed it to homogenity. When they introduced assymetries in learning speed across agents, larger populations began developing more stable and structured languages. Michel et al. [MRM+22] provided a theoretical explanation for the initial negative results. In standard training, receivers co-adapt to all senders they interact with, and at equilibrium the population-level objective takes the same functional form regardless of population size. They proposed partitioning agents (connecting them into sender-receiver pairs) to prevent this co-adaptation, which restored population size effects on compositionality and enabled successful communication with previously unseen partners.

Recent work has extended population frameworks in several directions. Lian et al. [LVB24] introduced NeLLCom-X, which implements role-alternating agents that can both speak and listen. They found that group size affects language optimization: larger groups develop more efficient languages with clearer trade-offs between redundant features. Mahaut et al. [MFDB25] explored populations of pre-trained visual networks with different architectures. Despite these differences, the networks developed shared communication protocols. Most population research, however, has focused on static communities where the same agents interact throughout training. Less attention has been paid to populations with turnover, where agents enter and leave over time.

3.2 Agent Turnover

Compared to static populations, emergent communication with agent turnover has received relatively little attention. The closest method in the language evolution literature is iterated learning, where the output of one generation of learners becomes the input for the next. In iterated learning experiments, languages pass through a transmission bottleneck as each generation learns from incomplete samples of the previous generation's language, creating pressure toward compressible, compositional structure [KGS14]. However, iterated learning typically models discrete, non-overlapping generations with explicit teaching phases, where one generation's language is recorded and used to train the next. This is the case in the studies by Kirby et al. [KGS14] and Ren et al. [RGL+20]. This approach

is different than the continuous, implicit transmission that occurs when agents of different ages coexist and learn together in a shared population.

Cogswell et al. [CLL+20] introduced a framework that simulates generational transmission in deep reinforcement learning environments. In their study, a population of senders and receivers trains together on a cooperative referential game. Every E epochs, a replacement policy selects agents to reset. The reinitialized agents lose all learned parameters, effectively entering the population as newcomers with no knowledge of the existing language. Because the population continues training, these newcomers must learn from experienced agents. This creates implicit cultural transmission: language passes between generations not through explicit teaching but through the pressure to coordinate successfully on the task.

Cogswell et al. tested three replacement strategies: uniform random selection, epsilon-greedy selection favoring the worst performing agents, and oldest-first replacement. All three improved compositional generalization compared to static populations. The authors attributed this to the following: newcomers who happen to use language consistent with the population's conventions receive higher rewards, implicitly selecting for languages that are easy to learn. Li and Bowling found a similar effect of agent turnover in their study [LB19]. In their framework, there is only one speaker which communicates with multiple generations of listeners. This encourages the senders to form languages that are easier to teach to newcomers.

3.3 Role-Alternating Agents

Most emergent communication research employs separate sender and receiver agents that cannot switch roles. A sender network learns to produce messages while a distinct receiver network learns to interpret them. This asymmetric design contrasts with human language use, where individuals can both produce and understand speech. Galke et al. [GRR22] identify role-alternation as a key missing ingredient in neural emergent communication simulations. They argue that in humans, the ability to alternate between speaking and listening roles underlies the emergence of linguistic structure, and that omitting this constraint from simulations may explain why neural agents fail to replicate certain findings from human experiments. Several approaches have emerged to address this gap.

Graesser et al. [GCK19] implemented symmetric agents that can act as both speaker and listener. They discovered that with only two such agents, each develops its own idiolect rather than a shared language. The agents achieve high accuracy when communicating with each other, but when tested against themselves (self-play), accuracy drops to chance level. Each agent learns a distinct protocol to which the other adapts, but neither can understand its own messages. At least three agents are necessary for a shared symmetric protocol to emerge spontaneously. Michel et al. [MRM+22] proposed a different approach called partitioning: coupling each sender with a specific receiver to form an "agent." During training, each receiver learns only from its associated sender rather than adapting to all senders in the population. This creates implicit coupling between speaking and listening components without requiring full architectural integration, and it prevents receivers from learning multiple sender-specific conventions.

Lian et al. [LVB24] implement full role-alternating agents where speaking and listening components share parameters and agents periodically communicate with themselves (self-play) to maintain self-understanding. This design enables realistic group communication where any agent can interact with any partner in either role. However, Lian et al. do not study population turnover, and

combining role-alternation with generational transmission presents unique considerations. Cogswell et al. [CLL+20] study turnover using two separate populations: Q-bots that observe tasks and ask questions, and A-bots that observe objects and provide descriptions. Each agent remains fixed in its role throughout training. At each replacement event, their system reinitializes one Q-bot and one A-bot selected according to a replacement policy—but these two agents have no special relationship to each other beyond being replaced at the same time. Any Q-bot can be paired with any A-bot during training, so the "generation" of a Q-bot is independent from the generation of the A-bots it communicates with. Role-alternating agents might offer a more intuitive approach to modeling turnover: each agent represents a complete communicator, and replacement cleanly removes one individual from the population rather than one sender and one unrelated listener. The experiments in this thesis employ role-alternating agents for this reason: when an agent dies and a new one is born, the population loses and gains a complete communicator rather than half of one.

3.4 Population Heterogeneity

Most emergent communication simulations employ homogeneous populations where all agents share identical architectures, learning rates, and optimization procedures. Rita et al. [RSG+22] proposed that this homogeneity may explain why population size fails to improve language properties in neural simulations. Their argument centers on population dynamics: when agents are functionally interchangeable, the optimization landscape remains unchanged regardless of population size. Adding more identical agents does not introduce the differential pressures that might structure language. This reasoning suggests that heterogeneity within populations, rather than population size itself, may be the operative factor.

Rita et al. tested this hypothesis by manipulating the relative learning speeds of speakers and listeners. They found that absolute learning speed had minimal effect on language properties, but the ratio between speaker and listener update rates proved consequential. When speakers updated faster than listeners, languages exhibited lower entropy, higher compositionality, and better synchronization across speakers. The authors interpreted this asymmetry as creating implicit selection pressure: fast-updating speakers face slow-adapting listeners who act as bottlenecks, filtering out languages that are difficult to acquire. When Rita et al. distributed learning speeds heterogeneously across larger populations—sampling each agent's update probability from a distribution—they observed the expected sociolinguistic pattern: larger populations developed more stable and structured languages with reduced variance across experimental runs. These findings provide initial evidence that heterogeneity enables population size effects, though the mechanism remains specific to their particular manipulation of learning dynamics.

Other forms of heterogeneity have received less attention. Mahaut et al. [MFDB25] studied populations of pre-trained visual networks with different architectures and training histories, finding that such heterogeneous agents could develop shared communication protocols through referential games. New agents learned existing community protocols more efficiently than communities could develop new protocols from scratch. While this architectural heterogeneity differs from learning dynamics heterogeneity, both lines of work suggest that agent diversity need not impede—and may facilitate—the emergence of shared conventions. This thesis introduces heterogeneity through age-dependent plasticity (Section 4.5). In populations with turnover, agents naturally vary in age at any given time. If plasticity decreases with age, younger agents function similarly to Rita et al.'s fast learners while older agents provide stable learning targets. Unlike artificially imposed

asymmetries, this heterogeneity emerges from the population structure itself—a consequence of turnover dynamics rather than an arbitrary parameter choice. The experiments in Section 5.4 test whether this biologically-motivated form of heterogeneity produces stabilizing effects analogous to those documented by Rita et al.

4 Methodology

This section describes the experimental framework used to investigate the research questions[1]. The implementation builds on the EGG toolkit, extending it with mechanisms for population dynamics and age-based plasticity. The following subsections progress from the underlying framework through game design, agent architecture, and population mechanics to the metrics used for evaluation. Specific parameter values and experimental configurations are presented in Section 5.

4.1 The EGG Framework

EGG (Emergence of lanGuage in Games) is a PyTorch-based toolkit developed by Facebook AI Research for implementing emergent communication games [KCBB19].[2] The toolkit provides modular components and has become the most widely used framework in emergent communication research [BM24]. Numerous studies have used EGG to investigate various questions in the field [CKDB19, KB20, CSA+22, LVB24].

EGG implements agents, games, and training components as PyTorch modules, making them easy to combine and extend. The framework provides default Sender and Receiver classes that handle message generation and processing. Users only need to implement the core perception logic (how agents encode inputs) and the loss function. For the discrete communication channel, EGG offers wrappers for both Gumbel-Softmax relaxation and REINFORCE optimization. Communication can be single-symbol or variable-length sequences generated by RNN cells. The Trainer module manages the training loop, validation, and checkpointing. The Callbacks system enables custom logging and metrics without modifying core code.

This thesis extends EGG to support population-based training with agent turnover and age-dependent plasticity. Where possible, the implementation uses EGG's existing components in their intended ways. The aim is to make the code as easy to read, reproduce and extend as possible.

4.2 Game Design

The game is a variant of the referential discrimination game (Section 2.2), following EGG's `objects_game` [KCBB19] based on Lazaridou et al. [LHTC18]. Each round, a sender observes a target object and transmits a message to a receiver, who must identify the target among distractors.

Input Space. Objects are discrete feature vectors $x = (x_1, \ldots, x_F)$ with F attributes. Attribute f takes values in $\{1, \ldots, d_f\}$, and the dimension vector $D = (d_1, \ldots, d_F)$ specifies the number of values per attribute. The input space $\mathcal{X}$ contains all valid attribute combinations.

[1] Code available at https://github.com/Jano04/EGG

[2] Code available at https://github.com/facebookresearch/EGG

Game Round. Each round proceeds as follows: (1) sample a target $x^* \in \mathcal{X}$ uniformly; (2) sample n_d distractors from $\mathcal{X} \setminus \{x^*\}$ without replacement to form candidate set C; (3) shuffle C to randomize target position; (4) the sender observes x^* and produces message m; (5) the receiver observes m and C, then outputs a probability distribution over candidates. Communication succeeds when the receiver assigns highest probability to the target.

Communication Channel. Messages are variable-length sequences from vocabulary $V = \{0, 1, \ldots, |V| - 1\}$, where symbol 0 denotes end-of-sequence (EOS). With maximum length L:

$$m = (m_1, \ldots, m_T), \quad m_t \in V, \quad T \leq L \tag{3}$$

The sender generates symbols autoregressively via an LSTM, sampling from policy $\pi_\theta(m_t \mid x^*, m_{<t})$. Gumbel-Softmax relaxation [JGP16] with temperature τ enables gradient-based training with discrete symbols.

Receiver Discrimination. The receiver processes the message through an LSTM (via EGG's `RnnReceiverGS`) to obtain embedding h_m. For each candidate $c_j \in C$, it computes score $s_j = h_m^\top f_\phi(c_j)$, where f_ϕ is the receiver's learned encoding function. Scores convert to probabilities via softmax:

$$\rho_\phi(c_j \mid m, C) = \frac{\exp(s_j)}{\sum_{c \in C} \exp(s_c)} \tag{4}$$

Training Objective. The receiver minimizes cross-entropy loss:

$$\mathcal{L} = -\log \rho_\phi(x^* \mid m, C) \tag{5}$$

Gumbel-Softmax relaxation allows gradients to flow through the entire sender-receiver pipeline, enabling joint optimization. Accuracy measures the proportion of rounds where the receiver's argmax matches the target:

$$\text{Acc} = \frac{1}{|\mathcal{B}|} \sum_{(x^*, C) \in \mathcal{B}} \mathbf{1}\left[\arg\max_{c \in C} \rho_\phi(c \mid m, C) = x^*\right] \tag{6}$$

where $\mathcal{B}$ denotes the evaluation batch.

Population Training. The single-pair game extends to populations following prior work [RSG+22, MRM+22, CSA+22]. A population consists of N agents, each capable of acting as sender or receiver (Section 4.3). Each training batch samples N sender-receiver pairs uniformly with replacement from alive agents. Self-play (same agent as both sender and receiver) is permitted. The training loss averages across pairs:

$$\mathcal{L}_{\text{pop}} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{(s_i, r_i)} \tag{7}$$

where (s_i, r_i) denotes the i-th sampled pair. During evaluation, the game tests all N^2 sender-receiver combinations to measure population-wide mutual intelligibility.

4.3 Agent Architecture

Each agent in the population can act as both sender and receiver. This symmetric design follows NeLLCom-X [LVB24], which is also based on EGG.

Symmetric Structure. Agent i consists of two modules: a sender module parameterized by θ_i and a receiver module parameterized by ϕ_i. Both modules encode feature vectors from the same input space $\mathcal{X}$, but serve different roles during communication. When agent i acts as sender, it uses θ_i to encode the target and initiate message generation. When acting as receiver, it uses ϕ_i to encode candidates and compute discrimination scores. The modules do not share parameters—each agent maintains independent sender and receiver capabilities.

Sender Module. The sender module encodes the target object into a hidden representation:

$$h_s = \tanh(W_s x^* + b_s) \tag{8}$$

where $W_s \in \mathbb{R}^{H \times F}$ is a weight matrix, $b_s \in \mathbb{R}^H$ is a bias vector, and H is the hidden dimension. This encoding initializes EGG's `RnnSenderGS` wrapper, which generates messages autoregressively via LSTM. At each step t, the LSTM produces a hidden state from which logits over the vocabulary are computed and a symbol is sampled (using Gumbel-Softmax during training, greedy decoding during evaluation). The message generation process is described in Section 4.2.

Receiver Module. The receiver module encodes each candidate object into the same hidden space:

$$h_r^{(j)} = \tanh(W_r c_j + b_r) \tag{9}$$

where c_j is the j-th candidate and $W_r \in \mathbb{R}^{H \times F}$, $b_r \in \mathbb{R}^H$ are learnable parameters. EGG's `RnnReceiverGS` wrapper processes the incoming message through an LSTM to produce embedding h_m. The receiver then computes discrimination scores via dot product between the message embedding and each candidate encoding:

$$s_j = h_m^\top h_r^{(j)} \tag{10}$$

These scores are normalized to probabilities via softmax, as described in Section 4.2.

Per-Agent Optimizers. Each agent maintains its own Adam optimizer that updates both its sender and receiver parameters jointly. This design serves two purposes. First, it enables per-agent learning rate control, which is essential for age-based plasticity (Section 4.5). Second, it ensures that only agents participating in a given batch receive gradient updates—when a batch samples N sender-receiver pairs, only the optimizers for agents appearing in those pairs are stepped. Agents not sampled retain their parameters unchanged. When a new agent is born (Section 4.4), its optimizer is reinitialized with fresh state, ensuring the newcomer begins learning without inheriting momentum from previous training.

4.4 Population Dynamics

Following Cogswell et al. [CLL+20], the population can undergo periodic turnover: agents die and new agents enter with no knowledge of the existing language. Newcomers learn from experienced agents through interaction alone, creating implicit cultural transmission without explicit teaching. Turnover is optional - disabling it yields a static population for baseline comparisons.

Agent States. Each agent exists in one of three states:

- **Alive**: Participates in training, can be sampled as sender or receiver, receives gradient updates, and ages each epoch.
- **Dead**: No longer participates in training. Parameters are frozen at their values at death.
- **Dormant**: Never activated. Parameters hold random initialization values, awaiting future birth.

The model allocates $N_{\max}$ agent slots at initialization, where $N_{\max} \geq N$. The first N agents begin alive; the remaining $N_{\max} - N$ begin dormant. Without turnover, all agents remain alive throughout training. With turnover enabled, dead agents preserve their frozen parameters, enabling cross-generational analysis by pairing agents from different time periods.

Age Tracking. Each agent's age counts training epochs since birth (or since warmup completion for the founding population). At the end of each epoch, all alive agents increment their age:

$$a_i^{(e+1)} = \begin{cases} a_i^{(e)} + 1 & \text{if agent } i \text{ is alive} \\ a_i^{(e)} & \text{otherwise} \end{cases} \tag{11}$$

Dead agents freeze at their death age; dormant agents remain at age zero until activated.

Warmup Phase. An optional warmup phase delays turnover until the founding population establishes a functional language. Training proceeds normally until validation accuracy reaches threshold $\omega \in [0, 1]$. Once accuracy $\geq \omega$, warmup completes: age tracking begins, turnover activates (if enabled), and the system captures a baseline snapshot of the founding population for later comparison. Setting $\omega = 0$ disables warmup, starting aging and turnover immediately.

Turnover Mechanism. The parameter k (kill epoch) controls turnover rate. When $k > 0$, a death occurs every k epochs after warmup completes. Setting $k = 0$ disables turnover entirely. Each death event proceeds as follows: a death policy selects one alive agent to die, the system immediately activates the next dormant agent (by index order), and the new agent receives freshly initialized parameters and a reset optimizer. This maintains exactly N alive agents throughout training.

The default death policy selects the oldest agent (ties broken randomly). Alternative policies - random selection, performance-based selection, and age-weighted probabilistic selection - are available but not used in the main experiments.

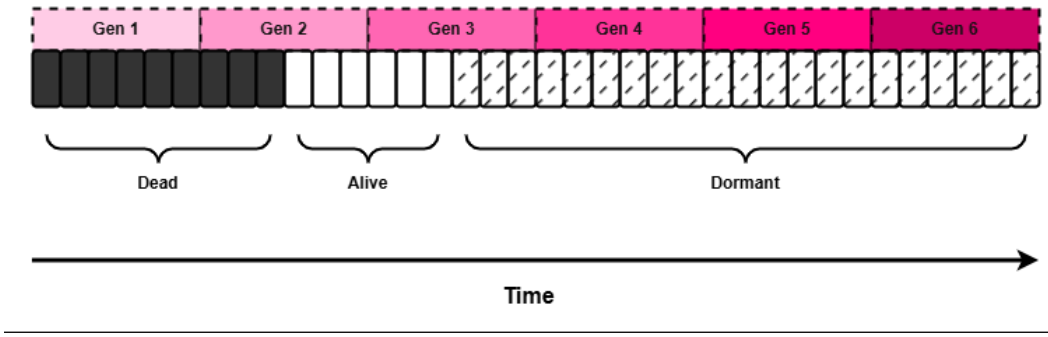

Figure 1: Example of Agent States and generational overlap. Here, half the alive agents are from generation 2, and half are from generation 3.

Generations. A generation groups N consecutive agents by birth order. Generation 1 contains the founding population (agents 0 through $N-1$), established at warmup completion. After N deaths, the population has completely turned over and generation 2 begins. The generation number is:

$$g = \left\lfloor \frac{n_{\text{death}}}{N} \right\rfloor + 1 \tag{12}$$

where n_{death} is the cumulative death count. This provides a natural unit for aggregating metrics: cross-generational accuracy, for instance, compares generation 1 agents against generation g agents to quantify cumulative drift. Notice that the generations can (and most of the time do) overlap. At a given time, there can exist agents from both generations g and $g+1$.

4.5 Age-Based Plasticity

Age-based plasticity models the decline in learning capacity observed in human language acquisition [HTP18, Gop20]. Young agents learn quickly and explore broadly; old agents learn slowly and maintain stable representations. Plasticity is optional—disabling it creates a homogeneous population where all agents use fixed learning parameters regardless of age.

Plasticity Function. Plasticity $p(a) \in [0, 1]$ maps agent age to learning capacity. Age is normalized to $[0, 1]$:

$$\bar{a} = \frac{a}{a_{\max} - 1} \tag{13}$$

where a is current age and $a_{\max}$ is a hyperparameter controlling when plasticity reaches its minimum. Note that $a_{\max}$ is independent of actual agent lifespan (determined by turnover rate k); this allows plasticity dynamics to be tuned separately from population turnover.
The primary function is a sigmoid modeling a critical period:

$$p_{\text{sigmoid}}(a) = \frac{1}{1 + \exp(\beta(\bar{a} - \mu))} \tag{14}$$

where β controls transition steepness and $\mu \in [0, 1]$ sets the critical point. A linear alternative $p_{\text{linear}}(a) = \max(0, 1 - \bar{a})$ provides uniform decline.

Controlled Parameters. Plasticity modulates two training parameters. Gumbel-Softmax temperature τ controls exploration during message generation:

$$\tau(a) = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot p(a) \tag{15}$$

Learning rate η controls adaptation speed:

$$\eta(a) = \eta_{\min} + (\eta_{\max} - \eta_{\min}) \cdot p(a) \tag{16}$$

Young agents (high p) explore diverse messages and adapt quickly to emerging conventions. Old agents (low p) produce consistent messages and resist change, providing stable learning targets for newcomers.

Update Schedule. After each age increment, τ and η update for all alive agents based on their current ages. Newborn agents receive maximum values ($\tau_{\max}$, $\eta_{\max}$); plasticity decreases as they age. When plasticity is disabled, all agents use fixed τ and η throughout training.

4.6 Metrics and Evaluation

This section describes the metrics introduced in Section 2.3. The framework captures periodic snapshots of population state and computes metrics quantifying language properties and drift.

Snapshots. A snapshot records population state at a specific training epoch. Each snapshot stores: messages produced by all alive agents for the validation set $\mathcal{X}_{\text{val}}$, the corresponding inputs (needed for cross-generational evaluation), agent ages and birth epochs, and sender/receiver weights. Snapshots are captured when warmup completes (the generation 1 baseline) and immediately before each agent death. Optional periodic snapshots at fixed intervals support finer-grained analysis. Storing weights enables post-hoc cross-generational testing without rerunning training.

Language Similarity. Language similarity (also called synchronization [RSG+22]) measures agreement among agents on how to describe inputs [MRM+22, RSG+22]. For each input $x \in \mathcal{X}_{\text{val}}$, all agents produce messages, and pairwise normalized edit distances are computed:

$$d_{\text{norm}}(m_i, m_j) = \frac{d_{\text{edit}}(m_i, m_j)}{\max(|m_i|, |m_j|)} \tag{17}$$

where d_{edit} is Levenshtein distance and $|m|$ denotes message length. Language similarity averages across all agent pairs and inputs:

$$S_L = 1 - \frac{1}{|\mathcal{P}| \cdot |\mathcal{X}_{\text{val}}|} \sum_{(i,j) \in \mathcal{P}} \sum_{x \in \mathcal{X}_{\text{val}}} d_{\text{norm}}(m_i(x), m_j(x)) \tag{18}$$

where $\mathcal{P}$ is the set of agent pairs and $m_i(x)$ denotes agent i's message for input x. Values near 1 indicate that agents use similar messages for the same inputs.

Compositionality. Three metrics assess compositional structure, computed using EGG's built-in implementations. Topographic similarity [LHTC18] measures Spearman's rank correlation ρ_s between input-space distances (Hamming) and message-space distances (edit distance). Values near 1 indicate that similar inputs receive similar messages. Positional disentanglement (posdis) measures whether message positions encode specific input attributes; bag-of-symbols disentanglement (bosdis) measures whether symbol presence encodes attributes regardless of position. Both use the mutual information gap metric from Chaabouni et al. [CKB+20]. All compositionality metrics are computed per agent and averaged across the population.

Message Statistics. Vocabulary usage measures the fraction of available symbols (excluding EOS) that appear in population messages. Message length statistics (mean and standard deviation) indicate channel capacity utilization.

Cross-Generational Accuracy. Cross-generational accuracy quantifies language drift by testing communication between agents from different time periods. Weights are loaded from the generation 1 snapshot and a later snapshot, then communication accuracy is evaluated across all sender-receiver pairings between the two generations, averaged over both communication directions:

$$A_{\text{gen}} = \frac{1}{2}\left(A_{G_1 \to G_t} + A_{G_t \to G_1}\right) \tag{19}$$

where $A_{G_1 \to G_t}$ denotes accuracy when generation 1 agents send to generation t receivers, and vice versa. A value of 1 indicates perfect mutual intelligibility across generations; lower values indicate drift.

Cross-Generational Similarity. While cross-generational accuracy measures whether agents can still communicate, cross-generational similarity measures how much the message form itself has changed. This metric extends language similarity across generations: for each input, normalized edit distances are computed between messages from generation 1 agents and current agents, then averaged. This provides a direct measure of surface-form drift independent of communicative success.

Previous-Generation Metrics. The cross-generational metrics above compare each generation to the original population. Previous-generation metrics compare consecutive generations instead, measuring drift rate rather than cumulative drift. Previous-generation accuracy is computed as:

$$A_{\text{prev}}(g) = \frac{1}{2}\left(A_{G_{g-1} \to G_g} + A_{G_g \to G_{g-1}}\right) \tag{20}$$

where $g \geq 2$ is the generation number. Previous-generation similarity follows the same structure, using normalized edit distance between messages from generations $g-1$ and g. Values near 1 indicate that consecutive generations remain mutually intelligible.

5 Experiments and Results

5.1 Experimental Setup

All three experiments use the referential game from Section 4.2 and share the same input space, communication channel, agent architecture, and training protocol. They differ only in population dynamics: Experiment 1 uses static populations, Experiment 2 introduces agent turnover, and Experiment 3 adds age-based plasticity.

Task and Data. Objects are discrete feature vectors with $F = 4$ attributes, each taking 5 values, giving an input space of $|\mathcal{X}| = 5^4 = 625$ objects. Each round, the receiver sees the target alongside $n_d = 5$ distractors. Agents communicate through variable-length messages with vocabulary size $|V| = 20$ (including EOS) and maximum length $L = 10$. The dataset contains 10,240 training tuples, 1,000 validation tuples, and 1,000 test tuples, generated with a fixed seed for reproducibility.

Training. Each agent uses the symmetric architecture from Section 4.3 with hidden dimension $H = 128$ and 64-dimensional LSTM embeddings. Training uses per-agent Adam optimizers with learning rate $\eta = 10^{-3}$ and weight decay 10^{-5}. Each training batch contains 1024 tuples and samples N sender-receiver pairs uniformly with replacement.

Evaluation. All metrics are computed on the validation set. During evaluation, all N^2 sender-receiver combinations are tested, with accuracy averaged across pairs. All experiments use three random seeds per condition; experiment-specific parameters appear in the Design subsections.

5.2 Experiment 1: Baseline Population Sizes

This experiment establishes baseline behavior in static populations without turnover or age-based plasticity. The results provide reference metrics for evaluating the effects of these mechanisms in Experiments 2 and 3.

5.2.1 Design

Experiment 1 tests ten population sizes: $N \in \{2, 4, 6, 8, 10, 12, 14, 16, 18, 20\}$. All populations train for 500 epochs with turnover disabled ($k = 0$) and uniform plasticity (fixed temperature τ and learning rate η for all agents throughout training). Each configuration runs with three random seeds, yielding 30 runs total. All other parameters follow Section 5.1.

5.2.2 Results

All population sizes successfully learn to solve the referential task and develop shared languages. Table 1 reports the final test accuracy and language similarity at epoch 500, averaged across seeds. Accuracy exceeds 97% for all configurations, with smaller populations ($N \leq 8$) achieving slightly higher values and lower variance. Language similarity reaches 0.70–0.83 across conditions, indicating that agents within each population converge on similar message conventions.
Figure 2 shows learning dynamics across population sizes. Smaller populations converge faster, reaching high accuracy within 50–100 epochs. Larger populations exhibit slower initial learning

Table 1: Final test accuracy and language similarity by population size (mean ± std across 3 seeds, epoch 500).

N	Accuracy (%)	Language Similarity
2	99.0 ± 0.4	0.76 ± 0.05
4	99.2 ± 0.1	0.77 ± 0.01
6	99.1 ± 0.1	0.79 ± 0.10
8	99.1 ± 0.0	0.80 ± 0.09
10	98.6 ± 0.4	0.75 ± 0.04
12	98.4 ± 1.1	0.74 ± 0.07
14	98.8 ± 0.1	0.83 ± 0.02
16	98.5 ± 0.5	0.82 ± 0.04
18	97.8 ± 1.2	0.70 ± 0.10
20	98.1 ± 0.2	0.71 ± 0.02

and occasional instabilities visible as temporary accuracy drops, but all eventually converge. These instabilities likely reflect the increased coordination difficulty when more agents must simultaneously agree on a shared protocol. Language similarity initially mirrors this trend - smaller populations are quicker to develop high similarity, but the relationship fades away after $\tilde{1}50$ epochs. There is also high variance between different seeds. Some populations show temporary peaks followed by partial decline in similarity, suggesting ongoing language negotiation even after high accuracy is achieved.

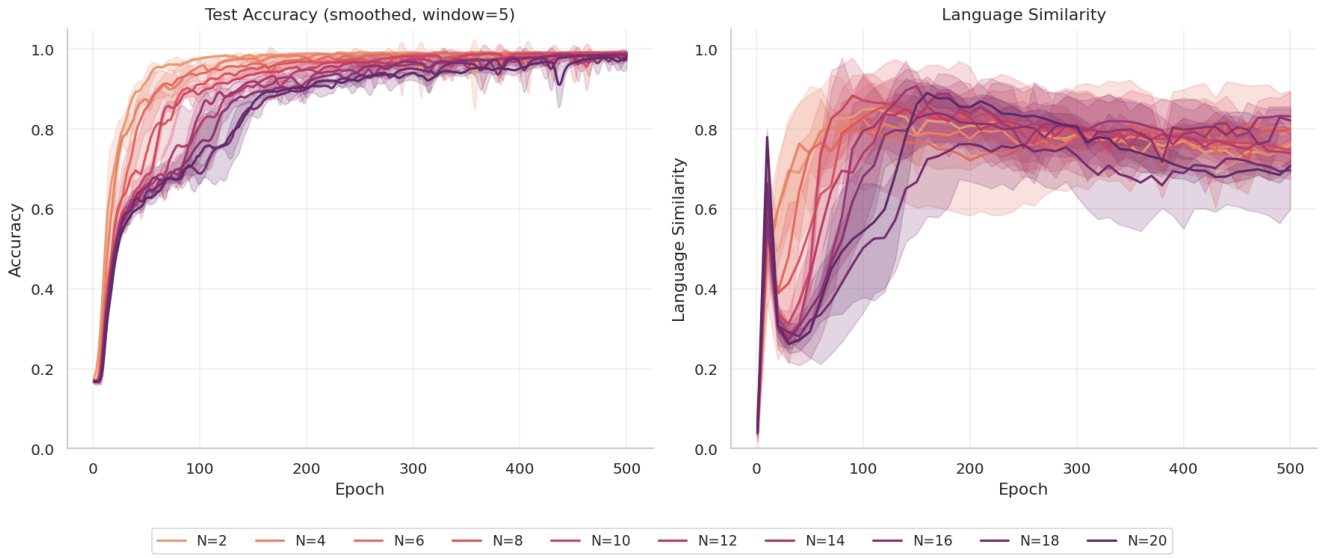

Figure 2: Learning dynamics across population sizes. Left: test accuracy over training epochs (5-epoch rolling mean). Right: language similarity over training epochs. Lines show means across 3 seeds; shaded regions indicate ±1 standard deviation. All populations converge to high accuracy and develop shared languages, though larger populations show more variance and slower initial convergence.

Neither accuracy nor language similarity shows a consistent monotonic relationship with population

size. While there is a slight tendency for larger populations to have lower final accuracy and higher variance, the differences are modest. This contrasts with findings from human experiments showing systematic effects of community size on language structure [RMLA19], but aligns with computational work suggesting that population size alone does not substantially affect emergent language properties in neural agents [CSA+22, RSG+22, MRM+22].
These baseline results confirm that static populations across the tested range can successfully develop functional communication. The metrics established here will serve as reference points for evaluating how turnover and age-based plasticity affect language emergence and stability in the following experiments.

5.3 Experiment 2: Agent Turnover

This experiment characterizes how agent turnover affects communication success and language stability. Turnover introduces two challenges: newcomers must learn the existing language from experienced agents, and the language itself may drift as the population composition changes over time.

5.3.1 Design

Experiment 2 comprises two sub-experiments. The **rate study** varies turnover frequency while holding population size constant, testing how quickly agents can be replaced before communication breaks down. The **population study** varies population size while holding turnover rate constant, testing whether larger or smaller populations better maintain language stability under turnover.
Table 2 summarizes the experimental configurations. The rate study uses a fixed population of $N = 10$ agents and varies the turnover rate $k \in \{1, 2, 5, 10, 20\}$, where k specifies the number of epochs between consecutive agent deaths. The population study uses a fixed turnover rate of $k = 10$ and varies population size $N \in \{2, 4, 8, 16\}$. Each configuration runs with three random seeds.

Table 2: Experiment 2 configurations. Both sub-experiments run for 5 complete generations.

Sub-experiment	Variable	Values	Fixed	Warmup	Seeds
Rate study	k	1, 2, 5, 10, 20	$N = 10$	Pre-trained	3
Population study	N	2, 4, 8, 16	$k = 10$	Until 95% acc.	3

The relationship between epochs, deaths, and generations requires clarification. A death occurs every k epochs, so after E epochs the population has experienced E/k deaths. A generation represents one complete population replacement cycle of N deaths. Thus generation $g = \lfloor \text{deaths}/N \rfloor + 1$. Both sub-experiments run for 5 generations, but this corresponds to different total epoch counts. For the rate study with $N = 10$: 5 generations requires $5 \times N \times k = 50k$ epochs, ranging from 50 epochs at $k = 1$ to 1000 epochs at $k = 20$. For the population study with $k = 10$: 5 generations requires $5 \times N \times k = 50N$ epochs, ranging from 100 epochs at $N = 2$ to 800 epochs at $N = 16$.
The turnover rate k creates a confounding variable in the experimental design: varying k simultaneously alters the learning opportunity for newcomers and the total optimization time for the population. A population with frequent turnover (low k) faces high disruption but undergoes fewer total gradient updates per generation, whereas a stable population (high k) accumulates more

parameter updates - and potentially more drift - over the same generational span. This trade-off makes it difficult to attribute observed drift patterns to a single cause, and motivates examining multiple complementary metrics.

The rate study loads agents from a pre-trained checkpoint where the population has already converged, isolating turnover effects from initial learning dynamics. The population study includes integrated warmup: training proceeds normally until accuracy reaches 95%, at which point turnover begins and generation counting starts. All other parameters follow Section 5.1.

5.3.2 Results

Larger populations require more training to establish an initial shared language before turnover can begin. Figure 3(a) shows warmup duration for the population study. Populations of $N = 2$ reach the 95% accuracy threshold in approximately 70 epochs, while populations of $N = 16$ require approximately 290 epochs. This pattern aligns with the baseline findings from Experiment 1 that larger populations converge more slowly.

Once turnover begins, its rate strongly affects communication success. Figure 3(b) shows test accuracy over epochs for the rate study. With $k = 20$ (one death every 20 epochs), accuracy remains stable above 90%. As turnover rate increases, accuracy degrades progressively: $k = 10$ maintains accuracy around 80%, $k = 5$ shows more disruption at 65–70%, $k = 2$ drops to 45–50%, and $k = 1$ causes severe degradation to 30–40%. The pattern reflects a repeating cycle: a newcomer enters with random weights, accuracy drops, the newcomer partially learns the language, accuracy recovers, then another newcomer enters. With very frequent turnover, the population never fully recovers between replacement events.

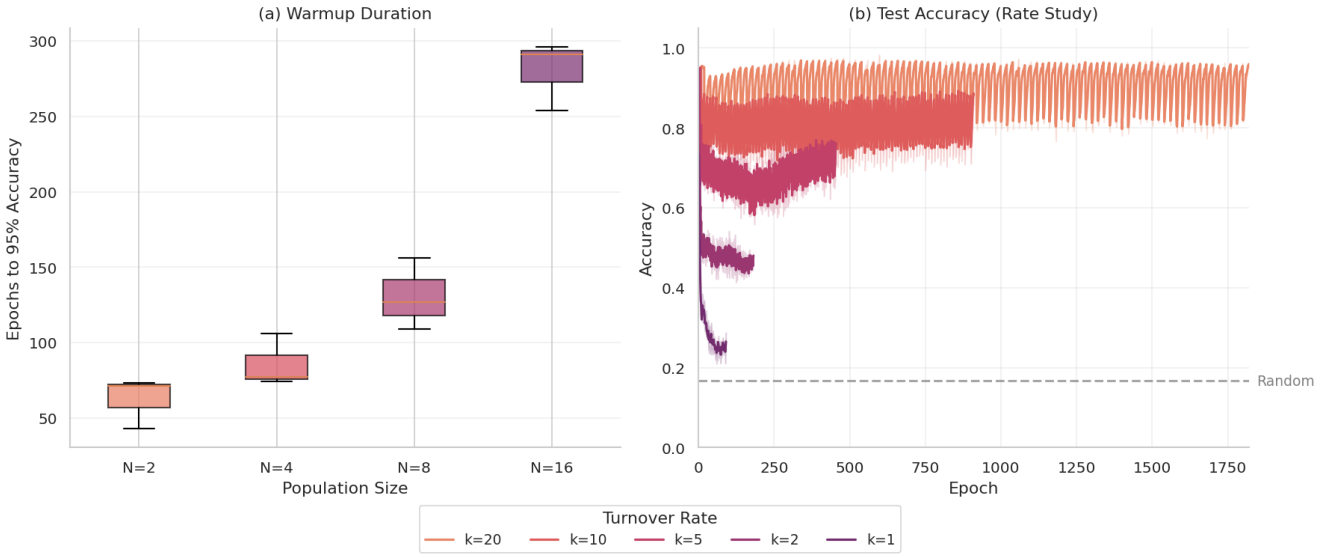

Figure 3: Initial convergence and turnover effects. (a) Warmup duration scales with population size: larger populations need more epochs to establish a shared language before turnover begins. (b) Turnover rate strongly affects accuracy: frequent turnover ($k = 1$) causes severe disruption, while infrequent turnover ($k = 20$) maintains high accuracy.

These accuracy differences create a methodological challenge for comparing drift across conditions.

Populations operating at different accuracy levels exist in fundamentally different functional regimes. A population maintaining 90% accuracy has successfully integrated newcomers and preserved communication, while one at 40% accuracy has effectively lost its shared language. Comparing cross-generational metrics between these regimes requires caution, as differences may reflect the functional state rather than drift dynamics per se.

Beyond immediate accuracy disruption, turnover causes cumulative language drift. Figure 4 shows cross-generational metrics for the rate study. Cross-generational accuracy (panel a) measures whether current agents can communicate successfully with the founding generation. Starting near 95%, it declines steadily as deaths accumulate. All turnover rates eventually converge to values between 25–35%, approaching the random baseline of 16.7%. However, slower turnover ($k = 20$) maintains higher cross-generational accuracy for longer, while faster turnover shows more rapid initial decline. Cross-generational similarity (panel b) measures whether current agents produce the same messages as founding agents for identical inputs, capturing language form rather than communicative success. This metric shows parallel decline, confirming that the message structure itself changes as the population turns over.

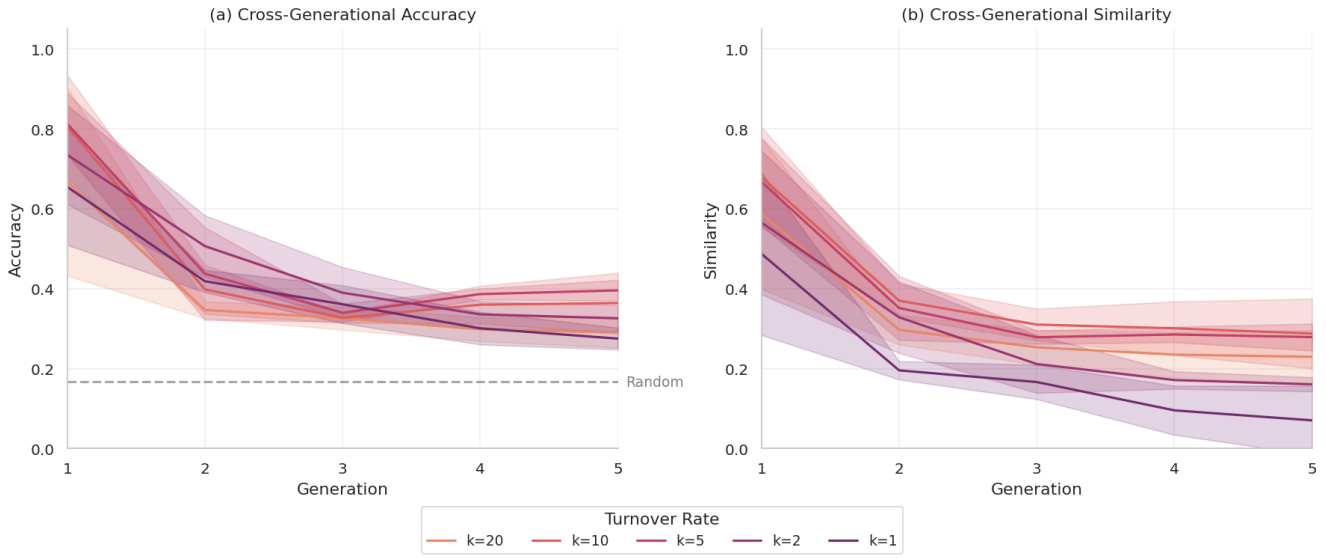

Figure 4: Language drift in the rate study ($N = 10$, varying k). (a) Cross-generational accuracy declines over generations for all turnover rates, eventually approaching random baseline. Slower turnover maintains intelligibility longer. (b) Cross-generational similarity shows parallel decline, confirming that message forms diverge from the founding language. Dashed line indicates random baseline accuracy.

The results reveal how the k trade-off manifests in practice. Higher k values produce higher within-generation accuracy because newcomers have time to learn, but they also allow more total epochs for the language to shift through optimization. Lower k values prevent effective learning, causing immediate accuracy collapse, but span fewer total epochs. When normalized by generation rather than epoch, all conditions show similar drift trajectories, suggesting that the number of replacement events matters more than the total training time.

The population study reveals how population size affects drift dynamics. Figure 5 plots cross-generational metrics by generation, normalizing the x-axis to enable comparison across population

sizes. For cross-generational accuracy (panel a), smaller populations ($N = 2$, $N = 4$) maintain higher values through early generations, while larger populations ($N = 16$) show faster initial decline. In smaller populations, each agent represents a larger fraction of the shared language, so newcomers must conform more strictly to existing conventions. In larger populations, individual agents have less influence, allowing more variation to accumulate. By generation 5, all population sizes show substantial drift, though smaller populations retain modestly higher cross-generational accuracy. Cross-generational similarity (panel b) shows a similar pattern.

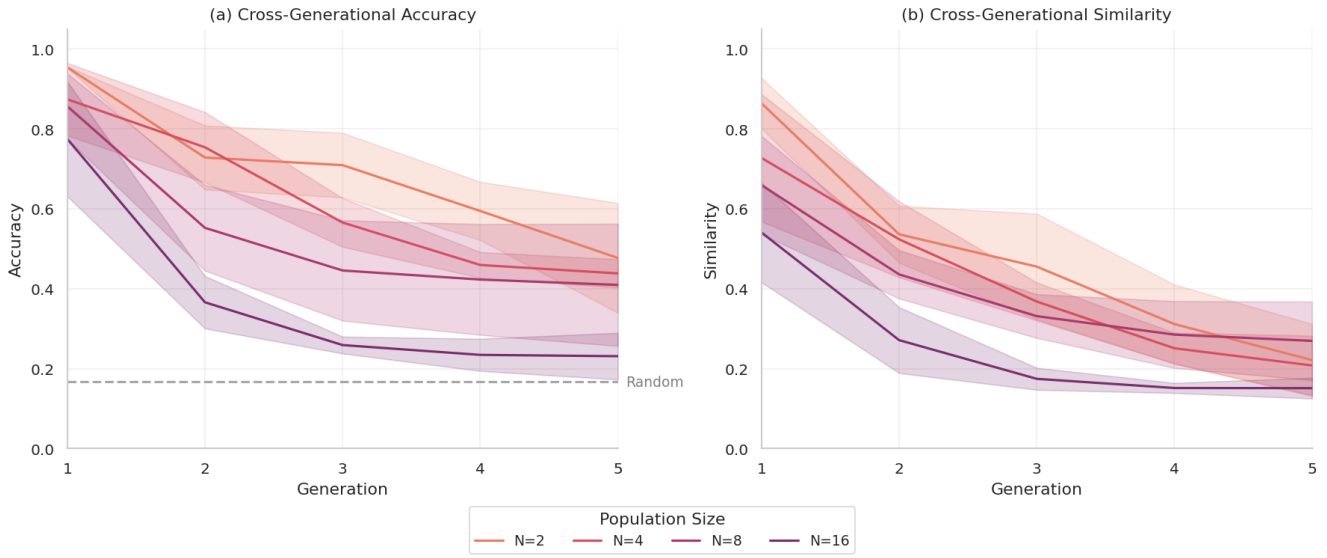

Figure 5: Language drift in the population study ($k = 10$, varying N). (a) Cross-generational accuracy by generation: smaller populations maintain higher accuracy longer. (b) Cross-generational similarity shows the same pattern. Dashed line indicates random baseline accuracy.

Interpreting cross-generational accuracy requires care. A value above the random baseline does not unambiguously indicate that the original language persists. Two alternative explanations exist: the language may have drifted while preserving some features of the original structure, or independently evolved languages may achieve above-chance accuracy due to convergent solutions imposed by the task constraints. Cross-generational similarity helps disambiguate these cases. When both metrics decline together, as observed here, the language form has genuinely changed. When accuracy remains high but similarity drops, agents may have found functionally equivalent but structurally different encodings. The parallel decline of both metrics in these experiments suggests true drift rather than convergent solutions.

These results establish that turnover creates fundamental challenges for language stability. Newcomers struggle to learn the existing language quickly enough, causing immediate accuracy drops that scale with turnover frequency. Over longer timescales, the language drifts regardless of turnover rate or population size. The founding generation's language gradually gives way to successor languages that each generation can use internally but that become increasingly incompatible with their predecessors.

5.4 Experiment 3: Full Aging Model with Plasticity

This experiment tests how does age-dependent plasticity, where younger agents learn faster but are more chaotic and older agents learn slower but are more stable, affects language stability under turnover. The hypothesis predicts that populations with age-based plasticity will show reduced drift compared to homogeneous populations with uniform levels of plasticity.

5.4.1 Design

Experiment 3 compares four plasticity conditions using the turnover framework from Experiment 2. Table 3 summarizes these conditions. The *adults* condition uses uniformly low plasticity (temperature $\tau = 0.5$, learning rate $\eta = 10^{-3}$), modeling a population where all agents resist change. The *children* condition uses uniformly high plasticity ($\tau = 5.0$, $\eta = 2.5 \times 10^{-2}$), modeling a population where all agents adapt quickly. The *age-based* condition implements the plasticity function from Section 4.5: newborn agents begin with high plasticity that decreases toward low plasticity as they age. The *baseline* condition replicates a turnover configuration from Experiment 2 without the plasticity manipulation.

Table 3: Experiment 3 plasticity conditions. All conditions use $N = 10$ agents, turnover rate $k = 10$, and run for 400 epochs (4 generations) after loading from a shared warmup checkpoint.

Condition	Temperature τ	Learning Rate η	Notes
Baseline	1.0 (fixed)	3×10^{-3} (fixed)	Plasticity disabled
Adults	0.5 (fixed)	10^{-3} (fixed)	Uniform low plasticity
Children	5.0 (fixed)	2.5×10^{-2} (fixed)	Uniform high plasticity
Age-based	0.5 – 5.0	10^{-3} – 2.5×10^{-2}	Varies with age, Gen 1 agents start as adults

All conditions use population size $N = 10$ and turnover rate $k = 10$. Training runs for 400 epochs post-warmup, spanning 5 complete generations (40 agent deaths). Each condition loads from the same pre-trained warmup checkpoint (baseline settings), ensuring identical starting conditions. The $a_{\max} = 100$ parameter is set to 100, which is equal to agents' lifespan. Each configuration runs with three random seeds.

The experimental design tests two predictions. First, populations with homogeneous structures and extreme plasticity should perform worse in terms of communication success. All-adult populations may be too slow to learn and fail to maintain a language under turnover. All-children populations may be too chaotic to transfer any knowledge to newcomers. Second, heterogeneous populations with mixed plasticity should outperform homogeneous populations with uniform plasticity in Cross-Generational metrics.

5.4.2 Results

Figure 6 shows test accuracy and language similarity over training epochs. Firstly, in all conditions, we can observe the cycle cause by turnover (described in Section 5.3). The baseline, adult, and age-based populations all maintain relatively similar accuracy levels, with adults slightly behind

the other two. Their accuracies are in the range of 70-90% depending on the phase of the cycle. The children condition however, already starts from very low initial accuracy - around 25%. This suggests that the population is so chaotic that it immediately loses the language formed during the warmup. The population regains some of the accuracy but plateaus in the 60-70% range and does not show significant improvement past generation 2. Language similarity shows that all conditions maintain reasonably synchronized languages, though they all show high variance across seeds. Adults are noticeably worse than other populations. Notably, the drop in accuracy for the children population does not occur in the language similarity curve.

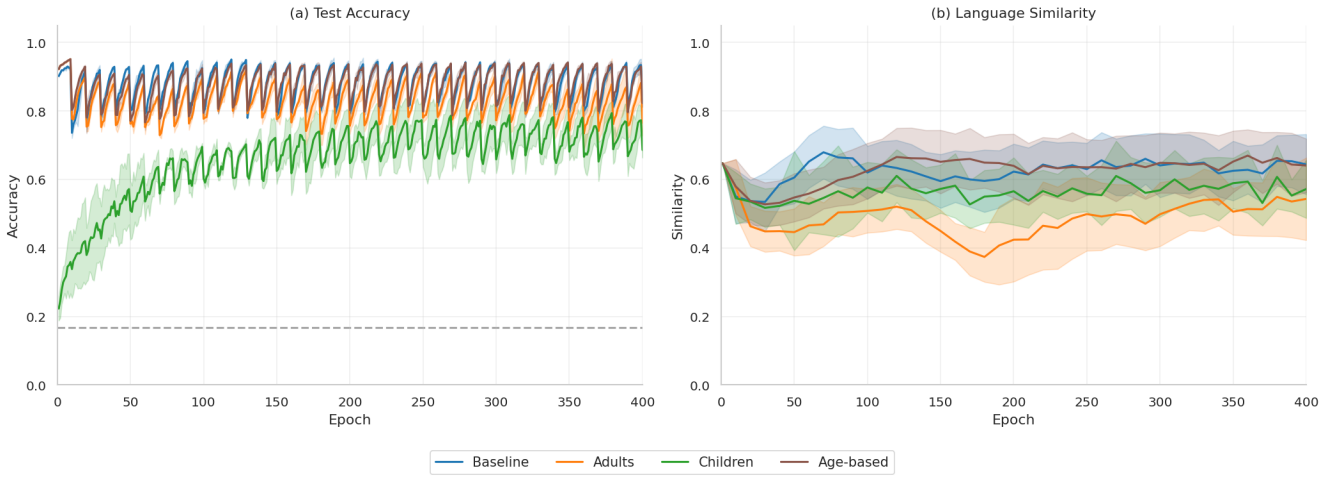

Figure 6: Accuracy and Language Similarity across plasticity conditions. (a) Test accuracy over epochs. The children condition experiences a big drop right after the warmup and converges to the lowest accuracy. Age-based and baseline maintain (very similar) high accuracy, with adults only slightly behind. (b) Language similarity over epochs. All conditions maintain relatively synchronized languages. The most significant difference within the conditions is that adults maintain the lowest language similarity throughout all generations. All conditions show large variance. Lines show means across 3 seeds; shaded regions indicate ±1 standard deviation.

Cross-generational metrics reveal substantial differences between conditions. Figure 7 shows cross-generational accuracy and similarity measured against the founding generation (Gen 1). Firstly, all curves exhibit certain similarities. As expected, in all cases, the drift metrics values decrease throughout time. The decrease is more significant at first, and diminishes over time. The age-based condition maintains significantly higher cross-generational accuracy than all other conditions. Its curve is flatter, which suggests less language drift from generation to generation. Cross-generational similarity shows a similar pattern. The children population experiences a substantial drop in both metrics, almost immediately after turnover is activated, similar to the drop in static metrics. The adults condition's curve resembles the shape of the baseline curve, only more pronounced, with lower values (higher drift) the majority of the time. The exception is the very beginning of turnover - at first, adults preserve more of the original language (likely due to their superior stability), but this effect is quickly lost, as more newcomers are introduced in the place of the original agents and are not able to learn as effectively.
To capture the rate of language change between consecutive generations, Figure 8 shows cross-generational accuracy and similarity measured against the previous generation rather than Gen 1.

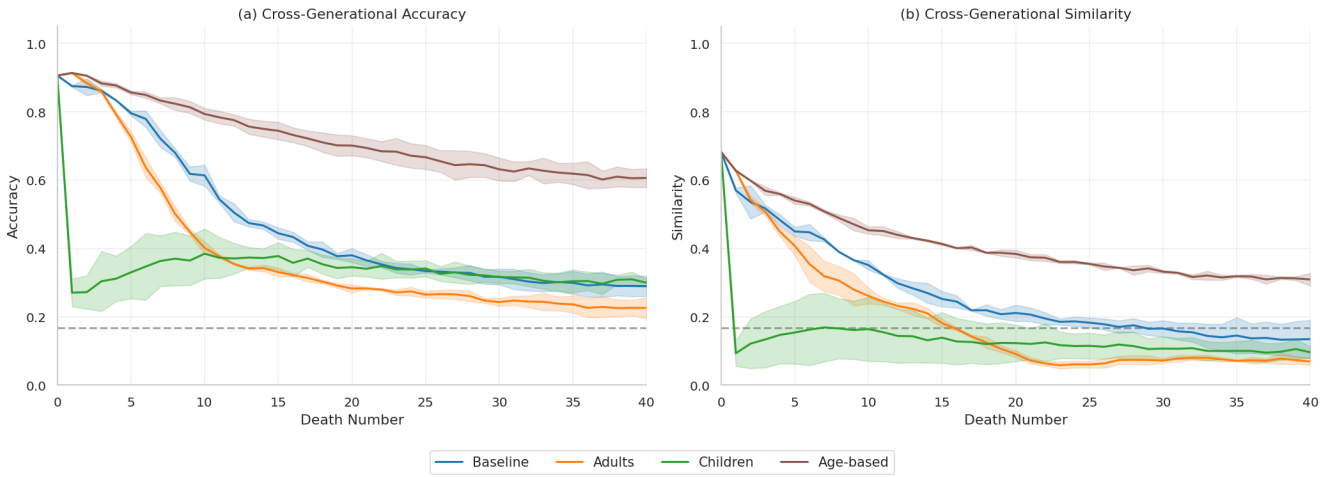

Figure 7: Language drift across plasticity conditions, measured against the original generation (Generation 1). (a) Cross-generational accuracy by death number. The age-based condition maintains substantially higher accuracy with Generation 1 throughout training. (b) Cross-generational similarity by death number. Age-based preserves message similarity far better than uniform plasticity conditions. Lines show means across 3 seeds; shaded regions indicate ±1 standard deviation. Dashed line indicates random baseline accuracy (16.7%).

The metrics confirm the trend visible in Figure 7: more drift in the beginning, less drift further in the simulation. Age-based condition maintains 80-90% accuracy with its previous generations at all times, indicating low drift. These metrics reveal that the age-based population preserves the language throughout generations not just cumulatively but at each generational transition.
Table 4 summarizes all metrics after 5 generations. The cross-generational metrics compare the final generation to the founding generation, while the previous-generation metrics are averaged across all generational transitions (generations 2 through 5) to capture the overall drift rate.

Table 4: Metrics after 5 generations (mean ± std across 3 seeds). Cross-Gen metrics compare Gen 5 to Gen 1. Prev metrics are averaged across all generational transitions.

Metric	Baseline	Adults	Children	Age-based
Test Accuracy (%)	82.2 ± 0.8	76.6 ± 3.0	68.6 ± 1.8	**82.3 ± 0.9**
Language Similarity	0.705 ± 0.02	0.601 ± 0.10	0.615 ± 0.08	**0.709 ± 0.03**
Cross-Gen Accuracy	0.290 ± 0.02	0.226 ± 0.03	0.299 ± 0.01	**0.606 ± 0.02**
Cross-Gen Similarity	0.134 ± 0.05	0.069 ± 0.01	0.096 ± 0.01	**0.309 ± 0.02**
Avg Prev Accuracy	0.789 ± 0.12	0.657 ± 0.16	0.608 ± 0.15	**0.883 ± 0.05**
Avg Prev Similarity	0.479 ± 0.11	0.348 ± 0.12	0.396 ± 0.18	**0.597 ± 0.09**

The comparison to baseline reveals the central finding of this experiment. Both homogeneous conditions perform worse than baseline on cross-generational metrics: adults show 22% lower cross-generational accuracy and 49% lower similarity, while children show 28% lower similarity. However, the age-based condition—which combines the properties of both—substantially outperforms baseline:

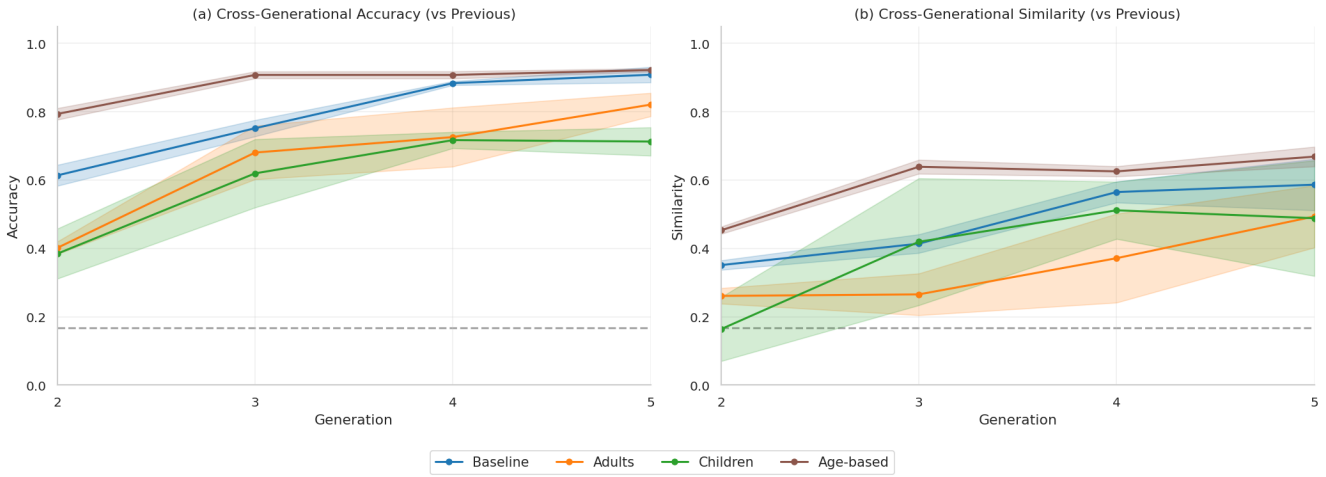

Figure 8: Language drift across plasticity conditions, measured against the previous generation (Generation $g-1$) (a) Cross-generational accuracy by generation. Age-based maintains over 80% accuracy with the previous generation at all times. (b) Cross-generational similarity by generation. Age-based shows the highest and most consistent similarity with its predecessor. Large variance within seeds for children and adults. Lines show means across 3 seeds; shaded regions indicate ± 1 standard deviation. Dashed line indicates random baseline accuracy (16.7%).

109% higher cross-generational accuracy (0.606 vs 0.290) and 131% higher cross-generational similarity (0.309 vs 0.134). The average previous-generation accuracy tells a similar story: children (0.608) and adults (0.657) both fall below baseline (0.789), but age-based (0.883) exceeds it by 12%. This pattern supports the hypothesis that age-based heterogeneity matters. The adults condition fails because agents cannot adapt quickly enough - newcomers struggle to learn from rigid teachers, and the population fragments. The children condition fails because agents change too rapidly - the language drifts faster than any stable convention can form. Neither extreme works well in isolation. But when combined in the age-based condition, each compensates for the other's weakness. Older agents with low plasticity provide stable, consistent learning targets. Younger agents with high plasticity can efficiently acquire this stable language before their own plasticity decreases. The result is a population that maintains both high within-generation accuracy and low cross-generational drift.

6 Discussion

6.1 Interpretation of Results

The central question of this thesis asked whether age-dependent plasticity improves language stability in populations with agent turnover. The experiments provide strong support for this hypothesis. Age-based plasticity substantially reduces language drift compared to uniform plasticity conditions, with the age-based population achieving more than double the cross-generational accuracy of both baseline and the homogeneous alternatives.

Experiment 2 established that turnover creates two distinct challenges for language stability. First, newcomers disrupt communication immediately upon entering the population. The severity of this

disruption scales with turnover frequency: populations with frequent replacement never fully recover between events, while populations with infrequent replacement maintain high accuracy. Second, the language itself drifts over time regardless of turnover rate. Cross-generational accuracy and similarity both decline as generations accumulate, confirming that the founding language gradually gives way to successor variants. Smaller populations preserve cross-generational communication longer than larger ones, likely because each agent carries more influence over the shared protocol.
Experiment 3 revealed that neither extreme of uniform plasticity works well under turnover. The adults condition fails because agents cannot adapt quickly enough - newcomers are too slow to learn, even despite having stable teachers and the language gradually decays. The children condition fails for the opposite reason: there are no stable teachers to learn from and even though children can learn quickly, the language drifts faster than stable conventions can form. Additionally, there is an immediate collapse in accuracy right after the warmup - indicating that children alone are simply not able to maintain the developed language. Only later, they are able to slowly recover. Both homogeneous conditions underperform the baseline on cross-generational metrics, establishing that the problem is not simply having too little or too much plasticity, but having the wrong distribution of it.
The age-based condition combines both extremes and substantially outperforms all alternatives. Cross-generational accuracy reaches 0.606 - more than double the baseline (0.290). The previous-generation metrics confirm this is not merely a coincidence: age-based populations maintain 80-90% accuracy with their immediate predecessors throughout training, higher than homogeneous conditions. This pattern supports the hypothesis that the age-based structure drives the stabilizing effect. Older agents with low plasticity provide stable, consistent learning targets. Younger agents with high plasticity efficiently acquire this stable language before their own plasticity decreases. Neither component works in isolation, but together they create a population that balances adaptability with stability.
Age-based plasticity does not eliminate drift entirely. By generation 5, cross-generational accuracy has declined to approximately 0.6. The founding language is not fully preserved; rather, the rate of change is reduced. This suggests that age-based plasticity is one mechanism (perhaps among several others) that might contribute to language stability.

6.2 Implications

These findings contribute to emergent communication research in several ways. First, they demonstrate that population heterogeneity can substantially affect language dynamics, extending prior work by Rita et al. [RSG+22] on learning rate asymmetries. In this case, heterogeneity is age-dependent. The failure of both homogeneous extremes, combined with the success of the heterogeneous condition, suggests that the interaction between stable and plastic agents is more important than the absolute level of either property.
Second, the results highlight the value of measuring drift at multiple timescales. Cross-generational metrics against the founding generation capture cumulative drift, while previous-generation metrics capture drift velocity. The age-based condition excels on both: it preserves more of the original language and changes less at each generational transition. This dual advantage suggests that the stabilizing mechanism operates continuously rather than merely slowing initial divergence.
For theories of language evolution, the results provide computational evidence that age-based plasticity can substantially reduce language drift under turnover. The effect sizes observed here -

age-based populations achieving higher cross-generational accuracy than other conditions - suggest this mechanism could play a meaningful role in maintaining cross-generational intelligibility in humans. The findings strengthen Gopnik's case [Gop20] that in human communities children's high plasticity and adults' stability may together preserve language across generations. They also extend Verhoef's and de Boer's [VdB11] to deep neural network agents and find similar patters as those in their work.

6.3 Limitations

Several limitations constrain the interpretation of these results. The experimental design creates a puzzle between turnover rate and total training time. Varying the kill epoch parameter k simultaneously changes how much time newcomers have to learn and how many optimization steps occur per generation. This makes it difficult to isolate whether observed differences stem purely from learning dynamics or simply the number of interactions between agents. Future work could aim to better decouple these factors through alternative experimental designs.

The metrics used to measure drift have inherent limitations. Cross-generational accuracy conflates true language preservation with the possibility that task constraints push independently evolved languages toward similar solutions. Cross-generational similarity measures surface form but not semantic content. The parallel decline of both metrics suggests genuine drift, but this interpretation rests on assumptions that cannot be fully verified.

The scale and complexity of these experiments differ substantially from human language communities. Population sizes of 2 to 10 agents are far smaller than human speech communities. The task, with 625 possible inputs and 6 candidate objects, may be too simple to require the compositional structure that age-based plasticity might help preserve. More complex tasks could reveal larger effects or different dynamics entirely.

The implementation of plasticity as temperature and learning rate modulation provides only a simple approximation of biological processes. Human age-related changes in language learning involve qualitatively different mechanisms including neural maturation, explicit teaching by other community members, and various social factors. Plasticity function, temperature and learning rate are somewhat arbitrary - other functional forms might produce different results, although the research seems to confirm findings of Rita et al. [RSG+22], where the relative values of parameters matter more than their absolute values in the context of heterogeneity. In other words, although the minimum and maximum plasticity levels are chosen arbitrarily, the fact that their blend produces significantly better results than all the homogeneous conditions is still scientifically relevant.

Finally, three random seeds per condition provides limited statistical power. The consistent ordering of conditions across all metrics and the clear separation between age-based and other conditions increase confidence in the findings, but more replications would strengthen these conclusions.

6.4 Future Work

Several directions merit further investigation. The current experiment tests one configuration of plasticity parameters. Systematic exploration of the temperature and learning rate ranges could identify optimal settings and clarify how extreme the difference between young and old agents needs to be for the stabilizing effect to emerge. The plasticity trajectory (controlled by the sigmoid's

steepness and critical point) also calls for investigation: does a gradual decline work better than a sharp transition?

Extending experiments to larger populations would test whether the stabilizing effect of age-based plasticity scales to more realistic community sizes. The current experiments use populations of 2 to 20 agents but human communities involve many more speakers. Larger populations might amplify the heterogeneity effect, as more agents occupy different points along the plasticity spectrum at any given time, or might dilute it if individual stable agents have less influence on the collective language.

The current framework investigates language drift mostly as degradation, but language change also serves an adaptive function. Human languages evolve to accommodate new concepts, technologies, and social structures. Future work could test whether age-based populations adapt better to changing environments by periodically introducing new input combinations that require learning new conventions. The question becomes whether heterogeneous populations will be better at simultaneously maintaining high accuracy among currently alive agents and efficiently inventing new words for novel concepts.

The problem identified in Experiment 2 between turnover rate and training time could be addressed through alternative drift metrics. Populations with slow turnover ($k = 20$) experience fewer deaths over the same epoch count as fast turnover ($k = 2$), but accumulate more gradient updates per death. Measuring drift per death would isolate the effect of each replacement event, while measuring drift per epoch would isolate the effect of continued optimization. Comparing these metrics across turnover rates could disentangle whether drift stems primarily from the disruption of agent replacement or from the gradual parameter changes that occur during training.

Tasks requiring compositional generalization would test whether age-based plasticity preferentially preserves structured languages, since prior work suggests compositional languages are easier for newcomers to learn. Alternative plasticity implementations - such as varying network capacity with age or combining plasticity with explicit imitation mechanisms - could test whether the observed effects depend on the specific temperature and learning rate manipulation used here.

7 Conclusion

This thesis asked whether age-dependent plasticity improves language stability in populations with agent turnover. The answer is yes: populations where plasticity decreases with age preserve cross-generational communication far better than populations with uniform plasticity.

Three experiments addressed this question. Experiment 1 confirmed that populations of 2 to 20 agents reliably develop shared languages in static conditions, establishing a baseline for subsequent tests. Experiment 2 introduced turnover and documented its effects: newcomers disrupt communication upon arrival, and the language drifts over time regardless of turnover rate or population size. Smaller populations and slower turnover delay drift but do not prevent it. Experiment 3 tested the central hypothesis by comparing age-based plasticity against uniform alternatives. Populations with all high-plasticity agents failed because the language changed faster than conventions could stabilize. Populations with all low-plasticity agents failed because newcomers could not learn quickly enough. The age-based condition, which combined both properties, achieved not only lower drift than only children or only adults, but also than a balanced homogeneous baseline.

The mechanism behind this effect follows a simple logic. Older agents with low plasticity anchor the

language by providing consistent, stable targets for newcomers to learn from. Younger agents with high plasticity acquire this stable language efficiently before their own plasticity declines. Neither property works alone. Low plasticity without high plasticity means newcomers cannot adapt. High plasticity without low plasticity means no one provides a stable target. The combination creates a population that both preserves existing conventions and integrates new members.

These results support the hypothesis, drawn from cognitive science, that developmental changes in learning ability help maintain language across generations. Gopnik proposed that children's plasticity and adults' stability together enable languages to persist despite continuous turnover in human communities. The experiments here provide computational evidence for this view: in simulated populations, age-based plasticity reduces drift in ways that uniform plasticity cannot.

Several questions remain open. The experiments used small populations, simple tasks, and arbitrary plasticity mechanisms. Whether the same principles apply at larger scales, with more complex languages, or through different implementations of age-related learning requires further investigation.

The central contribution of this thesis is a demonstration that how plasticity is distributed across a population matters for language stability. This finding connects emergent communication research to theories of human language evolution and suggests that population structure deserves attention alongside the properties of individual agents.

Acknowledgments

I would like to thank Dr. Tessa Verhoef for her supervision, guidance, and thoughtful engagement throughout this project. I would like to also thank Dr. Flor Miriam Plaza del Arco for her valuable feedback and support as second supervisor. I am grateful to the Leiden Institute of Advanced Computer Science for providing the computational resources that made this research possible, and to the REL Compute team for maintaining the infrastructure on which all experiments were conducted.

References

[BB18] Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game, 2018.

[BK06] Henry Brighton and Simon Kirby. Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, 12(2):229–242, 2006.

[BM24] Brendon Boldt and David Mortensen. A review of the applications of deep learning-based emergent communication, 2024.

[CKB+20] Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. Compositionality and generalization in emergent languages. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Online, July 2020. Association for Computational Linguistics.

[CKDB19] Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. Anti-efficient encoding in emergent communication, 2019.

[CLL+20] Michael Cogswell, Jiasen Lu, Stefan Lee, Devi Parikh, and Dhruv Batra. Emergence of compositional language with deep generational transmission, 2020.

[CSA+22] Rahma Chaabouni, Florian Strub, Florent Altché, Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. Emergent communication at scale. In *International Conference on Learning Representations*, 2022.

[FAdFW16] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning, 2016.

[GCK19] Laura Harding Graesser, Kyunghyun Cho, and Douwe Kiela. Emergent linguistic phenomena in multi-agent communication games. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3700–3710, 2019.

[Gop20] Alison Gopnik. Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1803):20190502, 06 2020.

[GRR22] Lukas Galke, Yoav Ram, and Limor Raviv. Emergent communication for understanding human language evolution: What's missing? In *Emergent Communication Workshop at ICLR 2022*, 2022.

[HTP18] Joshua K. Hartshorne, Joshua B. Tenenbaum, and Steven Pinker. A critical period for second language acquisition: Evidence from 2/3 million english speakers. *Cognition*, 177:263–277, 2018.

[JGP16] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[KB20] Eugene Kharitonov and Marco Baroni. Emergent language generalization and acquisition speed are not tied to compositionality, 2020.

[KCBB19] Eugene Kharitonov, Rahma Chaabouni, Diane Bouchacourt, and Marco Baroni. EGG: a toolkit for research on emergence of lanGuage in games. In Sebastian Padó and Ruihong Huang, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 55–60, Hong Kong, China, November 2019. Association for Computational Linguistics.

[KGJV83] Scott Kirkpatrick, C Daniel Gelatt Jr, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[KGS14] Simon Kirby, Tom Griffiths, and Kenny Smith. Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28:108–114, 2014. SI: Communication and language.

[KMLB17] Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog, 2017.

[KO21] Jooyeon Kim and Alice Oh. Emergent communication under varying sizes and connectivities. *Advances in Neural Information Processing Systems*, 34:17579–17591, 2021.

[Lab07] William Labov. Transmission and diffusion. *Language*, 83(2):344–387, 2007.

[LB19] Fushan Li and Michael Bowling. Ease-of-teaching and language structure from emergent communication, 2019.

[LB20] Angeliki Lazaridou and Marco Baroni. Emergent multi-agent communication in the deep learning era, 2020.

[LCK19] Jason Lee, Kyunghyun Cho, and Douwe Kiela. Countering language drift via visual grounding. *arXiv preprint arXiv:1909.04499*, 2019.

[Len67] Eric H Lenneberg. The biological foundations of language. *Hospital Practice*, 2(12):59–67, 1967.

[Lew08] David Lewis. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA, 2008.

[LHTC18] Angeliki Lazaridou, Karl Moritz Hermann, Karl Tuyls, and Stephen Clark. Emergence of linguistic communication from referential games with symbolic and pixel input, 2018.

[LPB17] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language, 2017.

[LPT20] Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. Multi-agent communication meets natural language: Synergies between functional and structural language learning. *CoRR*, abs/2005.07064, 2020.

[LSS+20] Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. Countering language drift with seeded iterated learning. In *International Conference on Machine Learning*, pages 6437–6447. PMLR, 2020.

[LVB24] Yuchen Lian, Tessa Verhoef, and Arianna Bisazza. Nellcom-x: A comprehensive neural-agent framework to simulate language learning and group communication, 2024.

[MFDB25] Matéo Mahaut, Francesca Franzon, Roberto Dessì, and Marco Baroni. Referential communication in heterogeneous communities of pre-trained visual deep networks, 2025.

[MRM+22] Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, and Angeliki Lazaridou. Revisiting populations in multi-agent communication, 2022.

[RGL+20] Yi Ren, Shangmin Guo, Matthieu Labeau, Shay B. Cohen, and Simon Kirby. Compositional languages emerge in a neural iterated learning model, 2020.

[RMLA19] Limor Raviv, Antje Meyer, and Shiri Lev-Ari. Larger communities create more systematic languages. *Proceedings of the Royal Society B*, 286(1907):20191262, 2019.

[RSG+22] Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. On the role of population heterogeneity in emergent communication, 2022.

[TLM+19] Olivier Tieleman, Angeliki Lazaridou, Shibl Mourad, Charles Blundell, and Doina Precup. Shaping representations through communication: community size effect in artificial learning systems. *arXiv preprint arXiv:1912.06208*, 2019.

[VdB11] Tessa Verhoef and Bart de Boer. Language acquisition age effects and their role in the preservation and change of communication systems. *Linguistics in Amsterdam*, 4, 01 2011.

[vdSSBvH22] Frans van der Slik, Job Schepens, Theo Bongaerts, and Roeland van Hout. Critical period claim revisited: Reanalysis of hartshorne, tenenbaum, and pinker (2018) suggests steady decline and learner-type differences. *Language Learning*, 72(1):87–112, 2022.

[WRUW03] Kyle Wagner, James A Reggia, Juan Uriagereka, and Gerald S Wilkinson. Progress in the simulation of emergent communication and language. *Adaptive Behavior*, 11(1):37–69, 2003.

A Reproducibility

A.1 Code Availability

The source code is available at https://github.com/Jano04/EGG. The implementation extends the EGG framework [KCBB19] with modules for population dynamics and age-based plasticity in `egg/zoo/aging/`. The repository README provides installation instructions and usage details.

A.2 Hardware

All experiments ran on REL Compute infrastructure at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University. Specifically, I used the `vibranium.liacs.nl` server:

- CPU: 24 Intel Xeon Silver 4214 cores @ 2.20GHz (48 threads)
- GPU: 2× NVIDIA GeForce RTX 3090 (24GB memory each)
- RAM: 256GB
- OS: Rocky Linux 9

A.3 Software Environment

The `requirements.txt` file lists all Python dependencies. Install with: `pip install -e .`

A.4 Running Experiments

Python grid files in `egg/zoo/aging/grids/` define experiment configurations. Execute them using EGG's built-in grid search tool:

```
python -m egg.nest.nest_local --game egg.zoo.aging.train \
    --py_sweep=egg.zoo.aging.grids.<grid_name>
```

The repository README provides detailed instructions.

A.5 Data Generation

The training procedure generates the dataset procedurally using a fixed seed (111 by default), ensuring identical data across runs. No external datasets are required.

A.6 Default Hyperparameters

Table 5 lists hyperparameters shared across all experiments. Section 5 documents experiment-specific variations.

Table 5: Default hyperparameters.

Parameter	Value
Batch size	1024
Learning rate	10^{-3}
Weight decay	10^{-5}
Hidden dimension	128
Embedding dimension	64
Vocabulary size	20
Max message length	10
Optimizer	Adam
Data seed	111

A.7 Random Seeds

Each configuration ran with 3 random seeds. Results report mean ± standard deviation. The `--deterministic` flag enables fully deterministic execution.