How Cultural Transmission and Individual Learning Shape the Emergence of Compositionality

MSc Thesis *(Afstudeerscriptie)*

written by

Francijn Barbara Keur
(born October 21st, 2000 in Leiden, the Netherlands)

under the supervision of **Dr. Raquel G. Alhama**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam.*

Date of the public defense:	**Members of the Thesis Committee:**
February 9th, 2026	Dr. Raquel G. Alhama (supervisor)
	Dr. Benno van den Berg (chair)
	Dr. Fausto Carcassi
	Dr. Marieke Schouwstra

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

Abstract

This thesis examines the impact of several factors affecting cultural transmission, such as transmission mode, population size, social network structure, and rate of replacement, on the emergence of a preference for compositionality. It also investigates the impact of individual learning mechanisms, such as the impact of new evidence and the strength of a simplicity bias. Using an agent-based model with Bayesian learners, I replicate earlier findings that a preference for compositionality emerges due to a trade-off between *compressibility* and *expressivity*. When agents learn from multiple teachers, a preference for compositionality can emerge even with a relatively weak simplicity bias. Population size and network structure do not seem to impact the emergence of preference for compositionality, though this might be due to the limited possibility for input variability in the model.

Contents

Chapter 1 – Introduction

It has been proposed that studying the evolution of human language could be the hardest problem in science (Kirby & Christiansen, 2003). One major difficulty is the fact that spoken and signed languages have not left behind any direct fossils to study (Arnon et al., 2025; Fitch, 2010). To overcome these problems, the field of language evolution uses a broad set of diverse methodologies from various scientific disciplines. Examples are computer simulations, laboratory experiments with artificial languages, fieldwork investigating recently emerged sign languages, or comparative studies with animal communication systems (Fitch, 2010; Raviv & Boeckx, 2025; Steels, 2011b). By combining insights from these different fields, researchers can gain more insight into how language emerged and evolved.

A widely proposed theory is that language evolved through interactions of three processes that operate on different time scales: individual language learning and use, cultural evolution, and biological evolution (Arnon et al., 2025; Kirby et al., 2007). More precisely, genes shape the learning mechanisms that in turn influence the outcomes of cultural evolution, and these outcomes modify the fitness landscape, as shown in Figure 1.1.

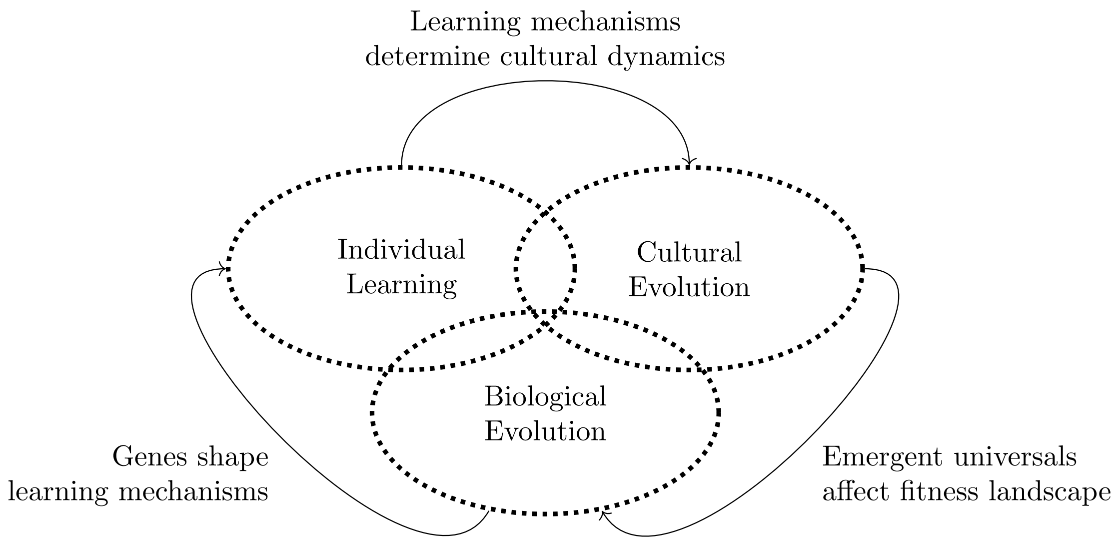


Figure 1.1: Language evolution is affected by the interactions between three processes operating on different time scales: Individual learning, cultural evolution, and biological evolution. Adapted from Figure 1 in Kirby et al. (2007).

This thesis focuses on the emergence of one specific property of language, *compositionality*, which has been defined as follows:

> *Compositionality*: The meaning of a complex expression is compositional if it is determined by its structure and the meanings of its constituents.(Pickel & Szabó, 2025).

Compositionality, sometimes more broadly referred to as linguistic structure, is seen as one of the core features of language: "Anything that deserves to be called a language must contain meaningful expressions built up from other meaningful expressions" (Pickel & Szabó, 2025). Together with recursion, compositionality allows us to generate infinitely many meaningful sentences from a finite set of units, and to generalize to novel, unseen situations (Partee et al., 1995). Consider, for example, the phrase "indigo coffee mug". While it is likely that you have never heard this precise combination of words before, you can still understand it, thanks to your knowledge of the individual word meanings and the compositional nature of language.

Learning more about how compositionality evolved not only teaches us more about what pressures shaped language to be the powerful communication system that it is today, but it might also inform us how we can guide artificial communication systems (e.g., neural agents or robots) towards compositionality, and thus more effective communication.

Over the past 25 years, a growing body of work has focused on the emergence of compositionality, with an emphasis on the cultural evolution of language (Arnon et al., 2025; Tamariz & Kirby, 2016). This stands in contrast to earlier work that tried to explain linguistic structure through biological evolution (e.g., Pinker & Bloom, 1990). From this cultural-evolutionary perspective, linguistic structure emerges through repeated cycles of *cultural transmission*, as language is used between individuals and across generations. Multiple modes of cultural transmission can be distinguished, such as horizontal transmission (interactions within one generation), vertical transmission (cross-generational interactions between biologically related members), and oblique transmission (cross-generational interactions between unrelated members) (Cavalli-Sforza & Feldman, 1981; Gong, 2010). Next to these transmission types, other aspects of transmission dynamics—such as population size and social network structure (who interacts with whom)—also seem to play an important role in shaping languages (e.g., Gong et al., 2011; Josserand et al., 2021, 2024; Lupyan & Dale, 2010; Raviv et al., 2019b, 2020).

However, how precisely the different transmission modes and other aspects of transmission dynamics interact with the emergence of compositionality remains unknown. Moreover, as illustrated in Figure 1.1, the results of cultural evolution are influenced by individual learning mechanisms. To obtain a comprehensive understanding of how compositionality emerges, this thesis investigates both the effects of transmission dynamics and the role of individual learning mechanisms on the emergence of a preference for compositionality. Accordingly, the research questions addressed in this thesis are divided into two categories:

Cultural Transmission:

(i) How do horizontal, vertical, and combined modes of transmission impact the emergence of compositionality?

(ii) How does population size impact the emergence of compositionality in a setting with both horizontal and vertical transmission?

(iii) How does social network structure impact the emergence of compositionality?

(iv) How does the rate of agent replacement affect the emergence of compositionality?

Individual Learning Mechanisms:

(v) How does regulating the impact of new evidence modify the emergence of compositionality under different transmission modes?

(vi) How does the strength of a simplicity bias influence the emergence of compositionality?[1]

The remainder of this thesis is structured as follows. First, I review the relevant literature in Chapter 2. Then, Chapter 3 introduces the Computational Framework, the agent-based model with Bayesian learners, that serves as the foundation for all experiments in this thesis. Chapter 4 presents experiments addressing research questions on Cultural Transmission (i-iv). Chapter 5 contains the experiments addressing the research questions on the individual learning mechanisms (v and vi). Chapter 6 revisits research questions (ii) and (iii), integrating insights from both Chapter 4 and Chapter 5. Lastly, Chapter 7 presents the general discussion and conclusions.

[1]The simplicity bias is in this case a preference for more compressible languages, which will be explained in more detail in Subsection 3.3.1

Chapter 2 – Literature Review

In this literature review, I will discuss previous work using two different methodological approaches commonly used in the area of language evolution, namely computer simulations and artificial language learning in the lab. The reason for focusing on these two methods is that both allow for systemic manipulation of variables in controlled conditions, which is less straightforward to do with, for example, fieldwork investigating newly emerging sign languages (Sandler et al., 2025). I begin by broadly outlining these two methods and will discuss their strengths, weaknesses, and related fields. Then I will go into more detail on previous work regarding compositionality and the three factors influencing cultural transmission: i) horizontal and vertical transmission (and their combination), ii) population size, and iii) network structures. Each factor is addressed in a separate section; the section on transmission modes also includes some previous work concerning individual learning mechanisms.

2.1.1 Artificial Language Learning in the Lab

Over the past two decades, researchers have investigated newly emerging language systems in controlled laboratory settings. In these kinds of studies, participants had to learn an artificial language or play a communication game, which were designed to simulate aspects of language evolution. To limit the impact of a participant's linguistic knowledge, many different communication media were used. Examples of these novel ways of communicating include, for example, silent gesture experiments (e.g., Motamedi et al., 2019), communicating with a sliding whistle (Verhoef, 2012; Verhoef et al., 2014), a drawing pad (e.g., Roberts et al., 2015), or an artificial language (e.g., Kirby et al., 2008). An advantage of these lab experiments is the highly controllable setting, where a researcher can carefully design experiments to investigate specific cognitive, social, or cultural pressures that might impact language change. Another advantage is that since these experiments are done with human participants, real-life cognitive biases and learning mechanisms are included. A downside is, however, that all participants are language-capable and have a "language-ready" brain, which might constrain the generalizability of the results. Additionally, these artificial settings cannot fully mimic how language changes in the real world. Lastly, practical constraints such as time, funding, and ethical considerations further limit the scale and scope of these kinds of experiments.

2.1.2 Computational Simulations

Computer simulations are widely used in the field of language evolution (Raviv & Boeckx, 2025). One of the advantages of computational work is the fact that researchers have to be very explicit about the assumptions they make (Smaldino, 2017). Sometimes it is argued that a formal model is too simple; often, many details that we see in the real world are purposely omitted or ignored. Smaldino (2017) argues that the simplicity of a formal model might just be one of its strengths: this simplicity allows for a clear articulation of all the parts of the system, as well as the relationships between these parts. An incorrect model might help to suggest which features cannot be left out. A simple model can be used as a starting point, from which it can be expanded to make the model more realistic. Furthermore, computer simulations allow one to easily adjust and investigate different parameter settings, scale up certain aspects of a simulation, such as population size or the time span over which it runs, allowing one to explore settings that can not be achieved in the laboratory (Raviv & Boeckx, 2025). Earlier models in the field were often symbolic (e.g., Griffiths & Kalish, 2007; Kirby & Hurford, 2002), while recently the focus has shifted towards deep-learning models (see Lazaridou & Baroni, 2020, for an overview). In this review, the discussion of deep-learning models is limited to multi-agent systems that explicitly investigate the linguistic properties of emergent communication (Chaabouni et al., 2020, 2022, e.g.,), rather than task-driven multi-agent reinforcement learning work that mostly focuses on solving coordination problems (e.g., Foerster et al., 2016; Lowe et al., 2017). Additionally, while work on multi-agent systems in the field of, for example, game theory, signaling games, or epistemology might be relevant (e.g., Shoham & Leyton-Brown, 2008), these formal models are not discussed to keep the scope focused on work that directly investigates the emergence of compositionality. Finally, it should be noted that computational models have some limitations: their simplicity and abstract nature harm the ecological validity, and they lack empirical validation (Smaldino, 2017); insights should be complemented by data from laboratory experiments, corpus data, or fieldwork. Other fields that are related, but beyond the scope of this review, are, for example: swarm robotics (e.g., Brambilla et al., 2013), or evolutionary computing (e.g., Eiben & Smith, 2015).

This thesis will employ the latter methodological approach: computational simulations. The primary reason for this is convenience, as designing and conducting experiments to address the research questions of this thesis is both time-consuming and costly. However, I would like to emphasize the importance of including empirical validation of computational models. If your model is not behaving similarly to empirical observations, it indicates that the computational model needs refinement. Adjusting, for example, assumptions or the model's simplicity could bring the model closer to real-world dynamics. Then, a more refined model can be used to investigate complex scenarios that would be difficult or impossible to test experimentally.

2.2 Transmission Modes

As briefly mentioned in Chapter 1, one can make the distinction between three modes of cultural transmission: i) Horizontal transmission—interactions between members of the same generation; ii) Vertical transmission—interactions between members of different generations who are biologically related; and iii) Oblique transmission—interactions between members of different generations who are not biologically related (Cavalli-Sforza & Feldman, 1981; Gong, 2010). However, the distinction between vertical and oblique transmission is often blurred, since computational models do not always include information about biological relatedness, and in lab experiments, biological relatedness is typically not included either, as participants are usually unrelated. For this reason, I will adopt a simplified distinction: horizontal transmission, referring to peer-to-peer interactions within the same generation, and vertical transmission, referring to interactions between different generations, regardless of biological relatedness.

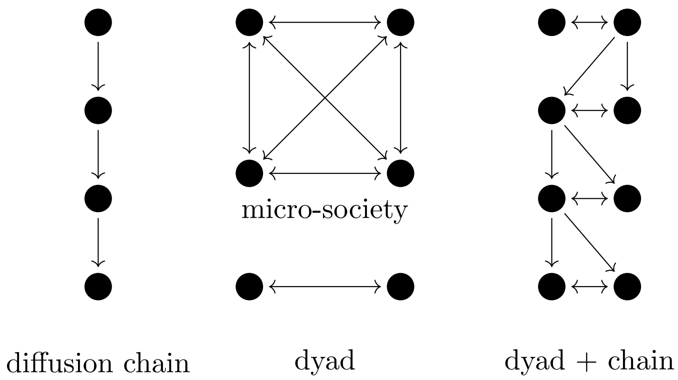


Figure 2.1: Illustration of cultural transmission modes: vertical transmission (across generations), horizontal transmission (within generation), and their combination.

2.2.1 Vertical Transmission

Iterated learning is based on the principle that language is passed on from generation to generation (i.e., vertical transmission), where a learner learns from data produced by the previous generation, who in turn learned the same way. Since the 1990s, a series of computer simulations has been used to show that compositionality can arise due to cultural transmission alone, once certain biases or learning strategies are in place. These models left out any natural selection or biological evolution (Kirby, 2017; Kirby et al., 2014). While many different models and learning mechanisms were considered (Batali, 2002; Brighton, 2002; Brighton et al., 2005; Kirby, 2000, 2002; Kirby & Hurford, 2002; Smith et al., 2003) overall results showed similarities: if a simulation started with unstructured language, compositionality would eventually arise as the language was passed on through generations (Kirby, 2017).[2] What seemed to impact the rate at which compositionality arises is the so-called *information bottleneck*, or in other words, the amount of data a learner observes. If a learner observed very little data, compositionality would arise faster than when more data was available. In the real world, this bottleneck can be compared to the limited data a child observes when they learn a language. All in all, this wide range of models suggests that the emergence of compositionality could be entirely explained by vertical transmission, if one assumes that certain learning strategies or inductive biases are already in place (Kirby, 2017).

To clarify the effect of innate constraints and biases on the emergence of compositionality, Griffiths and Kalish (2007) introduced the Bayesian approach to iterated learning. This allows one to make explicit assumptions about what the *prior inductive biases* of the learner are, while over time, the learner will be affected by the data they are exposed to (Kirby, 2017). In these Bayesian models, Bayes' rule is used to update hypothesis of an agent which is based on their prior inductive bias and observed data:

$$P(h|d) \propto P(d|h)P(h)$$

[2]Examples of these different model types and learning mechanisms are neural networks, symbolic models based on grammar induction and exemplar learning.

where d is observed data, h the hypothesis, $P(h)$ the prior and $P(d|h)$ the likelihood of observing the data given the hypothesis. In these models, the data, d, can be thought of as observed words or sentences, while each hypothesis h represents a language (Kirby, 2017). Since the process of iterated learning can be described as a Markov Chain, one can investigate its long-term behavior by calculating the transition matrices and stationary distributions (Griffiths & Kalish, 2007). Griffiths and Kalish (2007) found that the way in which the agent selects their hypothesis, so sampling from the distribution or MAP (maximum a posteriori), impacts the stationary distribution. In the case where agents sample their hypothesis (and learn from only one teacher), the stationary distribution converges to a distribution precisely the same as the prior, where the size of the bottleneck has no effect. For MAP-learners, the stationary distribution was close to, but not completely the same as the prior. Since both mechanisms lead to a stationary distribution close to, or exactly the same as the prior, they concluded that iterated learning reflects the innate biases of the learners (Griffiths & Kalish, 2007). Later work showed that exact convergence to the prior breaks down when learners learn from multiple teachers, if the agent is assuming to learn a single language (Smith, 2009). Kirby et al. (2007) showed that a weak bias can be strengthened through the process of iterated learning. Assuming that agents pick an hypothesis that disproportionally favors hypotheses with a higher posterior probability (of which MAP is the most extreme option), their experiment in a Bayesian Iterated Framework showed that, regardless of the strength of the initial bias for regularity, the stationary distribution showed a similar preference for regularity. In settings where the information bottleneck was smaller, i.e., the agent learned from fewer datapoints, the stationary distribution favors more regularity in comparison to a larger bottleneck.

Kirby et al. (2008) introduced an experimental iterated learning paradigm, which has been used in several different studies since then (see Tamariz and Kirby (2016) for an overview). These lab experiments allow us to relate the results from computer simulations to the learning abilities and inductive biases of real human learners. Participants were asked to learn an "artificial language", where made-up written labels were matched to visual stimuli. Each visual stimulus has 3 features: shape, color, and motion, which all vary over 3 possible values[3], resulting in a total of 27 stimuli. Initially, the labels paired with visual stimuli lacked systematic structure. The set-up makes use of "diffusion-chains": the first participant learns from data generated by the researchers while subsequent participants learn from their predecessor's output (see Figure 2.1). In total, the diffusion chain consisted of 10 participants. Based on the results from the computational models, it was expected that as the language was passed over generations, the labels would become more learnable and structured (reflecting the inductive biases of the participants). Results confirmed that the labels became more learnable and structured over time; however, the language also became ambiguous, as the same label was used for several stimuli. In the second experiment, discussed in the same paper, the researchers manually filtered ambiguous picture-label pairs and found that the final language showed signs of compositionality.

To summarize, research using the iterated learning paradigm (both computational models and experiments) has shown that compositionality can arise due to vertical transmission of the language in combination with inductive biases of learners.

[3]Square, circle, triangle; black, blue, or red; horizontal, bouncing, or spiraling motion for the shape, color, and motion, respectively

2.2.2 Horizontal Transmission

One of the pioneers in computational models investigating the impact of horizontal transmission on language is Luc Steels, who introduced the *Language Game Paradigm* (Steels, 1995, 1997, 2012). In a language game, virtual or robotic agents who are grounded in an environment create a shared lexicon through interaction and negotiation. During one iteration of the language game, an agent from the population is selected as speaker, and another as hearer. The speaker's goal is to convey information on something in the environment to the hearer. Usually, agents have some kind of grammar, which can contain meanings, categories, or structures. If communication is successful, the used linguistic elements are reinforced, while if communication fails, the elements are altered or removed from the speaker's internal grammar, or the hearer updates their grammar to contain the new elements (Van Eecke et al., 2018). The language game paradigm has been used to investigate emergent vocabularies of proper names (e.g., Steels, 1995), the emergence of concepts (e.g., Steels, Belpaeme, et al., 2005), and the emergence of grammar (Van Eecke et al., 2018). Luc Steels and colleagues also introduced 'Fluid Construction Grammar' (FCG), which is a computational platform for developing grammars from a constructional perspective (De Beule & Steels, 2005; Steels, 2011a, 2017). Broadly speaking, FCG is a system that is based on form-meaning mappings, which are called constructs. From these constructs, one can simulate production (meaning to form) and comprehension (form to meaning) (Beuls & Van Eecke, 2023). Agents start with an empty set of constructs and build their grammar through interactions with other agents.[4] Some earlier experiments combined the language game paradigm and FCG and explicitly investigated the emergence of compositional structure. For instance, De Beule and Bergen (2006) investigated whether compositional languages could occur within one generation and in a setting where language change is driven by communicative success. When an agent came across a new meaning, they could either create a new form for the entire meaning (holistically) or, if they already had a form for part of the meaning, they could come up with a compositional representation. Compositionality emerged in settings where agents had to communicate about atomic and more complex expressions. It also appeared when agents were themselves 'cognitively' limited, i.e., they ignored complex expressions until a certain level of communicative success was reached. Similarly, De Beule (2008) showed that compositionality, recursion, and hierarchy can emerge within the same generation, due to agents striving for communicative success, arguing that vertical transmission is not necessary.

More recently, language emergence has been studied in settings with neural agents (see Lazaridou & Baroni, 2020; Peters et al., 2025; Rita et al., 2024, for overviews). Often, these simulations make use of a version of the Lewis Signaling Game (Lewis, 1969).[5] In this simple communication game, one agent is the sender and the other the receiver. After the sender observes a state, they generate a message and send it to the receiver. The receiver then takes some action based on this message. This simple version of the game resembles the dyad setting, as depicted in Figure 2.1. An advantage of these neural models is that communication is much more open-ended compared to the older symbolic models, where either some grammar was already in place (e.g., Vogt, 2007), or agents relied on predefined

[4] A detailed explanation of the precise workings of FCG is beyond the scope of this thesis, but see Steels (2017) for an introduction and Beuls and Van Eecke (2023) for a description of the state of the art.

[5] While Steels' language games and Lewis' signaling games are closely related, Steels does not explicitly cite Lewis, and most deep-learning research on emergent communication builds on Lewis' framework, with only an occasional reference to Steels' work.

form-meaning mappings. Neural agents develop a protocol from scratch, often in an unsupervised setting, and thus allow for a more flexible testbed for studying language emergence (Lazaridou & Baroni, 2020). Furthermore, the framework allows for exploring more complex environments, such as using images as input or having agents navigate in a 2D or 3D environment (Lazaridou & Baroni, 2020). Agents generally learn via reinforcement learning or Gumbel-softmax (Lazaridou & Baroni, 2020). Of key interest in many of these studies is whether the communication protocol that emerges has natural language-like properties, where many studies focused on compositionality specifically (e.g., Chaabouni et al., 2020; Kottur et al., 2017; Lazaridou et al., 2018; Resnick et al., 2020). Often, the level of compositionality of the protocol is measured using *topographic similarity* (Brighton & Kirby, 2006), which captures the correlation between the distances in the meaning and signaling space. Recently, alternative metrics have been proposed to capture non-trivial compositionality (Korbak et al., 2020), or more intricate measures of compositionality that for example take positional disentanglement into account (Chaabouni et al., 2020).[6] While the emergent protocols do not always show signs of compositionality (Kottur et al., 2017), various studies have shown that factors such as introducing strangers (Cogswell et al., 2019; Ren et al., 2020), or communicating in a noisy environment (Kuciński et al., 2021) increase the level of compositionality of the emerged communication protocol.

The field of Experimental Semiotics focuses specifically on the emergence of new communication systems with the use of laboratory experiments (Galantucci, 2009; Galantucci & Garrod, 2011). Some of these experiments are focused on the vertical transmission of language (e.g., Kirby et al., 2008, 2015), while others study the horizontal transmission by having participants communicate in dyads or micro-societies (Tamariz & Kirby, 2016).

Fay et al. (2010), for example, arranged participants in either micro-societies or isolated pairs to play a "Pictionary" game, where pairs communicated about items using only drawings. In the micro-society condition, participants switched partners after a couple of rounds. For both conditions, they observed that signals started iconic, but became simpler and more symbolic over time. Furthermore, they highlight the difference between an iterated learning setting, where the final communication system reflects participants' inductive biases, while in a micro-society setting, the final communication system is based on negotiation with other members of the community. In a similar vein, Theisen et al. (2010) studied participants playing a graphical communication game in dyads, focusing solely on that condition. They were particularly interested in the level of arbitrariness and systematicity of the signals over time, finding that participants created systematic signals that started iconic but became more arbitrary over time.

Kirby et al. (2015) proposed that two complementary pressures are essential for the emergence of compositionality: a pressure for *compressibility*, arising from vertical transmission, and a pressure for *expressivity*, arising from horizontal transmission (this study is discussed in more detail in Subsection 2.2.3). Building on this idea, Raviv et al. (2019a) suggested that vertical transmission might not be necessary. Compressibility pressures could be introduced by either an expanding meaning space, or communicating in a larger group. To investigate this, Raviv et al. (2019a) conducted an experiment where pairs of participants within a micro-society performed a communication task, using an artificial language, over an expanding set of visual stimuli. Results confirmed the hypotheses; compressibility

[6] Something is non-trivially compositional when the meaning of the complex whole is a complex function of the meaning of its parts. An example would be "bad dancer" compared to the trivially compositional meaning of "blond dancer" (Berthet et al., 2025).

pressures introduced by communicating in a larger group and over an expanding meaning space were enough to elicit the emergence of compositional languages. Additional analysis revealed that group size had a bigger impact on the emergence of compositionality than the expanding meaning space. This experiment demonstrated that findings from earlier computational studies on horizontal transmission and compositionality such as De Beule (2008) and De Beule and Bergen (2006) also apply in real-world settings.

To conclude, in contrast to the work using the iterated learning paradigm, work investigating the impact of horizontal transmission is often driven by communicative success. This line of research shows that compositionality can arise within one generation, particularly when communication happens in larger groups.

2.2.3 Horizontal and Vertical Transmission Combined

Vogt (2005a) combined the language game and iterated learning in a computational model. The population was divided into teachers and learners, where the teachers were selected as the speakers and the learners as the hearers for the language game. Furthermore, at each "generation" in the model, the old learners became teachers and new agents were introduced as learners. The model uses grammar induction - where agents induce generalizations, i.e., compositional rules if they are available. Results showed that in a population with only 2 agents, compositionality would arise without an imposed bottleneck, but for compositionality to arise in a population of 6 agents, a bottleneck had to be enforced, which caused agents to generalize more often and thus generate more compositional rules. In a follow-up study (Vogt, 2005b), learners could also learn from other learners, instead of learning from the previous generation only. This setting created an *implicit bottleneck*, since learners had to come up with words for concepts they had not yet learned. Compositionality was more stable in settings with horizontal transmission (learners learning from learners) in comparison to the vertical transmission setting, where no explicit bottleneck was imposed. A downside of these models is the grammar induction mechanism, causing agents to create compositional rules whenever possible, effectively hard-coding the emergence of compositionality into the model itself.

Brace et al. (2015) and Kirby and Tamariz (2021) both also investigated what happens if agents learn from agents who are still learning the language themselves. The former focused on the emergence of compositionality, while the latter looked into the emergence of combinatoriality. In contrast to earlier iterated learning models, where agents often learned from a single mature teacher (e.g., Brighton, 2002; Kirby, 2001; Smith et al., 2003), here, agents can learn from both mature and immature teachers, where immature teachers are still learners themselves. Learning from mature teachers corresponds to vertical transmission of the language, since they learn from an older generation, while learning from immature learners corresponds to horizontal transmission. Using neural networks as agents, Brace et al. (2015) shows that compositional languages arise if a learner learns from more mature than immature teachers. The emergence of compositionality is unstable when immature teachers outnumber mature teachers. Furthermore, the greater the number of mature teachers, the higher the level of compositionality. This suggests that in their model, vertical transmission of a language is necessary for a stable emergence of compositionality. The Bayesian iterated learning model from Kirby and Tamariz (2021) showed that when agents learn from a combination of mature and immature teachers, combinatorial languages arose around 10 times faster in comparison to the setting where agents consistently learned from the oldest in the population.

Lastly, several studies using neural-agents included iterated-learning into the model, by, for example, a designated learning algorithm (Ren et al., 2020) or by having agents learn from the previous generation before they interact with each other (Guo et al., 2020). Results showed similar patterns: the emergent communication protocol had a higher level of compositionality when vertical transmission was included.

Theisen-White et al. (2011) was the first study combining horizontal and vertical transmission in the lab. Participants were asked to play a graphical communication game, similar to the ones previously described (Theisen et al., 2010). A difference was that after the first "generation" of participants, the next generation was shown the final drawings of the previous generation. While the drawings in the horizontal transmission only setting were compositional, adding vertical transmission enhanced the level of compositionality even further, suggesting that both forms of transmission could be a pressure that leads to the emergence of compositionality.

Kirby et al. (2015) hypothesized that a pressure for compressibility, caused by the vertical transmission of the language, in combination with a pressure for expressivity, caused by horizontal transmission of a language, leads to the emergence of compositionality. Pressure for compressibility only leads to degenerate languages, as observed by Kirby et al. (2008), and a holistic language would emerge when there is only pressure for expressivity. Their hypotheses were confirmed by a lab experiment similar to the experiment described earlier, with the modification that within one generation, participants interacted in dyads (see Figure 2.1). Furthermore, an additional Bayesian iterated learning model confirmed their hypotheses as well. In their conclusion, they argued that both horizontal and vertical transmission are necessary for compositionality to emerge.

In conclusion, some models and laboratory experiments have investigated the emergence of compositionality in a setting including both horizontal and vertical transmission. Overall, the different models and experiments showed similarities: if both forms of transmission were included, compositionality would emerge faster compared to when only one form of transmission was present.

2.3 Population Size

A cross-linguistic study by Lupyan and Dale (2010) suggests that larger communities have languages that are less morphologically complex, compared to languages used in smaller communities. Furthermore, Meir et al. (2012) show that a newly emerging sign language used in a larger community (Israeli Sign Language) is more structured compared to a newly emerging sign language used in a small community (Al-Sayyid Bedouin Sign Language). However, in the real world, population size cannot be examined separately from network structure and the number of non-native speakers (Raviv et al., 2019b). To disentangle these properties, some computational models and a limited number of experimental studies systematically examined the impact of group size on the emergence of linguistic structure.

The agent-based model by Vogt (2007), which is similar to the previously discussed model in Subsection 2.2.3 (Vogt, 2005a), suggests that larger groups promote faster emergence of compositionality. He argues that since there is an increase in input variability, there is also an increase in the likelihood of observing regular patterns between form and meaning, and thus a higher probability of compositionality to arise. In contrast, recent deep learning models of language emergence did not reliably exhibit this phenomenon (Galke et al., n.d.). Chaabouni et al. (2022) found that a larger population does not increase the agent's ability to generalize, or in other words, the protocol that emerged in a larger population was not more compositional than the one in a smaller population. Rita et al. (2021) argued that previous work did not consider a heterogeneous population: having differences in agents might be the reason for an increase in compositionality in a larger population. They explore this hypothesis by sampling different learning rates for each agent and thus creating a heterogeneous population. Results show that in a heterogeneous population, a larger group leads to a more stable and compositional language.

Raviv et al. (2019b) investigated in an experimental setting whether group size impacted the level of structure of the language that was used during a communication game. Participants had to communicate about visual stimuli using an artificial language. There were two conditions: a group size of four and a group size of eight. Results showed that the languages used in communities of size eight became more compositional over time than those in groups of four. In some small groups, structure never emerged: participants communicated with holistic signals. Raviv et al. (2019b) argue that since there is more input variability in a larger group, there is more pressure to converge to a compositional language.

2.4 Social Network Structures

Within a population, interactions between individuals are embedded in social networks, which can be represented by a graph. The nodes of the graph represent the individuals, and the interactions between the individuals are denoted by the edges. These social networks can have specific characteristics that could limit the transmission of linguistic features (Castellano et al., 2009; Group" et al., 2009). An example of such characteristics would be average shortest path length, clustering coefficient, or the average degree (i.e., the number of edges a node has). Based on these characteristics, one can distinguish several network types, such as a fully-connected, a small-world (Watts & Strogatz, 1998), a scale-free (Barabási & Albert, 1999), a hierarchical (Ravasz & Barabási, 2003), or a random network (Erdos, Rényi, et al., 1960). Furthermore, for example Reali et al. (2018) used a network structure inspired by real social networks based on mobile communication records from Portugal and the UK (Schläpfer et al., 2014).[7]

Previous work in the area of language evolution often makes the distinction between *esoteric* communities and *exoteric* communities, where the former are tightly-knit and smaller societies with barely any non-native speakers, while the latter are sparser and larger, with more contact with outsiders and thus more non-native speakers (Lupyan & Dale, 2010; Wray & Grace, 2007). Languages in esoteric societies are found to be more morphologically complex, while languages in exoteric communities are often simpler, since there is more pressure for the language to be learnable for (adult) strangers (Lupyan & Dale, 2010; Nettle, 2012; Wray & Grace, 2007). Small-scale, esoteric communities can be represented as a fully-connected network, while large-scale, exoteric communities can be represented as small-world or scale-free networks (Gong et al., 2011; Mudd, 2022).

Several computational models suggest that language change is impacted by social network structure.

[7] Details of these networks and the algorithms used to create them are discussed in Experiment 3.

New linguistic innovations spread faster in a dense network (e.g., fully-connected) compared to a sparser network (small-world or scale-free) (Fagyal et al., 2010; Gong et al., 2008, 2011). Others, for example, investigated the impact of social network structure on category formation (Gong et al., 2011; Zubek et al., 2017). Josserand et al. (2024) disentangled what features of a network have the most impact on language variability, using a Bayesian model where agents were arranged on many different networks. Specifically, they investigated the impact of the average shortest path length, clustering coefficient, global assortativity, degree distribution, and size on interindividual variation and intraindividual variation. Interindividual variation denotes the variation within the entire population, while intraindividual variation shows how consistent an agent is in their own language use. They found that intraindividual variation is not affected by network structure, but that interindividual variation is mainly impacted by clustering coefficient and path length.

Relatively little research has examined the impact of social network structures on the emergence of compositionality specifically. Lou-Magnuson and Onnis (2018) investigated whether smaller communities enhance the development of morphological structure, and whether larger communities inhibit this development, and so possibly increase syntactic structure. In their agent-based model, they use the number of reanalyses done by the agent, i.e., the number of times an agent reassigns a different meaning to a known form, as a measure of the increase of morphological complexity. Results showed that networks with a higher level of transitivity, thus a more tight-knit community, resulted in a language with higher morphological complexity. Moreover, there seemed to be a connectivity threshold beyond which, once the network was sufficiently connected, further increases in connectivity no longer led to greater morphological complexity. An additional experiment compared the impact of four different network types, with varying levels of transitivity and clustering, on the morphological complexity. They find that a network with high transitivity and so-called hubs - a few agents are connected to many other agents - complex morphology is inhibited. Using a symbolic iterated learning model where agents were placed on a network, Nakamura et al. (2015) found that when the amount of interaction with neighbors is higher than the amount of interaction with a parent, languages do not become compositional. The two different network structures they compare, scale-free and a 2D-lattice, do not seem to make a big difference. Overall, these computational models suggest that social network structure matters for language change, but its impact on compositionality specifically remains unclear.

In an experimental study, where again participants had to communicate using an artificial language, Raviv et al. (2020) found that network structures, such as scale-free, small-world, and fully-connected, did not impact the level of compositionality of the emerged language. Languages became more systematic over time, regardless of the network structure in which the participants were arranged. A limitation of this study was the relatively small group size of only 8 participants, resulting in an average shortest path of 1.5 between the participants in both the small-world and scale-free networks (compared to an average shortest path of 1.0 in a fully connected network), allowing linguistic variations to spread quickly through the entire population.

In summary, quite some work investigated the impact of several factors affecting cultural transmission, such as transmission mode, population size and social network structure, on the emergence of compositionality. However, some specific questions remain unanswered, such as the relative contribution of vertical and horizontal transmission on the emergence of compositionality, or for example how the rate of generational turnover affects the emergence of compositionality. Additionally, less work has focused on the individual learning mechanisms and their impact on the emergence of compositionality. For each of the research questions as introduced in Chapter 1, I provide a more detailed motivation in their respective experiment section.

Chapter 3 – Computational Framework

The computational framework used in this thesis, and as described in this section, is an altered version of the model used in Simon Kirby's simulating language course, which is a simplified version of the model by Kirby et al. (2015).[8] The description of the framework roughly follows the ODD (Overview, Design Concepts, Details) protocol for agent-based simulations (Grimm et al., 2006, 2010). My code is available here: https://github.com/francijnk/MoLThesis.

3.1 Purpose

The purpose of this computational framework is to investigate how different factors that impact cultural transmission (horizontal & vertical transmission, population size, network structure, and rate of generational replacement) affect the emergence of a preference for compositional languages. Furthermore, by altering individual learning mechanisms (regulating the impact of new evidence and the strength of the prior), we can investigate how this impacts the long-term outcomes of cultural transmission. Parameters, their descriptions, and default values can be found in Table 3.1.

Table 3.1 *Parameters, descriptions, and default values*

Parameter	Symbol	Description	Default
Population size	*pop_size*	Number of agents initialized at the start of the simulation.	25
Horizontal Transmission	h	Number of rounds the language game is played during horizontal transmission.	20
Vertical Transmission	v	Number of rounds the language game is played during vertical transmission.	20
Network Type	$Network$	Network type representing how the population is connected. Options are fully-connected, small-world, scale-free, hierarchical, mobile-phone, or dynamic.	fully-connected
Mode	M	Sampling mode used by the agents during production. Options are Sample or MAP (Most likely hypothesis).	Sample
Error Probability	ϵ	The probability an agent makes the occasional mistake when producing a signal.	0.05
Iterations	$N\_iterations$	Number of iterations per simulation run.	10000
Run ID	ID	ID for the run to set the seed.	1

[8]Their model can be found here: https://github.com/smkirby/simlang2022-23/blob/main/lab6.ipynb.

3.2 Entities & Variables

There is a single entity in the model: the agent. The population is a set of agents, and is characterized by its network structure. More information on the different network types and their properties can be found in Experiment 3. An agent is characterized by its unique ID and a distribution over possible languages. Similar to Kirby et al. (2015), in this model, a language is a system that expresses meanings using forms. The forms and meanings chosen are very simplistic, yet they allow one to discriminate between the different language types: *degenerate*, *holistic*, and *compositional.* The set of meanings is defined as $\mathcal{M} = \{02, 12, 03, 13\}$ where the first component of a meaning can either be a 0 or a 1, and the second component can be either a 2 or a 3. The set of forms is defined as any combination of a's and b's, thus $\mathcal{F} = \{aa, ab, ba, bb\}$. All the possible combinations of forms and meanings result in 256 possible mappings, or thus languages, which can be divided into the 4 different types. A language is *degenerate* if each meaning has the same form, such as $\{(02, aa); (12, aa); (03, aa); (13, aa)\}$. A language is *compositional* if a part of the meaning is consistently mapped to either an 'a' or a 'b', for example, $\{(02, aa); (12, ba); (03, ab); (13, bb)\}$, where the first part of the meaning consistently maps to the first symbol of the signal, and the second part of the meaning consistently maps to the second symbol of the signal. A *holistic* language is a language where every meaning has a distinct form, but the mapping is not compositional, thus $\{(02, aa); (12, bb); (03, ab);\ (13, ba)\}$. Languages are labeled as *other* if they are only partially holistic, degenerate, or compositional. Notice that in this setup, some types occur more frequently in the space of all languages than others. See Table 3.2 for the frequency of each language type. Additionally, while I use the term *language* here, conceptually these are not different languages, such as English and Dutch, but abstract communication systems that differ in their form-meaning mappings.

Table 3.2 *Frequency of types in all 256 languages*

Language Type	Frequency
Degenerate	4
Compositional	8
Holistic	16
Other	228

3.3 Process overview and scheduling

The model consists of an initialization phase and two sub-processes: the language game and data collection.

3.3.1 Initialization

During the initialization phase of the model, *pop_size* agents are created. All agents are initialized with the same distribution over languages, which is characterized by a preference for compressible hypotheses. This preference for compressability can be seen as a (domain-general) simplicity bias (Chater & Vitányi, 2003; Culbertson & Kirby, 2016; Kirby et al., 2015). Chater and Vitányi (2003) surveys studies exploring the idea that simplicity drives cognitive processes. Culbertson and Kirby (2016) reviews research showing that a weak, domain-general simplicity bias, when amplified by communication and cultural transmission, can produce language-specific patterns such as compositionality, regularization, and consistent word order. But how does one decide what a 'simpler' grammar is? Following Kemp and Regier (2012) and Perfors et al. (2011), grammars are simpler when they contain either fewer rules or fewer and simpler rewrite rules. To obtain the compressibility of each language, I calculate its coding length using the method described by Kirby et al. (2015). First, the form-meaning mappings, i.e., languages, need to be formalized as rewrite grammars. In this small form–meaning space, there are two types of rewrite rules: $X \to YZ$ or $X : M \to F$, where M is a set of possible meanings (or a single meaning feature if it represents the set) and F is a form. The first rule indicates that X can be rewritten as YZ (order-specific), and that the meaning of X is the combination of the meanings of YZ. For the second rule, it means that X rewrites to the form F and takes any of the meanings in M. The rewrite grammar can then be reduced to its minimally redundant form, which is the minimal string of characters from which you can reconstruct the grammar. To be more specific, to obtain the minimal redundant form, the left-hand side (LHS) of the rule is concatenated with the right-hand side (RHS). In case the LHS is of the form $X : M \to F$ and M contains more than one meaning, the meanings are comma-separated. Distinct rules are separated by a full stop. The coding length of a language in bits is calculated from the minimally redundant forms as follows:

$$L(l) = -\sum_{i=1}^{|l|} \log_2 p(l_i)$$

where $p(l_i)$ is the probability of the ith character in the code for l. To be precise, if the character a appears twice in the entire minimal redundant form of length 7, its probability is $\frac{2}{7}$. Then the compression-based prior for a language, $G_0(l)$, is obtained as follows: $G_0(l) \propto 2^{L(l)}$. Essentially, this leads to a preference for degenerate languages, since these are highly compressible, while holistic languages are the least compressible and thus have the lowest initial probability. For each type of language mentioned in Section 3.2, the corresponding formal grammar, minimal redundant form, coding length, and proportion of the prior distribution can be found in Table 3.3. See Table 3.4 for the values of the prior per language type.

To initialize the population, a network structure is created, where each node corresponds to a unique agent ID. Edges in the network represent who can play a language game with whom. More information

Table 3.3 *Formal grammars, minimal redundant form, coding lengths, and normalized prior for different languages.*

	Degenerate Language	**Compositional Language**	**Holistic Language**
Formal grammar	$S : \{02, 12, 03, 13\} \rightarrow ab$	$S \rightarrow AB$ $A : 0 \rightarrow a$ $A : 1 \rightarrow b$ $B : 2 \rightarrow a$ $B : 3 \rightarrow b$	$S : 02 \rightarrow aa$ $S : 12 \rightarrow bb$ $S : 03 \rightarrow ab$ $S : 13 \rightarrow ba$
Minimal redundant form	S02,12,03,13ab	SAB.A0a.A1b.B2a.B3b	S02aa.S12bb.S03ab.S13ba
Coding length $L(l)$	40.55	59.20	67.29
Normalized prior $G_0(l)$	0.100	0.0000002423	0.000000000892

Table 3.4 *Prior probabilities for each language type*

Language type	**Prior**
Compositional	0.000001938715
Degenerate	0.9987714
Holistic	0.00000001426784
Other	0.001226692

on the different network types and their properties is provided in Experiment 3.

3.3.2 Language Game

The number of times the language game is played in each step of the model is dependent on the parameter values of *horizontal_transmission* and *vertical_transmission*, and will be explained in more detail in Experiment 1. One iteration of the language game can be described as follows: a speaker and a learner are selected from the population, as well as an arbitrary meaning. Then the speaker samples a language from their distribution, and produces a signal that, according to the sampled language, corresponds to the selected meaning.[9] Note that in this model, each meaning can be referred to with only 1 signal, i.e., no synonymity is possible. With a small probability ϵ, an arbitrary signal is produced to simulate the occasional mistake. To include a pressure for expressibility, it is assumed that we are dealing with a pragmatic speaker. This means that, before selecting a signal, the speaker considers whether a learner will guess the intended meaning after observing the signal. This consideration is based on the language the speaker themselves selected, hence I refer to the learner in this case as 'hypothetical learner'. If the hypothetical learner does not guess the intended meaning, a random signal is selected. This process penalizes ambiguous languages that use the same form for different meanings. The probability that the signal is produced that corresponds to the selected meaning in the sampled language is given by:

$$P(f|m, l) = \begin{cases} (1 - \epsilon), & \text{If hypothetical learner chose the correct meaning.} \\ \frac{1}{|\mathcal{F}|}, & \text{Otherwise.} \end{cases}$$

[9] In this set-up, the agents are sample-learners. In case the Mode is set to MAP learners, agents pick the most probable hypothesis.

Where f is one of the forms, m is the selected meaning, l is the sampled language, and ϵ is the probability of error.

The hypothetical learner chooses a meaning based on which meaning(s) correspond to the received signal. If no meanings correspond to the signal, a random meaning is chosen, and if several meanings correspond to the signal, the hypothetical learner uniformly samples one of the possible meanings. Notice that this 'hypothetical learner' is only a way for the speaker to produce a signal; the actual selected learner does not have anything to do with this process.

Once the signal is selected, it is time for the learner to update their distribution, based on the meaning/signal pair they observed. Let $P(\mathbf{l})$, where $\mathbf{l} = \{l_1, l_2, ...l_{256}\}$, denote the current distribution of the agent over all possible languages. The likelihood of the observed form meaning pair depends on whether or not it is in a language:

$$P(f|m,l) \propto \begin{cases} (1-\epsilon), & \text{if } f,m \in l \\ (\frac{\epsilon}{|\mathcal{F}|-1}), & \text{if } f,m \notin l \end{cases}$$

where ϵ is the error-probability and f is the produced form and m is the selected meaning.[10] Then, the posterior gets updated for each possible language $l_i \in \mathbf{l}$ using Bayes rule:

$$P(l_i|f,m) \propto P(f|m,l_i) \cdot P(l_i)$$

After updating for all languages, the posterior is normalized. In the next posterior update, the most recent posterior is used as the prior to allow for gradual learning. Updates are performed in log-space to avoid numerical underflow.

At each iteration, agents' posterior distributions over languages are collected. These values are aggregated across the entire population by language type and then normalized. This data gives insight into which language types are most prominent in the population.

All simulations were run using the Snellius computing cluster, using CPU nodes with 128 cores per node and 224 GM of RAM.

[10] Note that the likelihood when producing a form is slightly different from the likelihood that is used for updating. This was the case in their implementation, and I decided not to change it.

Chapter 4 – Cultural Transmission

Language evolves culturally, as it is repeatedly transmitted between individuals within and across generations. It remains unclear, however, how various factors affecting cultural transmission, such as transmission mode, population size, social network structure, and rate of generational turnover, precisely affect the emergence of compositionality. This Chapter presents a series of experiments investigating the impact of these factors. Each experiment begins with a brief motivation, followed by details on the experimental setup, results, and a discussion.

Experiment 1: Horizontal and Vertical Transmission

Vertical transmission of language, extensively studied via iterated learning both computationally and experimentally, has demonstrated that it promotes the emergence of compositionality when inductive biases or learning mechanisms are already in place (e.g., Kirby & Hurford, 2002; Kirby et al., 2008). The horizontal transmission of language, particularly in larger groups, promotes the emergence of compositionality within a single generation (De Beule & Bergen, 2006; De Beule & Steels, 2005; Raviv et al., 2019a). Combining these modes of transmission tends to accelerate the emergence of compositionality (e.g., Kirby et al., 2015; Vogt, 2007). However, existing studies combining horizontal and vertical transmission have limitations: often the population within one generation consists of only 2 agents, reducing the impact of horizontal transmission (e.g., Guo et al., 2020; Kirby et al., 2015). Studies using larger populations within one generation either rely on grammar induction, effectively hard-coding the emergence of compositionality (Vogt, 2005b), or use of neural networks, which contain implicit inductive biases (Brace et al., 2015).

The experiment presented here investigates the effect of horizontal and vertical transmission in larger populations within each generation, without relying on grammar induction mechanisms, and uses agents with explicit inductive biases. Additionally, the setup allows for separating the contributions of each of the transmission modes.

4.1.1 Experimental Setup

To investigate the contribution of horizontal and vertical transmission to the emergence of compositionality, the two transmission modes are included in the model as follows.

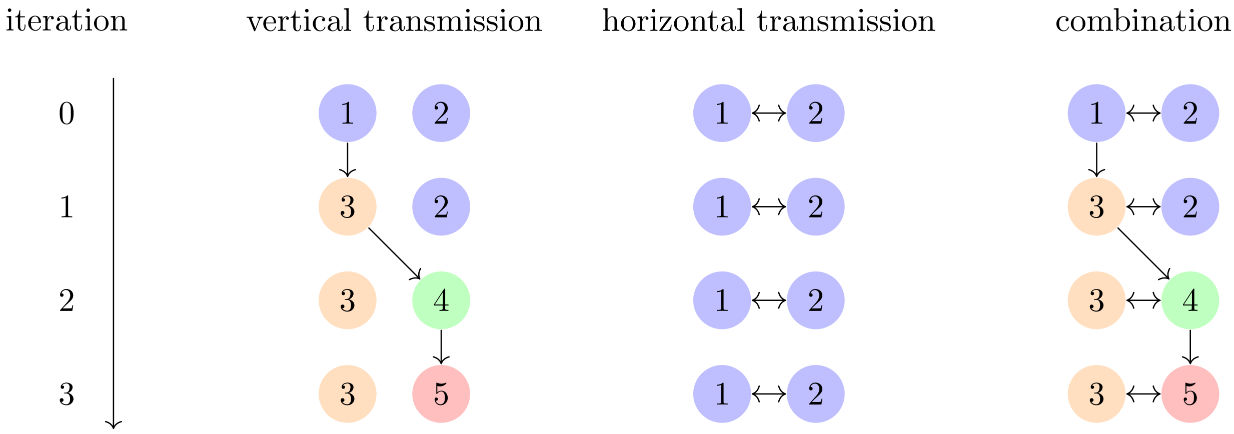
Horizontal transmission is simulated by randomly selecting a speaker and a learner from the population, which is repeated h times. Then, each selected speaker and learner will play the language game, as described in Subsection 3.3.2, h times. In one iteration of the model, a total of h^2 horizontal language games are played.[11]

Vertical transmission is simulated by removing r randomly chosen agents from the population, after which the same number of new agents are added. New agents are initialized with the *compression-based* prior, as described in Subsection 3.3.1. After initialization, the new agents play v language games as learner, where a randomly selected agent from the 'old' population is the speaker. In one iteration of the model, $r \times v$ vertical language games are played.

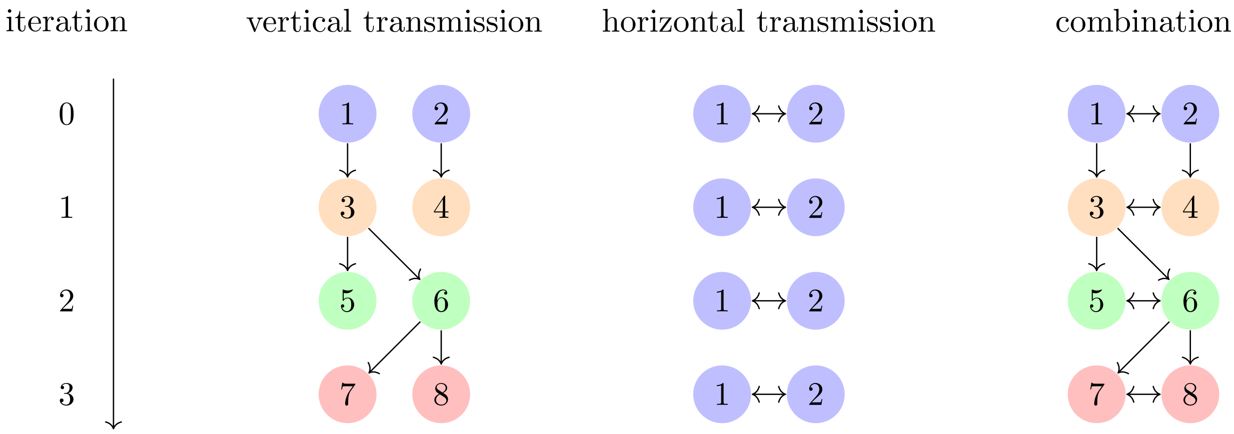
By varying the amount of h and v, representing the number of language games in a horizontal or vertical setting, the impact of these variables on the emergence of a preference for compositionality is examined. Furthermore, results are reported for simulations where $r = 1$, where at each iteration only one agent is replaced, and $r = pop\_size$, where at each iteration the entire population is replaced. See Figure 4.1.1 for an example of the different conditions. The population size was set to 25, and h and v were assigned values of 0, 10, or 20. Note that there is pressure for expressivity due to the "pragmatic speaker", as the speaker agent tries to be understood, explained in Subsection 3.3.2. Data is collected after the horizontal language games, as the vertical language games can be seen as a preparation for the next iteration.[12]

[11]While it is quite unrealistic, h was used twice to reduce the number of parameters in the model.

[12]This ensures that the data captures the impact of both the horizontal and vertical transmission. If data were collected after vertical transmission, the posteriors of the new agents would only be impacted by the data observed from their vertical language games.



(a) $r = 1$



(b) $r = pop\_size$

Figure 4.1.1: Graphs depicting the different conditions when the population size is 2. In the top figure, if vertical transmission > 0, at each iteration, one agent from the population is replaced ($r = 1$), while in the bottom figure, at each iteration, the entire population is replaced ($r = pop\_size$). Numbers identify each unique agent, and colors identify the iteration at which the agent was introduced to the population. Arrows between agents show who plays the language game with whom.

4.1.2 Results

The results presented here are averaged over a population of 25 agents and aggregated across 50 independent runs per condition. Agents in these simulations are sample-learners, meaning they sample from their posterior distribution during production rather than selecting the maximum a posteriori hypothesis (MAP).[13] Simulations were run for 10000 iterations.

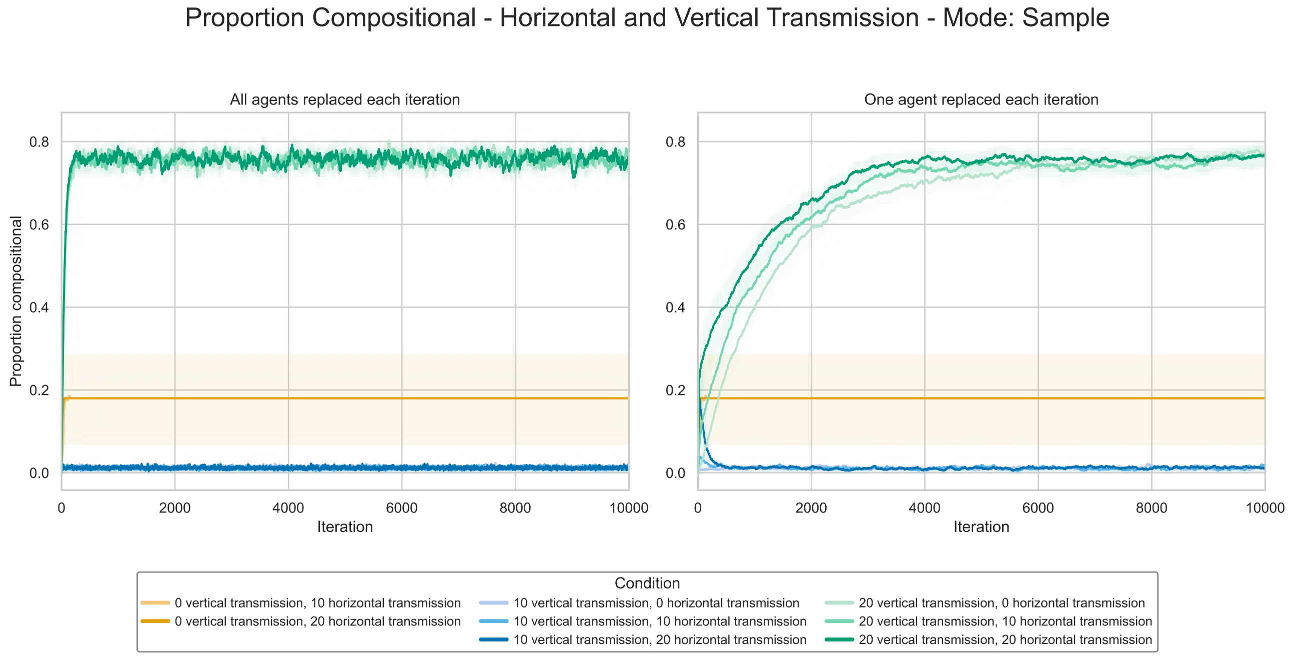


Figure 4.1.2: Proportion of posterior for compositional languages averaged for the entire population of sample-learners, averaged over 50 run per condition. The shaded area represents 95% confidence intervals. If there is vertical transmission, in the left Figure, all agents are replaced at each iteration, while in the Figure on the right, only one agent is replaced per iteration.

Horizontal transmission alone does not lead to a strong preference for compositionality, as shown in Figure 4.1.2. On average, only around 18% (CI: 7% - 29%) of the posterior probability mass is focused on compositional languages in these settings. While the preference is weaker than in the condition with $v = 20$ and $h > 0$, it is still substantially higher than chance: given that there are 8 compositional languages in the entire space of 256 languages, the chance level is 3.125%. As shown in Figure 4.1.3, most of the posterior distribution lies on "other" languages if there is only horizontal transmission. Furthermore, these simulations exhibit more variability compared to the other simulations, as indicated by the larger confidence interval. In the case of vertical transmission, the preference for compositionality depends on the number of vertical transmission rounds. When the new agent does not observe enough data, no preference for compositionality emerges with a proportion of compositionality of 1%, regardless of the level of horizontal transmission. In this condition, most of the posterior lies on degenerate languages (see Figure 4.1.3), suggesting that there was not enough pressure for expressivity. If, on the other hand, the new agent observed more data, a clear preference for compositionality emerges. In the scenario where one agent is replaced per iteration, horizontal transmission in these conditions accelerates the emergence of a preference for compositionality, but the final level of compositionality converges to a similar point, of around 77% (CI: 75% - 79%) regardless of the amount of horizontal

[13] Plots for MAP-learners can be found in Appendix B.

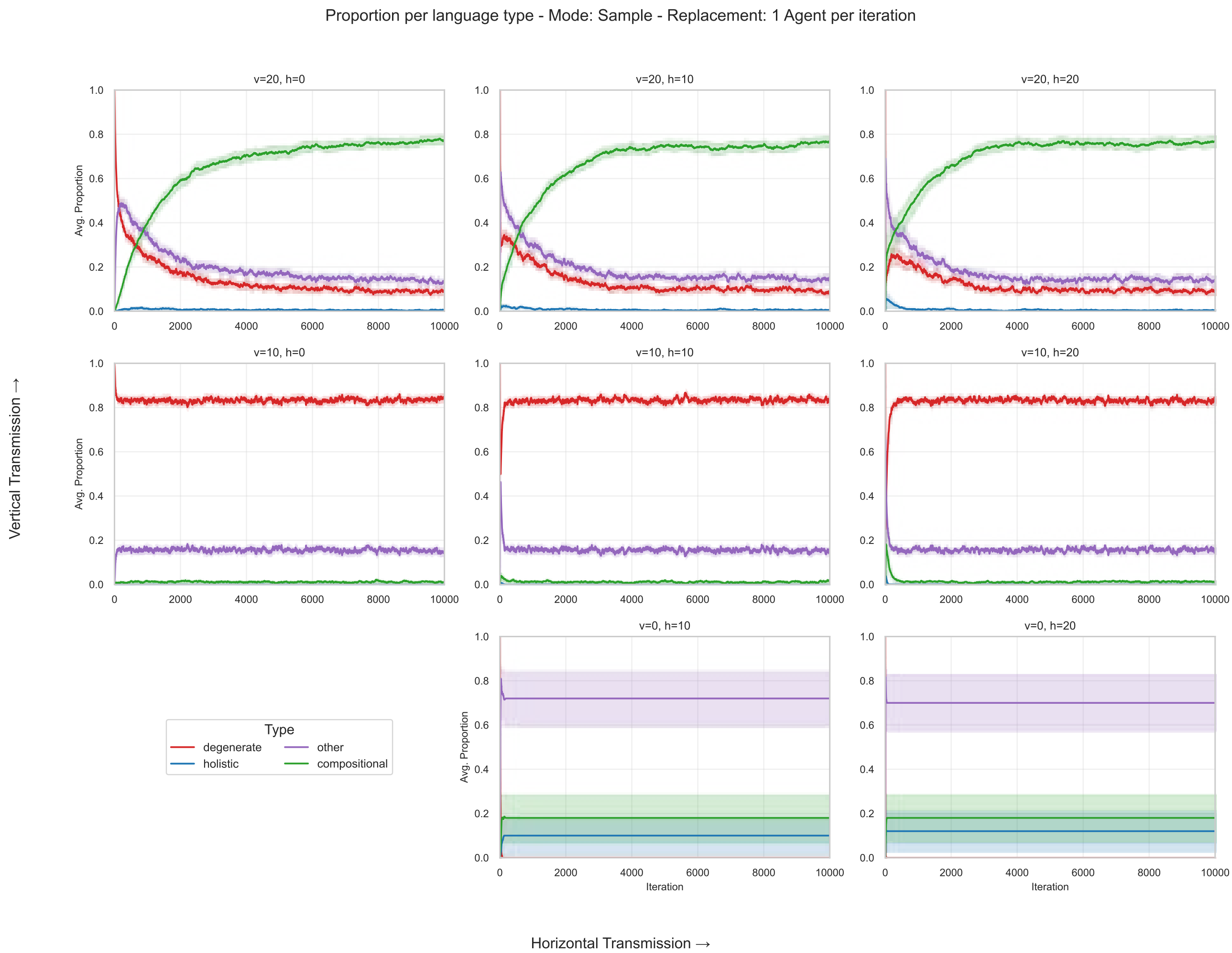


Figure 4.1.3: Proportion of posterior per language type, for varying levels of horizontal and vertical transmission. Values are averaged for the entire population of sample-learners and averaged over 50 runs per condition. Per iteration, one agent is replaced (if there is vertical transmission). The shaded area represents 95% confidence intervals.

transmission. Finally, when the entire population is replaced each generation, convergence to a high preference for compositionality is much faster compared to single-agent replacement, though both settings eventually converge to similar levels of compositionality of around 75%.

4.1.3 Discussion

Similar to previous work (e.g., Kirby et al., 2015), the right trade-off between pressure for *compressibility*, caused by the compression-based prior of agents and *expressivity* caused by the interactions between agents, results in a preference for compositionality. This is most evident in settings with vertical transmission, where there is a stable pressure for compressibility, since new agents with a fresh compressibility prior are introduced, and thus it the pressure for compressibility does not fade away over time. When a new agent has too few interactions ($v = 10$), no preference for compositionality arises since the pressure for expressivity is too weak. When the number of interactions increases ($v = 20$), a strong preference for compositionality emerges.

Horizontal transmission alone does not lead to a strong preference for compositionality, which differs from results in laboratory experiments (Raviv et al., 2019a). As shown in Figure 4.1.3, during simulations with no vertical transmission, a preference for "other" languages emerged. It could be that the "other" languages are partly compositional, suggesting that horizontal transmission leads to the emergence of more structured languages, but not fully compositional ones. However, this difference could also be explained by the models' assumed learning dynamics. Currently, the pressure for compressibility diminishes over time if no new learner is introduced, since only agents with a "fresh" initial prior have a strong preference for compressible languages. As an agent interacts with other agents, the observed data overwrites the prior of the agent, and so the compressibility pressure. However, if someone assumes that humans have a stable cognitive bias in favor of simpler, and thus more compressible languages, it seems unlikely that this bias would vanish entirely due to interactions with others. In Experiment 5, I investigate this further by examining the effect of regulating the impact of new evidence.

The amount of vertical interactions, even when only one agent is replaced per iteration, has a greater impact on the preference for compositionality compared to the amount of horizontal interactions. To be specific, in simulations with $v = 10$, regardless of the number of horizontal interactions, most of the posterior probability mass is concentrated on degenerate languages. Though based on the increased horizontal interactions, we would expect a higher pressure for expressivity in these settings.

A sensitivity analysis, reported in Appendix A, revealed that indeed, over time, most of the variance in the model output is caused by the amount of vertical interactions. Why precisely this is the case remains unclear. Future work could point out whether this effect is model-specific or if it generalizes to other modeling frameworks as well.

Finally, in the setting where one agent is replaced per iteration, more horizontal transmission leads to faster emergence of a preference for compositionality, which is in line with previous work (e.g., Brace et al., 2015). In the setting where all the agents are replaced, we cannot observe this, as the increased number of vertical interactions completely overrules any differences between horizontal settings.

Experiment 2: Population Size

Previous experiments, both with human participants and computational agents, showed that population size affects the emergence of compositionality: in larger populations, compositionality arises faster and more prominently than in smaller ones (e.g., Raviv et al., 2019b; Rita et al., 2021; Vogt, 2007). However, it remains unclear whether we can observe similar effects in a Bayesian setting and whether this effect is present in a setting with both horizontal and vertical transmission. The current experiment explores this question by testing for population size effects in the model.

4.2.1 Experimental Setup

Increasing the population size in the simulations can be done easily by adjusting the parameter *pop_size*. It should be noted that as population size increases, the number of interactions remains stable during both horizontal transmission and vertical transmission when one agent is replaced per iteration. However, in the setting with vertical transmission and when the entire population is replaced, then the total number of language games played is larger when the population is larger. Simulations were run for $v = 20$, $h = 20$, population sizes 2, 10, 25, 50, 100, and 200, and for 10000 iterations. Furthermore, simulations were run in which either the entire population was replaced each iteration or a single agent was replaced at each iteration.

4.2.2 Results

The results shown in Figure 4.2.1 and Figure 4.2.2 indicate that a larger population does not lead to a stronger or faster emergence of a preference for compositionality.[14] When the entire population is replaced at each iteration, all population sizes follow a similar trajectory, converging to comparable levels of compositionality ranging from 75% (CI: 73% - 76%) for a population of size 50 to 79% (CI: 71% - 87%) for a population of size 2. In the condition where one agent is replaced per iteration, a preference for compositionality emerges for all population sizes. The larger the population is, the longer it takes to converge. For population sizes 100 and 200, it is unclear whether they would eventually converge to the stable level of around 77%, as they currently reach a final value of around 69% and 55%, respectively.

[14] Results reported here are for sample-learners; corresponding plots for MAP-learners can be found in Appendix B.

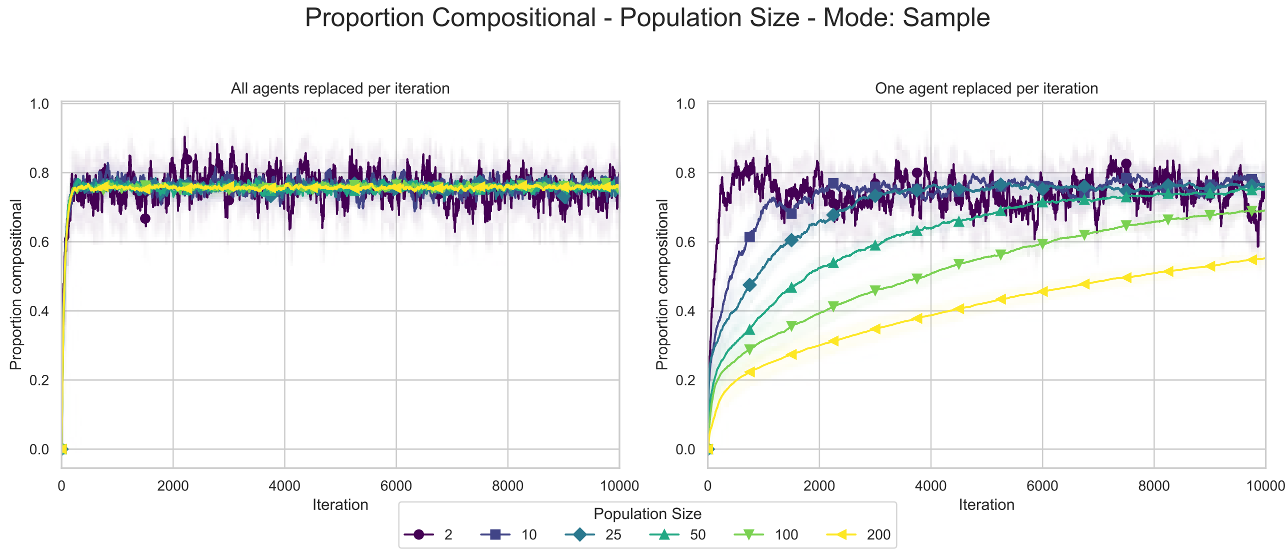


Figure 4.2.1: Posterior probability for compositional languages averaged for the entire population of sample-learners, averaged over 50 runs, for different population sizes. The shaded area represents the 95% confidence interval. In the left figure, every agent is replaced each iteration, and in the figure on the right, only one agent is replaced each iteration.

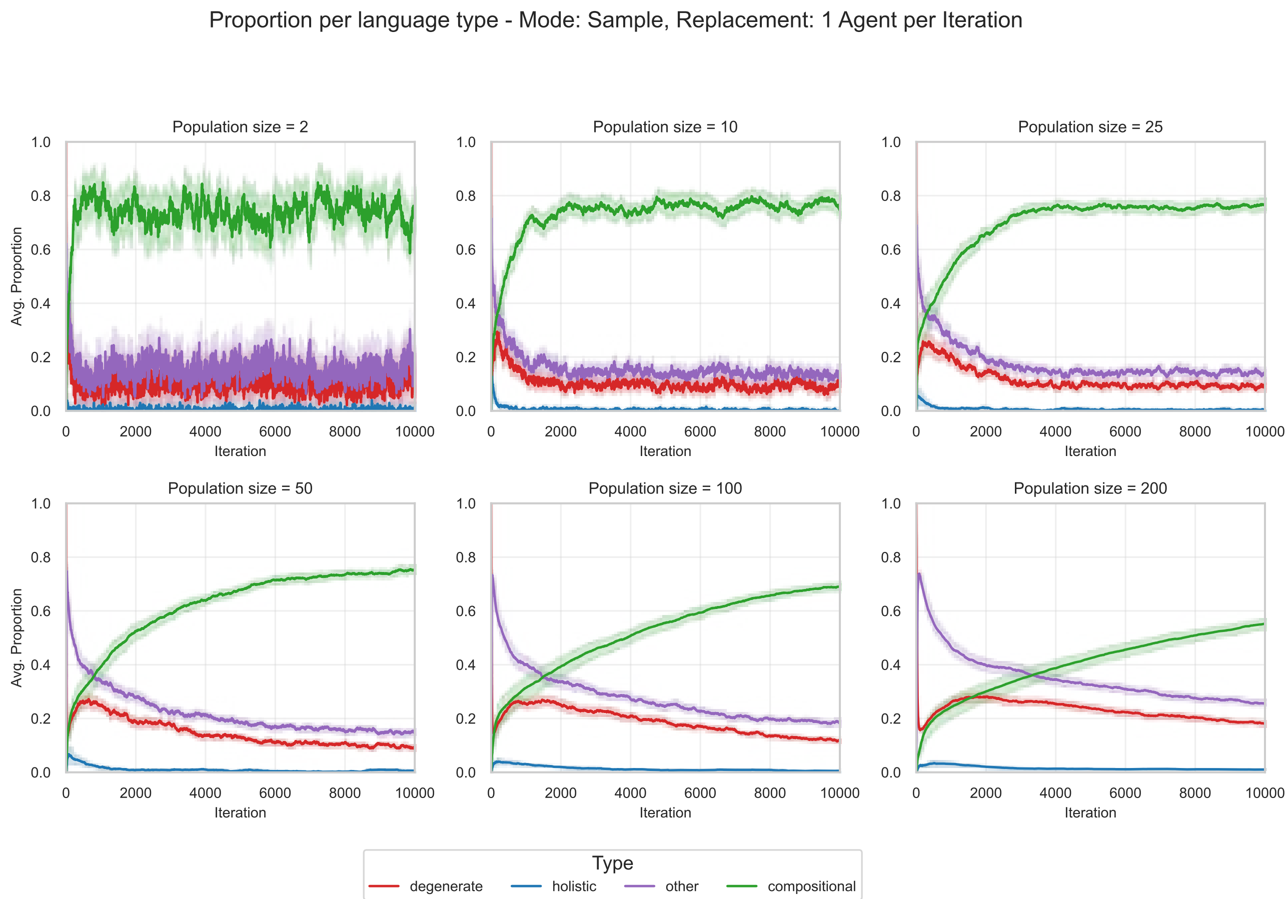


Figure 4.2.2: Proportion of posterior per language type, for various population sizes. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

4.2.3 Discussion

In contrast to previous work (Raviv et al., 2019b; Rita et al., 2021; Vogt, 2007), a larger population did not result in a faster emergence or a higher overall level of compositionality. After a closer inspection of the models' dynamics, these results are actually somewhat unsurprising. In this model, the emergence of compositionality is caused by a combination of two pressures: 1) a pressure for *compressibility*, which is present due to the initial prior of the agents, and 2) a pressure for *expressivity*, which is present due to the pragmatic agents who penalize ambiguous languages, as explained in Subsection 3.3.2. The strength of the expressivity pressure depends on how often agents interact, as regulated by parameters h and v. When these two pressures are balanced in a specific way, a preference for compositional languages emerges. Notably, Experiment 1 showed that the number of horizontal interactions does not have a substantial effect on the emergence of compositionality when vertical interactions are present; so, in the following explanation, I will focus on the pressure of expressivity caused by vertical interactions only.

When population size increases, every agent still has the same compressibility prior, and hence, the average pressure for compressibility in the entire population remains constant. To gain insight into why we observe no population size effects, it is useful to make the distinction between two simulation settings: (i) one where all agents are replaced per iteration, and (ii) one where only one agent is replaced each iteration. In setting (i), we have the same balance between compressibility and expressivity regardless of population size: during each iteration, every agent is replaced, and so every agent has 20 vertical interactions. Since the balance of expressivity and compressibility does not change as we increase the population size, we observe no difference in the emergence of compositionality. In setting (ii), we replace one agent per iteration, and so as we increase the population size, the average number of vertical interactions per agent per iteration decreases. As a result, it takes longer to reach the balance between expressivity and compressibility that leads to the emergence of compositionality as the population size increases. Even though convergence is slower in larger populations, the system ultimately reaches the same level of compositionality, as the long-term outcome depends on the dynamics between compressibility and expressivity.

Additionally, in the real world, a larger population leads to more input variability, which needs to be conquered to achieve successful communication (Raviv et al., 2019b). Linguistic variants that are predictable and structured, i.e., compositional, can be a helpful strategy to overcome this high level of variability. In a computational model with neural agents, Rita et al. (2021) observe that an increase in input variability only occurs when the population is heterogeneous, so the agents differ in, for example, their learning mechanisms. When heterogeneity was included, larger populations led to more structured languages, consistent with experimental findings (Raviv et al., 2019b).

In the current model, at the start of the simulation, the population is homogeneous: they do not differ in their learning mechanisms, initial prior, or sampling mode. Over time, the agents will differ in distributions, but this might not be enough to elicit the population size effect as observed in previous work. In Experiment 7, I will investigate whether we can observe population size effects in a heterogeneous population.

Experiment 3: Social Network Structures

In addition to transmission mode and population size, the network structure of the social community might also impact the emergence of compositionality (Raviv et al., 2020). The distinction can be made between esoteric and exoteric communities, where the former are tightly-knit, and often smaller communities mostly consisting of native speakers, and the latter are sparser and larger, with more adult non-native speakers (Raviv et al., 2020). Languages used in these different communities seem to have different properties; i.e., languages used in exoteric communities tend to have a less complex morphological system compared to languages used in esoteric communities (Lupyan & Dale, 2010). We would expect a higher level of compositionality in exoteric communities, compared to the tightly-knit esoteric communities (Reali et al., 2018).

In the real world, however, it is hard to investigate the impact of the individual parameters that characterize these different communities, such as population size, network structure, or the number of outsiders, since these are naturally confounded.

Previous work on the effect of social network structure and the emergence of compositionality is conflicting. On the one hand, several computational models suggest that tightly-knit networks promote morphological complexity, while others find minimal effects of network topology on the emergence of compositionality (Lou-Magnuson & Onnis, 2018; Nakamura et al., 2015). Experimental work with small populations organized into various network structures has reported no clear influence of network type on the emergence of structure in the languages (Raviv et al., 2020). Additionally, to the best of my knowledge, little work has yet investigated language change in a dynamic network, i.e., one in which nodes and edges are added and removed.

The experiment presented here investigates the effect of social network structure on the preference for compositionality in a setting with both horizontal and vertical transmission. The network types considered in this experiment are a fully-connected network, a small-world network (Watts & Strogatz, 1998), a scale-free network (Barabási & Albert, 1999), a hierarchical network (Ravasz & Barabási, 2003), a dynamic network, and a network based on mobile-phone data (Schläpfer et al., 2014).

4.3.1 Experimental Setup

Each agent has their own place in the network and is represented by a node. The edges in the network represent who can play the language game with whom. The edges are undirected, since communication is often bi-directional. Furthermore, the networks do not contain self-loops, meaning that agents cannot communicate with themselves. The levels of horizontal and vertical transmission ($h = 20, v = 20$) and the population size ($pop\_size = 25$) are stable during the simulations. Simulations were run for 10000 iterations, where at each iteration, one agent in the network is replaced.[15] The networks are implemented using the Python package NetworkX (Hagberg et al., 2008).

[15] Except for in the dynamic network, where potentially more agents are replaced each iteration.

Fully-Connected Network

As the name already suggests, in this network, every agent is connected to every other agent. See Figure 4.3.1 for an example of a fully connected network. This network could represent small-scale, i.e., esoteric, communities as it is more likely that everyone communicates with everyone (Gong et al., 2011). This network structure is rarely found in the real world, but it might be similar to early hunter-gatherer societies (Raviv et al., 2020).

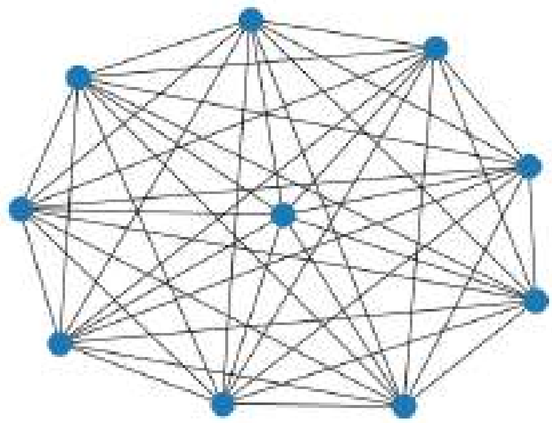


Figure 4.3.1: Fully-connected network with 10 agents.

Small-World Network

A small-world network is a sparsely connected network with relatively little distance between unconnected nodes (small average path length) and high clustering. Every node has approximately the same number of edges. It could represent more sparsely connected, exoteric communities (Mudd, 2022; Raviv et al., 2020). The small-world network is generated by starting with a ring network of n nodes. Then each node is connected with k nearest neighbors. For each edge (u, v) in the network, it is replaced by a new edge (u, w) with probability p, where w is a uniformly chosen different node (Watts & Strogatz, 1998). The small-world network used in the model is initialized with $n = 25$, $k = 4$, and $p = 0.1$, similar to Gong et al. (2011). It is ensured that the network is connected; otherwise a new network is generated. See Figure 4.3.2 for an example of a small-world network.

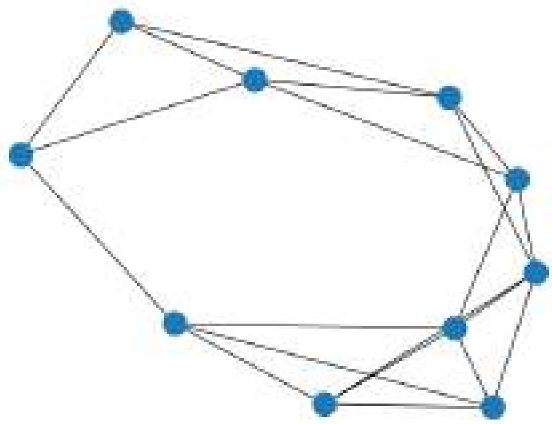


Figure 4.3.2: Example of a small-world network with 10 agents, where initially each node is connected to $k = 4$ of its nearest neighbors (2D-Lattice), and the rewiring probability is $p = 0.1$, similar to the small-world network used by (Gong et al., 2011).

Scale-Free Network

A scale-free network is also a sparse network with relatively short paths, but the connections roughly follow a power-law distribution: a few nodes have many connections, while others are more isolated (Barabási & Albert, 1999). A node with many connections is sometimes referred to as a “hub”. These hubs lead to a faster spread of linguistic innovations (Gong et al., 2011; Raviv et al., 2020), and so result in less variation, potentially leading to a lower level of compositionality compared to the small-world network. A scale-free network can be created by preferential attachment: when a new node is added to the network, the probability p of it being connected to node i in the network depends on the connectivity k_i, such that $p(k_i) = k_i / \sum_j k_j$ (Barabási & Albert, 1999). To create the scale-free network, the standard scale-free graph generator function of the NetworkX package is used (Hagberg et al., 2008). See Figure 4.3.3 for an example of a scale-free network.

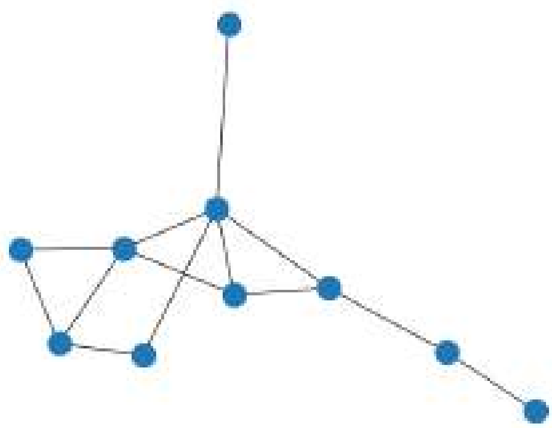


Figure 4.3.3: Example of a scale-free network with 10 agents.

Hierarchical Network

This network combines two properties that are often observed in the real world: scale-freeness and high clustering (Ravasz & Barabási, 2003). In a sense, it can thus be seen as a combination of a scale-free and a small-world network, since a hierarchical network has both hubs and high clustering. As described by Ravasz and Barabási (2003), a hierarchical network starts as a small cluster of a fully-connected network with 5 nodes. Next, four replicas of this fully-connected network are connected by the non-central nodes to the central node of the initial network, resulting in a network of size 25. Then, to create a larger network, again, four replicas of this network are created, and the 16 non-central nodes are connected to the central node of the original network, resulting in a network of size 125. This process can be repeated indefinitely. An example of a hierarchical network of size 25 is shown in Figure 4.3.4. While the hierarchical structure in this specific network is highly artificial and so unlikely to represent any real-world situation, it is argued that larger societies have a hierarchical structure (e.g., Peter & Sergey, 2009), which is mimicked by this network.

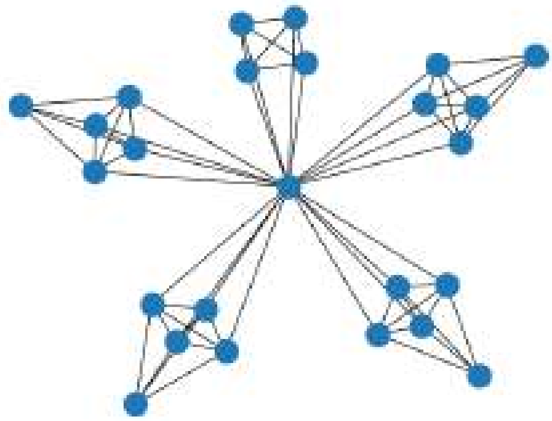


Figure 4.3.4: Example of a hierarchical network with 25 agents.

Network Based on Mobile Phone Data

This network is based on mobile phone records in Portugal and the UK (Schläpfer et al., 2014), and is similar to the network used by Reali et al. (2018). Mobile phone records showed that the average number of contacts per phone, i.e., the average degree k, grows super-linearly with population size n: $k\ n^{\beta-1}$ (Reali et al., 2018; Schläpfer et al., 2014). The specific networks created were similar to networks used by Reali et al. (2018), where $\beta = 1.677$ and clustering coefficient $C \approx 0.25$. To create a network with a specific average degree and clustering coefficient, I sample graphs similar to Reali et al. (2018). First, I create a random Erdős-Rényi graph (Erdős & Rényi, 1961), using the graph generator available in NetworkX, where the number of nodes is equal to the population size, and the probability of edge creation is set to 0.25. The following loop is then run iteratively to obtain the final graph: Two random nodes are selected, and if there's an edge, it is removed; otherwise, an edge is added. Then, the average degree and clustering coefficient of the network are calculated. Based on the current average degree and clustering coefficient, the hamiltonian is calculated as follows

$$h = \beta_1(AD_{current} - AD_{target})^2 + \beta_2(C_{current} - C_{target})^2$$

where AD is the average degree and C is the clustering coefficient. The graph with updated edges is accepted as the current graph with probability $p = e^{h_{prev} - h_{new}}$, where h_{prev} is the hamiltonian of the old graph and h_{new} is the hamiltonian of the new graph. After some trial and error, the number of iterations was set to 3000, $\beta_1 = 500$, and $\beta_2 = 10000$, which resulted in a final graph with the desired average degree and clustering coefficient.

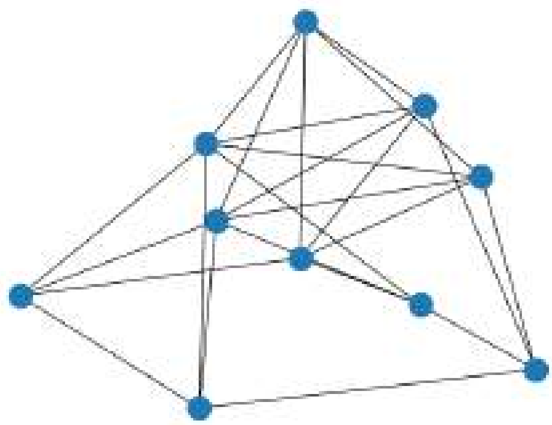


Figure 4.3.5: Example of a network based on mobile phone data with 10 agents.

Dynamic Network

The dynamic network is the only network that changes in structure over the course of the simulation. The network is initialized as a small-world network. Then, at each iteration, there is edge removal and addition, as well as node removal and addition. During edge removal and addition, for each agent, the probability of an edge being removed or added is 0.5. If an edge is added for agent u, 10 agents from the population are selected whose distribution over languages is most similar to the distribution of agent u. This is measured using the Jensen-Shannon distance, which is the square root of the Jensen-Shannon Divergence (Lin, 2002), for which I used the function included in the SciPy package (Virtanen et al., 2020). From these 10 agents, one random agent is selected to create the new edge towards. If an edge is removed, a random neighbor is chosen to be disconnected. During node removal, at least one node, or thus agent, is removed from the network. Which agent is removed from the network is based on communicative success: after the horizontal interactions, each agent is evaluated. During evaluation, an agent plays the language game as the speaker with every other agent in the population as the learner. If the guessed meaning by the learner is the same as the intended meaning, the communicative success score of the speaker agent is increased by 1. After evaluation of all agents, the communicative success scores are normalized and then reversed, such that the least successful agents have a higher probability.[16] From this probability distribution, an agent is sampled to be removed from the network.

[16]Specifically, $p_{\text{new}_i} = p_{max} - p_{\text{old}_i}$, where p_{new_i} is the new probability for agent i, p_{max} is the highest probability value in the distribution and p_{old_i} is the old value for agent i.

Additionally, agents who are not connected to any other agent are removed from the network. At each iteration, at least one new agent is introduced to the network. In case additional agents were removed since they were completely disconnected, the same number of new agents is added to maintain a stable population size. A random agent is selected to be the first parent, and a random neighbor of parent 1 is chosen to be parent 2. Then, the new agent is added to the graph, and edges are created to connect to its parents. The new agent will learn for v rounds from one of its parents.

4.3.2 Properties of the Networks

Numerous properties of the networks were tracked and are reported in Table 4.1.

The *average degree* is the average number of edges per node in the network. If the average degree is high, the network is more tightly knit compared to a network with a lower average degree.

The *clustering coefficient* is a measure that captures how connected a node's neighbors are to each other, and is calculated as follows:

$$c_u = \frac{2T(u)}{deg(u)(deg(u)-1)}$$

where T is the number of triangles through node u, and $deg(u)$ is the degree of node u. The clustering coefficient reported in Table 4.1 is the average clustering coefficient of the entire graph.

The average shortest path measures, as the name suggests, the length of the shortest path between any two nodes in the network:

$$average\_sp = \sum_{\substack{s,t \in N \\ s \neq t}} \frac{d(s,t)}{n(n-1)}$$

where N is the set of nodes in the network, n is the number of nodes, and $d(s,t)$ is the shortest path from node s to t. This captures how fast information travels through the network.

N_edges is the total number of edges in the network. Lastly, the network density is the percentage of possible edges that are realized in the network.

Table 4.1 *Network characteristics: means and standard deviations.*

Type	Average Degree	Clustering Coefficient	Average Shortest Path	N Edges	Network Density
Fully-connected	24.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	300.0 ± 0.0	100%
Mobile-phone	8.827 ± 0.038	0.261 ± 0.006	1.643 ± 0.006	110.34 ± 0.479	37%
Hierarchical	5.28 ± 0.0	0.902 ± 0.0	2.033 ± 0.0	66.0 ± 0.0	22%
Small-world	4.0 ± 0.0	0.383 ± 0.058	2.81 ± 0.183	50.0 ± 0.0	17%
Dynamic	3.279 ± 0.418	0.158 ± 0.072	2.61 ± 0.249	38.713 ± 5.626	13%
Scale-free	2.624 ± 0.28	0.178 ± 0.061	2.382 ± 0.146	32.8 ± 3.504	11%

Based on these properties, the networks are ordered from tightly-knit (top) to sparse (bottom). Generally, denser networks converge more rapidly (Fagyal et al., 2010; Gong et al., 2008, 2011), but sparser networks might promote the emergence of compositionality, since throughout the entire population there is a higher input variability since new innovations spread slower in the network. To be able to communicate successfully, and thus overcome the higher input variability, there is a higher pressure for compositionality (Raviv et al., 2020).

4.3.3 Results

The results shown in Figure 4.3.6 and Figure 4.3.7 indicate that overall, network structure has little to no effect on the emergence of a preference for compositionality. Across five of the network types- fully-connected, small-world, scale-free, hierarchical, and the network based on mobile phone data- the proportion of compositional languages in the population increases at a similar pace. It converges to a similar level, of around 75% of the final distribution. Only in the dynamic network do we observe a faster increase, but a lower final preference for compositionality, reaching approximately 68% (CI: 55% - 81%). Compared to the other networks, in the dynamic network, the proportion of holistic languages in the final distribution is quite high, around 30% compared to 0% for the other networks. There is an observable difference in convergence speed between the different network types. The scale-free network converges the slowest, especially compared to the fully-connected and mobile-phone networks.

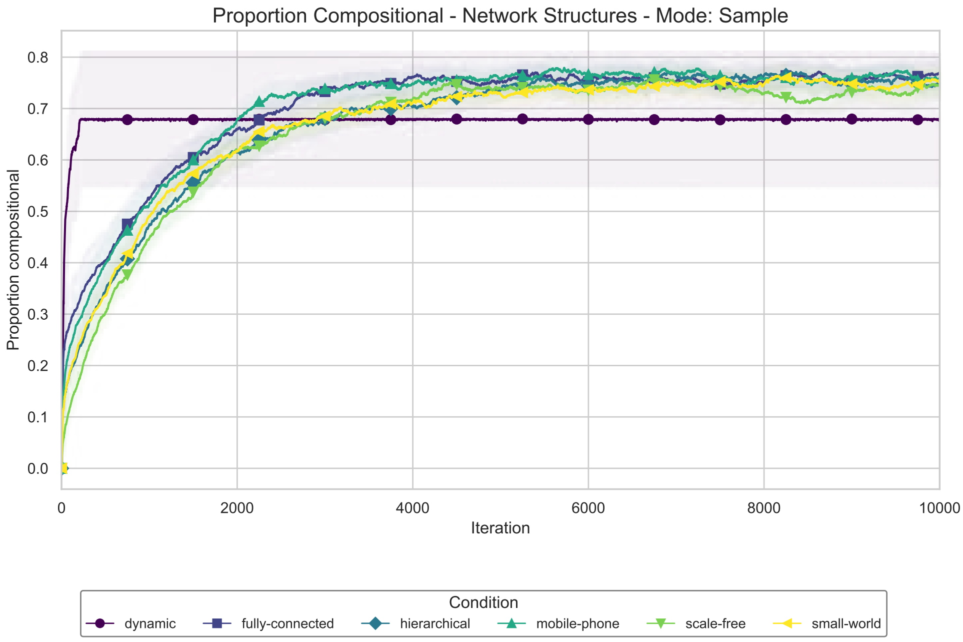


Figure 4.3.6: Posterior probability for compositional languages averaged for the entire population of sample-learners, averaged over 50 runs, for different network types. The population size is 25. Vertical and horizontal transmission rounds = 20. Shaded area represents 95% confidence intervals.

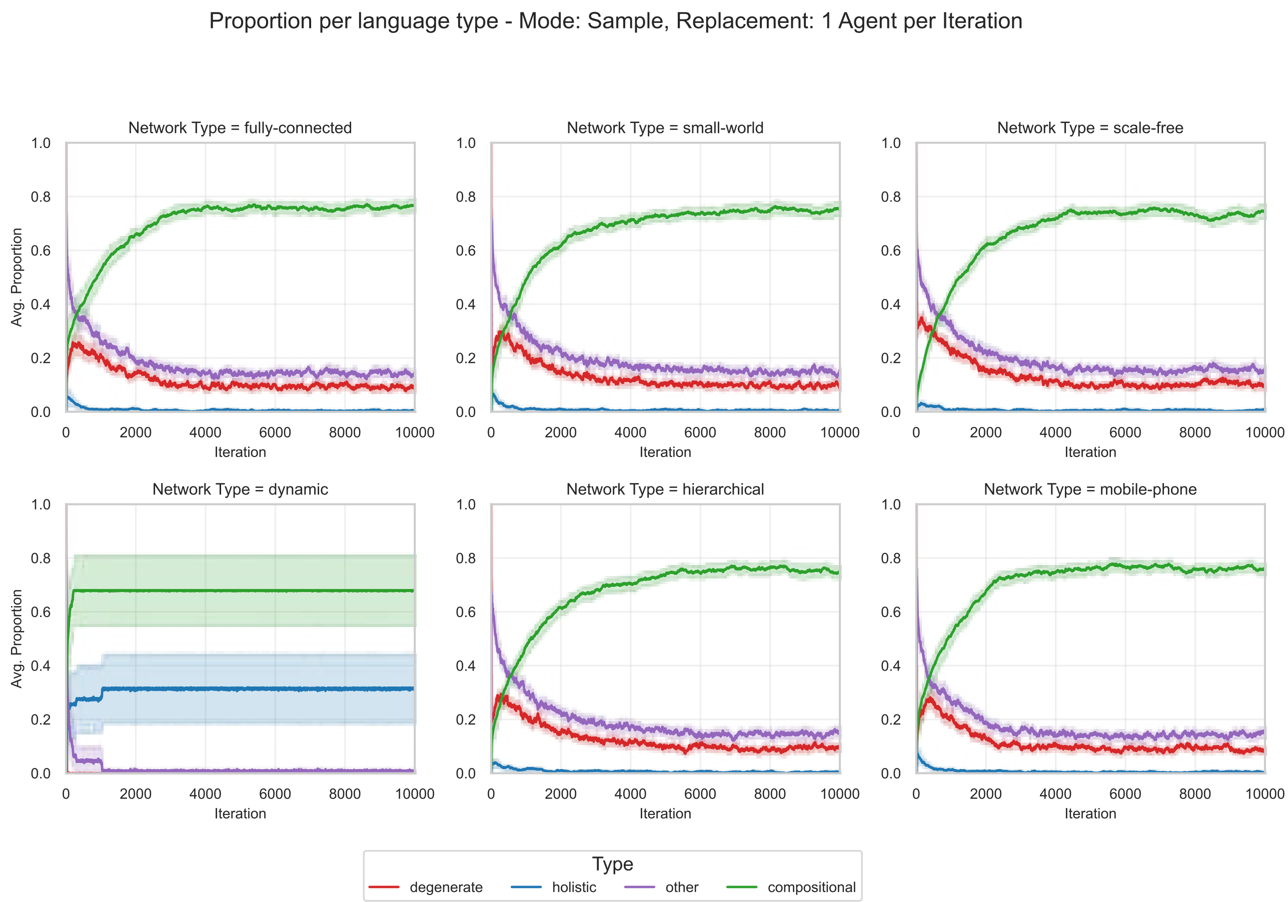


Figure 4.3.7: Proportion of posterior per language type, for various network types. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

4.3.4 Discussion

The difference in convergence speed between the network types can be explained by the network density. Sparser networks (e.g., scale-free or small-world networks) converge more slowly than tightly-knit networks, such as the fully-connected or mobile-phone networks. As there are fewer edges in the network, it takes longer to spread information through the entire population. These findings are in line with previous work (Fagyal et al., 2010; Gong et al., 2008, 2011). A notable exception is the dynamic network, which converged fastest while it is quite sparse. This could be explained by the different proportion of vertical transmission compared to the other networks, as in this setting, potentially more agents are replaced per iteration. Furthermore, in the dynamic network, a larger proportion of the posterior distribution focused on holistic languages compared to the other network types. This can be explained by the extra pressure for expressivity, which is introduced by replacing agents with low communicative success instead of a randomly selected agent. Since degenerate, and thus most compressible languages, are not efficient for communication, agents with higher communicative success may have a posterior distribution with a weaker preference for compressibility, which facilitates the rise of holistic languages.

The finding that network structure has little impact on the final level of preference for compositionality is consistent with previous experimental results using smaller-scale networks (Raviv et al., 2020) and extends these findings to larger networks with a population size of 25. Additionally, it is also in line with previous computational work (Nakamura et al., 2015). Nevertheless, conclusions should be drawn with caution, as some limitations of the current model settings might explain the observed results.

As discussed in Subsection 4.2.3, the emergence of a preference for compositionality in this model is caused by the combination of pressure for *expressivity* and *compressibility.* The different network types do not change the pressure for compressibility, since all agents still have the same compressibility prior; hence, the trade-off between expressivity and compressibility is not affected. Similarly, for vertical interactions, the network structure does not change anything either: a new agent takes the place of the removed agent in the network, and has 20 interactions with the removed agent. The only exception is the dynamic network, where possibly more than one agent is replaced per iteration. When an agent is completely unconnected, it is removed, and an extra new agent is added. This results in a higher level of vertical interactions, and thus a higher pressure for expressivity compared to the other networks. Furthermore, in the dynamic network, agents are removed from the network with low communicative success, leading to extra pressure for expressivity. Only in the case of horizontal transmission is the network structure limiting expressivity in a way, since it restricts who can communicate with whom. As observed in Experiment 1, if there's any vertical transmission, the contribution of the horizontal transmission to the final stationary distribution is limited. Since vertical transmission is included in the current simulations, we do not observe any difference between the network types except for the dynamic network, because the network structure does not alter the trade-off between expressivity and compressibility. To potentially observe an impact of network structure on the emergence of a preference for compositionality, a setting without vertical transmission should be considered. However, as mentioned in Experiment 1, in a setting with only horizontal transmission, the compressibility pressure fades away due to the impact of new evidence the agents observe. In Experiment 5, I will introduce a mechanism to regulate the impact of new evidence, allowing for a setting with horizontal transmission and a stable pressure for compressibility. Then, in Experiment 8, I examine the impact of network structure in a setting with only horizontal transmission while regulating the impact of new evidence.

Experiment 4: Rate of Replacement

In all previously reported experiments, if there was any vertical transmission, either one agent was replaced per model iteration or the entire population was replaced. Here, I explore settings around these two extremes to gain more insight into the impact of replacement dynamics on the emergence of a preference for compositionality. Additionally, the replacement mechanism is adjusted to make it more ecologically plausible: instead of selecting an agent at random, an agent with low communicative success is replaced, which can be seen as a sort of "survival of the fittest" mechanism.

4.4.1 Experimental Setup

To gain more insight into the effects of agent replacement, I modified the model so that at each iteration, agents are replaced probabilistically based on a global parameter r. The number of replacement decisions matches the population size. If, for example, $r = 0.5$, each of the *pop_size* replacement opportunities has a 50% chance of resulting in an agent replacement.[17] When a replacement happens, an agent from the old population with low communicative success is selected and removed from the population. To be specific, after the horizontal interactions, each agent is evaluated. During evaluation, an agent plays the language game as the speaker with every other agent in the population as the learner. If the guessed meaning by the learner is the same as the intended meaning, the communicative success score of the speaker agent is increased by 1. After evaluation of all agents, the communicative success scores are normalized and then reversed, such that the least successful agents have a higher probability.[18] From this probability distribution, an agent is sampled to be removed from the network. Then, the new agent learns for 20 (vertical transmission) rounds from an arbitrary agent in the remaining population. Simulations were run with replacement probabilities ranging from 0.0 to 1.0. See Table 4.2 for the average number of agents replaced per iteration for each value of r.

Table 4.2 *Average Number of Agents Replaced per Iteration*

r	Agents replaced (mean)
0.00	0
0.01	0.25
0.05	1.25
0.10	2.50
0.25	6.25
0.50	12.50
0.75	18.75
0.90	22.50
1.00	25.00

For both horizontal transmission and vertical transmission, the number of interactions was set to 20. The population size was 25, and simulations were run for both sample and MAP-learners, for 10000 iterations.

[17] This does not mean that each agent has an independent probability r of being replaced. Rather, each replacement event is determined probabilistically, and when replacement occurs, the specific agent is chosen based on communicative success.

[18] Specifically, $p_{\text{new}_i} = p_{max} - p_{\text{old}_i}$, where p_{new_i} is the new probability for agent i, p_{max} is the highest probability value in the distribution and p_{old_i} is the old value for agent i.

4.4.2 Results

As shown in Figure 4.4.1, different replacement probabilities result in varying preference for compositionality in the final distribution.[19] Replacement probabilities of 0.5, 0.75 and 0.9 result in the highest preference for compositionality, of around 81% (CI: 71% - 92%), 85% (CI: 76% - 95%) and 84% (CI: 73% - 94%), respectively. When the replacement probability is lower or higher, the final preference for compositionality is lower. To be specific, when approximately 1 agent, 2.5 agents, or all agents are replaced each iteration, the final value of compositionality is around 65%. When either 1 agent per approximately 4 iterations is replaced, or 6.25 per iteration, the final preference for compositionality is around 72%. There is a clear trend in convergence speed for $r > 0.0$: as the replacement probability increases, the convergence speed increases as well. This is in line with the previous experiments, where convergence speed was higher when the entire population was replaced compared to when only one agent per iteration was replaced.

As shown in Figure 4.4.2, simulations where $r > 0.0$ result in an average final distribution divided between compositional and holistic languages, where the relative balance between the two language types depends on the replacement probability.

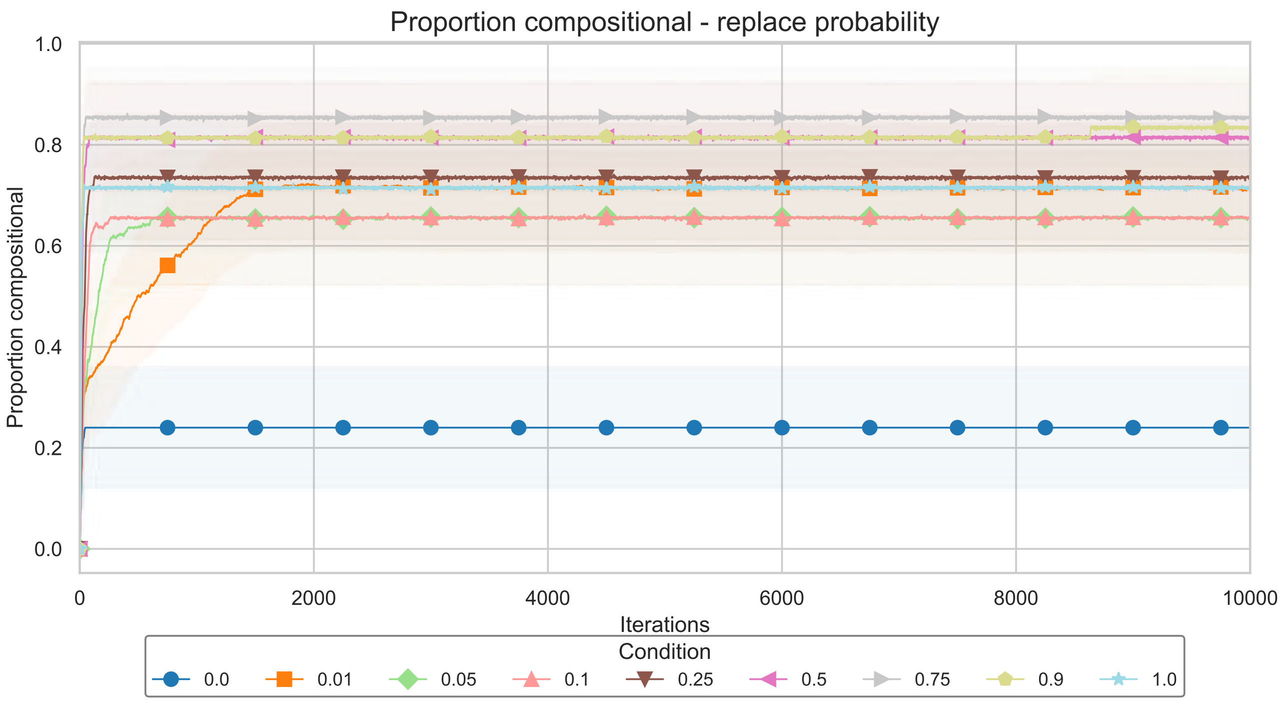


Figure 4.4.1: Posterior probability of compositional languages across 9 replace probabilities, averaged over 50 runs. Simulations were run for 10000 iterations. Population size = 25.

[19]The results reported here are for sample-learners; corresponding plots for MAP-learners can be found in Appendix B.

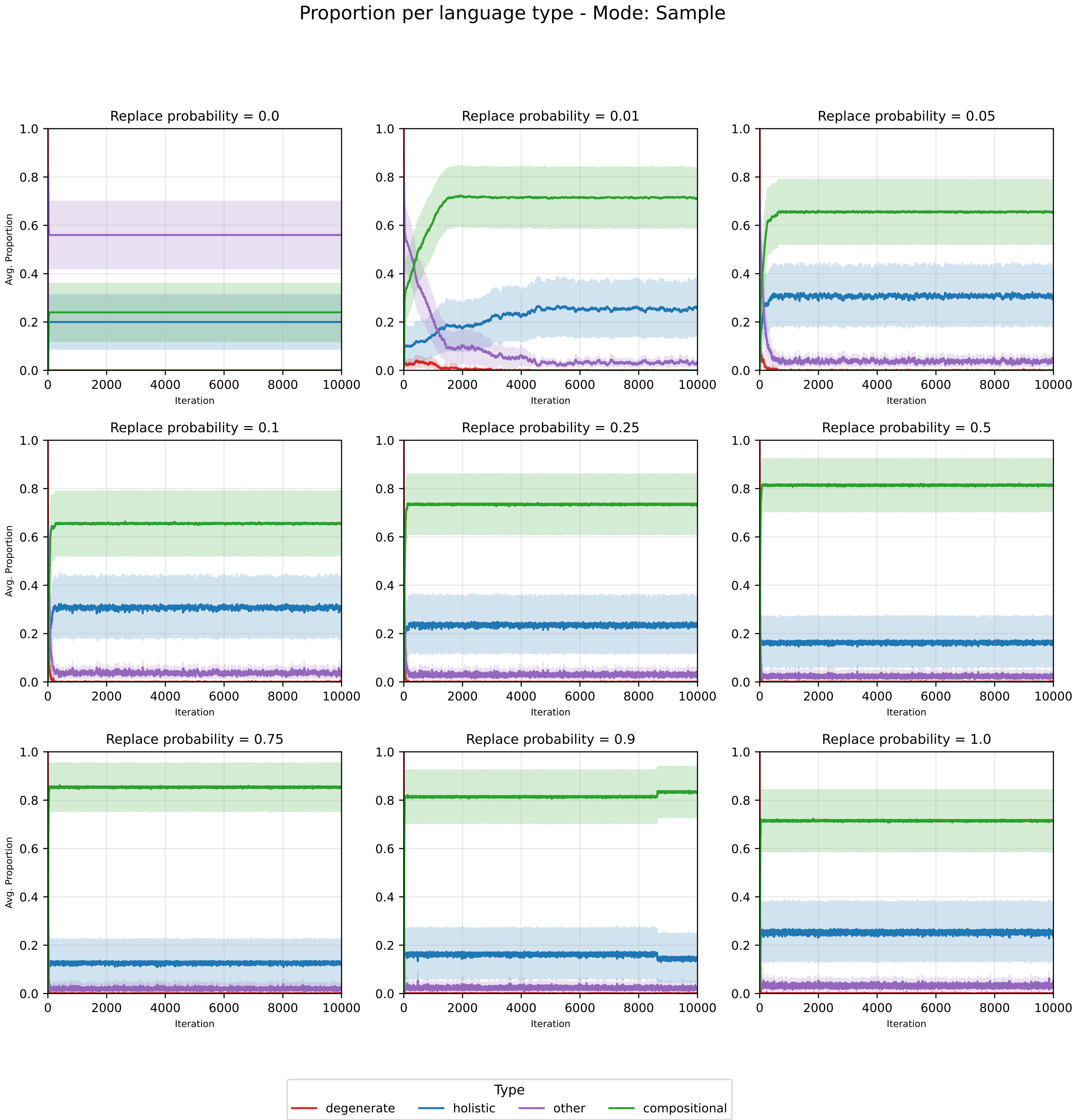


Figure 4.4.2: Proportion of posterior per language type, for various replacement probabilities. Values are averaged across the entire population of sample-learners and averaged over 50 runs. The shaded area represents 95% confidence intervals.

4.4.3 Discussion

Overall, the rate of replacement seems to impact the level of preference for compositionality. Again, similar to the previous experiments, this preference for compositionality emerges due to a trade-off between *expressivity* and *compressibility*. Since in this model the agent to be replaced is selected based on low communicative success, replacement of an agent not only introduces pressure for compressibility, as a new agent with a strong compressibility prior is added to the population, but it also introduces pressure for expressivity. Because of this, it is hard to explain why we see this specific division between compositional and holistic languages in the various replacement settings. Compared to previous experiments in which the replaced agents were selected at random rather than based on communicative success, the proportion of holistic languages in the final distribution is higher. The extra pressure for expressivity thus leads to the rise of holistic languages.

The difference in convergence speed can be explained by the fact that as more agents are replaced, more (vertical) interactions take place, and so the distributions of agents are updated more often, resulting in faster convergence compared to when fewer agents per iteration are replaced.

Chapter 5 – Individual Learning Mechanisms

Cultural transmission is not only impacted by factors such as transmission modes, population size, or network structure, but it is also impacted by learning mechanisms of the individual (Arnon et al., 2025; Kirby et al., 2007). Accordingly, the experiments reported in this chapter investigate how changes in the assumptions about agents' learning mechanisms affect the emergence of a preference for compositionality.

Experiment 5: Regulating the Impact of New Evidence

In all the previously reported experiments, the standard version of Bayesian updating is used. By the nature of this update rule, eventually, the initial compressibility prior is completely washed out by the evidence an agent observes. Since we see this initial compressability prior as a "cognitive bias" (Culbertson & Kirby, 2016; Kirby et al., 2015), it is conceptually quite strange that this cognitive bias completely diminishes once an agent observes enough data. By making a small change to the update rule, the prior can remain a stable bias through the entire simulation.

5.1.1 Experimental Setup

By adjusting the update rule to a version where, at every update, a little bit of the compressibility prior is added into the posterior, the preference for compressibility never completely fades away. The new update rule is as follows:

$$P_t(l_i|f,m) \propto (1-\alpha) \cdot P(f|m,l_i) \cdot P_{t-1}(l_i) + \alpha \cdot P_0(l_i)$$

Where $\alpha \in [0.0, 0.00025, 0.0005, 0.00075, 0.001]$, $P(f|m,l_i)$ is the likelihood, $P_0(l_i)$ is the initial compressibility prior, and $P_{t-1}(l_i)$ is the posterior value of language l_i before updating. The posterior is updated for every $l_i \in \mathbf{l}$. After updating, the posterior distribution is normalized.

Similar to Experiment 1, h and v were assigned values of 0, 10, or 20. The population size was 25, and simulations were run for 10000 iterations for both MAP and sample-learners.

5.1.2 Results

Results presented in Figure 5.5.1 show that when a little bit of the compressibility prior is injected at each update, i.e., $\alpha > 0$, horizontal transmission alone leads to a preference for compositionality, while this is not the case when $\alpha = 0$. Specifically, 10 horizontal interactions lead to a preference for compositionality, of around 99% when $alpha > 0$. More horizontal interactions ($h = 20$) lead to a slightly lower level of compositionality, ranging from 90% (CI: 83% - 98%) for $\alpha = 0.001$ to 97% (CI: 92% - 100%) for $\alpha = 0.00075$. When none of the prior was injected during updating, the preference for compositionality was around 18% (CI: 7% - 28%) for both $h = 10$ and $h = 20$. When there is a sufficient amount of vertical transmission, injecting increasing amounts of the prior at each update leads to a decrease in the final preference for compositionality. To be specific, when there is no prior injection, the final level is around 77%, regardless of the level of horizontal transmission, compared to a final value of around 59% when the most prior is injected. Similar to Experiment 1, if there is too little vertical transmission ($v = 10$), no preference for compositionality emerges, reaching a final value around 0% for all conditions, as there is not enough pressure for expressivity. Furthermore, similar to Experiment 1, once vertical transmission is included, the impact of horizontal transmission is minimal.

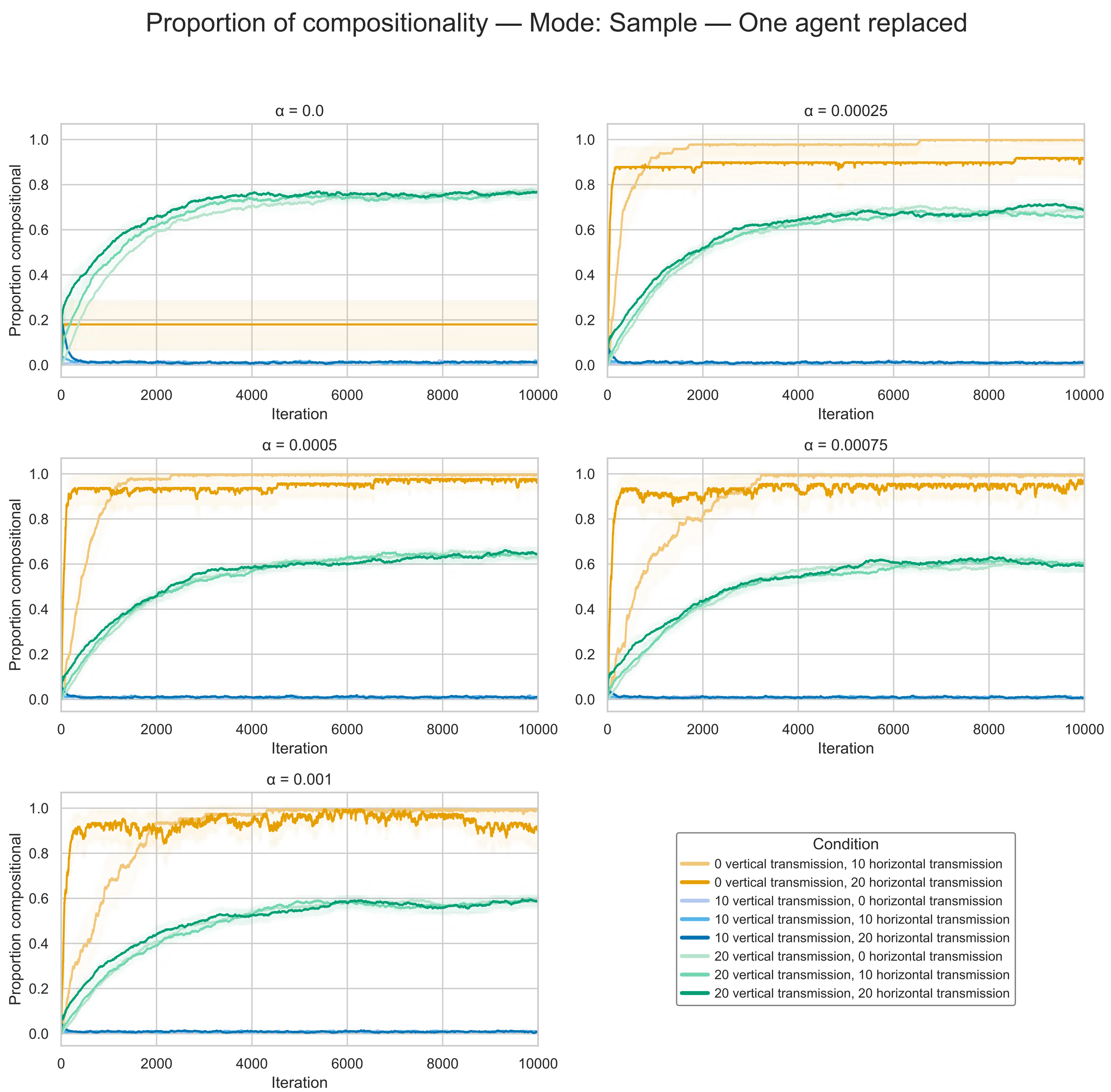


Figure 5.5.1: Proportion of posterior for compositional languages averaged for the entire population of sample-learners, averaged over 50 runs, for various values of α. The shaded area represents 95% confidence intervals. If there is vertical transmission, one agent is replaced per iteration.

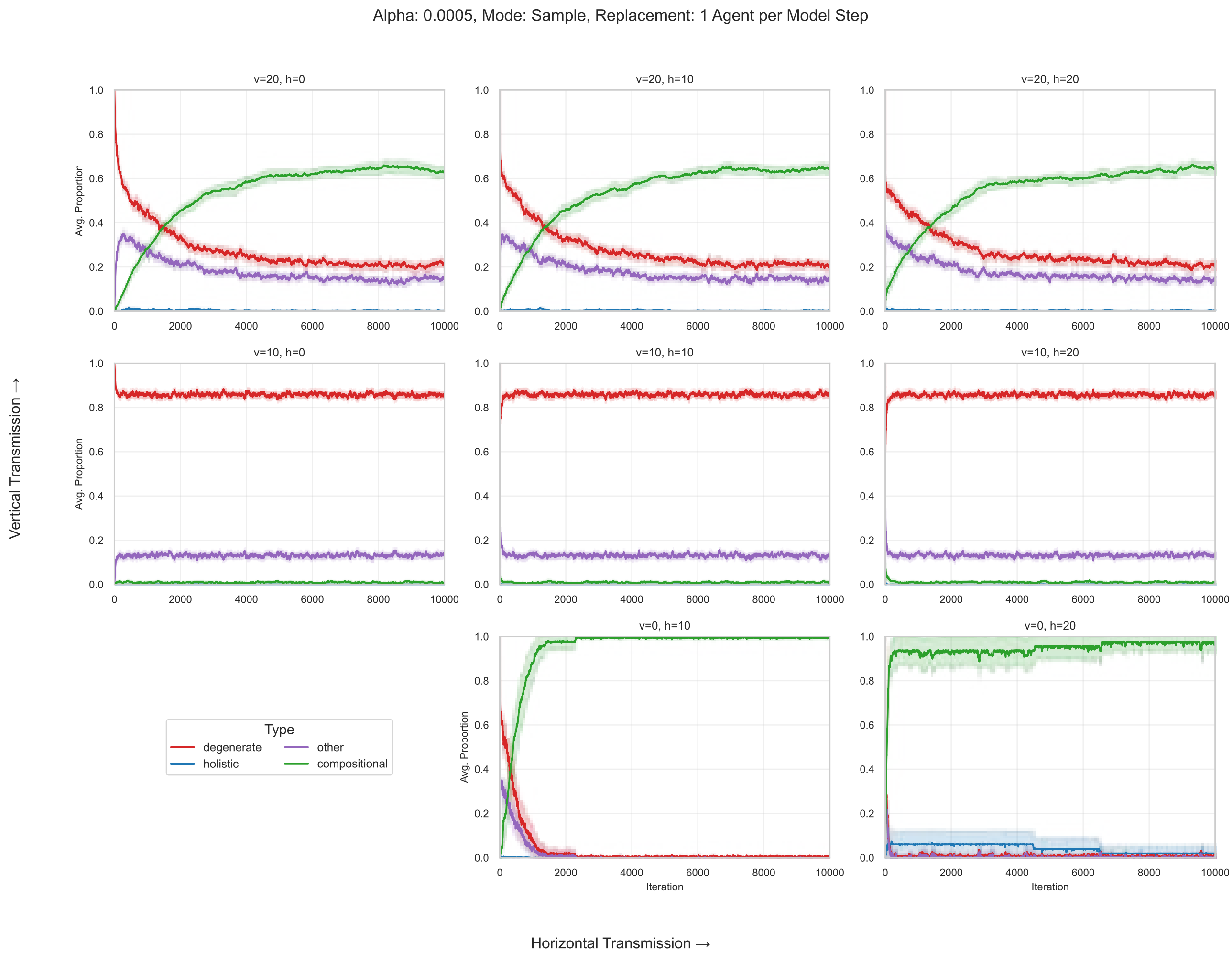


Figure 5.5.2: Proportion of posterior per language type, for various levels of horizontal and vertical transmission. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. $\alpha = 0.0005$. The shaded area represents 95% confidence intervals.

5.1.3 Discussion

When a small amount of compressibility prior is added to the posterior at every update, resulting in a stable bias for compressibility throughout the simulation, a preference for compositionality emerges within a single generation. The finding that compositionality can emerge within a single generation is in line with previous experimental and computational work (Raviv et al., 2019a; Vogt, 2005a). The stable bias does not have a big impact on the results when vertical transmission is included in the simulations.

These results highlight the importance of how assumptions about individual learning mechanisms interact with the outcomes of cultural evolution. To be specific, maintaining a stable bias changes the dynamics of the emergence of compositionality when only horizontal transmission is considered.

Similar to the previous experiments, the preference for compositionality reflects a trade-off between *compressibility* and *expressivity*. As the impact of the evidence decreases, since more of the compressibility prior is interjected at each update, the pressure for compressibility is stronger. This shift in the trade-off between compressibility and expressivity explains the decrease in final value for compositionality when larger values of α are used, as more of the final distribution is centered on degenerate languages in these settings.

Although the current learning mechanism is not directly based on empirical findings in human cognition, it demonstrates that even minor changes to the learning mechanisms can lead to significant differences in outcomes of cultural transmission. For future work, it would be interesting to model a learning mechanism directly based on empirical work.

Experiment 6: Strength of the Compressibility Prior

For all previously reported experiments, agents' initial priors were determined by the compressibility of each language in its minimally redundant form (see Subsection 3.3.1 for details). This prior strongly favors degenerate languages, reflecting a domain-general bias toward 'simplicity' (Chater & Vitányi, 2003; Culbertson & Kirby, 2016; Kirby et al., 2015).

Kirby et al. (2007) showed that iterated learning can lead to strong linguistic universals, even when assuming a weak bias. Through vertical cultural transmission, this weak bias gets amplified over generations. In their model, the regularity of the final language depended completely on the size of the information bottleneck, and not on the strength of the prior. These results were only observed when agents are not precisely sample learners, i.e., they pick a hypothesis with a higher probability disproportionally more often, of which MAP-learning is the most extreme option.

Whereas Kirby et al. (2007) directly assumed a bias in favor of structural regularity, the model used in this thesis combines a compressibility bias with a pressure for expressivity. The current experiment investigates whether the strength of this compressibility bias impacts the emergence of a preference for compositionality.

5.2.1 Experimental Setup

The compressibility bias used in all previous experiments is quite strong, since around 99% of the prior is focused on degenerate languages. To weaken the bias, I introduce β, which allows us to adjust how strongly coding length influences the initial base distribution:

$$G_0(l) \propto 2^{\beta \cdot L(l)}$$

Here, $\beta \in [0, 1]$, and $L(l)$ is the coding length of language l. When $\beta = 0$, the initial prior is uniform, so coding length has no effect. As β increases, more compressible languages receive higher prior probability. By varying β, we can test whether the strong simplicity bias is necessary for the emergence of a preference for compositional languages, or whether similar patterns emerge under weaker assumptions. See Figure 5.6.1 for the prior per language type for various values of β. The population size was 25, and vertical and horizontal transmission were set to 20. Simulations were run for both sample and MAP learners, for 10000 iterations.

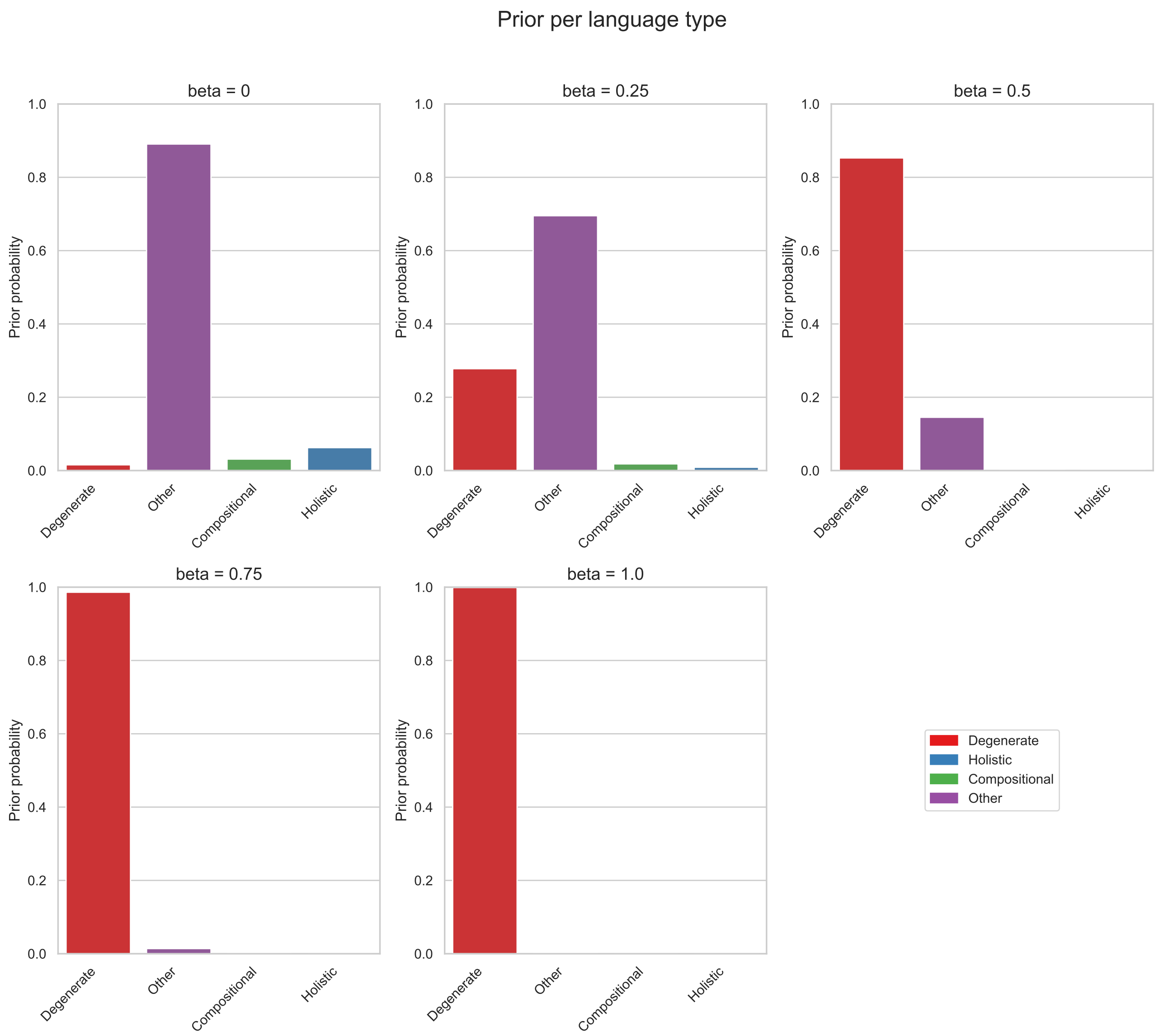


Figure 5.6.1: Prior probability across language types for different values of β. When $\beta = 0$, the prior is uniform, and "other" languages dominate because they are most frequent in the set of all languages. As β increases, the prior increasingly favors more compressible (simpler) language types.

5.2.2 Results

As shown in Figure 5.6.2, when agents are sample-learners, a weaker compressibility bias (e.g., $\beta \geq 0.5$), in combination with pressure for expressivity, leads to the emergence of a preference for compositionality. Specifically, when one agent is replaced per iteration, and $\beta = 0.5$, $\beta = 0.75$ or $\beta = 1.0$, the final value for a preference for compositionality are 75% (CI: 73% - 77%), 82% (CI: 80% - 84%) and 77% (CI: 74% - 79%), respectively. Even when the compressibility bias is relatively weak ($\beta = 0.25$), and so most of the prior does not lie on degenerate hypotheses but on 'other' hypotheses, we see the emergence of a preference for compositionality, where around 60% (CI: 58% - 63%) of the final distribution lies on compositional languages, when one agent is replaced per iteration. As shown in Figure 5.6.3, as the strength of the compressibility prior decreases, a larger proportion of holistic languages are in the posterior. This is the case since holistic languages are effective for communication, in comparison to degenerate or other languages, but they are not compressible. On the other hand, if the compressibility prior is extreme ($\beta = 1$), a larger proportion of the posterior lies on degenerate languages compared to the other settings.

When we look into the results for MAP-learners (Figure 5.6.4), the results are even stronger: when $\beta = 0.25$, so the compressibility prior is much weaker than the original one, a strong preference for compositionality arises, of around 85% (CI: 83% - 87%). Additionally, assuming a weaker bias ($0.25 \geq \beta \geq 0.75$), leads to an overall higher final level of compositionality of above 80%, while the original compressibility prior leads to a final level of compositionality of around 76%. This is because in this setting, more of the posterior is focused on degenerate languages, as shown in Figure 5.6.5.

When there is no compressibility bias at all, the results of the sample and MAP-learners are similar, where the biggest part of the posterior is focused on holistic languages.

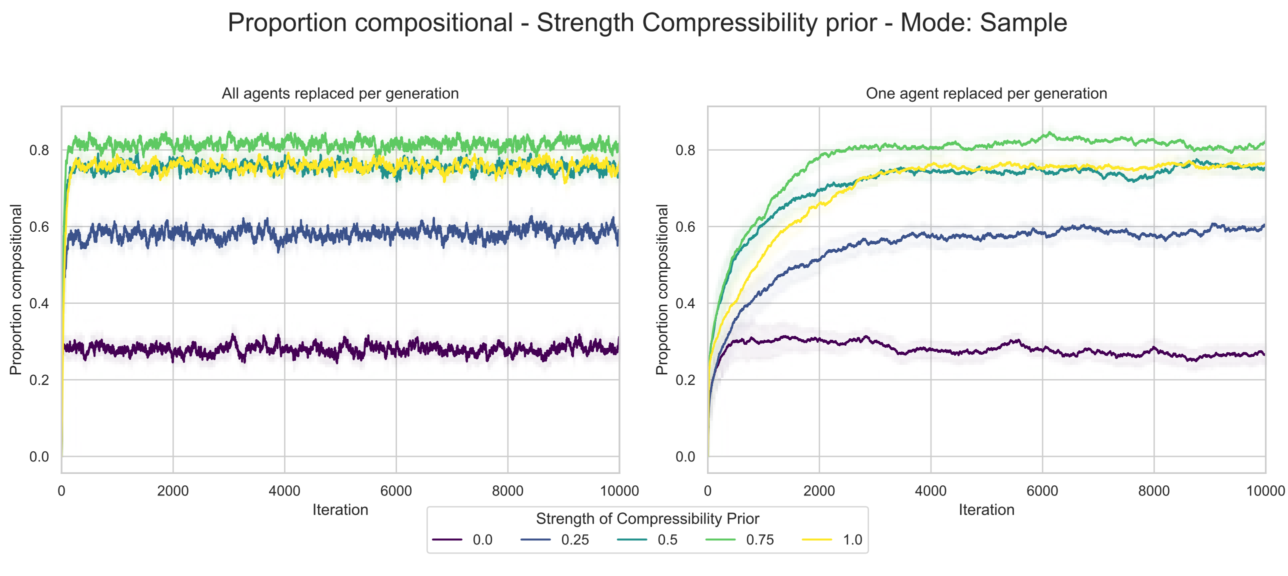


Figure 5.6.2: Proportion compositionality in posterior distribution for various strengths of the compressibility bias. Agents are sample-learners. Population size = 25, horizontal transmission rounds = 20, vertical transmission = 20.

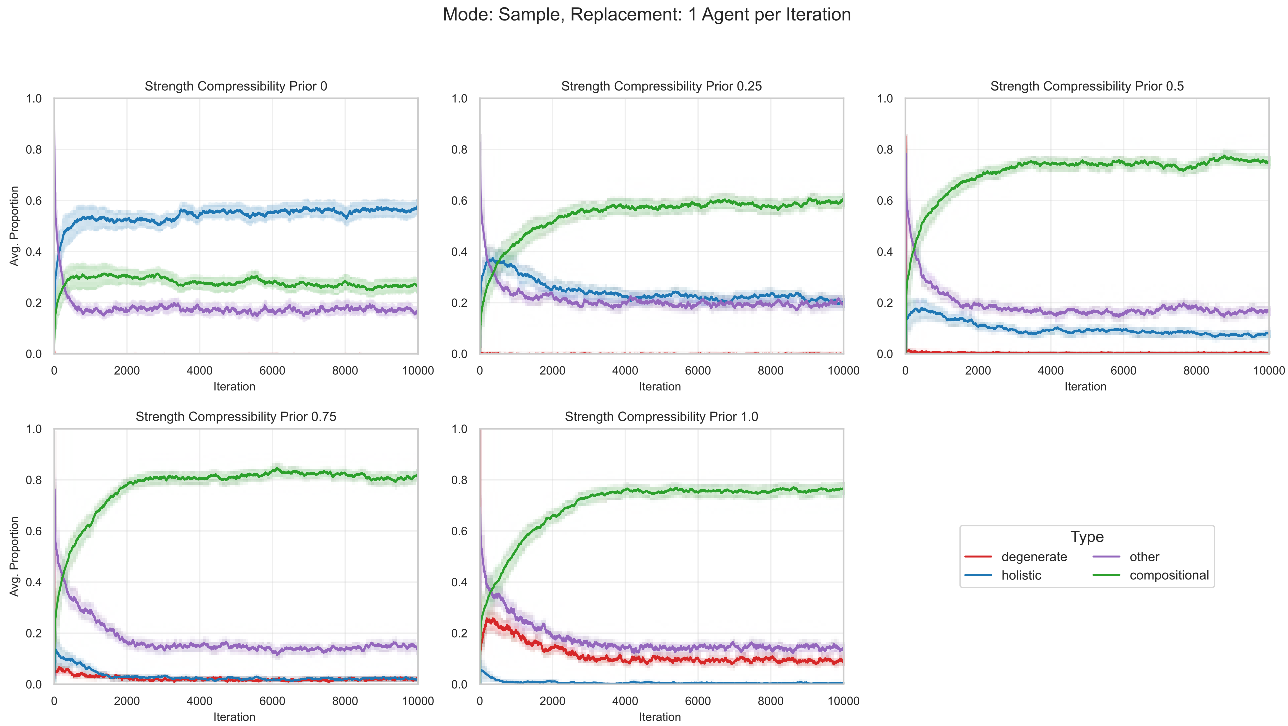


Figure 5.6.3: Proportion of posterior per language type, for various compressibility prior strength. Agents are sample-learners. If $\beta = 0$, the prior is uniform across all languages, while if $\beta = 1.0$, the compressibility prior is the strongest. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

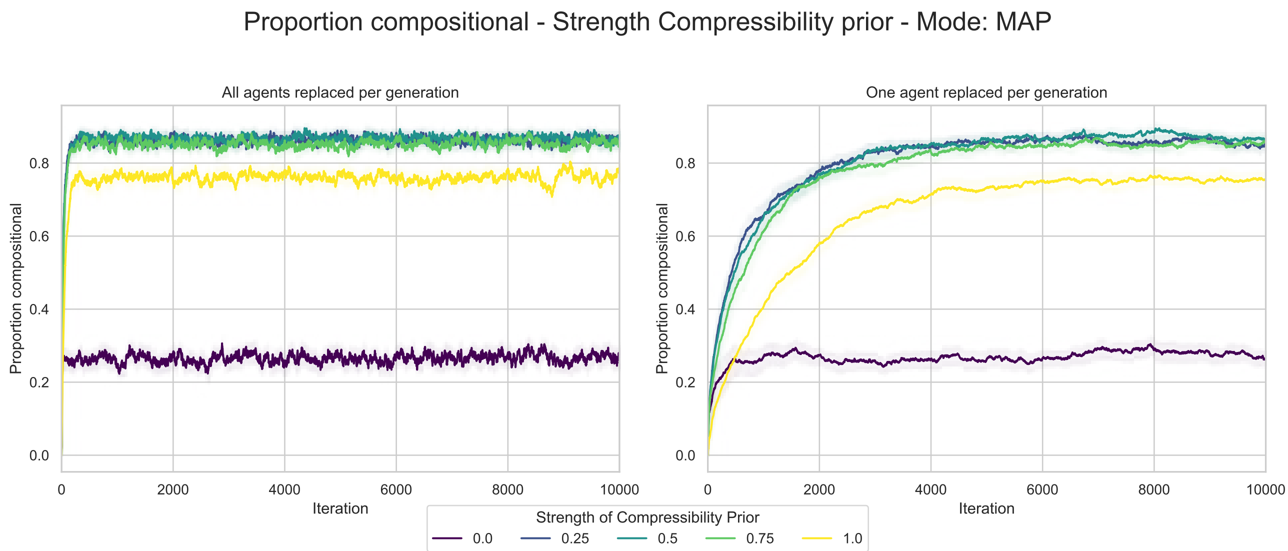


Figure 5.6.4: Proportion compositionality in posterior distribution for various strengths of the compressibility bias. Agents are MAP-learners. Population size = 25, horizontal transmission rounds = 20, vertical transmission = 20.

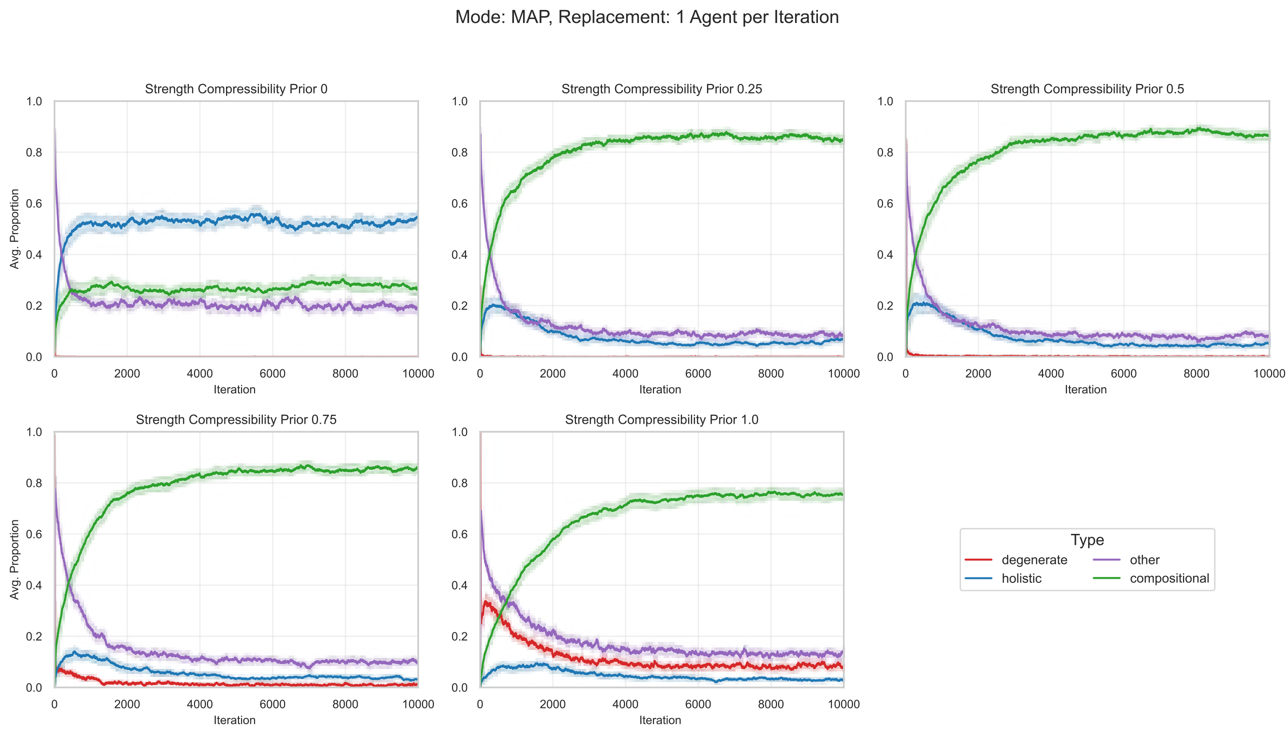


Figure 5.6.5: Proportion of posterior per language type, for various compressibility prior strength. Agents are MAP-learners. If $\beta = 0$, the prior is uniform across all languages, while if $\beta = 1.0$, the compressibility prior is the strongest. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

5.2.3 Discussion

As in Kirby et al. (2007), we see that weak biases in combination with vertical transmission can lead to strong linguistic universals, in this case, the emergence of compositionality. In contrast to their findings, we observe these results with both MAP and sample learners, rather than only with MAP learners. This difference can be explained by the fact that in our model, agents learn from multiple teachers rather than from just one, as this affects the stationary distribution (Smith, 2009). This finding is particularly interesting because it shows that, regardless of sampling mechanism, weak biases can lead to linguistic universals when learning happens from multiple teachers, which is a more ecologically valid assumption. While it remains difficult to determine precisely what inferential process humans use, it is generally accepted that humans acquire language from multiple teachers. Overall, the results indicate that weak individual biases for compressibility, especially in a setting with MAP-learners, a preference for compositionality can arise when combined with a pressure for expressivity and vertical transmission. How precisely this bias for compressibility arises remains unclear, but a plausible explanation is the need to communicate in an expanding meaning space under constraints of limited memory. For future work, it would be interesting to explore settings between these two extreme sample modes. In such a setting, an agent samples a language with a higher probability disproportionally more often. This would give more insight into how precisely the sampling mode interacts with the emergence of compositionality.

Chapter 6 – Cultural Transmission Revisited

In this chapter, I introduce two follow-up experiments based on population size (Experiment 2) and on social network structures (Experiment 3). Specifically, for the population size experiment, I consider a setting with a heterogeneous population, and for the experiment on network structures, I look into a setting with horizontal transmission only while maintaining a stable compressibility prior (Experiment 5). Together, these experiments give insight into whether earlier results are caused by specific modeling assumptions or if the results are robust across different settings.

Experiment 7: Heterogeneous Population

This experiment is a follow-up of Experiment 2, where in comparison to previous work (e.g., Raviv et al., 2019b; Rita et al., 2021; Vogt, 2007), we did not see a faster or more prominent emergence of compositionality in larger populations. As suggested by Rita et al. (2021), the absence of this population size effect might be caused by the homogeneity of the population. In a heterogeneous population, the increase in the population leads to an increase in input variability, which is not the case in a homogeneous population. Overcoming this larger input variability might be the reason why we see a faster and more prominent emergence of compositionality in larger populations. In the current experiment, I investigate whether this population size effect can be observed when the population is heterogeneous.

6.1.1 Experimental Setup

To create a heterogeneous population, agents are initialized with initial priors varying in their compressibility strength. Specifically, an agent's compressibility strength, and thus their prior, is determined by generating a β value between 0 and 1 (as introduced in Experiment 6). The prior of the agent is then calculated as follows:

$$G_0(l) \propto 2^{\beta \cdot L(l)}$$

where $L(l)$ is the coding length of a language. Simulations were run for population sizes 2, 10, 20, 50, 100, and 200. The number of horizontal and vertical interactions was set to 20, and simulations were run for both MAP and sample learners for 10000 iterations.

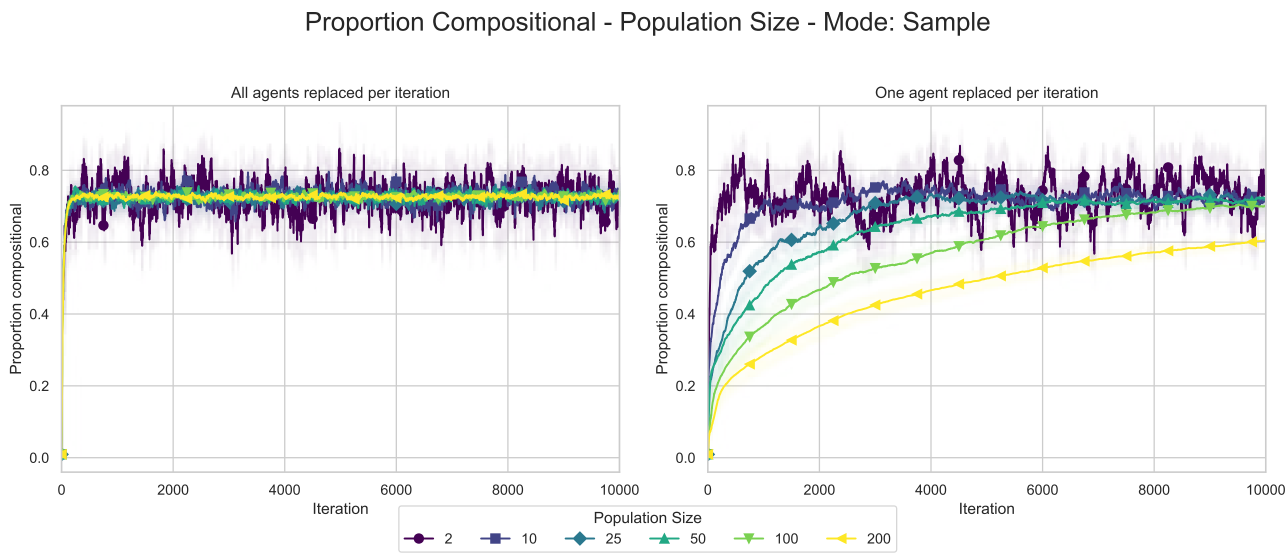


Figure 6.7.1: Posterior probability for compositional languages averaged for the entire population of sample-learners, averaged over 50 runs, for different population sizes. The shaded area represents the 95% confidence interval. In the left figure, every agent is replaced each iteration, and in the figure on the right, only one agent is replaced each iteration.

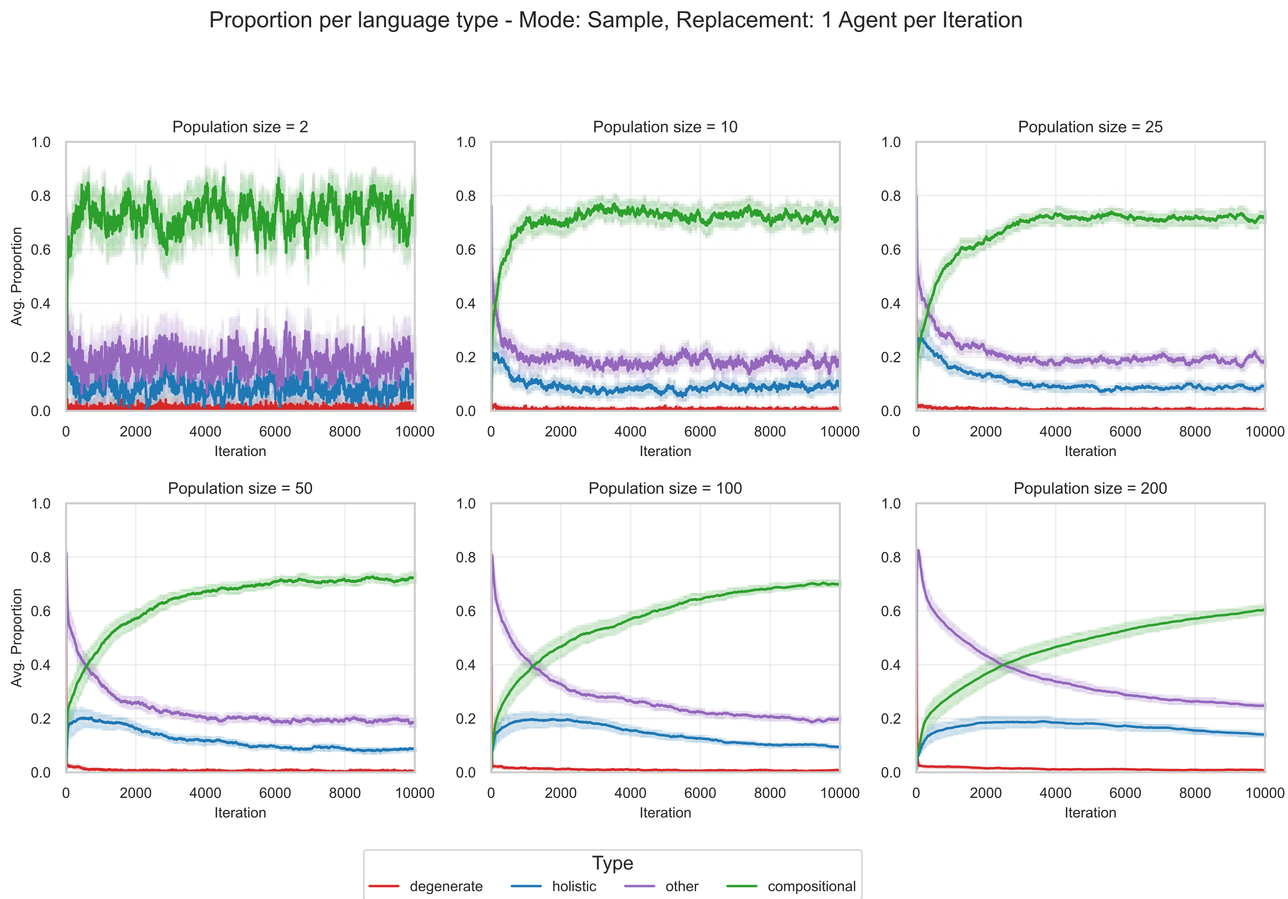


Figure 6.7.2: Proportion of posterior per language type, for various population sizes. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

6.1.2 Results

As shown in Figure 6.7.1 and Figure 6.7.2, larger populations do not lead to a faster or more prominent emergence of compositionality, even when the population is heterogeneous. Specifically, the final values of the proportion of the posterior focused on compositional languages range from 70% to 74% when the entire population is replaced each iteration, and from 61% to 77% when one agent is replaced each iteration. Compared to the results of Figure 4.2.1, populations of size 100 and 200 show a faster increase in preference for compositionality, but overall convergence slows as population size increases when one agent is replaced per iteration. If the entire population is replaced, there is no difference in convergence speed between the homogeneous and heterogeneous populations. However, the two settings differ slightly in the final value of preference for compositionality, where the heterogeneous population results in a slightly lower preference of around 72%, compared to approximately 77% in the homogeneous population.

6.1.3 Discussion

These results suggest that having a heterogeneous population is not sufficient to produce the population size effect on the emergence of compositionality as reported in previous work (e.g., Raviv et al., 2019b; Rita et al., 2021; Vogt, 2007). The absence of this effect may be explained by three factors. First, the form-meaning space in this model is very small, which limits the possibility for increased input variability. Each of the four meanings can be matched with any of the four forms, leading to only 16 different form-meaning pairs. Hence, having a larger and diverse population will likely not lead to increased input variability, and so there is no additional pressure for compositionality to emerge in larger populations. Second, the pressure for communicative success in this model is implemented by pragmatic agents, who only penalize ambiguous languages rather than communicative failures. Agents are not actively striving for communicative success; they only try to avoid ambiguity. Hence, the expressivity pressure in the current model might be too limited to overcome the input variability leading to the emergence of compositionality. Lastly, it might be the case that this specific manipulation does not make the population heterogeneous enough to cause a difference in the input variability.

This suggests that, next to heterogeneity, other factors such as pressure for communicative success and a large enough meaning-signaling space might be necessary to elicit a bigger preference for compositionality in larger populations.

Experiment 8: Social Network Structures Revisited

This experiment is a follow-up of Experiment 3, and focuses on the impact of social network structures in a setting with only horizontal transmission specifically. The goal of this experiment is to investigate whether the network structures inhibit the pressure for expressivity, affecting the emergence of a preference for compositionality.

In the original simulations with horizontal transmission only, the compressibility bias was completely washed out by new evidence. By regulating the impact of new evidence, the agents in this simulation have a stable compressibility bias, as introduced in Experiment 5.

6.2.1 Experimental Setup

Simulations are run for the different network types, as introduced in Experiment 3. To be specific, I compare a fully-connected network, a small-world network, a scale-free network, a hierarchical network, and a network based on mobile-phone data. The dynamic network is excluded, as it requires the introduction of new agents (i.e., vertical transmission) when an agent in the network becomes completely disconnected due to edge removal. Simulations were run for both MAP and Sample learners, with 20 horizontal transmission rounds, a population size of 25, and 10000 iterations. To maintain a stable compressibility bias throughout the simulations, a little bit of the prior is injected at each update, as introduced in Experiment 5, with $\alpha = 0.0005$.

6.2.2 Results

As shown in Figure 6.8.1, all network types show a similar trajectory in the emergence of preference for compositionality. The final mean level of compositionality was high in every condition, ranging from 87% (CI: 80% - 96%) for the small-world network to 96% (CI: 92% - 99%) in the scale-free network. The final values for the fully connected network, mobile-phone network, and hierarchical network were 90% (CI: 81% - 98%), 92% (CI: 84% - 99%), and 94% (CI: 89% - 99%), respectively. While the final values of compositionality are similar, there is an observable difference in convergence speed between the different network types. Specifically, sparser networks (e.g., scale-free) take about 1000 iterations to reach a stable level of compositionality compared to more tightly-knit networks (e.g., fully-connected or mobile-phone), which stabilize around 125 iterations. Shown in Figure 6.8.2, the remaining part of the posterior lies on holistic languages, but this is consistently less than 10% for all the network types.

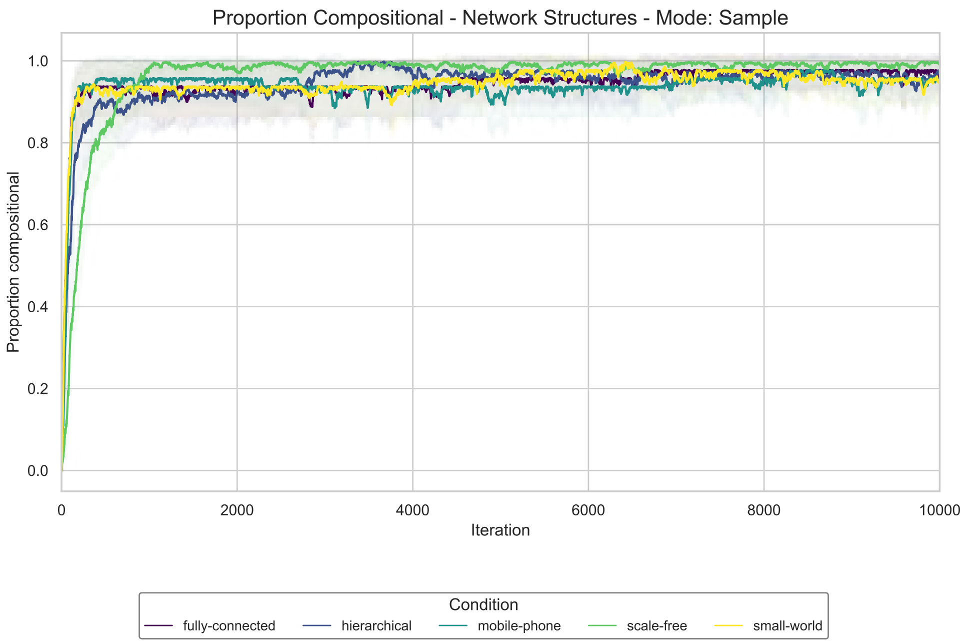


Figure 6.8.1: Posterior probability for compositional languages averaged for the entire population of sample-learners, averaged over 50 runs, for different network types. The population size is 25. Vertical and horizontal transmission rounds = 20. Shaded area represents 95% confidence intervals.

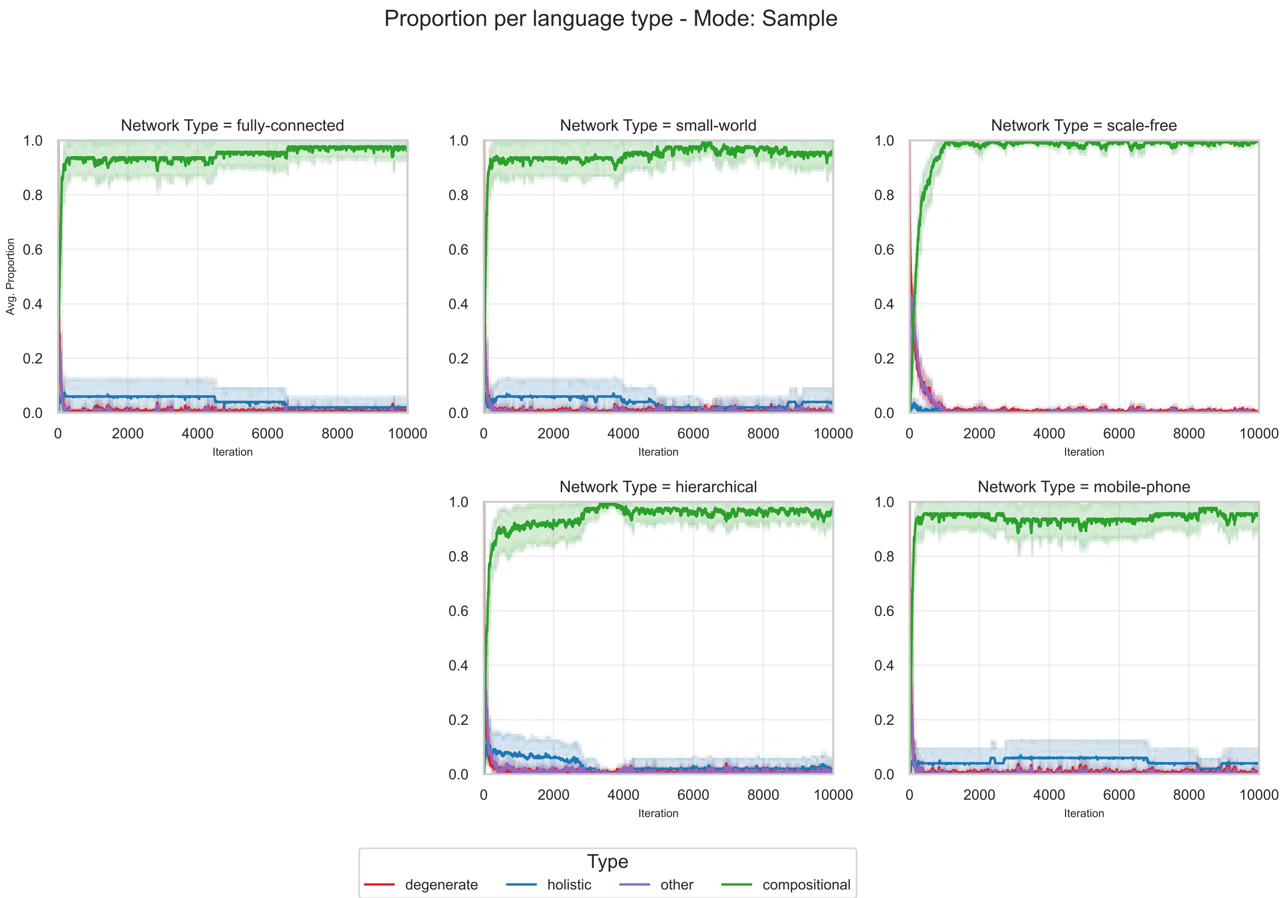


Figure 6.8.2: Proportion of posterior per language type, for various network types. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

6.2.3 Discussion

Consistent with Experiment 3, and previous experimental work (Raviv et al., 2020), we found no impact of social network structure on the final level of compositionality. In this Experiment, compared to Experiment 3, there was only horizontal transmission, and so for all interactions in the simulation, the network structure limited who could interact with whom. Since we still observe no impact of network structure on the emergence of a preference for compositionality, we can conclude that the results are not caused by the setup in Experiment 3, which included vertical transmission whereby the network structure did not limit the interactions in any way. Overall, we can conclude that network structure can influence the amount of compositionality in the short term, as different networks exhibit different convergence speeds. In sparser networks, it takes longer for compositionality to emerge compared to more tightly-knit ones. In the long term, however, network structures do not completely prevent the emergence of compositionality, as preference for compositionality always arises regardless of network structure.

However, these conclusions should be drawn with caution, since previous work has shown that network structure impacts input variability. Specifically, networks with longer average paths and low clustering result in higher variability than more tightly-knit networks (Josserand et al., 2024). Raviv et al. (2020) suggest that more input variability might be a driver for the emergence of compositionality, since it is harder to communicate successfully with more variation, and so there is more pressure for a systematic language to arise. As mentioned before, in this model, the possibility for variability is limited, since there are only 4 meanings and 4 signals. So, it could be the case that the networks do not lead to differences in input variability, and therefore, there is also no difference in the level of compositionality. Similarly, Raviv et al. (2020) argue that the lack of input variability could cause the lack of difference in level of compositionality in their experiment. Future work should carefully consider experimental setups that allow for more input variability, such as expanding the meaning and signaling spaces, or by adopting a more open-ended signaling space rather than a fixed set of signals.

Chapter 7 – General Discussion & Conclusions

7.1 Discussion

This thesis addressed two sets of research questions concerning the emergence of compositionality. The first set of experiments, presented in Chapter 4, examined how various factors impacting cultural transmission, such as vertical and horizontal transmission, population size, social network structure, and rate of generational replacement, shape the emergence of compositionality. The second set of questions, presented in Chapter 5, focused on the role of individual learning mechanisms and their impact on the emergence of compositionality. Specifically, the first experiment examined how regulating the impact of new evidence affects the emergence of compositionality, and the second investigated how the strength of the compressibility bias impacts the emergence of compositionality. Finally, Chapter 6 presented two follow-up experiments, building upon Experiment 2 and Experiment 3.

In general, all experiments showed that compositionality emerges from a trade-off between *expressivity* and *compressibility*, as proposed by Kirby et al. (2015). When this trade-off is disrupted, for example by introducing new agents with a strong compressibility prior or increasing the amount of interactions between agents during vertical transmission, the final proportion of compositional languages in the posterior distribution of the population changes, as demonstrated in Experiment 1. Similarly, by introducing new agents at a different rate, which results in differences in compressibility pressure, the emergence of compositionality is affected, as demonstrated in Experiment 4. Conversely, if the experimental setting does not affect the pressure for compressibility or expressivity, as in Experiment 2 with various population sizes and in Experiment 3 with different network structures, the emergence of compositionality is not affected.

This trade-off between *expressivity* and *compressibility* is a well-known phenomenon that seems to play a more general role in shaping the semantic structure of lexical systems. For example, this trade-off is has been observed across several semantic domains, including kinship systems (Kemp & Regier, 2012) and color terms (Zaslavsky et al., 2018), among others. The tension between these two pressures is thus not only important for the emergence of compositionality, but more generally for the organization and evolution of communication systems.

Results from Experiment 1 and Experiment 4 highlighted that introducing new learners (i.e., vertical transmission) encourages the emergence of compositionality. Experiment 1 and the sensitivity analysis showed that the number of vertical interactions has a greater impact on the final level of compositionality than the number of horizontal interactions. In other words, if a new agent with a still strong preference for compressibility is added to the population, the initial number of interactions with its assigned teacher has a large impact on the entire population compared to the interactions among existing agents. The precise reason for this remains unclear, and it may be specific to this computational model.

The experiments on population size (Experiment 2) and social network structure (Experiment 3) demonstrated the importance of carefully selecting specific modeling settings and considering potential interactions before conducting an experiment. In both cases, the observed results could be explained by limitations of the modeling settings: the heterogeneity of the population in Experiment 2 and the inclusion of vertical transmission in Experiment 3. This illustrates how modeling assumptions can impact experimental outcomes, underscoring the need for critical evaluation of these assumptions. As argued by Smaldino (2017), critical evaluation of modeling assumptions can be helpful for modeling refinements. As such, the follow-up experiments (Experiment 7 and Experiment 8) dealt with the initial limitations, yet they still showed similar results, where both population size and social network structure did not seem to impact the emergence of compositionality. Both experiments further revealed a limitation of the model: the small meaning signaling space restricted the possibility for an increase in input variability.

The meaning-signaling space was originally chosen for its simplicity. Its small size allows for running full inference over all possible languages, resulting in a conceptually and computationally simpler model, while it is still complex enough to distinguish between the different language types (degenerate, holistic, compositional, and other). However, precisely its simplicity is thus also a limitation, since it does not allow for much variation in the level of input variability. Simply expanding the meaning signaling space to allow for more input variability is not as straightforward as it seems. To illustrate, increasing the meaning signaling space by a single dimension, e.g., the set of meanings $\mathcal{M} = \{03, 13, 23, 04, 14, 24, 05, 15, 25\}$ and the set of signals $\mathcal{F} = \{aa, ab, ac, ba, bb, bc, ca, cb, cc\}$, results in $|\mathcal{F}|^{|\mathcal{M}|} = 9^9$ possible languages. This large set of possible languages makes full inference infeasible, and addressing this requires a model that is both computationally and conceptually more complex. One possible solution is to utilize the Chinese Restaurant Process, similar to Kirby et al. (2015). However, implementing this was beyond the scope of this thesis, as I prioritized conducting the experiments on individual learning mechanisms and the follow-up experiments, broadening and strengthening the experiments on the factors impacting cultural transmission. Moreover, Woensdregt et al. (2021) argued that scaling these kinds of Bayesian models to a realistic size is intractable. They proved that learning signal-meaning mappings, i.e., *language learning*, is NP-hard, even when other parts of the model are severely restricted. Specifically, the set of possible languages an agent considers grows exponentially with the number of meanings and signals, and cannot be searched efficiently.[20] Therefore, this type of model of language evolution is intractable, and it is impossible to scale these models beyond toy-size, limiting the generalizability of the findings.

As a potential solution, Woensdregt et al. (2021) suggest constraining the input domain while being explicit about the reason for the specific restrictions. This could make the computations tractable, enabling simulation at a more realistic scale. Future work could build on this suggestion by designing a model that not only allows for simulating at a more realistic scale, but also for more input variability.

Another promising direction would be to conduct experiments in an *emergent communication* setting,

[20] In contrast to the example where the number of possible languages is $|\mathcal{F}|^{|\mathcal{M}|}$, they allow for synonyms, resulting in $2^{|\mathcal{M}|\times|\mathcal{F}|}$ possible languages. Hence, considering a set of only 50 meanings and 25 signals results in 1.9×10^{376} possible languages.

using neural networks as agents. An advantage is that this is a more open-ended framework, where agents come up with a communication protocol themselves, instead of having a predefined set of signals or languages. In these kinds of models, there is no need to search the hypothesis space as in the Bayesian models, so the meaning and signaling space is easier to scale to a more realistic size. Furthermore, parameters such as alphabet size and message length can be used to directly manipulate the potential for input variability.

In the bigger picture, this could help to investigate whether the level of input variability is directly related to the level of compositionality. As suggested by Raviv et al. (2019b), more input variability leads to a higher pressure for compositionality. However, future work should not only consider external factors impacting input variability, such as population size or social network structure, but also internal factors such as one's cognitive abilities. To illustrate, an individual who is able to distinguish a greater number of phonemes experiences higher input variability. If a species lacks the ability to distinguish between different signals, there is less pressure for compositionality to arise, and so this might explain why non-human communication systems are less complex than ours.

Furthermore, iconicity may play an important role in the level of input variability and so potentially the level of compositionality. Mudd et al. (2020) showed that smaller communities can more often communicate successfully by using an iconic link, where the form of the signal directly resembles the meaning, as they share more social context than larger communities. Using an agent-based model Mudd et al. (2020) showed that because of this, smaller communities show a higher level of lexical variability than larger communities. It would be interesting to see how precisely iconicity interacts with not only the level of input variability but also with the emergence of compositionality.

The experiments on individual learning mechanisms, reported in Chapter 5, revealed that a preference for compositionality can arise across various types of learning mechanisms. Experiment 5 showed that when a little bit of the initial compressibility prior is injected at each update, a preference for compositionality can emerge, even within one generation. Furthermore, Experiment 6 showed that for various strengths of the compressibility prior, for both sample and MAP learners, a preference for compositionality emerged. On the one hand, these experiments show that a preference for compositionality emerges robustly across various conditions. On the other hand, they show that it is important to be explicit about the learning mechanisms used, since compositionality emerging within one generation differed between the two update mechanisms. Future work could explore more realistic learning mechanisms directly informed by empirical findings from language acquisition research. Additionally, since children and adults differ in language learning strategies (e.g., Newport, 2020), these differences could be interesting to model as well.

Finally, grounding computational work in laboratory experiments is important to critically assess whether the results are generalizable to real-life settings. I am particularly interested in investigating whether the different transmission modes lead to differences in the rate of emergence and final level of compositionality. As previous work has shown, compositionality can emerge due to the pressure of compressibility caused by vertical transmission of the language in combination with the pressure of expressivity caused by communication in dyads (Kirby et al., 2015). However, compositionality can also arise within just one generation, when communication takes place in a micro-society, which results in a different pressure for compressibility (Raviv et al., 2019b). To date, no experiment has investigated whether these two types of compressibility pressures yield differences in either the rate of emergence or the final level of compositionality. An experiment similar to Raviv et al. (2019b) and Kirby et al. (2015), in which participants play a communication game using an artificial language, could be used to address this gap. To be precise, by comparing three different conditions: vertical transmission in dyads, horizontal transmission in a micro-society, and vertical transmission in a micro-society, the effect of the various compressibility pressures can be investigated. While comparing the effect of these two types of compressibility pressure is closely related to the work in this thesis, the model in its current form is not suitable to investigate this specific question, as for the compressibility pressure from a larger population, some form of memory constraints and a larger signaling space is needed.

7.2 Conclusions

In this thesis, I investigated how different factors impacting cultural transmission, such as transmission mode, population size, network structure, or rate of generational replacement, impact the emergence of a preference for compositionality. Additionally, since individual learning mechanisms affect the outcomes of cultural transmission, I considered various settings regarding these learning mechanisms, such as the strength of a simplicity prior and the strength of the impact of new evidence. In line with previous work, the experiments revealed that the emergence of a preference for compositionality is the result of a trade-off between a pressure for *compressibility* and *expressivity.* I showed that the amount of vertical interactions between a new agent and an older agent has a substantially greater impact on the emergence of compositionality than horizontal interactions within the same generation. Furthermore, I demonstrated that whether a preference for compositionality can arise within one generation depends on the assumed learning mechanism. This highlights the importance of carefully considering and justifying certain modeling assumptions. Moreover, I showed that even a relatively weak prior bias for compressibility, in combination with a pressure for expressivity, leads to the emergence of compositionality for both MAP and sample learners, if an agent learns from multiple other agents. Lastly, population size or social network structure does not seem to impact the emergence of a preference for compositionality in the model used in this thesis. This could potentially be explained by the limited possibility for input variability. In conclusion, when investigating the emergence of a preference for compositionality, one should not only consider aspects that impact cultural transmission and the trade-off between a pressure for *expressivity* and *compressibility* but also the role of the assumed individual learning mechanisms.

Bibliography

Arnon, I., Carmel, L., Claidière, N., Fitch, W. T., Goldin-Meadow, S., Kirby, S., Okanoya, K., Raviv, L., Wolters, L., & Fisher, S. E. (2025). What enables human language? A biocultural framework. *Science*, *390*(6775), eadq8303. https://doi.org/https://doi.org/10.1126/science.adq8303.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509.

Batali, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. *Linguistic evolution through language acquisition: Formal and computational models*, 111–172.

Berthet, M., Surbeck, M., & Townsend, S. W. (2025). Extensive compositionality in the vocal system of bonobos. *Science*, *388*(6742), 104–108. https://doi.org/10.1126/science.adv1170.

Beuls, K., & Van Eecke, P. (2023, March). Fluid construction grammar: State of the art and future outlook. In C. Bonial & H. Tayyar Madabushi (Eds.), *Proceedings of the first international workshop on construction grammars and nlp (cxgs+nlp, gurt/syntaxfest 2023)* (pp. 41–50). Association for Computational Linguistics. https://aclanthology.org/2023.cxgsnlp-1.6/.

Brace, L., Bullock, S., & Noble, J. (2015). Achieving compositional language in a population of iterated learners. In P. Andrews, L. Caves, R. Doursat, S. Hickinbotham, F. Polack, S. Stepney, T. Taylor, & J. Timmis (Eds.), *Advances in artificial life, ecal 2015* (pp. 349–356). MIT Press. https://eprints.soton.ac.uk/376655/.

Brambilla, M., Ferrante, E., Birattari, M., & Dorigo, M. (2013). Swarm robotics: A review from the swarm engineering perspective. *Swarm Intelligence*, *7*, 1–41. https://doi.org/10.1007/s11721-012-0075-2.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54. https://doi.org/10.1162/106454602753694756.

Brighton, H., & Kirby, S. (2006). Understanding linguistic evolution by visualizing the emergence of topographic mappings. *Artificial life*, *12*(2), 229–242. https://doi.org/https://doi.org/10.1162/artl.2006.12.2.229.

Brighton, H., Smith, K., & Kirby, S. (2005). Language as an evolutionary system. *Physics of Life Reviews*, *2*(3), 177–226. https://doi.org/10.1016/j.plrev.2005.06.001.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of modern physics*, *81*(2), 591–646.

Cavalli-Sforza, L. L., & Feldman, M. W. (1981). *Cultural transmission and evolution: A quantitative approach*. Princeton University Press.

Chaabouni, R., Kharitonov, E., Bouchacourt, D., Dupoux, E., & Baroni, M. (2020, July). Compositionality and generalization in emergent languages. In D. Jurafsky, J. Chai, N. Schluter, &

J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4427–4442). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.407.

Chaabouni, R., Strub, F., Altché, F., Tarassov, E., Tallec, C., Davoodi, E., Mathewson, K. W., Tieleman, O., Lazaridou, A., & Piot, B. (2022). Emergent communication at scale. *International Conference on Learning Representations*. https://openreview.net/forum?id=AUGBfDIV9rL.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22. https://doi.org/10.1016/S1364-6613(02)00005-0.

Cogswell, M., Lu, J., Lee, S., Parikh, D., & Batra, D. (2019). Emergence of compositional language with deep generational transmission. *arXiv preprint arXiv:1904.09067*. https://doi.org/10.48550/arXiv.1904.09067.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in psychology*, *6*, 1964. https://doi.org/10.1016/S1364-6613(02)00005-0.

De Beule, J. (2008). The emergence of compositionality, hierarchy and recursion in peer-to-peer interactions. In *The evolution of language* (pp. 75–82). World Scientific. https://doi.org/10.1142/9789812776129_0010.

De Beule, J., & Bergen, B. K. (2006). On the emergence of compositionality. In *The evolution of language* (pp. 35–42). World Scientific.

De Beule, J., & Steels, L. (2005). Hierarchy in fluid construction grammars. *Annual Conference on Artificial Intelligence*, 1–15. https://doi.org/10.1007/11551263_1.

Eiben, A. E., & Smith, J. E. (2015). *Introduction to evolutionary computing*. Springer.

Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, *5*(1), 17–60.

Erdős, P., & Rényi, A. (1961). On the strength of connectedness of a random graph. *Acta Mathematica Hungarica*, *12*(1), 261–267. https://doi.org/https://doi.org/10.1007/bf02066689.

Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L., & Lakkaraju, K. (2010). Centers and peripheries: Network roles in language change. *Lingua*, *120*(8), 2061–2079. https://doi.org/10.1016/j.lingua.2010.02.001.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive science*, *34*(3), 351–386. https://doi-org.proxy.uba.uva.nl/10.1111/j.1551-6709.2009.01090.x.

Fitch, W. T. (2010). Language from a biological perspective. In *The evolution of language* (pp. 13–34). Cambridge University Press.

Foerster, J., Assael, I. A., de Freitas, N., & Whiteson, S. (2016). Learning to communicate with deep multi-agent reinforcement learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf.

Galantucci, B. (2009). Experimental semiotics: A new approach for studying communication as a form of joint action. *Topics in Cognitive Science*, *1*(2), 393–410. https://doi-org.proxy.uba.uva.nl/10.1111/j.1756-8765.2009.01027.x.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in human neuroscience*, *5*, 11. https://doi.org/10.3389/fnhum.2011.00011.

Galke, L., Ram, Y., & Raviv, L. (n.d.). Emergent communication for understanding human language evolution: What's missing? *Emergent Communication Workshop at ICLR 2022*. https://doi.org/10.48550/arXiv.2204.10590.

Gong, T. (2010). Exploring the roles of horizontal, vertical, and oblique transmissions in language evolution. *Adaptive Behavior*, *18*(3-4), 356–376. https://doi.org/10.1177/1059712310377241.

Gong, T., Baronchelli, A., Puglisi, A., & Loreto, V. (2011). Exploring the roles of complex networks in linguistic categorization. *Artificial life*, *18*(1), 107–121. https://doi.org/10.1162/ARTL_a_00051.

Gong, T., Minett, J. W., & Wang, W. S.-Y. (2008). Exploring social structure effect on language evolution based on a computational model. *Connection Science*, *20*(2-3), 135–153. https://doi.org/10.1080/09540090802091941.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive science*, *31*(3), 441–480. https://doi.org/10.1080/15326900701326576.

Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., Goss-Custard, J., Grand, T., Heinz, S. K., Huse, G., et al. (2006). A standard protocol for describing individual-based and agent-based models. *Ecological modelling*, *198*(1-2), 115–126. https://doi.org/10.1016/j.ecolmodel.2006.04.023.

Grimm, V., Berger, U., DeAngelis, D. L., Polhill, J. G., Giske, J., & Railsback, S. F. (2010). The odd protocol: A review and first update. *Ecological modelling*, *221*(23), 2760–2768. https://doi.org/10.1016/j.ecolmodel.2010.08.019.

Group", T. ". G., Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., Holland, J., Ke, J., Larsen-Freeman, D., & Schoenemann, T. (2009). Language is a complex adaptive system: Position paper. *Language Learning*, *59*(s1), 1–26. https://doi.org/https://doi.org/10.1111/j.1467-9922.2009.00533.x.

Guo, S., Ren, Y., Havrylov, S., Frank, S., Titov, I., & Smith, K. (2020). The emergence of compositional languages for numeric concepts through iterated learning in neural agents. *The Evolution of Language: Proceedings of the 13th International Conference (Evolang13)*, 131. https://doi.org/10.48550/arXiv.1910.05291.

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference* (pp. 11–15). https://www.osti.gov/biblio/960616.

Herman, J., & Usher, W. (2017). SALib: An open-source python library for sensitivity analysis. *The Journal of Open Source Software*, *2*(9). https://doi.org/10.21105/joss.00097.

Iwanaga, T., Usher, W., & Herman, J. (2022). Toward SALib 2.0: Advancing the accessibility and interpretability of global sensitivity analyses. *Socio-Environmental Systems Modelling*, *4*, 18155. https://doi.org/10.18174/sesmo.18155.

Josserand, M., Allassonnière-Tang, M., Pellegrino, F., & Dediu, D. (2021). Interindividual variation refuses to go away: A Bayesian computer model of language change in communicative networks. *Frontiers in Psychology*, *12*, 626118. https://doi.org/10.3389/fpsyg.2021.626118.

Josserand, M., Allassonnière-Tang, M., Pellegrino, F., Dediu, D., & de Boer, B. (2024). How network structure shapes languages: Disentangling the factors driving variation in communicative agents. *Cognitive Science*, *48*(4), e13439. https://doi-org.proxy.uba.uva.nl/10.1111/cogs.13439.

Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, *336*(6084), 1049–1054. https://doi.org/https://doi.org/10.1126/science.1218811.

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation*, *5*(2), 102–110. https://doi.org/10.1109/4235.918430.

Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, & J. R. Hurford (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 303–323). Cambridge University Press.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. *Linguistic evolution through language acquisition: Formal and computational models*, 173–204.

Kirby, S. (2017). Culture and biology in the origins of linguistic structure. *Psychonomic bulletin & review*, *24*, 118–137. https://doi.org/10.3758/s13423-016-1166-7.

Kirby, S., & Christiansen, M. H. (2003). Language evolution: The hardest problem in science? In *Language evolution*. Oxford University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, *105*(31), 10681–10686. https://doi.org/10.1073/pnas.0707835105.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245. https://doi.org/10.1073/pnas.0608222104.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114. https://doi.org/https://doi.org/10.1016/j.conb.2014.07.014.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language* (pp. 121–147). Springer London. https://doi.org/10.1007/978-1-4471-0663-0_6.

Kirby, S., & Tamariz, M. (2021). Cumulative cultural evolution, population structure and the origin of combinatoriality in human language. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1843), 20200319. https://doi.org/10.1098/rstb.2020.0319.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102. https://doi.org/10.1016/j.cognition.2015.03.016.

Korbak, T., Zubek, J., & Raczaszek-Leonardi, J. (2020). Measuring non-trivial compositionality in emergent communication. *arXiv preprint arXiv:2010.15058*. https://doi.org/10.48550/arXiv.2010.15058.

Kottur, S., Moura, J., Lee, S., & Batra, D. (2017, September). Natural language does not emerge 'naturally' in multi-agent dialog. In M. Palmer, R. Hwa, & S. Riedel (Eds.), *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2962–2967). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1321.

Kuciński, Ł., Korbak, T., Kołodziej, P., & Miłoś, P. (2021). Catalytic role of noise and necessity of inductive biases in the emergence of compositional communication. *Advances in neural* *information processing systems*, *34*, 23075–23088. https://proceedings.neurips.cc/paper_files/paper/2021/file/c2839bed26321da8b466c80a032e4714-Paper.pdf.

Lazaridou, A., & Baroni, M. (2020). Emergent multi-agent communication in the deep learning era. *arXiv preprint arXiv:2006.02419*. https://doi.org/10.48550/arXiv.2006.02419.

Lazaridou, A., Hermann, K. M., Tuyls, K., & Clark, S. (2018). Emergence of linguistic communication from referential games with symbolic and pixel input. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1804.03984.

Lewis, D. K. (1969). *Convention: A philosophical study*. Wiley-Blackwell.

Lin, J. (2002). Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115.

Lou-Magnuson, M., & Onnis, L. (2018). Social network limits language complexity. *Cognitive Science*, *42*(8), 2790–2817. https://doi.org/https://doi.org/10.1111/cogs.12683.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, *30*. https://proceedings.neurips.cc/paper_files/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf.

Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one*, *5*(1), e8559. https://doi.org/10.1371/journal.pone.0008559.

Meir, I., Israel, A., Sandler, W., Padden, C. A., & Aronoff, M. (2012). The influence of community on language structure: Evidence from two young sign languages. *Linguistic Variation*, *12*(2), 247–291. https://doi.org/10.1075/lv.12.2.04mei.

Motamedi, Y., Schouwstra, M., Smith, K., Culbertson, J., & Kirby, S. (2019). Evolving artificial sign languages in the lab: From improvised gesture to systematic sign. *Cognition*, *192*, 103964. https://doi.org/10.1016/j.cognition.2019.05.001.

Mudd, K. (2022). *How social structure affects the persistence and features of shared sign languages* [Doctoral dissertation, Vrije Universiteit Brussel].

Mudd, K., de Vos, C., & De Boer, B. (2020). The effect of cultural transmission on shared sign language persistence. *Palgrave Communications*, *6*(1), 1–11. https://doi.org/10.1057/s41599-020-0479-3.

Nakamura, M., Matoba, R., & Tojo, S. (2015). Simulation of emergence of local common languages using iterated learning model on social networks. *International Journal on Advances in Intelligent Systems*, *8*(3&4), 374–384.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1597), 1829–1836. https://doi.org/10.1098/rstb.2011.0216.

Newport, E. L. (2020). Children and adults as language learners: Rules, variation, and maturational change. *Topics in cognitive science*, *12*(1), 153–169. https://doi.org/https://doi.org/10.1111/tops.12416.

Partee, B., et al. (1995). Lexical semantics and compositionality. *An invitation to cognitive science: Language*, *1*, 311–360.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001.

Peter, T., & Sergey, G. (2009). Evolution of complex hierarchical societies. *Social Evolution & History*, *8*(2), 167–198. https://cyberleninka.ru/article/n/evolution-of-complex-hierarchical-societies.

Peters, J., Waubert de Puiseau, C., Tercan, H., Gopikrishnan, A., Lucas de Carvalho, G. A., Bitter, C., & Meisen, T. (2025). Emergent language: A survey and taxonomy. *Autonomous Agents and Multi-Agent Systems*, *39*(1), 1–73. https://doi.org/10.1007/s10458-025-09691-y.

Pickel, B., & Szabó, Z. G. (2025). Compositionality. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2025). Metaphysics Research Lab, Stanford University.

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and brain sciences*, *13*(4), 707–727. https://doi.org/10.1017/S0140525X00081061.

Ravasz, E., & Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical review E*, *67*(2), 026112. https://doi.org/10.1103/PhysRevE.67.026112.

Raviv, L., & Boeckx, C. (2025, May). Introduction. In *The Oxford Handbook of Approaches to Language Evolution*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780192886491.002.0008.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151–164. https://doi.org/10.1016/j.cognition.2018.09.010.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B*, *286*(1907), 20191262. https://doi.org/10.1098/rspb.2019.1262.

Raviv, L., Meyer, A., & Lev-Ari, S. (2020). The role of social network structure in the emergence of linguistic structure. *Cognitive Science*, *44*(8), e12876. https://doi-org.proxy.uba.uva.nl/10.1111/cogs.12876.

Reali, F., Chater, N., & Christiansen, M. H. (2018). Simpler grammar, larger vocabulary: How population size affects language. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1871), 20172586. https://doi.org/10.1098/rspb.2017.2586.

Ren, Y., Guo, S., Labeau, M., Cohen, S. B., & Kirby, S. (2020). Compositional languages emerge in a neural iterated learning model. *8th International Conference on Learning Representations*. https://hal.science/hal-02914840.

Resnick, C., Gupta, A., Foerster, J., Dai, A. M., & Cho, K. (2020). Capacity, bandwidth, and compositionality in emergent language learning. *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 1125–1133. https://doi.org/10.48550/arXiv.1910.11424.

Rita, M., Michel, P., Chaabouni, R., Pietquin, O., Dupoux, E., & Strub, F. (2024). Language evolution with deep learning. *arXiv preprint arXiv:2403.11958*. https://doi.org/10.48550/arXiv.2403.11958.

Rita, M., Strub, F., Grill, J.-B., Pietquin, O., & Dupoux, E. (2021). On the role of population heterogeneity in emergent communication. *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.2204.12982.

Roberts, G., Lewandowski, J., & Galantucci, B. (2015). How communication changes when we cannot mime the world: Experimental evidence for the effect of iconicity on combinatoriality. *Cognition*, *141*, 52–66. https://doi.org/10.1016/j.cognition.2015.04.001.

Saltelli, A. (2002). Making best use of model evaluations to compute sensitivity indices. *Computer physics communications*, *145*(2), 280–297. https://doi.org/10.1016/S0010-4655(02)00280-1.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., & Tarantola, S. (2010). Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer physics communications*, *181*(2), 259–270. https://doi.org/10.1016/j.cpc.2009.09.018.

Sandler, W., Dachkovsky, S., Stamp, R., Raviv, L., & Boeckx, C. (2025). The evolution of the grammar of the body in sign languages. In *The oxford handbook of approaches to language evolution*. Oxford University Press.

Schläpfer, M., Bettencourt, L. M., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G. B., & Ratti, C. (2014). The scaling of human interactions with city size. *Journal of the Royal Society Interface*, *11*(98), 20130789. https://doi.org/10.1098/rsif.2013.0789.

Shoham, Y., & Leyton-Brown, K. (2008). *Multiagent systems: Algorithmic, game-theoretic, and logical foundations*. Cambridge University Press.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In *Computational social psychology* (pp. 311–331). Routledge.

Smith, K. (2009). Iterated learning in populations of bayesian agents. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *31*(31). https://escholarship.org/uc/item/6wh9d141.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial Life*, *9*(4), 371–386. https://doi.org/10.1162/106454603322694825.

Sobol, I. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates [The Second IMACS Seminar on Monte Carlo Methods]. *Mathematics and Computers in Simulation*, *55*(1), 271–280. https://doi.org/https://doi.org/10.1016/S0378-4754(00)00270-6.

Steels, L. (1995). A self-organizing spatial vocabulary. *Artificial life*, *2*(3), 319–332. https://doi.org/10.1162/artl.1995.2.3.319.

Steels, L. (1997). Self-organizing vocabularies. *Artificial Life V*, 179–184.

Steels, L. (2011a). Design patterns in fluid construction grammar. In *332 pp, amsterdam: John benjamins, 2011*. John Benjamins Publishing Company.

Steels, L. (2011b). Modeling the cultural evolution of language. *Physics of life reviews*, *8*(4), 339–356. https://doi.org/10.1016/j.plrev.2011.10.014.

Steels, L. (2012). *Experiments in cultural language evolution* (1st ed.). John Benjamins Pub. Co.

Steels, L. (2017). Basics of fluid construction grammar. *Constructions and frames*, *9*(2), 178–225. https://doi.org/10.1075/cf.00002.ste.

Steels, L., Belpaeme, T., et al. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, *28*(4), 469–488. https://doi.org/10.1017/S0140525X05430083.

Tamariz, M., & Kirby, S. (2016). The cultural evolution of language. *Current opinion in psychology*, *8*, 37–43. https://doi.org/10.1016/j.copsyc.2015.09.003.

Theisen, C. A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction studies*, *11*(1), 14–32. https://doi.org/10.1075/is.11.1.08the.

Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. *Proceedings of the annual meeting of the cognitive science society*, *33*(33).

Van Eecke, P., Steels, L., & Beuls, K. (2018). Generalisation and specialisation operators for computational construction grammar and their application in evolutionary linguistics research.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and cognition*, *4*(4), 357–380. https://doi.org/10.1515/langcog-2012-0019.

Verhoef, T., Kirby, S., & De Boer, B. (2014). Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. *Journal of Phonetics*, *43*, 57–68. https://doi.org/10.1016/j.wocn.2014.02.005.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Vogt, P. (2005a). The emergence of compositional structures in perceptually grounded language games. *Artificial intelligence*, *167*(1-2), 206–242. https://doi.org/10.1016/j.artint.2005.04.010.

Vogt, P. (2005b). On the acquisition and evolution of compositional languages: Sparse input and the productive creativity of children. *Adaptive Behavior*, *13*(4), 325–346. https://doi.org/10.1177/105971230501300403.

Vogt, P. (2007). Group size effects on the emergence of compositional structures in language. *European Conference on Artificial Life*, 405–414. https://doi.org/10.1007/978-3-540-74913-4_41.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918.

Woensdregt, M. S., Spike, M., de Haan, R., Wareham, T., van Rooij, I., & Blokpoel, M. (2021). Why is scaling up models of language evolution hard? *Proceedings of the annual meeting of the cognitive science society*, *43*(43). https://escholarship.org/uc/item/021734q4.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language & communication*, *18*(1), 47–67. https://doi.org/10.1016/S0271-5309(97)00033-5.

Wray, A., & Grace, G. W. (2007). The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form. *Lingua*, *117*(3), 543–578. https://doi.org/10.1016/j.lingua.2005.05.005.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient human-like semantic representations via the information bottleneck principle. *arXiv preprint arXiv:1808.03353*. https://doi.org/10.48550/arXiv.1808.03353.

Zubek, J., Denkiewicz, M., Barański, J., Wróblewski, P., Raczaszek-Leonardi, J., & Plewczynski, D. (2017). Social adaptation in multi-agent model of linguistic categorization is affected by network information flow. *PloS one*, *12*(8), e0182490. https://doi.org/10.1371/journal.pone.0182490.

Appendix A – Sensitivity Analysis: Horizontal & Vertical Transmission

To gain more insight into how *horizontal transmission, vertical transmission* each impact the emergence of compositionality, I performed a Sobol Sensitivity analysis (Saltelli, 2002; Saltelli et al., 2010; Sobol, 2001). Sobol sensitivity analysis is a variance-based method to quantify how much each parameter, as well as the interaction between parameters, contributes to the variability in the model's output. It calculates first-order effects (influence of individual parameters), higher-order effects (interactions between parameters), and total effects (combined impact including interactions). More specifically, first-order effects look at the variance of the output of the model, in this case proportion of the posterior focused on compositional languages, when only parameter X_i is varied. Total order effects capture the overall contribution of parameter X_i, including its interaction with other parameters. Second-order effect measures the output variance caused by the interaction between parameters X_i and X_j.

To run a Sobol sensitivity analysis, the model must be evaluated across a wide range of parameter settings. More specifically, I used the Saltelli sampling technique, as provided in the SALib package (Herman & Usher, 2017; Iwanaga et al., 2022). This is an extension of the standard Sobol scheme, which is a quasi-random, low-discrepancy sequence used to generate a uniform sample of a parameter space.

For possible parameter values, I chose $h \in [1, 50)$ and $v \in [0, 20)$, as the results from Experiment 1 suggested that vertical interactions have more impact on the emergence of compositionality than horizontal interactions. By increasing the range for horizontal interactions, I made sure that the bigger impact of vertical interactions was not caused by too little horizontal interactions.

To obtain the final parameter values, one has to choose a value for N, which is the base sample size. As N was set to 512, and the possibility for calculating the second order effects was set to true, the final number of parameter values for which the model was evaluated was $N \cdot (2D + 2)$, where D is the number of parameters considered. This results in 3072 different parameter settings.

Since the model is stochastic, I run each parameter setting five times and average the results to reduce noise. Population size was set to 25, one agent was replaced per iteration, the mode was set to sample, and the model ran for 10000 iterations.

Then, the Sobol analysis was performed using the function provided by the SALib package (Herman & Usher, 2017; Iwanaga et al., 2022).

A.0.1 Results

As shown in Figure A.1, in the first few iterations of the model, h, thus the number of horizontal interaction, has the biggest impact on the variance in the proportion of compositionality in the distribution. Over time, however, v explains all the variance in the output of the model, as shown by the first-order and total-order effects. As shown by the second-order effects, is the interaction between h and v relevant in the beginning of the simulation, but this diminishes over time. Overall, we can conclude that the final output of the simulations is driven by the amount of vertical interactions.

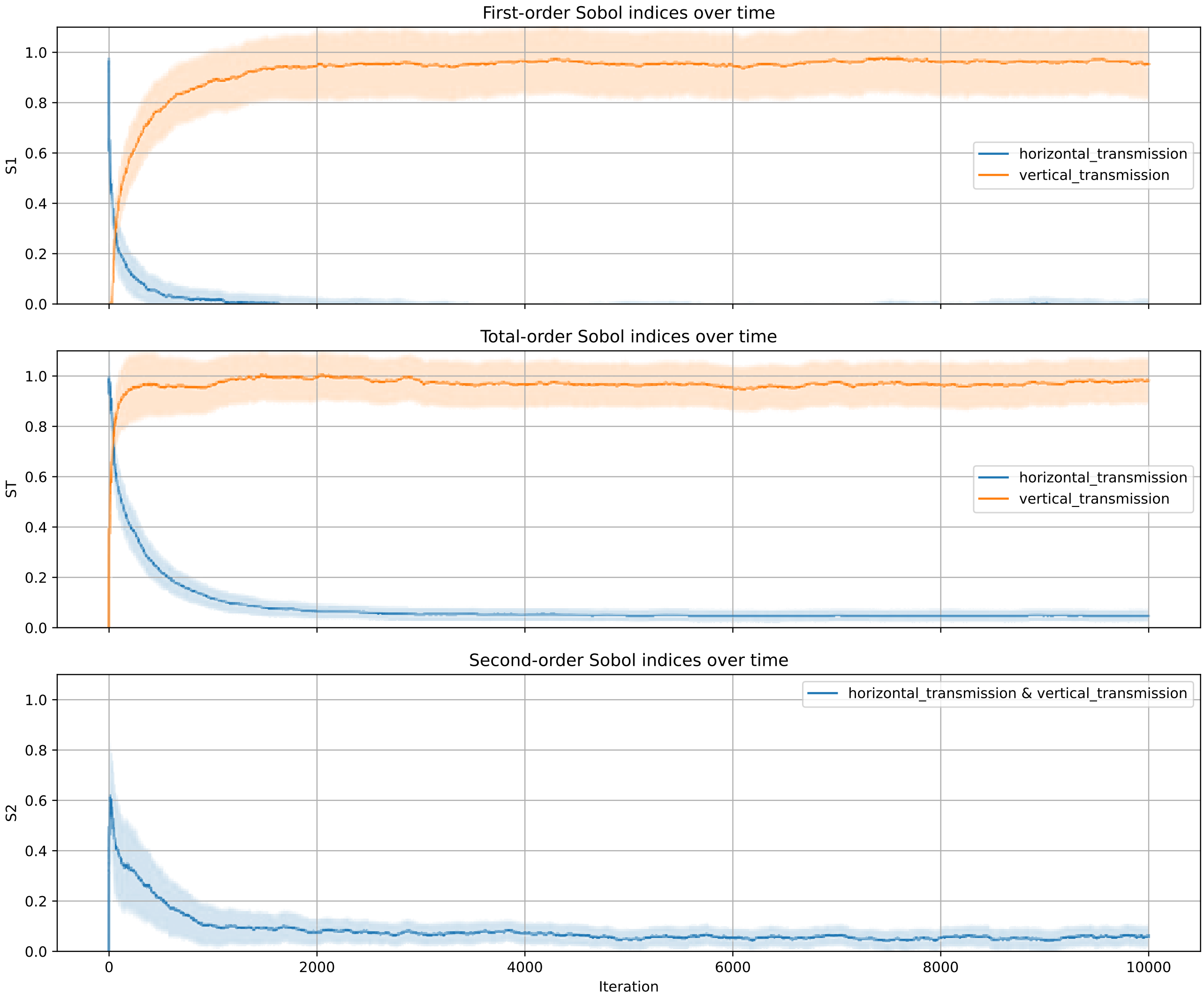


Figure A.1: First, total and second order effects of a Sobol sensitivity analysis of the parameters h and v.

Appendix B – Additional Figures Experiments Cultural Transmission

B.1 Transmission Modes

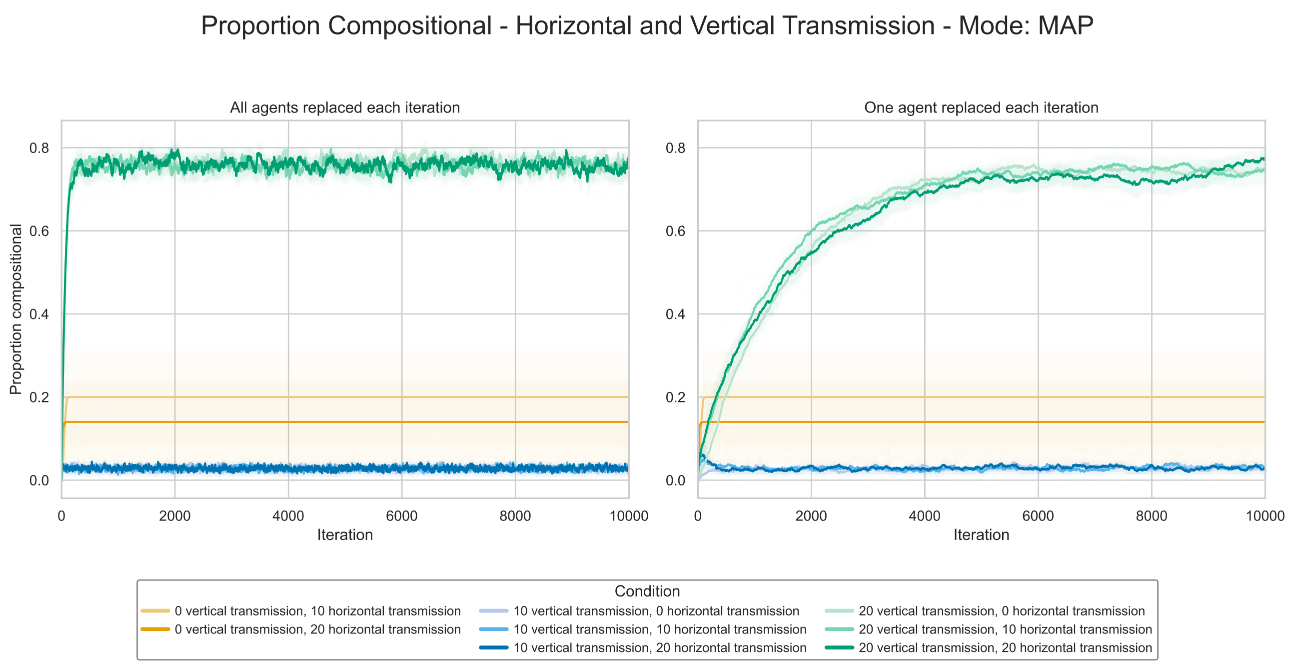


Figure B.1: Posterior probability for compositional languages for MAP-learners, averaged over 50 runs. Shaded area represents 95% confidence intervals.

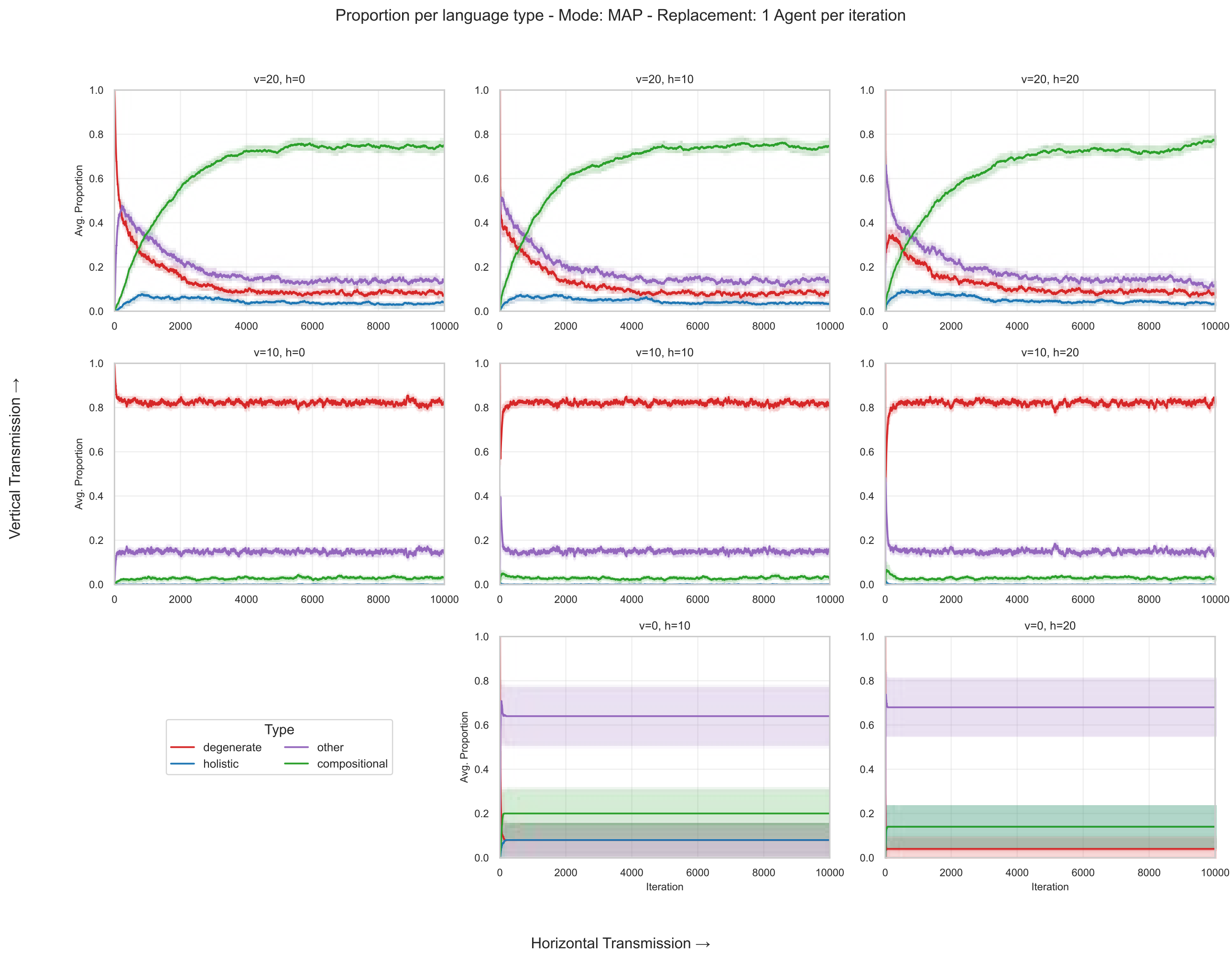


Figure B.2: Proportion of posterior per language type, for varying levels of horizontal and vertical transmission. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced (if there is vertical transmission). The shaded area represents 95% confidence intervals.

B.2 Population Size

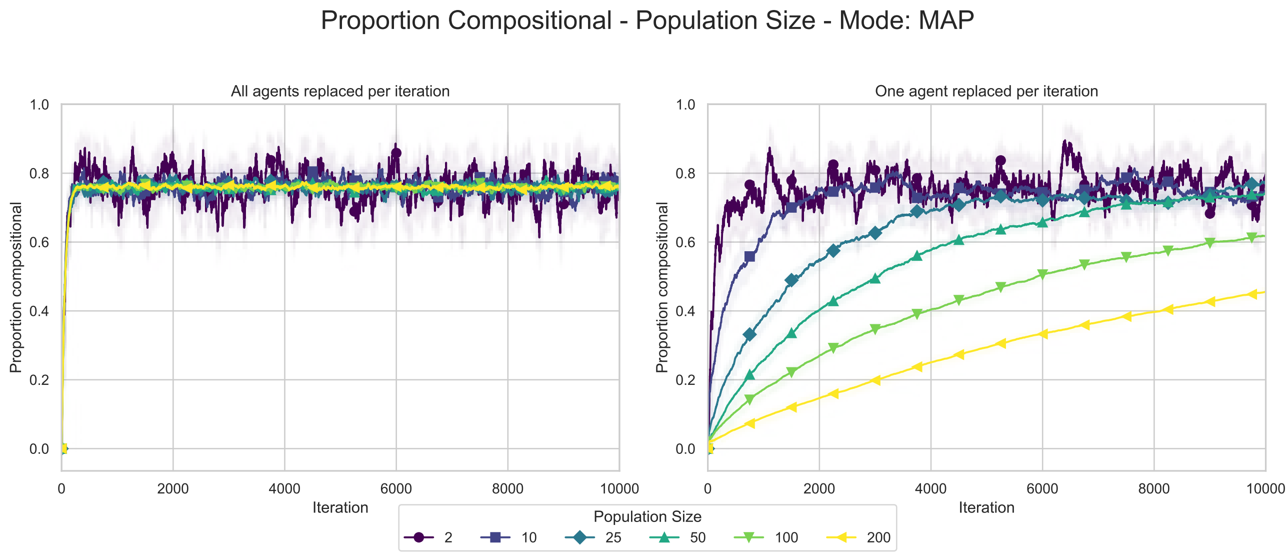


Figure B.3: Posterior probability for compositional languages averaged for the entire population of MAP-learners, averaged over 50 runs, for different population sizes. The shaded area represents the 95% confidence interval. In the left figure, every agent is replaced each iteration, and in the figure on the right, only one agent is replaced each iteration.

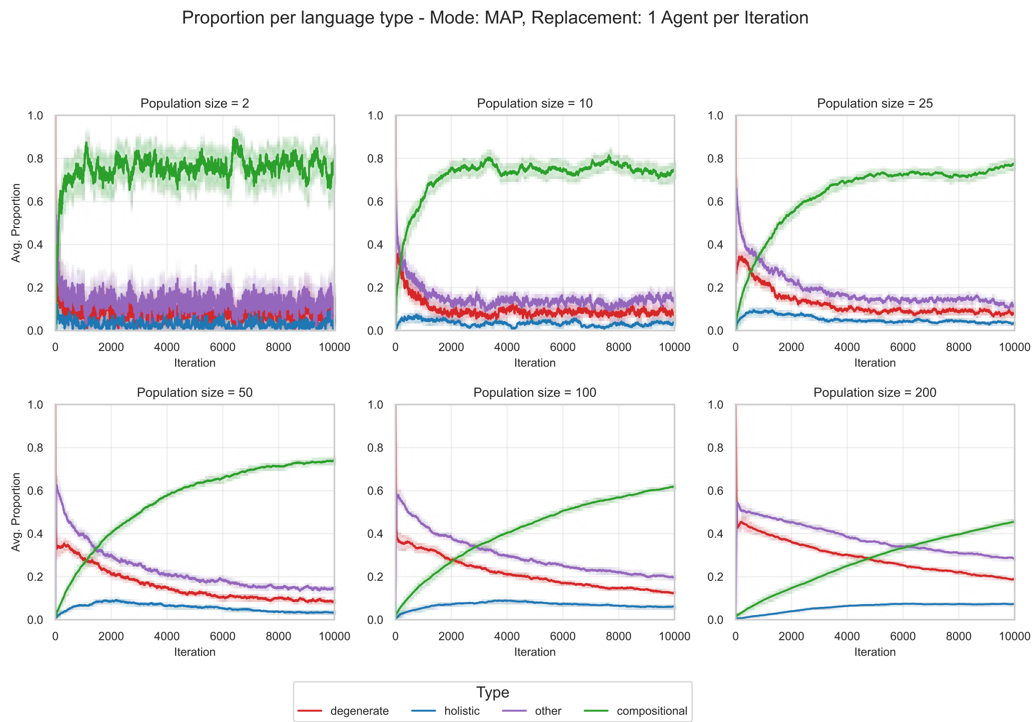


Figure B.4: Proportion of posterior per language type, for various population sizes. Values are averaged for the entire population of MAP-learners and averaged over 50 runs. Per iteration, the one agent is replaced. The shaded area represents 95% confidence intervals.

B.3 Social Network Structures

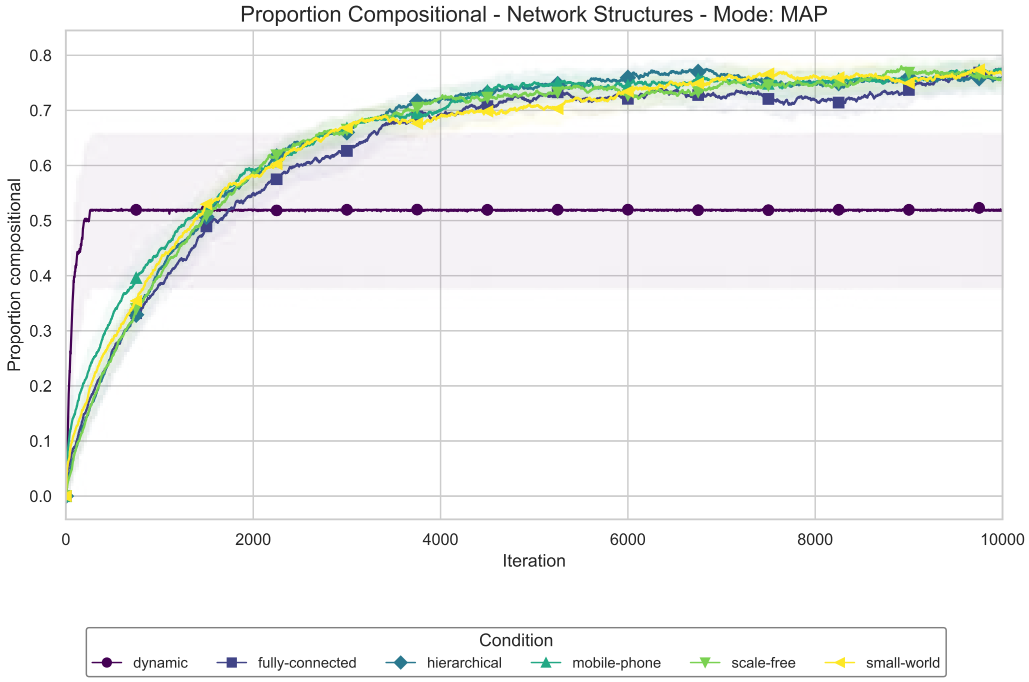


Figure B.5: Posterior probability for compositional languages averaged for the entire population of MAP-learners, averaged over 50 runs, for different network types. The population size is 25. Vertical and horizontal transmission rounds = 20. Shaded area represents 95% confidence intervals.

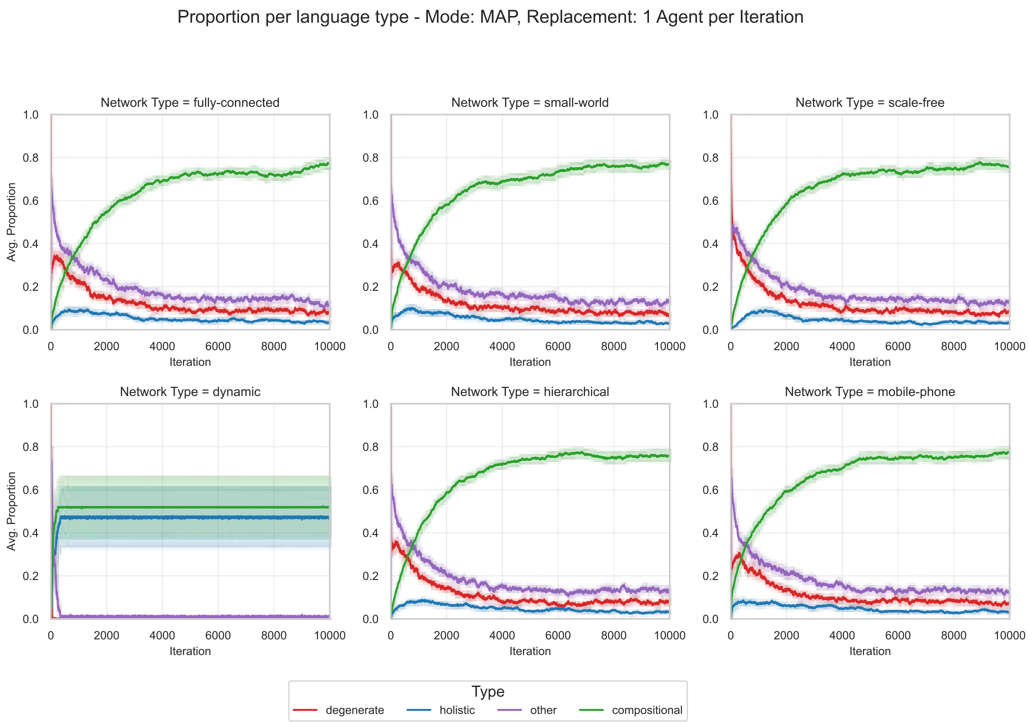


Figure B.6: Proportion of posterior per language type, for various network types. Values are averaged for the entire population of MAP-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

B.4 Rate of Replacement

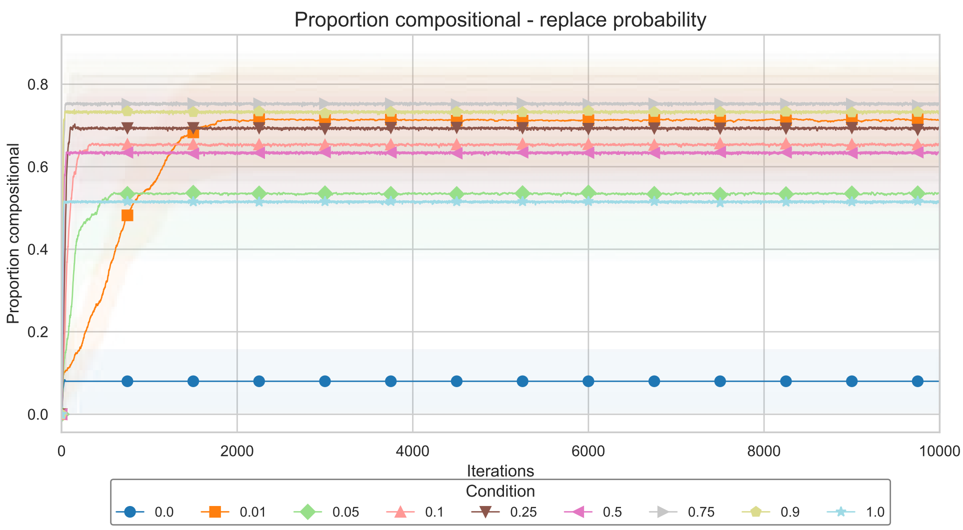


Figure B.7: Posterior probability of compositional languages across 9 replace probabilities, averaged over 50 runs. Simulations were run for 10000 iterations. Population size = 25.

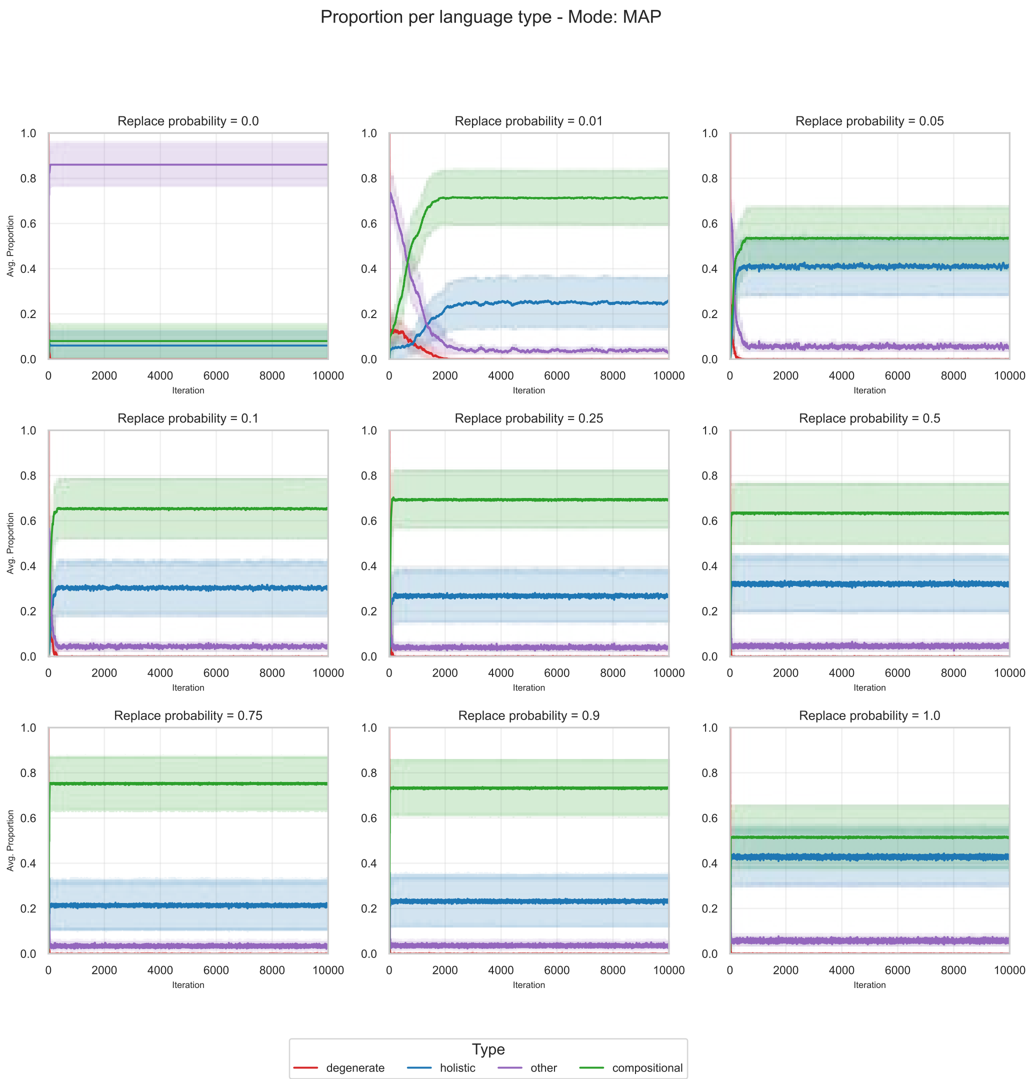


Figure B.8: Proportion of posterior per language type, for various replacement probabilities. Values are averaged across the entire population of MAP-learners and averaged over 50 runs. The shaded area represents 95% confidence intervals.

Appendix C – Additional Figures Experiments Individual Learning Mechanisms

C.1 Regulating the Impact of New Evidence

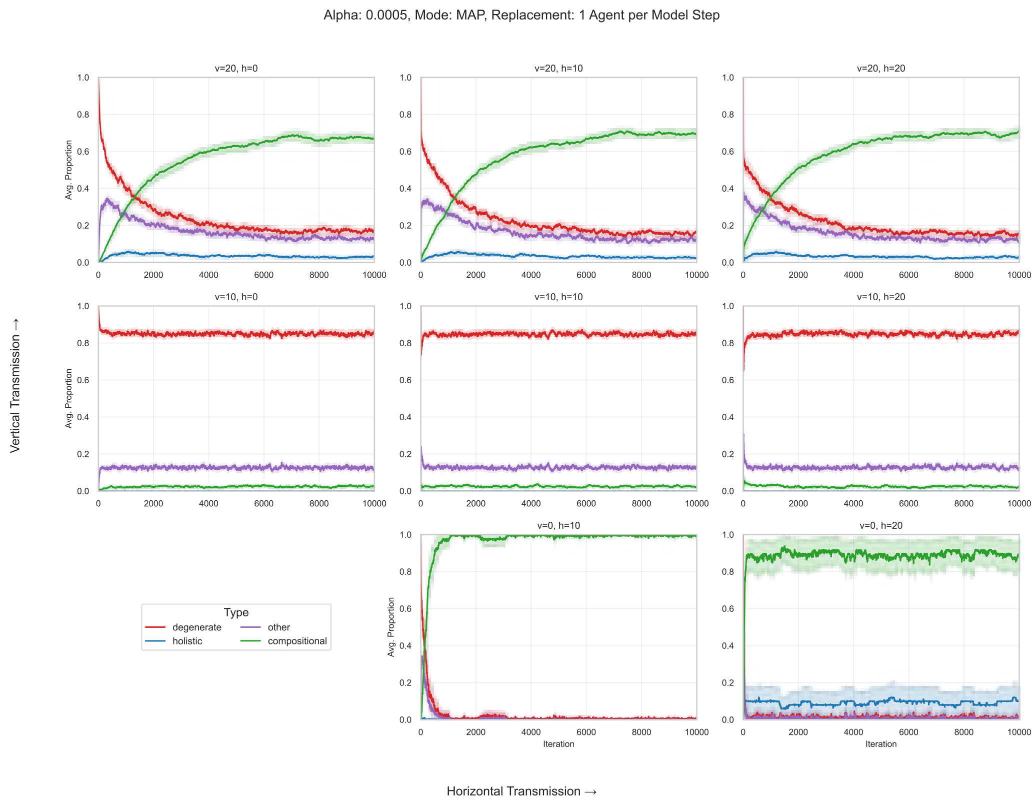


Figure C.1: Proportion of posterior per language type, for various levels of horizontal and vertical transmission. Values are averaged for the entire population of MAP-learners and averaged over 50 runs. Per iteration, one agent is replaced. $\alpha = 0.0005$. The shaded area represents 95% confidence intervals.

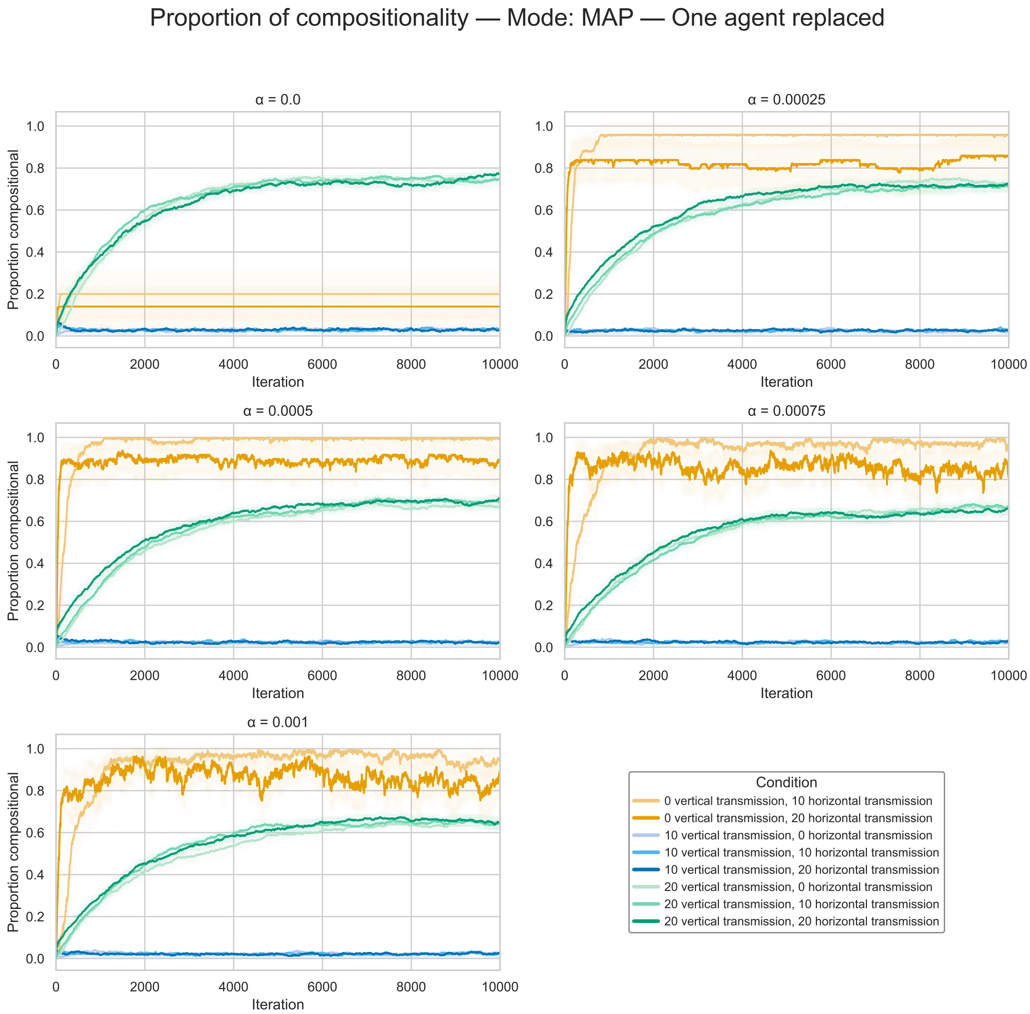


Figure C.2: Proportion of posterior for compositional languages averaged for the entire population of MAP-learners, averaged over 50 runs, for various values of α. The shaded area represents 95% confidence intervals. If there is vertical transmission, one agent is replaced per iteration.

Appendix D – Additional Figures Experiments Cultural Transmission Revisted

D.1 Population Size Effects in a Heterogeneous Population

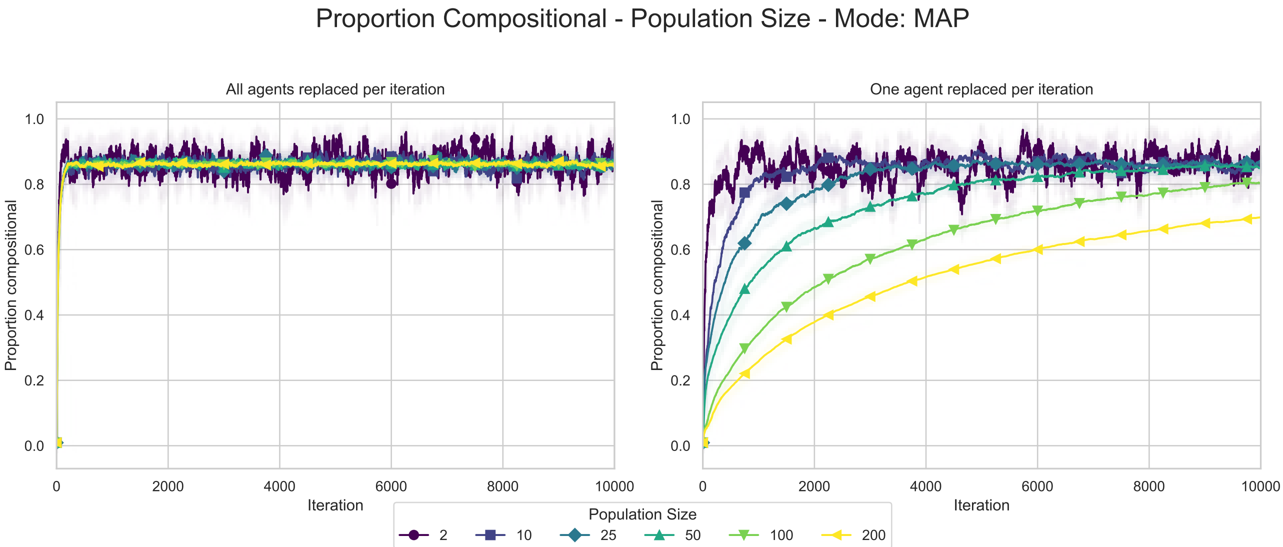


Figure D.1: Posterior probability for compositional languages averaged for the entire population of MAP-learners, averaged over 50 runs, for different population sizes. The shaded area represents the 95% confidence interval. In the left figure, every agent is replaced each iteration, and in the figure on the right, only one agent is replaced each iteration.

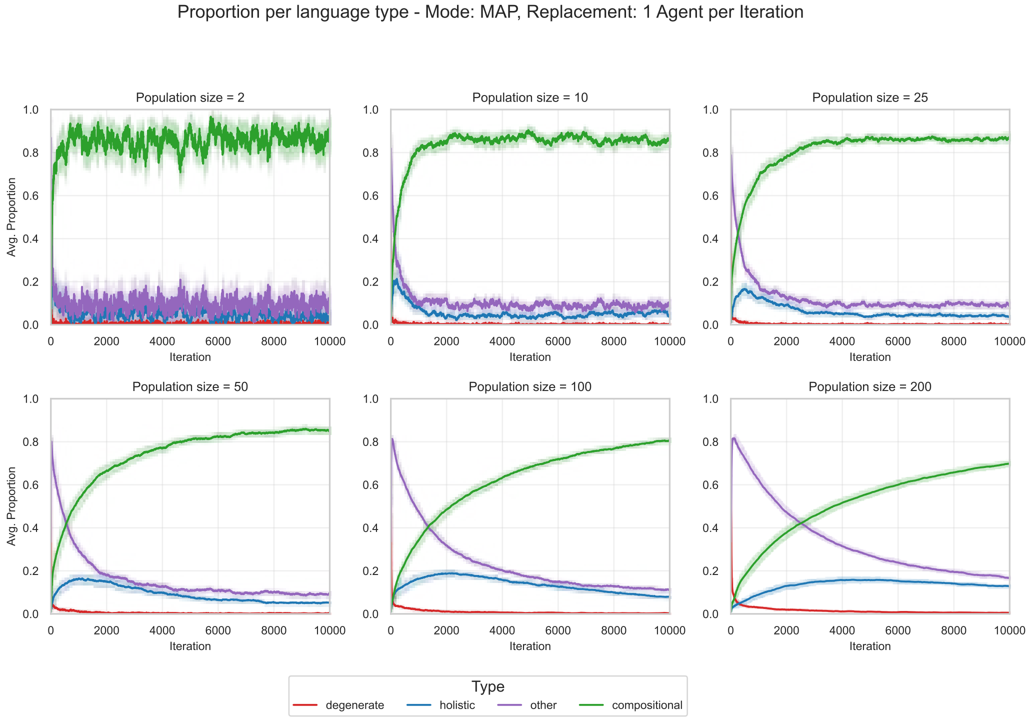


Figure D.2: Proportion of posterior per language type, for various population sizes. Values are averaged for the entire population of MAP-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

D.2 Social Network Structures Revisited

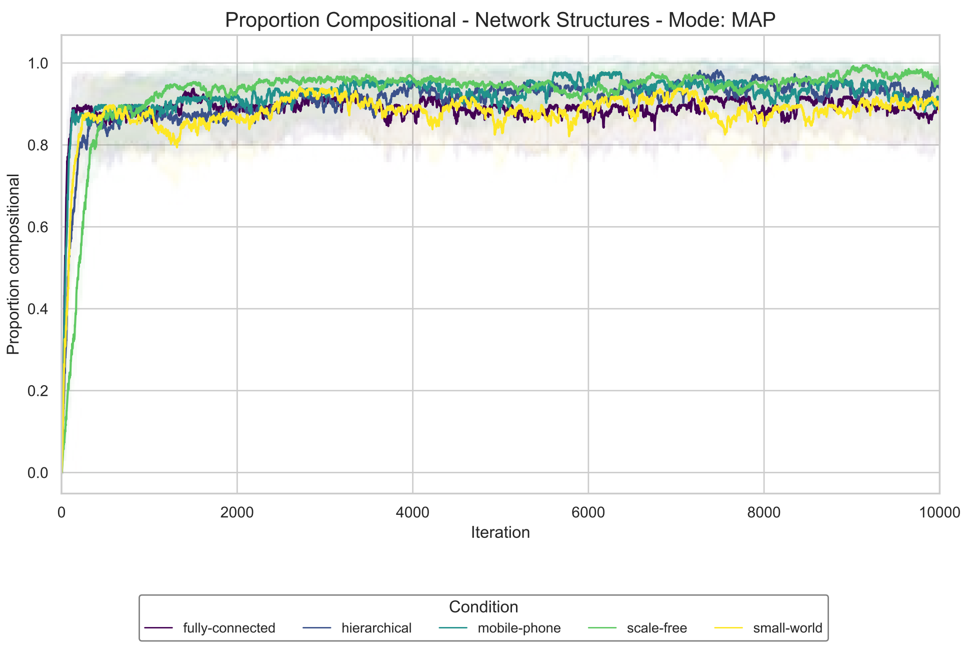


Figure D.3: Posterior probability for compositional languages averaged for the entire population of MAP-learners, averaged over 50 runs, for different network types. The population size is 25. Vertical and horizontal transmission rounds = 20. Shaded area represents 95% confidence intervals.

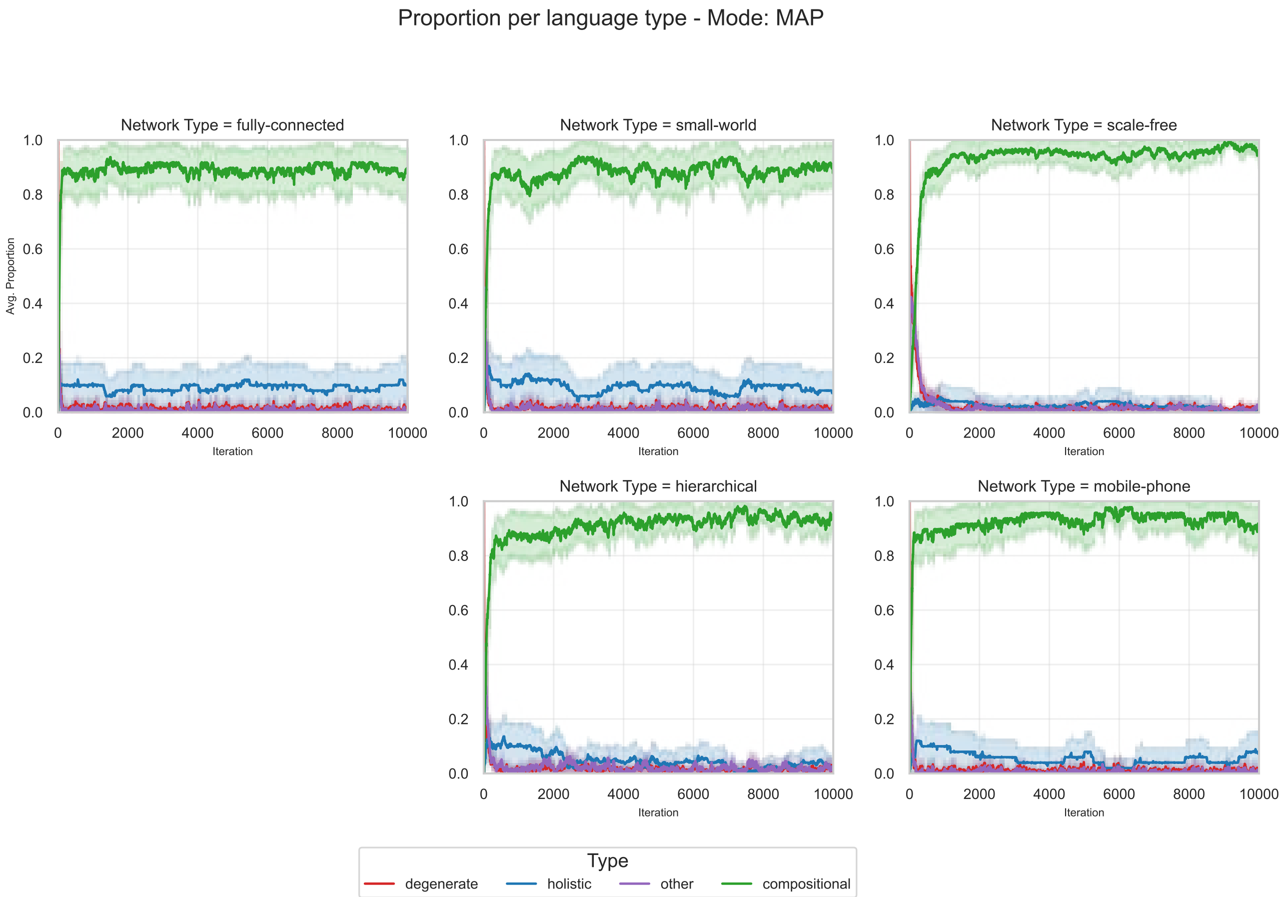


Figure D.4: Proportion of posterior per language type, for various network types. Values are averaged for the entire population of MAP-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

Appendix E – Explorative Experiments

Experiment 9: Holistic starting language

For the experiments previously reported, the very first generation of agents starts producing signals based on the language they sampled from their prior distribution. Since degenerate languages are highly favored in the prior, their starting hypothesis is most likely degenerate. Because we are dealing with 'pragmatic' agents (i.e., the agent selects their signal based on whether communication is successful or not), the chances are $\frac{1}{4} - \epsilon$ that the agent generates a signal from the degenerate language, and $\frac{3}{4} + \epsilon$ that the agent generates an arbitrary signal. So essentially, the initial signals used by the first generation of agents are quite random. It is quite unlikely, however, that language started out to be this random. Previously, it has been suggested that (proto)languages might have started holistically, similarly to non-human animal communication systems (Wray, 1998). This experiment aims to explore what happens to the emergence of compositionality in this model if the first generation of agents learn from a holistic starting language.

E.1.1 Experimental Setup

Similar to Kirby et al. (2015), the very first generation of agents learns for a set number of learning rounds from a randomly selected holistic language. An example of how the prior of a single agent is distributed across the different language types, after updating for $i = 5, i = 10$ or $i = 20$ learning rounds from a randomly selected holistic language, can be seen in Figure E.9.1. Note that in these settings, the initial prior of the agents remains unchanged, i.e., new agents still have a prior preference for highly compressible languages. Horizontal and vertical transmission were assigned values of 20, and the population size was 25. Simulations were run for MAP-learners and sample learners, for 10000 iterations.

E.1.2 Results

As shown in Figure E.9.2, the initial learning rounds, regardless of the amount, lead to a slower emergence of a preference for compositionality, in the setting where one agent is replaced per iteration. When the entire population is replaced, there is no observable difference between the conditions. When one agent is replaced per iteration, eventually all conditions converge to a similar level of compositionality, of around 75%. In Figure E.9.3, at the very beginning of the simulation, we see a

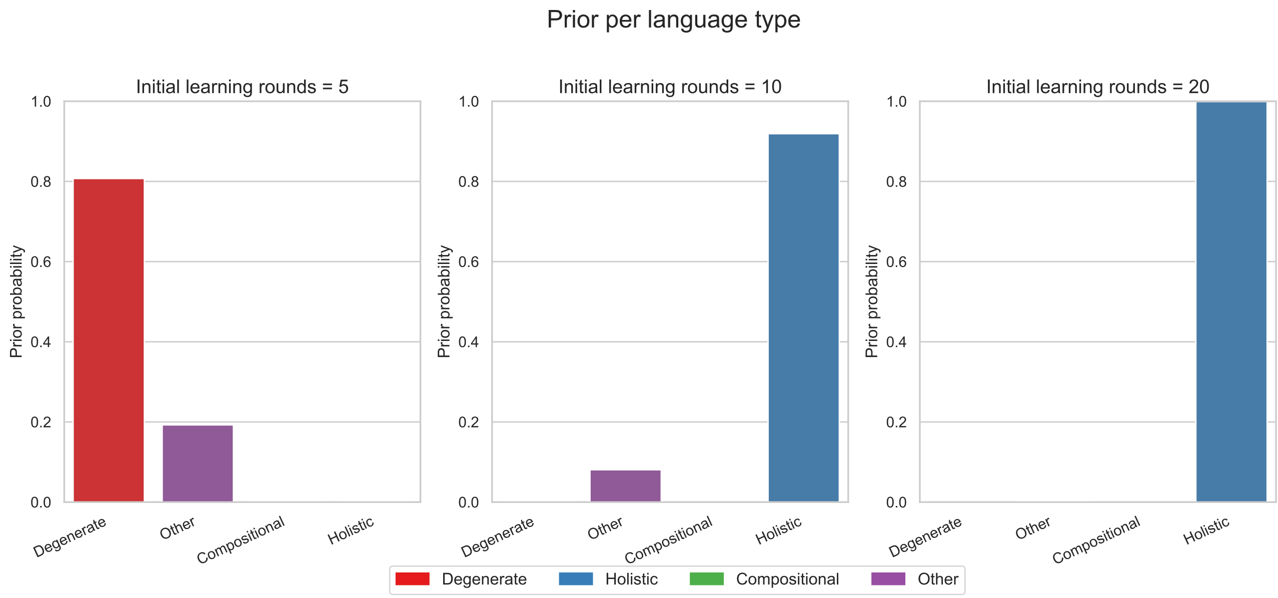


Figure E.9.1: Prior probability across language types after learning for varying numbers of initial rounds.

peak in the proportion of holistic languages, where the height of the peak is related to the number of initial learning rounds.

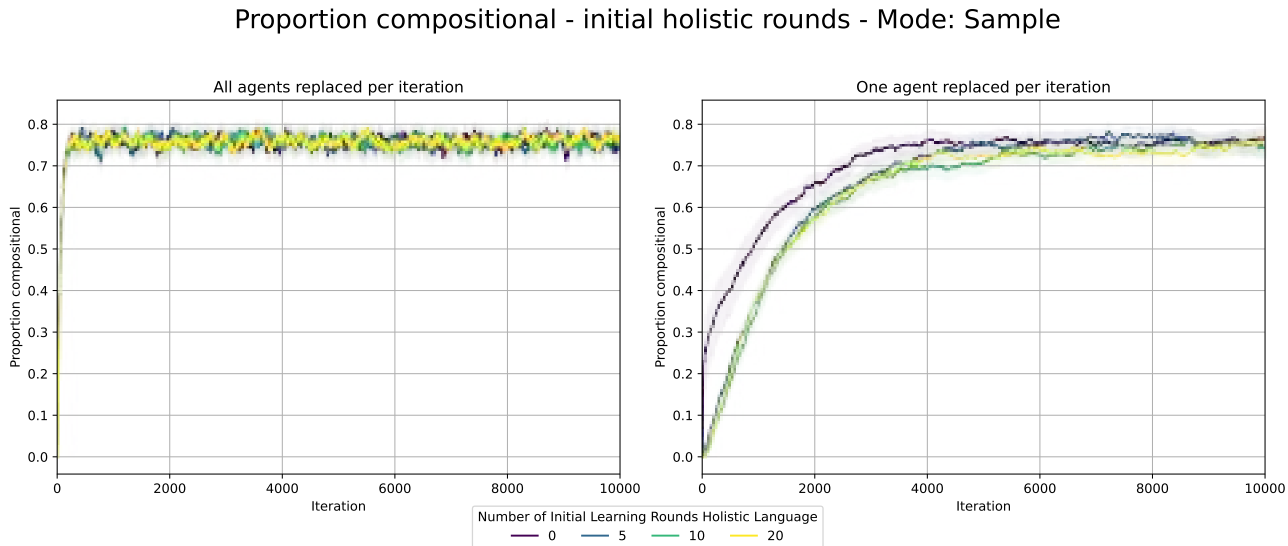


Figure E.9.2: Proportion compositionality in posterior distribution for various amounts of initial learning rounds from a holistic starting language. Agents are sample-learners. Population size = 25, horizontal transmission rounds = 20, vertical transmission = 20.

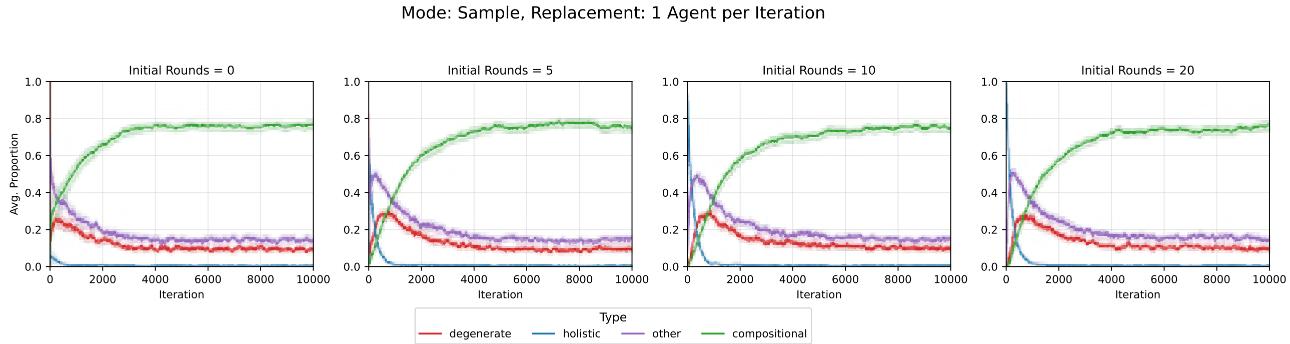


Figure E.9.3: Proportion of posterior per language type, for various amounts of initial learning rounds from a holistic starting language. Agents are sample-learners. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

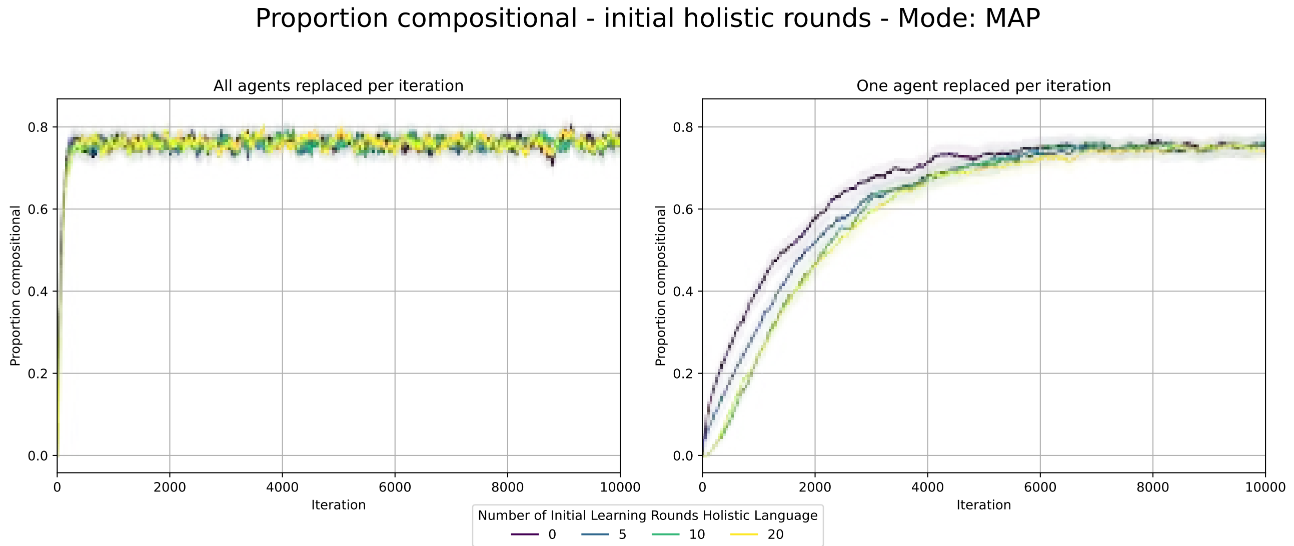


Figure E.9.4: Proportion compositionality in posterior distribution for various amounts of initial learning rounds from a holistic starting language. Agents are MAP-learners. Population size = 25, horizontal transmission rounds = 20, vertical transmission = 20.

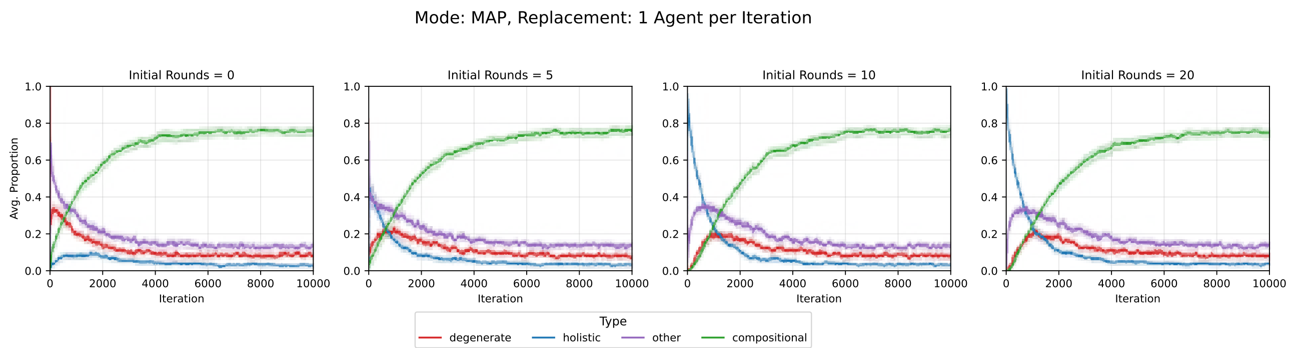


Figure E.9.5: Proportion of posterior per language type, for various amounts of initial learning rounds from a holistic starting language. Agents are MAP-learners. Values are averaged for the entire population of sample-learners and averaged over 50 runs. Per iteration, one agent is replaced. The shaded area represents 95% confidence intervals.

E.1.3 Discussion

On closer inspection, these results are unsurprising. As suggested by Griffiths and Kalish (2007), the stationary distribution is determined by the initial prior, the sampling strategy of the agent, and the number of teachers an agent learns from. Additionally, in this model, the pressure for expressivity (as implemented by the pragmatic agents) and the number of interaction rounds also have an impact on the stationary distribution. This suggests that, indeed, as we see in these results, adding initial learning rounds from a holistic language does not impact the long-term results or the stationary distribution, since it does not change any of these factors that have an impact on the stationary distribution. In hindsight, I should have considered the working of the model more thoroughly a priori, as the outcome of the experiment could have been anticipated from the model's assumptions and its impact on the long-term dynamics.

Experiment 10: Interaction between Prior Strength and Regulating the Impact of Evidence

This experiment investigates whether there is an interaction between the parameter that regulates the impact of new evidence (Experiment 5) and the strength of the compressibility prior (Experiment 6).

E.1.4 Experimental Setup

Simulations were run for α, the parameter that regulates the impact of new evidence, and $\alpha \in [0.0, 0.00025, 0.0005, 0.00075, 0.001]$. $\beta \in [0.0, 0.25, 0.5, 0.75, 1.0]$, the parameter that regulates the strength of the compressibility prior. Values for horizontal and vertical transmission were set to 20, and the population size was 25. Simulations were run for both sample and MAP-learners, for 5000 iterations.

E.1.5 Results

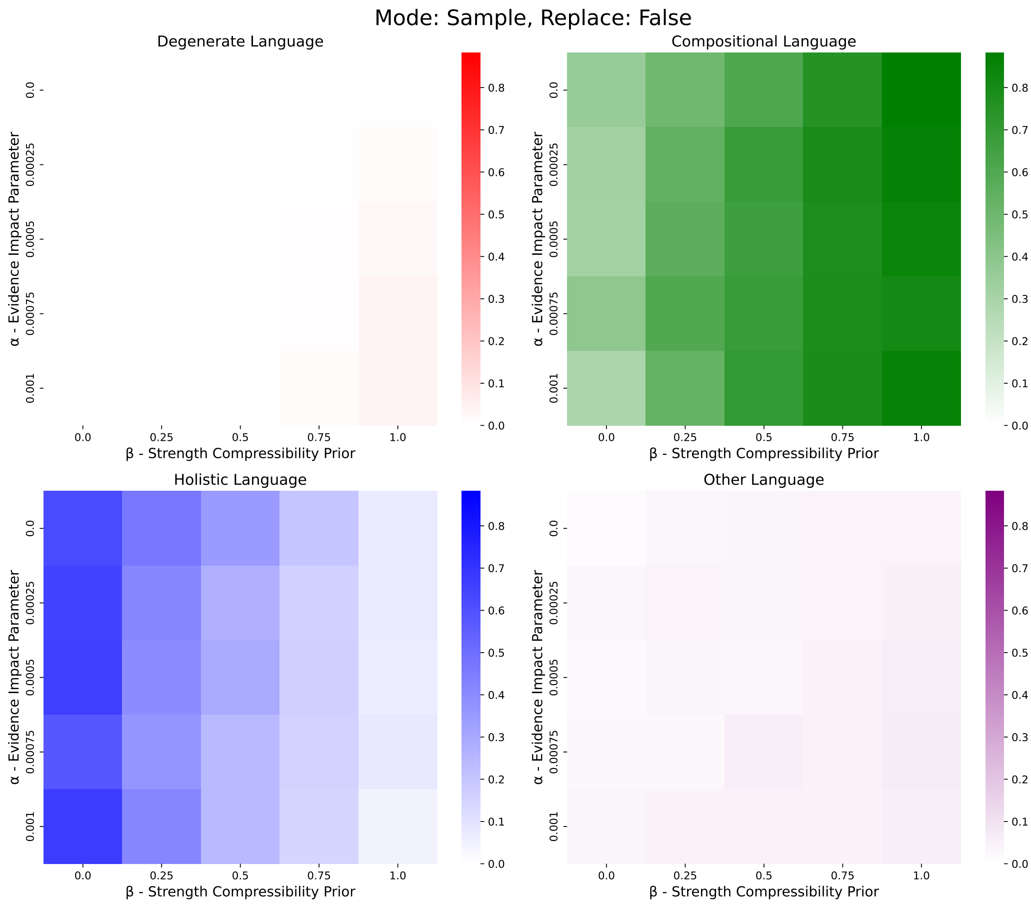


Figure E.10.1: Heatplots interaction strength compressibility prior and impact of new evidence.