MAX PLANCK INSTITUTE
FOR PSYCHOLINGUISTICS

WWW.MPI.NL

Language and society:

How social pressures shape grammatical structure

LIMOR RAVIV

Language and society:

How social pressures shape grammatical structure

© 2020, Limor Raviv

ISBN: 978-94-92910-12-7

Cover picture: Drawing by Nick Lowndes

Printed and bound by Ipskamp Drukkers b.v.

This research was supported by the Max Planck Society for the Advancement of Science, Munich, Germany.

The educational component of the doctoral training was provided by the International Max Planck Research School (IMPRS) for Language Sciences. The graduate school is a joint initiative between the Max Planck Institute for Psycholinguistics and two partner institutes at Radboud University – the Centre for Language Studies, and the Donders Institute for Brain, Cognition and Behaviour. The IMPRS curriculum, which is funded by the Max Planck Society for the Advancement of Science, ensures that each member receives interdisciplinary training in the language sciences and develops a well-rounded skill set in preparation for fulfilling careers in academia and beyond.

Language and society: How social pressures shape grammatical structure

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 7 mei 2020
om 12.30 uur precies

door

Limor Raviv
geboren op 21 oktober 1985
te Petah Tikva (Israël)

Promotor: Prof. dr. Antje S. Meyer

Copromotor: Dr. Shiri Lev-Ari (Royal Holloway University of London, Verenigd Koninkrijk)

Manuscriptcommissie:

Prof. dr. Asli Ozyurek

Prof. dr. Caroline F. Rowland

Dr. Mark Dingemanse

Language and society: How social pressures shape grammatical structure

Doctoral Thesis
to obtain the degree of doctor
from Radboud University Nijmegen
on the authority of the Rector Magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Thursday, May 7, 2020
at 12.30 hours

by

Limor Raviv
born on October 21, 1985
in Petah Tikva (Israel)

Supervisor: Prof. dr. Antje S. Meyer

Co-supervisor: Dr. Shiri Lev-Ari (Royal Holloway University of London, UK)

Doctoral Thesis Committee:

Prof. dr. Asli Ozyurek

Prof. dr. Caroline F. Rowland

Dr. Mark Dingemanse

To my amazing grandmother Rita

Table of Contents

1	General Introduction	11
2	Compositional structure can emerge without generational transmission	25
3	Larger communities create more systematic languages	69
4	The role of social network structure in the emergence of linguistic structure	105
5	What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability	157
6	General Discussion	215
	References	257
	English summary	281
	Hebrew Summary (סיכום בעברית)	283
	Dutch Summary (Nederlandse samenvatting)	285
	Acknowledgements	287
	Curriculum vitae	295
	Publications	297
	MPI Series in Psycholinguistics	299

1 General Introduction

"I find it fascinating that there are seven thousand different ways to do what we're doing right now."

John McWhorter

The evolution of language and the origin of linguistic diversity have been occupying people since the dawn of days. Multiple myths and folk tales attempt to explain how languages began, and why they became so different from each other. According to the biblical account, all humans once spoke the same language (Genesis 11:1-9). Having only one shared language allowed people to cooperate with one another and engage in large-scale endeavors to challenge God (i.e., The Tower of Babel). In return, God punished the people by "confusing"[1] their languages and scattering them around the world, ensuring that humankind would no longer be able to globally understand each other. This biblical story is probably the most well-known origin myth regarding the evolution of language and the source of linguistic diversity today, but is far from being the only one: very similar tales have been documented in Sumerian mythology, Ancient Greek mythology, Central American religions and tribal legends in Africa (Boas, Teit, Farrand, Gould, & Spinden, 1917; Kramer, 1968; Maher, 2017; Teit, 1917).

But was there really once a single language that all people could understand? The truth is that linguists just don't know. The biggest challenge in the field of language evolution is the lack of direct evidence regarding the origin and development of the first human languages: we do not have access to the minds and languages of our ancestors who lived hundreds of thousands of years ago. We therefore have no information on how the very first language (or languages) looked like: did early humans use gestures or vocalizations? Did they imitate sounds they heard in their environment, or babbled randomly until certain sounds took on meaning? Without a time-travel device, none of these questions can be answered directly. Therefore, all evidence favoring one theory or another comes from indirect and analogous observations, ranging from babies' language development trajectories, to the communication systems of primates and

[1] The name of the city *Babel* is derived from the Hebrew verb root B.B.L, which means "confuse".

other non-human animals, to computational models and experimental paradigms mimicking the transition from a state of no language or proto-language to a state of complex language.

Although the origin of languages remains mysterious, the thought of everyone in the world speaking the same language is an interesting one. If we view language as evolving primarily for the sake of communication, it seems very reasonable that we should all be able to understand each other and use one common tongue. It is easy to imagine how the past and present world would benefit from speaking a universal language. Indeed, in the past centuries there have been multiple attempts to impose a Lingua Franca (e.g., Latin), and even several attempts to artificially create such a language (e.g., Esperanto). But despite the theoretical (though unlikely) possibility of all languages deriving from a single origin, and despite the potential appeal of having a universal language, that is obviously not the case: There are about 7,000 different languages around the world today, and these languages vary greatly from one another[2].

The origins of linguistic diversity

Why are there so many different languages? What are the sources of this astonishing linguistic diversity? While some questions about the evolution of language are impossible to answer, it turns out that we can provide some answers to these two questions. People speculate that languages differ due to cultural differences, environmental differences, historical changes, etc. The main principle underlying these intuitions is that languages are constantly changing: even if we had a common language once, it would probably not be the same 200 years later. Moreover, if groups of people split and migrate to different areas and engage in different activities, these changes would probably not look the same across different groups. Languages develop under different conditions, and such differences may lead to the formation of different languages and to ongoing changes in their sounds, their lexicon and their grammar. In short, languages are dynamic and adapt to their environment.

One explanation for why languages differ is that they exist in diverse physical environments: languages are used in mountains and in open

[2] Of course, one can also focus on the similarities between languages (i.e., linguistic universals): all natural languages serve communicative goals and are constrained by humans' cognitive capacities; therefore, they inevitably share some features (Hockett, 1960).

plains, in deserts and in forests, in tropical climates and in freezing weather conditions. Importantly, these environmental differences can shape the languages that evolve in each region. Specifically, several geographic–linguistic correlational studies have tested the association between environmental features (such as climate and altitude) and linguistic features (such as sound systems). These studies report significant differences between languages evolving under different physical conditions, and highlight the physical affordances of various climates and landscapes. Specifically, our bodies react differently to different temperatures and air pressures: organs that are involved in the process of language production and comprehension (e.g., lungs, lips, vocal cords, ears) may respond differently to different conditions, making some linguistic elements harder or easier to produce and comprehend.

For example, languages that are spoken in dry and cold climates are less likely to develop a tonal distinction between words, like that found in Chinese (Everett, Blasi, & Roberts, 2015). This is because in such dry climates, it is slightly harder to control the vocal cords and, consequentially, the tone of our voice. As such, almost all languages that make use of complex tones (and therefore require precise manipulation of pitch) are found in more humid regions. Another example is the prevalence of ejective consonants almost uniquely in mountain areas (Everett, 2013). Ejective consonants, which are produced by compressing a pocket of air in the throat instead of the lungs, are very common in high elevation regions around the world, but are very rare in other landscapes. This geographic correlation is mainly attributed to the fact that ejectives are much easier to produce in high altitudes, since low air pressure dramatically reduces the physiological effort required for the compression of air in the pharyngeal cavity. Another study found that languages spoken in rugged terrains and higher elevations typically have more consonants, while languages spoken in regions with higher tree-cover, warmer temperatures, and more precipitation typically have fewer consonants (Maddieson & Coupé, 2015). This finding has been attributed to difficulties in comprehension, and more specifically, to the effectiveness of transmission of different types of sounds in different environments: consonants are transmitted less reliably in an environment with more disruption (e.g., denser vegetation). In sum, these studies show that languages can adapt to fit their environments, and can help explain patterns of linguistic diversity world-wide.

Social structure and language diversity

Languages are also used in different *social* environments. Languages evolve in different communities, with different population sizes, different social structures, and different social needs. If languages evolved first and foremost for providing for speakers' communicative goals, they should adapt to fit these social needs as well. In other words, since languages are adaptive systems that can be shaped by their environment, they are also bound to be shaped by relevant social factors in their environment. A simple, yet powerful, example for this idea comes from the well-known (albeit erroneous) myth of Eskimos having over 50 different words for snow, or of Dutch people having over 50 different words for rain. Such common statements are inaccurate (e.g., Inuktitut has far fewer than 50ways of describing snow and ice, and even these are mostly complex polysyntactic words created by combining a stem and affixes (Krupnik & Müller-Wille, 2010)), but they nevertheless support the intuitive idea that languages adapt to fit their speakers' communicative needs: if it is relevant and/or important for speakers to frequently differentiate and distinguish between different types of snow (or any other category, for that matter), then their language is likely to reflect that need by having suitable words (Regier, Carstensen, & Kemp, 2016). Another simple example comes from the constant addition of new words to languages' lexicons whenever new technologies or concepts are introduced: words like "internet", "fax", "blog" and "Brexit" were invented in response to relevant things people wanted to be able to talk about efficiently. Such lexical adaptations, while simple, help to demonstrate how languages can be affected by people's social needs.

Languages may also be affected by socio-demographic features of the community in which they are spoken. For example, languages may differ depending on the way people interact with each other, the frequency of these interactions, the number of people in the community, how far they live from each other, the degree of familiarity and hierarchy between speakers, etc. Such social characteristics can potentially influence languages on many levels: from their grammatical structure (e.g., how systematic and transparent languages are in terms of their morphology), to their stabilization patterns and rates of change (e.g., how fast innovations spread in the community and become norms), to their level of uniformity and convergence (i.e., how much dialectal variability exists in the community). Indeed, diachronic studies, typological analyses, and computational models suggest that languages are shaped by the social

properties of the culture in which they evolved, and attribute special roles to the variables of community size, network structure, and the degree of isolation vs. contact with outsiders (Baxter, 2016; Bentz, Dediu, Verkerk, & Jäger, 2018; Bentz & Winter, 2013; Dale & Lupyan, 2012; Fagyal, Swarup, Escobar, Gasser, & Lakkaraju, 2010; Gong, Baronchelli, Puglisi, & Loreto, 2012; Ke, Gong, & Wang, 2008; Lewis & Frank, 2016; Lou-Magnuson & Onnis, 2018; Lupyan & Dale, 2010; Meir, Israel, Sandler, Padden, & Aronoff, 2012; Milroy & Milroy, 1985; Nettle, 2012; S. G. Roberts & Winters, 2012; Trudgill, 1992, 2008, 2009; Vogt, 2007; Wichmann, Stauffer, Schulze, & Holman, 2008; Wray & Grace, 2007; Zubek et al., 2017)[3].

The literature on language change often draws a distinction between languages spoken in *Esoteric* vs. *Exoteric* communities (Milroy & Milroy, 1985; Roberts & Winters, 2012; Trudgill, 1992, 2002, 2009; Wray & Grace, 2007). Generally speaking, esoteric communities are small, closed societies that have little contact with outsiders, while exoteric communities are considerably bigger and more open, so that there is a higher degree of interaction with outsiders and more non-native speakers. For the sake of illustration, imagine the difference between a small and isolated community in the Amazons of Peru or in the Papua New Guinea, and a big and wide-spread community in Europe or in Central America. Such communities may vary greatly in their social structures and social needs, and, consequently, in their languages.

Many researchers speculate that esoteric and exoteric societies would develop different types of languages given that they are subjected to different communicative pressures (Meir et al., 2012; Trudgill, 1992, 2002; Wray & Grace, 2007). For example, members of esoteric communities are typically highly familiar with each other and share much common ground. Such intimate relations can potentially lead to more alignment and uniformity in the language community, but also to higher chances of developing rich and non-transparent grammatical structures that rely heavily on context. On the other hand, members of exoteric communities are far more likely to interact with strangers and outsiders they've never met before, and the community has a higher proportion of adult second-language learners who are not native speakers of the language. Given speakers' inability to rely on shared history when talking

[3] The literature introduced in this chapter will be presented and discussed multiple times throughout this dissertation. This repetition was unavoidable since all experimental chapters were submitted as individual journal publications.

to strangers, and given the well-known difficulty of adults in learning complex and opaque languages later in life (Birdsong, 2006; DeKeyser, 2013), exoteric communities may be under a stronger pressure to simplify their languages, and develop more transparent and systematic structures. However, while such claims are interesting and theoretically motivated, they have remained mostly untested until recently.

In a highly influential cross-linguistic study that looked at thousands of languages around the world, Lupyan and Dale (2010) tested whether exoteric and exoteric societies differ in how complex vs. simpler their languages are. To this end, they examined the correlation between three social-demographic features that typically differentiate esoteric and exoteric societies (i.e., the size of the population, the geographical spread, and degree of language contact) and 28 structural-linguistic features that are related to morphological complexity (e.g., the prevalence of inflectional morphology, the number of cases, the degree of syncretism, the presence of markers for coding plurality, evidentiality, possessives, etc.). Their results showed that morphological complexity was significantly related to all these features, but most strongly correlated with population size: languages with many speakers had simpler morphological structures overall. For example, big communities typically have languages that are less inflected, have simpler noun and verb agreement systems, have simpler inflectional verb morphology and fewer tenses, and often lack inflectional morphological for negation, evidentiality, and aspect (all of which are expressed lexically using individual words instead). Together, the findings of Lupyan and Dale (2010) provided empirical support for the idea that differences in social structure can help explain patterns of linguistic diversity, and suggested that there are important differences between language spoken in big communities and languages spoken in small communities.

However, based on this study alone it is still hard to say whether the number of people speaking the language is really what is driving the observed pattern of results. First, we cannot draw any causal conclusions from correlational studies – the association between community size and language structure cannot be taken as proof that differences in community size *lead* to differences in language structure. Second, community size is only one characteristic of a given society in real-world settings, and is naturally confounded with many other social features (e.g., network structure, the degree of language contact). While these confounding social features are the exact characteristics underlying the distinction between esoteric and exoteric societies, they make it highly problematic to evaluate

the *unique* contribution of each factor separately. For example, the fact that small communities also tend to be tightly-knit and highly connected (i.e., most people in the community know each other) can be mediating the effect of community size, and potentially serve as an alternative explanation for why small communities show higher levels of grammatical complexity. Similarly, the relative morphological simplicity documented in big communities can be driven by the fact that big communities also tend to have more adult non-natives speakers, who struggle more with learning a new language. In fact, this is the explanation that Lupyan and Dale (2010) offer for their results: they argue that community size is merely a proxy for the proportion of second-language learners, which is the true underlying reason for why we see a correlation between community size and language complexity. The idea is that bigger groups have simpler languages as a result of accommodating to adults' learning difficulties: if simpler languages are easier to learn, having many non-natives will lead to simplification of the language over time.

In any case, it is not possible to disentangle the individual roles of group size, network connectivity, and the proportion L2 learners using correlational studies. Moreover, such studies do not promote our understanding of the mechanisms behind social structure effects. Some agent-based simulations attempted to shed light on the individual contribution of social features by manipulating and examining one specific parameter at a time, and seeing how it affects various linguistic outcomes. These models suggest that the different social properties that characterize esoteric and exoteric societies (i.e., community size, network structure, and the proportion of non-native speakers in the population) are each associated with different communicative pressures, yet all seem to contribute in parallel to the differences reported between the languages of esoteric and exoteric communities. For instance, community size seems to be a relevant feature for the emergence of systematic grammars: one model found that compositionality (i.e., systematic and transparent mapping between parts of meaning and parts of speech) tended to emerge more extensively in larger populations of agents due to an increase in the number of words, which increased the likelihood of finding regular patterns between utterances and meanings (Vogt, 2007, 2009). Moreover, different network structures can potentially account for differences in the overall structure of languages. Specifically, sparse networks that have highly-connected agents (i.e., "hubs" or "leaders") tend to develop categorization in color terms much faster than sparse networks without such agents (Gong et al., 2012), while networks with high connectivity

between agents are more likely to develop languages with complex morphological structures (Lou-Magnuson & Onnis, 2018). Finally, a high proportion of adult non-native learners can lead to greater morphological simplification when assuming that adult agents have a prior bias against complexity (but child agents don't), and this trend seems to be modulated by population size: Dale & Lupyan (2012) reported that in small populations of agents, having only child learners (with only the bias to imitate) significantly increased the chances that the language will develop complex morphology, while in slightly bigger populations of agents, inflections disappear.

Although computational models support the hypothesis that the social environment can affect the development of languages, they offer only limited insights into the effect of different features on linguistic diversity, and are not sufficient to confirm the claim that different social structures lead to differences in linguistic structure. Specifically, such models are not tested against empirical data, and are often minimalistic and bear little resemblance to real human social dynamics and cognition. For example, the agents in most computational models have unlimited memory capacity (which humans clearly do not have), and often update their lexical inventories after every interaction by overriding all previous variants (which humans clearly do not do). As such, these models warrant further experimental validation.

In sum, there is little empirical evidence for the causal role of different social factors in explaining patterns of linguistic diversity, and such claims rely mostly on theoretical models and correlational studies. While some empirical work offers relevant (albeit indirect) evidence for the role of community size and adult second-language learning on linguistic complexity (Atkinson, Kirby, & Smith, 2015; Atkinson, Mills, & Smith, 2018; Atkinson, Smith, & Kirby, 2018), the exact effects of social structure on linguistic structure remain unconfirmed. Crucially, no experimental work has directly examined how differences in community size, network structure and the proportion of non-native speakers affect languages in laboratory settings. Nevertheless, carefully designed experiments are a promising way to examine the causal role of these social factors: Although population size, network structure and the proportion of L2 learners are confounded in real-world communities, using controlled experiments in laboratory settings can enable us to manipulate each of these factors separately, and examine changes in linguistic outcomes as a result.

The current thesis

The goal of my PhD project was to experimentally tease apart these confounding social parameters, and directly test their unique contribution to the formation of languages. I was inspired by behavioral studies with human participants that examine the creation of novel communicative systems in the lab (Christensen, Fusaroli, & Tylén, 2016; Fay, Arbib, & Garrod, 2013; Fay, Garrod, Roberts, & Swoboda, 2010; Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015; Roberts & Galantucci, 2012; Selten & Warglien, 2007), and developed a new paradigm for examining the formation and the nature of new artificial languages created by groups of interacting participants. The motivation for developing this paradigm was that it allows us to look at how languages evolve in real-time in a micro society depending on its specific social features. In particular, I manipulated the size of groups and the way participants were connected to each other (i.e., network structure) and examined how these changes affected the languages that evolved in each group. I also examined the underlying mechanism behind community size and network structure effects (i.e., differences in input variability), and tested the premise that more systematic languages are easier to learn. Originally, I also planned to examine the role of non-native speakers in the population. However, it was not possible to complete this study due to several technical and theoretical constraints. I outline the planned design for this study and provide a detailed discussion of these constraints in Chapter 6 (the General Discussion).

The basic design of the group communication game used in Chapters 2-4 is as follows: groups of participants came to the lab and were asked, over the course of several hours, to create a new artificial language to communicate with each other. Participants were not allowed to use Dutch, English, or any other language they knew, and could only use nonsense and Gibberish words they invented on-the-go (e.g., '*wape*', '*tes-ik*'). Participants' goal was to successfully interact about different types of simple dynamic scenes, which always included one of four novel shapes moving on the screen in some direction. The shape in each scene also had a unique texture. Participants needed to come up with words to describe these scenes, and earned points when they successfully understood each other (i.e., if they managed to choose the right scene from a set of possible scenes given a word). Participants in the same group interacted for several rounds in alternating pairs, so that they were paired with a different person in every round. In each round, paired participants took turns in guessing

and producing words – experiencing both being a producer and a comprehender. In each interaction, one participant produced a word, and the other participant had to choose the scene they thought their partner meant from a set of possible scenes. The number of scenes participants needed to refer to gradually increased over rounds, so participants had more and more meanings to communicate about over time. At the beginning of the experiment, the participants were guessing the names and making words up randomly. But over the course of several hours, they could start developing linguistic structure and regularities, such as creating a specific morpheme for each shape or each direction.

I then looked at the languages participants created throughout the course of the entire experiment, and characterized them based on four measures: (1) the degree of communicative success (i.e., how accurately participants understood each other); (2) the degree of convergence on a shared lexicon (i.e., whether different participants used the same words); (3) the degree of stability (i.e., how much languages changed over time); and (4) the degree of compositional linguistic structure in the language (i.e., whether similar scenes were labeled systematically using consistent morphemes/words). Looking at these four measures and how they changed over time allowed me to characterize the emerging languages, and provided valuable insights into the live formation of grammar as a result of communicative needs. Contrasting these measures across different experimental conditions sheds light on whether and how differences in social structure affect different linguistic properties. In order to promote open science, reproducibility and scientific transparency, the data and code for all analyses reported in this dissertation is openly available online for readers and reviewers.

Chapter 2 introduces the paradigm in detail, and establishes its validity and effectiveness in terms of language emergence during communication. In this chapter, I analyzed the behavior of different groups of four interacting participants as they created a new language in the lab, and examined the emerging languages in terms of communicative success, convergence, stability and linguistic structure. Most notably, I tested whether groups in this paradigm can develop systematic compositional structure (one of the hallmarks of natural languages) purely as a result of members' communicative needs. This was an important point because previous experiments on the evolution of language suggested that this type of linguistic structure can only emerge when there is also a learnability pressure, i.e., when languages are transmitted across multiple generations (Kirby et al., 2015). Therefore, the main goal of this

chapter was to show that, in contrast to previous assumptions, communicative pressures alone can shape languages' grammar in meaningful ways, even in the absence of learning pressures as a result of generation turnover. Specifically, I tested how two aspects of real-world communication, namely, interaction with multiple people and interaction about an expanding meaning space, can introduce a pressure for generalization and systematization that leads to the creation of compositional languages. I also examined the unique contribution of each of these aspects in order to determine the relevant communicative pressures that give rise to linguistic structure.

Chapter 3 directly tests the role of community size in the formation of languages using the group communication paradigm. In this chapter, I compared small groups of four participants, to larger groups of eight participants, and contrasted their behaviors along the same four measures described above. My main predication was that community size would have a significant and causal effect on languages' structure, so that larger groups would create more systematically structured languages – corresponding to the claim that big communities tend to have simpler and more regular languages (Lupyan & Dale, 2010; Wray & Grace, 2007). Although both group sizes are considerably smaller than real-world communities, I hypothesized that given the miniature nature of the experiment, doubling the number of people in the group would already make a significant difference in the languages these groups would create. Specifically, having more people to interact with should lead to more input variability (i.e., more lexical variants) and less shared history between group members. Given this greater communicative challenge, I hypothesized that members of larger groups would be under stronger pressure to converge on a shared language that is simple, predictable, and more systematically structured – leading them to create more compositional languages. I also tested two of the postulated mechanisms underlying community sizes effects, namely, the idea that differences in input variability and in shared history between small and larger groups lead to differences in their achieved levels of systematic structures.

Chapter 4 directly tests the individual role of social network structure using the same group communication paradigm and the same linguistic measures. In this chapter, I examined the formation of new languages that developed in different micro-societies that varied in their network structure. Community size was kept constant across conditions, so that all networks were comprised of eight participants, yet differed in their degree of connectivity (i.e., how many people each participant interacts with) and

homogeneity (i.e., whether all participants are equally connected). Specifically, I contrasted three different types of networks, which are typically used in computational models: (1) Fully connected networks, in which all members interact with each other; (2) Small-world networks, which are much sparser and have many members that never interact, although these "strangers" are nevertheless linked indirectly via a short chain of shared connections; And (3) Scale-free networks, which are as sparse as small-world networks, but whose members' distribution of connectivity roughly follows a power law so that one of the participants is highly connected to almost everyone in the network (a "hub") and others are much less connected. My main prediction was that sparser networks would develop more systematic languages, as a result of higher levels of input variability and diversity in such networks, which increase the pressure for generalization and systematization (Lou-Magnuson & Onnis, 2018; Wray & Grace, 2007). I also predicted that scale-free networks would develop even more compositional languages compared to small-world networks, since the existence of a "hub" can further promote the spread of compositional innovations (Fagyal et al., 2010; Gong et al., 2012; Zubek et al., 2017). Following the findings of Chapter 3, I also examined difference in input variability as a potential underlying mechanism behind network structure effects.

Chapter 5 is a pre-registered study that tested the causal relationship between systematic linguistic structure and language learnability. In this chapter, I tackled a crucial premise underlying all previous chapters, as well as theories on language evolution, second language learning and the origin of linguistic diversity: the highly influential assumption that more systematic languages (i.e., languages with more regular, compositional and transparent grammars) are easier to learn. For example, iterated language earning studies have shown that language learnability and linguistic structure both increase over the course of cultural transmission, and suppose that these two patterns are inherently linked: languages are argued to become more learnable *because* they become more structured (Cornish, 2010; Cornish, Tamariz, & Kirby, 2009; Kirby et al., 2008; Smith, 2011; Zuidema, 2003). Although direct empirical evidence for this argument is lacking, it serves as an essential component in the theoretical reasoning of such iterated learning models, and are also essential for the claim that community size effects are driven by adults' difficulty in learning complex and opaque languages (Dale & Lupyan, 2012; Lupyan & Dale, 2010). In addition, I tested whether languages created by big communities were easier to learn, i.e., whether the larger groups in

Chapter 3 created languages that would be better acquired by new individuals. To this end, I used an artificial language learning paradigm with stimuli adapted from the group communication paradigm. In this experiment, individuals needed to learn a new miniature language with labels for describing the same scenes used in Chapters 2-4. Importantly, participants were trained on different input languages, all of which were created by either big or small groups from Chapter 3, and varied in their degree of systematic structure and in their group size origin, while being relatively similar in their average word length and internal confusability. After training, participants were tested on their knowledge of the input language in a memory test (measuring participants' reproduction accuracy on the scene-label pairings) and in a generalization test (measuring participants' ability to label new, unseen scenes). I compared the acquisition of these different languages with two predictions in mind: (1) that linguistic structure would significantly affect language learnability, so that more compositional languages with systematic form-to-meaning mappings would be easier to learn; (2) that group size would have an additional effect on language learnability, so that that across all structure levels, participants who learned languages that were created by big groups would show higher reproduction accuracy.

Chapter 6 summarizes the main experimental findings of this thesis and discusses their theoretical implications. In this chapter, I also reflect on the original plan of this thesis, and introduce the methodological issues which prevented me from executing an additional experiment to test the role of second-language learners in the community. Finally, I speculate on other social factors that may contribute to linguistic diversity, and make recommendations for future work.

2 Compositional structure can emerge without generational transmission

Abstract[4]

Experimental work in the field of language evolution has shown that novel signal systems become more structured over time. In a recent paper, Kirby, Tamariz, Cornish, and Smith (2015) argued that compositional languages can emerge only when languages are transmitted across multiple generations. In the current paper, we show that compositional languages can emerge in a closed community within a single generation. We conducted a communication experiment in which we tested the emergence of linguistic structure in different micro-societies of four participants, who interacted in alternating dyads using an artificial language to refer to novel meanings. Importantly, the communication included two real-world aspects of language acquisition and use, which introduce compressibility pressures: (a) multiple interaction partners and (b) an expanding meaning space. Our results show that languages become significantly more structured over time, with participants converging on shared, stable, and compositional lexicons. These findings indicate that new learners are not necessary for the formation of linguistic structure within a community, and have implications for related fields such as developing sign languages and creoles.

[4] This chapter is based on *Raviv, L., Meyer, A. S., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. Cognition, 182, 151-164. doi:10.1016/j.cognition.2018.09.010*

Introduction

Amongst the most important questions in the field of language evolution are how and why linguistic structure emerged, and under which pressures it evolved (Bickerton, 2007). According to usage-based theories, language is an adaptive and culturally transmitted system that has evolved to fit speakers' cognitive biases and constraints (Deacon, 1997; Reali & Griffiths, 2009; Smith, 2011) and to maximize their communicative success (Beckner et al., 2009; Mirolli & Parisi, 2008). A critical phase in the process of language evolution is the transition from an unstructured proto-language to a state of a full-blown language that exhibits compositional structure (Jackendoff, 1999; Zlatev, 2008). Compositionality, i.e., the systematic recombination of small units to express different meanings, is considered one of the hallmarks of natural language, which differentiate it from animal communication systems (Hockett, 1960). Indeed, one of the things that makes natural languages so unique is their infinite expressive power, which is the direct result of compositionality: we can talk about an unlimited set of meanings thanks to our ability to recombine a limited set of sub-elements in systematic ways.

In the past two decades, two different strands of experimental work have attempted to investigate the factors involved in the emergence of linguistic systems from two distinct perspectives. First, Experimental Semiotics studies focused on the communicative and social nature of language evolution, and examined how interactions between pairs or groups influence convergence, iconicity and complexity of visual signals (e.g., Galantucci & Garrod, 2011; Garrod, Fay, Lee, Oberlander & McLeod, 2007). In Experimental Semiotics studies, the main pressure is a communicative pressure for expressivity: signals should be expressive, informative and communicatively efficient in order to allow for reliable discrimination between potential referents, and should be shared across participants to allow for mutual understanding. Second, Iterated Learning studies focused on how individuals' cognitive biases and constraints shape previously established signs over the repeated transmission to new generations of learners, and examined how signal systems change in terms of learnability and structure (e.g., Beckner, Pierrehumbert & Hay, 2017; Kirby, Cornish & Smith, 2008). In Iterated Learning studies, the main pressure is a learning pressure for compressibility: limitations on memory create a pressure for signals to become simpler, more compressed and more predictable, so that languages could be easily learned from a finite

set of exemplars, and generalizable to a new set of exemplars (Kirby, Griffiths & Smith, 2014; Kirby et al., 2008). Both these literatures have generated numerous novel findings with important implications for the evolution of language. For example, Experimental Semiotics paradigms have been used to examine the emergence of arbitrary signals from iconic signs (e.g., Garrod et al., 2007). Iterated Learning has typically been used to examine the creation of compositional regularities (e.g., Kirby et al., 2008), but has also been used to examine the evolution of case markers (e.g., Smith & Wonnacott, 2010) and color terms (e.g., Xu, Dowman & Griffiths., 2013).

In a recent and highly influential study, Kirby, Tamariz, Cornish, and Smith (2015) combined the paradigms of Experimental Semiotics and Iterated Learning and contrasted two experimental conditions: communication with transmission vs. communication without transmission. In the communication and transmission condition (the "chain" condition), pairs of participants communicated about a structured meaning space using an artificial language, and then their languages were transmitted to new pairs of participants over several generations. In the communication without transmission condition (the "closed group" condition), pairs interacted amongst themselves for several rounds, with no new learners being introduced over time. The results showed that when languages were transmitted over multiple generations of pairs, they developed compositional, morphology-like structures in which different affixes were systematically combined to express similarities in meanings. In contrast, when the same pairs communicated for repeated rounds without generational turnover, they created holistic, unstructured languages in which each item was assigned a unique label and feature overlap between items was not reflected in the labels.

Kirby et al. (2015) argued that the reason that compositionality did not emerge in the closed-group condition is because pairs were able to get highly familiar with the signs, so there was no reason for them to develop compressed, systematic structures instead of holistic languages. They interpret their results as showing that (1) compositionality arises only as a tradeoff between expressivity and compressibility pressures; and (2) that expressivity and compressibility pressures stem from two independent sources - communication and transmission – which operate at different timescales. Kirby et al. (2015) view these two processes as bringing about conflicting constraints: while horizontal intra-generational communication pushes languages to become maximally expressive, vertical cross-generational transmission pushes languages to become

maximally compressed. By providing a systematic mapping between meanings and signals, compositionality offers an equilibrium between the need to minimize the associated memory and cognitive costs while maximizing languages' expressivity. This idea suggests that the basic architecture of natural language can be explained by the interaction of conflicting weak cognitive biases and processing limitations, and by taking the pragmatic context in which languages evolve into account (Christiansen & Chater, 2016; Culbertson & Kirby, 2016).

Importantly, Kirby et al. (2015) fully equate expressivity and compressibility pressures with communication and transmission respectively. They argue that horizontal communication gives rise to expressivity pressures due to people's communicative goals: languages should be expressive given the need to interact and successfully discriminate between different meanings. Vertical transmission is argued to give rise to compressibility pressures due to people's memory limitations and cognitive biases: languages should be simple and easy to learn given that are being repeatedly learned over generations by new people. They predict that compositionality emerges only when both communication and transmission are at play, as a solution to these competing pressures. On one hand, a compressibility pressure operating in isolation (e.g., languages are only transmitted across generations of learners, but not used for communication) leads to underspecified languages with minimalistic lexicons, where multiple meanings are represented with a single word (as found in Kirby et al., 2008). While such simple systems were highly compressed and easy to learn, they were degenerated, ambiguous and lacked expressivity. On the other hand, an expressivity pressure operating in isolation (e.g., languages are only used for communication, but never transmitted to new learners) should potentially result in languages with massive lexicons, where each meaning is represented with a unique word. While such holistic systems would be maximally expressive, they would also be incompressible and therefore hard to learn and remember by new individuals. If languages need to be both expressive and compressed (i.e., because they are being used for communication as well as being transmitted to new learners), developing regularities in the form of compositional structure will maintain their informativity while reducing the memory load and increasing languages' learnability. This is because compositional languages allow for the expression of multiple different meanings using a recombination of the same basic elements. As such, a compositional language is highly compressed and simpler in comparison to a holistic language (where the

same set of meanings would require memorizing more unique words), while also being highly expressive and informative in comparison to a degenerated language (where the same set of meanings would be indistinguishable). In sum, Kirby et al. (2015) predict that both communication and transmission are necessary for the emergence of compositionality, and conclude that communication alone (i.e., without generation turnover) is not enough for compositionality to emerge. This finding has since been replicated with different meaning spaces (Carr, Smith, Cornish & Kirby, 2016; Winters, Kirby & Smith, 2015) and with artificial sign languages (Motamedi, Schouwstra, Smith & Kirby, 2016).

This conclusion has far-reaching implications for the literature on the evolution of language, as well as for the broader field of cultural evolution. First, it directly relates to work on creolization and emerging sign language by suggesting that one of the "design features" of natural language may need several generations to emerge. Supporting this idea, studies on the developing Nicaraguan sign language have shown that complex linguistic structure emerges over multiple cohorts of learners (Senghas, Kita & Özyürek, 2004), and work on pidgins has suggested that new child learners are required in order to develop recursion (Bickerton, 1983). Second, it affects the reasoning and predictions made about the structure of human lexicons over time: from understanding trends in metaphorical mappings (Xu, Malt & Srinivasan, 2017) to measuring the entropy and informativity of words (Bentz, Alikaniotis, Cysouw & Ferrer-i-Cancho, 2017). Going beyond language evolution and change, this conclusion has already influenced work on a wide range of human behaviors. For example, compressibility pressures during cross-generational transmission have been implied to play a role in explaining cross-cultural differences in folk tale complexity (Acerbi, Kendal & Tehrani, 2017), musical universals (Trehub, 2015), and the propagation and stabilization of behavioral conventions (Scott-Phillips, 2017).

In the current paper we suggest that communication in the real world includes not only expressivity pressures, but also several sources for compressibility pressures. In other words, while we agree with Kirby et al. (2015) that both expressivity and compressibility pressures are necessary for the emergence of compositionality, we believe that both pressures are already present during real-world communication. Therefore, we predict that in contrast to Kirby et al.'s (2015) conclusion, compositionality can emerge during communication in a closed group without generational transmission. This prediction is in line with several non-linguistic communication studies, which found that compositional

structure can emerge in signal systems during interaction alone. First, Selten and Warglien (2007) found that when pairs of participants communicated using strings of consonants (e.g., RZ) to refer to a structured meaning space of shapes and patterns, 12% of pairs developed compositional codes where they systematically combined unique consonants that were assigned according to shape and pattern. Even though compositional structure was not prevalent in the codes developed by participants, this study does provide evidence that compositionality *can* emerge during dyadic interaction without additional learners. Second, Theisen, Oberlander and Kirby (2010) found that some compositionality existed in drawings in dyadic interaction, with participants' drawings showing some re-use of smaller elements to express similarities in meanings (e.g., using squiggly lines to refer to activities/situations). However, the systematicity in these drawings was determined subjectively and existed already in the first round of interaction rather than developed with time over the course of communication. Third, Nölle et al. (2018) found that when pairs needed to communicated about items that were not immediately present in the moment of communication (simulating displacement), their silent gestures became more systematic so that some part-gestures were used at least twice to describe items that shared a meaning category. Finally, Verhoef, Walker and Marghetis (2016) report that visual signal systems (i.e., spatial lines generated by a vertical touch bar) for describing temporal concepts became significantly more compositional over the course of dyadic communication, with systematic re-use of visual signals to represent different meanings. An additional motivation for the idea that communication plays a role in the emergence of structure comes from a study that examined the negotiation of drawings in dyads and micro-societies over repeated interactions (Fay, Garrod, Roberts & Swoboda, 2010). While this study did not examine compositionality, it reported the refinement and the simplification of visual signs as a product of communication, so that drawings became more compressed and less iconic over time. Together, these findings suggest that communication can give rise to structure over time, even without generation turnover.

In the current study we assess whether compositional structure can reliably emerge in an artificial language during communication in a closed group, when the interaction includes two real-world properties of languages acquisition and use that can give rise to communicative compressibility pressures: namely, talking to multiple people, and interacting over an expanding lexicon. We argue that these two properties

introduce compressibility pressures that can drive the formation of compositionality in languages during interaction in a closed group, even without transmission to new learners. In general, compressibility pressures emerge due to participants' limited memory capacity: it is simply too hard to memorize many unique and unrelated labels in a relatively short time. Here we propose that such memory limitations can stem from different sources: compressibility pressures in transmission stem from biases and constraints on learning a given input language, while compressibility pressure in communication stem from the need to converge on a shared, expressive, and productive language with others. While communication in previous studies (e.g., Kirby et al., 2015; Selten & Warglien, 2007; Theisen et al., 2010) included communication with only one partner over a fixed set of meanings, communication in the real-world involves talking to many different people, and referring to an open set of topics. Kirby et al. (2015) touch upon both of these properties in their discussion, but they do so only in relation to transmission: they discuss the consequences of learning languages with larger lexicons (p.98), and predict that chains with bigger populations will develop more structure over time (p. 99). Here, we suggest that these two properties of language acquisition and language use can introduce compressibility pressures during communication, which are sufficient for the emergence of compositionality.

The first possible source of compressibility pressures in real-world communication is interaction with many different people. Models of language acquisition in early infancy stress the importance of receiving input from multiple speakers, who introduce variability in pronunciation, speaking rates, styles, and vocabulary (Kuhl, 2004). This input variability can highlight systematic differences and similarities in linguistic input, and help to separate relevant patterns and consistencies from irrelevant differences in the input. This idea is supported by language learning studies that demonstrate how an increase in input variability (e.g., learning from multiple speakers) can boost categorization, generalization and pattern detection in both infants and adults (Gomez, 2002; Lev-Ari & Shao, 2017; Lively, Logan & Pisoni, 1993; Maye, Werker & Gerken, 2002; Perry, Samuelson, Malloy & Schiffer, 2010; Rost & McMurray, 2009; 2010). In addition, communication seems to lead to the elimination of unpredictable variation (Fehér, Wonnacott & Smith, 2016). Indeed, talking to multiple people is considered a key factor in models of language contact and language change, pushing languages to develop more structure. Specifically, it has been argued that interaction with more

people results in more transparent and more simplified grammars (Nettle, 2012; Wray & Grace, 2007). According to these models, interacting with more people introduces more input variability and more noise, which need to be overcome before the community can reach convention. Thus, interacting with more people can favor systemization in languages by introducing more input variability and therefore a stronger need for generalizations.

In Kirby et al. (2015), communication included interaction with only one other person, so input variability was low and it was relatively easy to achieve convergence: Pairs were able to agree on a holistic, unstructured language that contained a unique label for each item. However, developing such a holistic lexicon is far more complicated when the unique labels of more than one partner need to be remembered, or when the lexicon should be shared across multiple people. When there are more people to interact with, input variability increases as each person introduces their own unique variations to each of the labels, which is taxing for memory. In addition, if labels are idiosyncratic and the language is unstructured, each label needs to be negotiated separately and independently with all partners. Therefore, the need to converge with multiple people introduces a memory limitation (i.e., compressibility pressure), pushing languages to become less holistic and develop more transparent and more predictable structures (e.g., by introducing compositionality), so that they can be easily shared across participants without negotiating each label separately. Supporting this claim, two computational models have shown that compositional languages can emerge over the course of multiple dyadic interactions in populations of five interacting agents (De Beule & Bergen, 2006; Gong, Ke, Minett & Wang, 2004). These models show that compositional languages are favored during repeated communication even within a single generation, and demonstrate how an increase in compositionality can facilitate communicative success and convergence between agents in the population.

A second possible source of compressibility pressures in real-world communication is interaction over an expanding lexicon, a notable property of language use and acquisition. Children need to communicate and refer to more and more things over time. Furthermore, growth in vocabulary size is associated with increased generalization in language: knowing more words can boost children's learning of lexical categories, morphological paradigms and syntactic structures (Blom, Paradis & Duncan, 2012; Goldberg, 1999; Perry, Axelsson & Horst, 2015;

Samuelson & Smith, 1999). Familiarity with more exemplars can help children detect significant patterns in the input and improve their ability to generalize the pattern to new, unfamiliar exemplars. Importantly, children's ultimate goal is to learn how to produce and comprehend an infinite set of meanings from a finite set of exemplars. This point is also a main theme in computational work by Kirby and colleagues, which stressed the importance of a "learning bottleneck" during transmission for the formation of compositionality (e.g., Kirby et al., 2008; Kirby & Hurford, 2002; Smith, Brighton & Kirby, 2003): agents are usually not exposed to the entire repertoire of the language, and learn only a subset of the system. Despite their partial exposure, learners are later required to produce labels to new unfamiliar events. For example, Kirby et al. (2008) trained participants on only half of the items in the language, but tested them on all items. This learning bottleneck created a learnability pressure and promoted generalization. In Kirby et al. (2008)'s seminal set of experiments, this property of transmission and learning was introduced as the main pressure pushing languages to develop systematic structures over generations of learners (i.e., compressibility pressures).

Such a bottleneck was absent in Kirby et al. (2015). In that study, pairs communicated about a fixed (and relatively small) number of items for several rounds, and got highly familiar with the entire meaning space of the language over time. Given sufficient time, memorizing a unique label for every item was feasible, and there was no pressure to develop a systematic and predictable way to label items. However, such a strategy will become problematic if the meaning space is much bigger, or if it expands over time: if participants develop holistic languages that have no inner structure, not only will they need to negotiate the labels for each new item separately and independently without the ability to rely on previously established labels, but they will also be faced with memory limitations once the language contains a large enough number of meanings. Thus, the need to discriminate between more and more items over time introduces a pressure for generalization and systemization similar to a "learning bottleneck". As participants are exposed to more and more items (and consequentially, more input variability), they should be able to detect repeating patterns in their input, which can promote the development of more productive and more predictable labeling methods. This idea is also supported by the findings of Nölle et al. (2018), who report that participants' gestures became more systematic when new meanings were introduced. The productive power of natural language, which stems from its compositional structure, is therefore motivated by the fact that some

elements in the input (real world or an artificial meaning space) are repeated in various contexts. Given this feature, compositionality will allow participants to efficiently express novel meanings and be immediately understood, due to the recombination of elements that have already been negotiated. In other words, interacting over an expanding meaning space (which is also structured to some extent) biases against holistic and unstructured systems.

Some preliminary findings suggest that compositionality can indeed arise in these conditions, which are more ecologically valid and relate more to the way language is used in the real world. In particular, we conducted a pilot study in which three closed micro-societies of four participants communicated about novel items (Raviv, Meyer & Lev-Ari, 2017). Participants interacted in alternating dyads using an artificial language, and needed to describe a set of items to each other in order to earn points in a communication game. Each item was one of four novel shapes, and appeared in a particular size ranging from 2 cm^2 to 9 cm^2. Additionally, each item had a unique fill pattern. At first, participants were exposed to only eight items and needed to name them using novel labels. Over the course of six rounds, we added more and more items to the game and examined changes in the languages created by the participants. As our goal was to create a paradigm where structure emerges in a closed group, we tried to maximize communicative compressibility pressures by including both pressures (i.e., communicating with multiple partners and an expanding meaning space), rather than teasing them apart. The results of this pilot study showed that linguistic structure (measured in the same way as in Kirby et al. 2015, see detailed description below) significantly increased over communication rounds, and some compositionality emerged even in the absence of generational transmission.

While these results were encouraging, they were based on three groups only. Additionally, while the analysis over all groups showed a significant increase in compositionality, a closer look suggested that this might have been the case for only two out of the three groups. Finally, it seemed that languages mostly developed compositional coding for the dimension of shape, but less or not at all for the dimension of size. This result is in line with the "shape bias" reported during novel word learning: children and adults are much more likely to categorize novel items based on their shape, and much less likely to do so based on size (Landau, Smith & Jones, 1988). Therefore, to replicate and confirm our findings, in Study 1 we ran twice as many groups of four participants each, and substituted the size dimension with a more salient dimension (i.e., motion) that turned the

items into dynamic, event-like scenes. The results of this study are reported in full below, and confirm that compositional structure can emerge during communication without generational turnover.

In Study 2, we evaluated the relative contribution of the two compressibility pressures using a meta-analysis that included data from the six groups in Study 1 as well as 18 additional groups of either four or eight participants, which were tested using the same paradigm. The results of this meta-analysis replicated the main finding of Study 1 and show that interaction with multiple partners was the main driver for the emergence of compositionality during communication.

Study 1

The goal of this study was to test whether compositional structure can emerge without generational transmission. In particular, we examined whether introducing two compressibility pressures, i.e., interaction with multiple partners and an expanding meaning space, would suffice for triggering the emergence of compositionality. We used a group communication game in which different micro-societies of four members interacted in alternating pairs, so that each participant interacted with the other three members of the group at least twice. Importantly, participants communicated using an artificial language that referred to an expanding meaning space of novel scenes: the number of scenes in the game increased over time, such that by the end of the experiment participants needed to communicate about almost triple the number of scenes as compared to the beginning. Each scene in this experiment was composed of a shape moving in a given direction across the screen. We tested whether compositionality emerged over time, that is, whether similar meanings were referred to using similar labels. In addition, we examined convergence, stability, and communicative success in the languages to characterize the emerging communication systems and to better understand how these properties change over time.

Method

Participants

24 adults (mean age: 23.2; 18 women) took part in the experiment reported here, comprising six closed groups with four members each. Though our pilot results suggested that three groups are sufficient to test the

emergence of compositionality, we doubled the sample size to ensure that the results are robust. All participants were native Dutch speakers and were recruited using the participant database of the Max Planck Institute for Psycholinguistics. Participants were paid between 20 and 26 Euros for their participation, depending on the amount of time they spent in the lab (ranging between 2:00 to 2:45 hours). In addition, four participants from the winning group received an additional 20 euros for collecting the highest number of points. Ethical approval was granted by the Faculty of Social Sciences of the Radboud University Nijmegen. The study was part of a bigger project whose goal is to test the effect of group size on compositionality, and thus included six additional groups of eight participants each. We report the results of the bigger project elsewhere (Raviv, Meyer & Lev-Ari, under review). Importantly, the compositionality results reported here hold if we analyze all 12 groups, or only the six other omitted groups (see also Study 2).

Stimuli

We created visual scenes that varied along two semantic dimensions: shape and angle of motion, creating a semi-structured, continuous meaning space. We created three different versions of the stimuli, which differed in the distribution of shapes and angles (for a full list of shapes and their associated angles see Appendix A). Each version contained exactly 23 scenes, and was presented to two different groups. Groups that played the same version were given the scenes in a reverse order during the communication phase.

All scenes appeared on the screen surrounded by a white 8 cm^2 frame, and the movement was restricted within those borders. Each scene included exactly one of four distinct shapes (sized 2.55cm^2), which moved repeatedly from the center of the frame in a straight line in a given angle. The four shapes were created to be novel and ambiguous, in order to prevent easy labeling with existing words. In addition, each moving shape was associated with a unique fill pattern, giving each scene an idiosyncratic, unstructured feature.

Our meaning space was therefore semi-structured: some semantic features (e.g., shape, direction of movement) repeated across different scenes, while some features (e.g., fill pattern) did not. This property of the meaning space was meant to simulate the real world, where some elements repeat in different combinations while others are unique. As

such, our meaning space promoted categorization and structure with respect to shape and motion, while it also allowed participants to adopt a holistic strategy in which scenes are individualized according to fill pattern. In addition, motion was a continuous rather than a categorical feature, so that participants were not encouraged to categorize it in any particular way: they could parse it in various ways, and could differ in the way they categorized what it "new" and what is a "recombination".

For each version of the stimuli, the 23 scenes were created in the following way: first, we selected 23 static items from an initial, fixed set of 28 static items, which contained seven tokens of each shape. Each token was associated with a unique blue-hued fill pattern. The 23 static items were randomly drawn from this fixed set with the constraint that each type of shape should appear between four to seven times. Then, each of the 23 static items was associated with an angle in order to create a scene. Angles were randomly selected from a set of 16 angles within the 360-degree-range (0º, 30º, 45º, 60º, 90º, 120º, 135º, 150º, 180º, 210º, 225º, 240º, 270º, 300º, 315º, 330º)[5], following the constraint that each type of shape had to be associated with at least one angle from each of the four quadrants. The rest of the items' angles were randomly drawn from this set of angles.

Procedure

The experiment was designed as a group communication game, with each group comprised of four different members. Participants were told they were about to create a new "Fantasy Language" in the lab, and use it in order to communicate with each other about different novel scenes. No talking or gesturing was allowed during the experiment, and participants were instructed to use only the "Fantasy Language" and their assigned laptops in order to communicate. The experimenters actively monitored participants' productions throughout the experiment to ensure they do not include known words. If a participant typed a label that contained a known word, they were required to change it. Notably, this method was highly successful, with only a few exceptions. Those exceptions were implicit in nature, and were not detected during the experiment by either the

[5] Due to a technical error, during the last test round two groups were presented with angles selected from a set of 36 angles separated by 10 degrees (i.e., 0, 10, 20, 30, 40, 50…). Given that participants have developed productive and systematic languages by that point, they did not notice this error and were easily able to name these scenes.

participants or the experimenters. Importantly, in most of these exceptions, the strings referred to the idiosyncratic fill pattern of the shapes, thus hindering rather than promoting compositionality. Participants' letter inventory was restricted, and included five vowel characters (a,e,i,o,u) and ten consonants (w,t,p,s,f,g,h,k,n,m) which participants could combine freely. We restricted the number of consonants as a means to limit participants' ability to construct known Dutch words. The consonants were chosen based on Dutch phonology, while not including letters like "r" and "l" in order to avoid the use of acronyms or shortcuts for indicating left and right. In addition to these letters (all in lower case), participants could also use a hyphen (but not the space bar).

The experiment had eight rounds in total and took about two hours to complete. It included three unique phases: a group naming phase (round 0), a communication phase (rounds 1-7) and a test phase (round 8). One or two experimenters were present during the entire duration of the experiment.

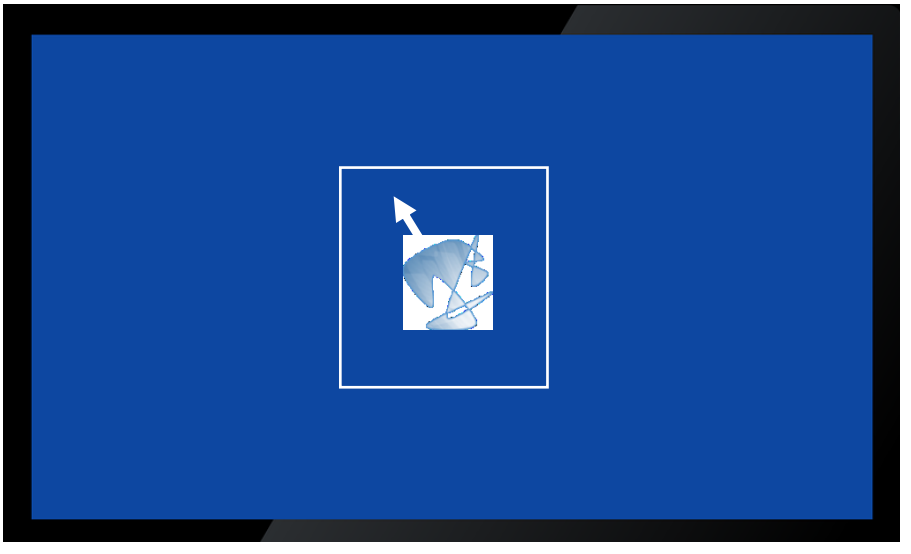
For the initial naming phase (round 0), eight scenes were randomly drawn from the set of 23 scenes chosen for this group (see Stimuli) with the constraint that each shape and each quadrant were represented at least once. During this phase, participants sat together in a room next to a single computer, and were exposed to the eight selected scenes that appeared on the computer screen one by one in a random order. For each scene, one of the participants was asked to use their creativity and type a description for it using one or more nonsense words. Participants took turns in describing the scenes (i.e., typing them using the computer keyboard), so the first scene was described by participant A, the second scene was described by participant B, and so on. Importantly, no use of Dutch or any other language was allowed, and participants were instructed to come up with novel, "gibberish" labels. Once a participant had typed a description for a given scene, it was presented on a screen along with the scene to the rest of their group members for about five to seven seconds. This procedure was repeated until all eight scenes have been presented and named, with each participant describing exactly two scenes. After all scene-label combinations had been created and presented once, we presented the scene-description pairings to participants twice more in a random order in order to establish common ground.

Following the group naming phase, participants were told that they had now created the initial vocabulary of the "Fantasy Language" and so they can start playing the actual game (the communication phase). The

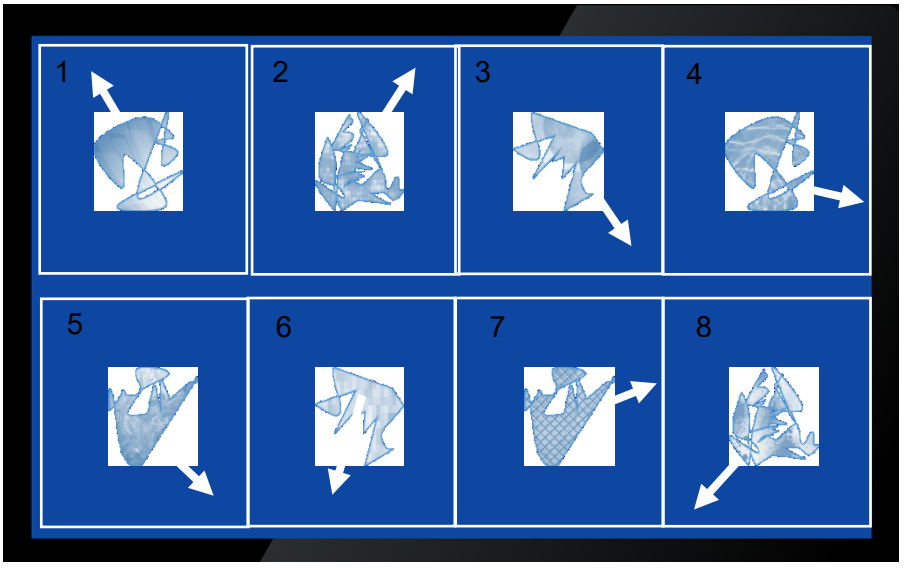
participants were told that the goal of the game was to be communicative and earn as many points as possible as a group, with a point awarded for every successful interaction. The experimenters stressed that this was not a memory game but a communication game, and that participants could choose to use the labels produced during the group naming phase, but they did not have to. If a participant had a better label for a given scene that would be understood by their partner, they could choose to use that label instead.

During the communication phase (rounds 1 to 7), group members interacted in alternating dyads, exchanging communication partners at every round such that each pair in the group interacted at least twice overall. At the beginning of each communication round, the group was split up into two pairs, who sat in different corners of the same room and were separated by a large room divider. Each participant was then assigned a laptop. In each communication round, paired participants played a total of 23 guessing games with each other, with participants alternating between the roles of producer and guesser. In a given game, the producer saw the target scene on their screen (see Figure 1A), and typed a description for it using their keyboard. Once the producer finished typing, they pressed Enter and the description appeared in a large font on their screen, without the target scene. They then rotated their screen using a rotating platform and presented only the description to their partner. The guesser was presented with a grid of eight different scenes on their screen (the target and seven distractors; see Figure 1B), with each scene associated with a number between 1 and 8. The guesser then pressed the number associated with the scene they thought their partner referred to using their laptop's keyboard. Note that the numbers 1-8 were only available to the guesser during this phase, but were blocked from use in participants' typed descriptions. The guesser then received feedback on their screen (see Figure 1C), which they rotated and shared with the producer, allowing participants to learn and align. If the interaction was successful, the pair was awarded with 1 point. At the end of each round, pairs saw the number of points that they accumulated in this round on their screens. Importantly, the total number of points earned by all pairs was added up to a group score, and participants' goal was to maximize their score as a group. Groups were explicitly motivated to earn points: they were told that they were competing against other groups, and that the group with the highest score will win an additional prize of 20 euros.

(A)



(B)



(C)

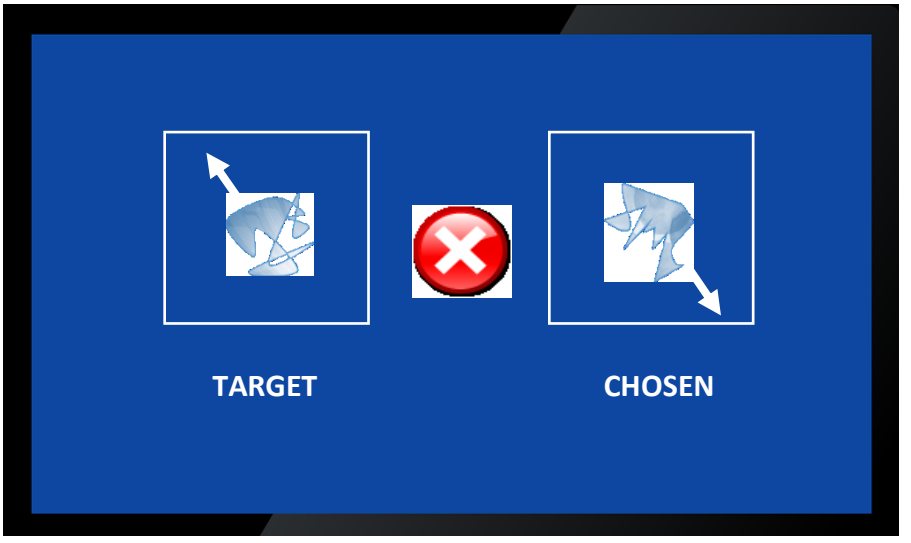


Figure 1: Example of the computer interfaces in a single game in the communication phase. Arrows illustrate the shapes' direction of movement on the screen. The producer saw the target scene on their screen (A) and typed a description for it using their keyboard. Once the guesser saw the description (presented on the producer's screen), they selected a scene from a set of eight possible scenes that was presented on their screen (B). Finally, participants were given feedback, including the target and the chosen scene (C).

Crucially, the number of different target scenes increased from round to round, creating an expanding meaning space. Round 1 included only the eight scenes described in the group naming phase, which repeated for a total of 23 games. In the next round, three new scenes were added to the eight familiar ones, resulting in 11 different target scenes. These appeared in random order for a total of 23 games, with the constraint that each familiar scene was presented at least once and that new scenes were presented at least twice. In round 3 we again added three more new scenes to the existing 11, and randomized these 14 scenes to fill 23 games according to the same principle. This continued for all following rounds until there were exactly 23 different scenes in round 6, each appearing once without repetition. No more scenes were introduced in round 7, allowing participants to communicate about the entire meaning space more than once.

After the last communication round, each participant completed a test phase in which they were presented individually with all scenes in a

random order, and were asked to type their descriptions using the "Fantasy Language". After the test, participants also filled out a questionnaire about their performance in the experiment, including questions such as "Did you notice any structure in the scenes used in Fantasy Language?", and "Did you try to adopt your partner's language?". Finally, all participants were debriefed by the experimenter.

Results

We examined the artificial languages developed in this experiment according to four measures: (1) communicative success, (2) degree of convergence, (3) language stability, and (4) compositional structure. While our main goal was to examine the emergence of compositionality (captured by the last-mentioned measure), looking at each of the four measures separately enabled us to better characterize the emerging communication systems and to understand how different linguistic properties changed over time.

For all analyses reported in the paper, we used mixed effects regression models. Note that in these types of communication experiments, groups are treated as individual units, similar to single participants in traditional psychology experiments. All models were generated using the lme4 and pbkrtest packages in R (Bates, Maechler, Bolker & Walker, 2015; Halekoh & Højsgaard, 2014; R Core Team, 2016). The pbkrtest package provides p-value using the Kenward-Roger Approximation, which gives more conservative p-values for models based a relatively small number of observations. All models converged with the maximal random effects structure. Unless noted otherwise, this structure included random intercepts for each of the six groups and each of the 23 scenes, and random slopes for all fixed effects with respect to different groups and different scenes. We report the fixed effects structure of each model separately. The raw data and the code for running all analyses can be found at https://osf.io/wht86/.

Communicative success

Communicative success was measured as response accuracy during the communication phase. We used a logit mixed-effects regression model to predict accuracy (coded as 1 or 0) in a given turn. The fixed effects were ROUND NUMBER and ITEM CURRENT AGE (both centered). All items started

with an age of 1 (the first exposure), except for the eight scenes that were introduced in the naming phase, which started with an age of 2 (as we considered round 0 to be the first exposure). Therefore, ITEM CURRENT AGE codes the number of rounds a participant has been exposed to a specific scene until that point in the game, and measures the effect of familiarity with a given scene on performance. In contrast, ROUND NUMBER measures the effect of overall language proficiency and degree of shared history on performance. The model showed that participants became significantly more successful as rounds progressed (β=0.2, SE=0.06, z=3.1, p<0.01; see Table 1 and Figure 2). No other effect was significant.

Table 1: Accuracy model

	Estimate	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-0.273937	0.2174	-1.26	0.207
Item Current Age	-0.000381	0.0213	-0.018	0.985
Round Number	**0.202047**	**0.0651**	**3.1**	**0.001 ****

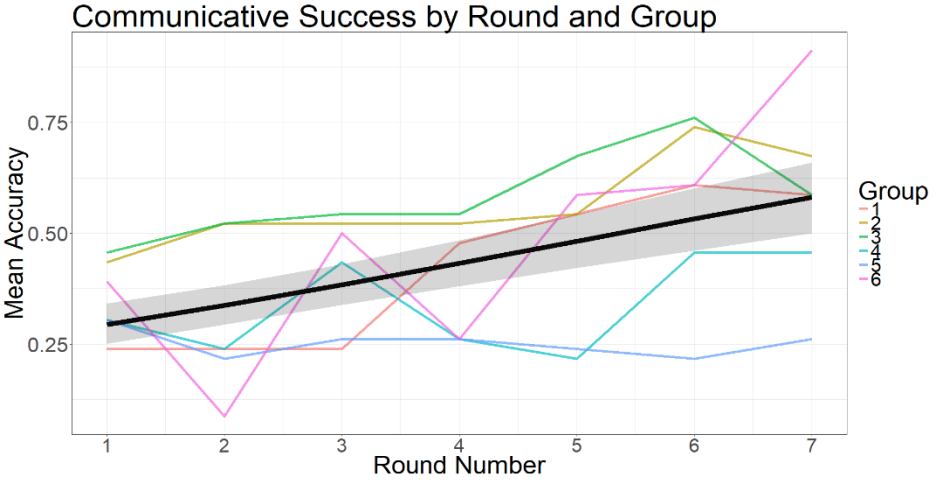


Figure 2: Summary statistics of mean accuracy by Round Number. The colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represents the model's standard error. Round Number ranged from 1 (the first communication round) to 7 (the last communication round).

Convergence

Convergence was measured by calculating the differences between the labels produced by different participants for the same scene in a given round: for each scene in round *n*, convergence was calculated by averaging over the normalized Levenshtein distances between all labels produced by different participants for that scene. The normalized Levenshtein distance between two strings is the minimal number of insertions, substitutions, and deletions of a single character that is required in order to turn one string into the other, divided by the number of characters in the longer string of the two. This distance was then subtracted from 1 to represent string similarity, reflecting the degree of shared lexicon in the group by examining how aligned participants were. Convergence was expected to increase over time so that different participants will use increasingly similar labels.

We used a mixed-effects linear regression model to predict convergence. The fixed effects were ROUND NUMBER and ITEM CURRENT AGE (both centered). The model showed a numeric increase in string similarities over rounds indicating an increase in convergence, but this was only marginal in our relatively conservative threshold for significance (β=0.02, SE=0.01, t=2, p=0.067; see Table 2 and Figure 3). No other effect was significant. The model thus suggests that the participants started developing a shared lexicon over time, and were marginally more converged as rounds progressed. Yet notably, participants were never fully aligned: even in the final round, the average similarity between labels produced by different participants for the same scenes was around 0.5 (see Figure 3), indicating that participants used labels which shared on average about half of their characters.

Table 2: Convergence model

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.38961	0.03813	10.218	< 0.001 ***
Item Current Age	0.0012	0.00476	0.2526	0.806
Round Number	**0.02655**	**0.01266**	**2.096**	**0.067 .**

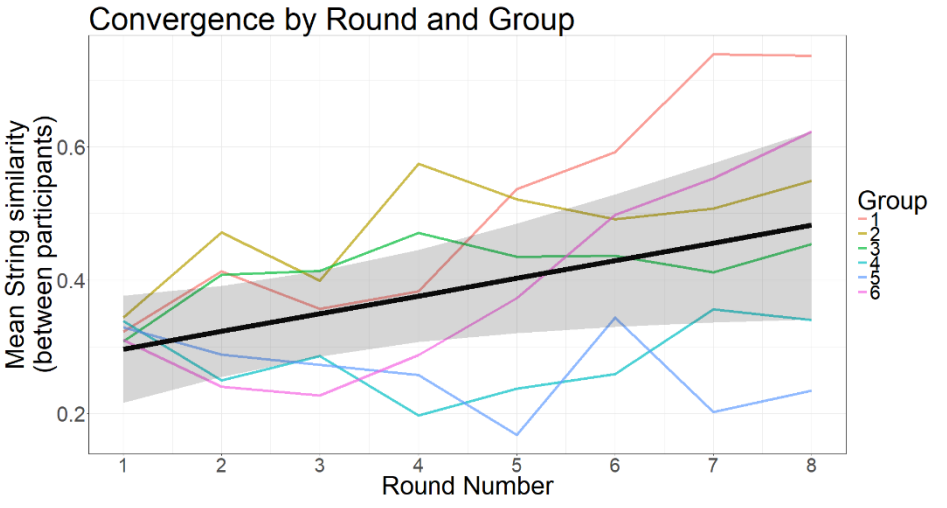


Figure 3: Summary statistics of mean convergence by Round Number. Higher string similarities between participants indicate greater convergence. The different colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represents the model's standard error. Round Number ranged from 1 (first communication round) to 8 (the final test round).

Stability

Languages' stability was measured by calculating the differences between the labels created by participants for the same scenes on consecutive rounds: for each scene in round *n*, stability was calculated by averaging over the normalized Levenshtein distances between all labels produced for that scene in round *n* and all labels produced for that scene in round *n+1*. This distance reflects the degree of change in participants' reproduction of the labels over time. Note that this parameter is referred to as "Learnability" in Kirby et al. (2008; 2015), since it reflected the degree of transmission errors between learned and produced labels in each generation in an iterated learning paradigm. Here, the string differences are not measured over consecutive generations of different learners, but rather over consecutive rounds of communication, with the same people producing the strings (and modifying them). This distance was then subtracted from 1 to represent string similarity, reflecting how consistent participants were in reproducing the labels over consecutive rounds. Since in our design participants were not asked to memorize and recall the scenes but rather use the label they find most effective, this parameter

indicates the degree of language stability (and not transmission fidelity). Stability was expected to increase over time as participants become more familiar with the language. We used a mixed-effects linear regression model to predict stability. The fixed effects were ROUND NUMBER and ITEM CURRENT AGE (both centered). The model showed a numeric increase in string similarities over rounds, such that stability marginally increased with time (β=0.028, SE=0.01, t=2.19, p=0.06; see Table 3 and Figure 4).

Table 3: Stability model

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.42706	0.03119	13.689	< 0.001 ***
Item Current Age	0.00215	0.00615	0.3497	0.735
Round Number	**0.02811**	**0.01279**	**2.1967**	**0.0609 .**

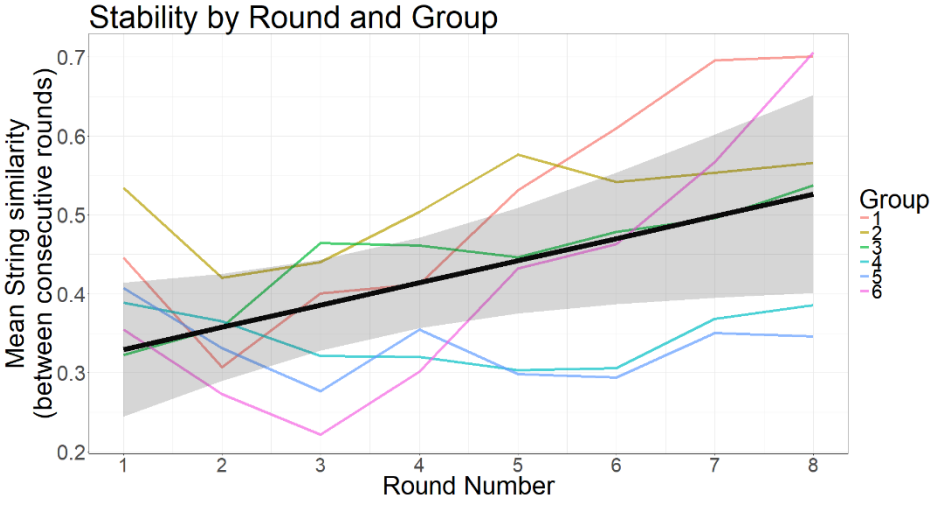


Figure 4: Summary statistics of mean stability by Round Number. Higher string similarity between consecutive rounds indicate greater stability. The different colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represents the model's standard error. Round Number ranged from 1 (a comparison of the first communication round to the naming round) to 8 (a comparison of the final test phase to the last communication round).

Interestingly, examining the rate of stabilization for scenes as they entered the game revealed that newer scenes stabilized faster (Figure 5). For example, scenes that entered the game in the second round had a stability score of 0.35, but scenes that entered the game in the third, fourth, fifth, and sixth round had scores of 0.38, 0.41, 0.47, and 0.49, respectively. That is, the later scenes entered the game, the less they changed, presumably because over time, participants have developed structured languages that provided a predictable and consistent way of describing new meanings. Thus, new labels are already coined in a manner that fits the structure of the language.

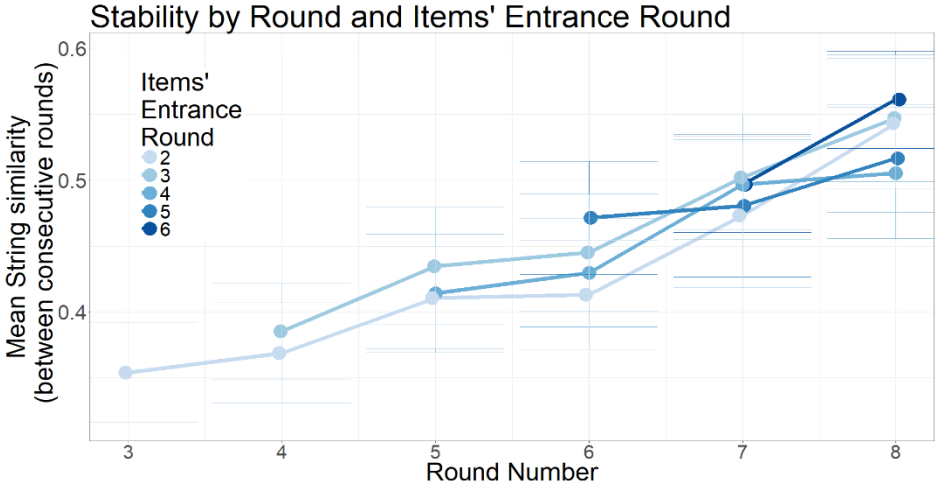


Figure 5: Summary statistics of mean stability by Round Number and Items' Entrance Round for all labels that were introduced after the initial round. Higher string similarities between consecutive rounds indicate greater stability. Items' Entrance Round reflects the point in time at which the item was introduced into the game, and ranged from 2 (the first items that entered the game in Round 2) to 6 (the last items that entered the game in Round 6). The blue hued lines represent the starting round of new labels, with darker hues for items that entered the game in a later stage. Round Number ranged from 3 (compared to Round 2) to 8 (the final test phase compared to the last communication round).

Compositional structure

Compositional structure was measured by calculating the correlations between labels' string distances and scenes' semantic distances in a given language. Semantic differences were calculated in the following way: first, scenes that differed in shape were given a difference score of 1, and scenes which contained the same shape were given a difference score of 0. Then, we calculated the absolute difference between scenes' angles, and divided it by the maximal possible distance between angles (180 degrees) to yield a continuous, normalized score between 0 and 1. Given that motion was a continuous dimension and that differences between angles are perceptually smaller than the categorical difference between shapes, shape was considered a perceptually favorable feature. Therefore, we treated the maximal difference in angles (180 degrees) in the same way as a difference between shapes. Finally, the difference scores for shape and angle were added. Semantic distances therefore ranged between 0.18 (the same shape moving in angles that are 10 degrees apart) and 2 (different shapes moving in angles that are 180 degrees apart). Labels' string distances were calculated using the normalized Levenshtein distances between all possible pairs of labels produced by participant p in round n, excluding pair-wise comparisons between labels produced for the same scene. The two sets of pair-wise distances (i.e., string distances and meaning distances) were then correlated using the Pearson product-moment correlation. This measure reflects the amount of structure in the mapping between words and meanings in different participants' languages over time, by examining the degree to which similar meanings are being expressed using similar strings.

In most iterated learning studies (e.g., Kirby et al. 2008; 2015), an increase in structure over time is demonstrated by an increase in the z-scores provided by the Mantel test for the correlations between meaning and string distances described above. However, this was problematic to do in the current design, since z-scores become larger as the number of observations increase. Since our meaning space was expanding over rounds, z-scores would have become inflated over rounds. Therefore, we chose to examine compositional structure by looking directly at the raw correlations. Running the analyses with z-scores rather than the raw correlation does not change the significance or direction of any of the reported effects.

It is also important to note that the structure measure used here and in Kirby et al. (2015) cannot differentiate between different types of

linguistic structures (e.g., compositionality vs. structured ambiguities, like in the case of systematic use of homonyms), and only indicates how much structure is present in the language. In previous iterated learning studies, evidence for compositionality (e.g., re-use of sub-strings) was based solely on individual examples of signal systems with such structures, as analyzed manually by the authors. Here, we also tried to justify our claim about the emergence of compositionality by using a segmentation algorithm developed by Stadler (under preparation), which provides statistical support for the systematic re-use for sub-strings in addition to subjective observations.

We used a mixed-effects linear regression model to predict the correlation between meanings and strings in participants' languages in a given round. Following Beckner et al. (2017), we included both the linear and the quadratic term for centered ROUND NUMBER. The model had random intercepts for producers nested within groups (but not for scenes, as structure score was calculated over all scenes in a given round), as well as by-producer random slops for the effect of ROUND NUMBER. The model showed that structure increased significantly over rounds (β=1.19, SE=0.1, t=4.4, $p<0.01$; see Table 5). The quadratic term for ROUND NUMBER was not significant, indicating that structure increased in a linear manner. The model thus confirmed that the languages in this experiment became significantly more compositional over time despite the lack of generational transmission. As Figure 6 shows, there was a high degree of compositional structure in this experiment, with some groups reaching correlations as high as 0.6.

Table 5: Structure model

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.44257	0.0375	11.78	< 0.001 ***
Round Number (linear)	**1.19169**	**0.2166**	**5.501**	**0.002 ** **
Round Number (quadratic)	-0.20625	0.1984	-1.039	0.341

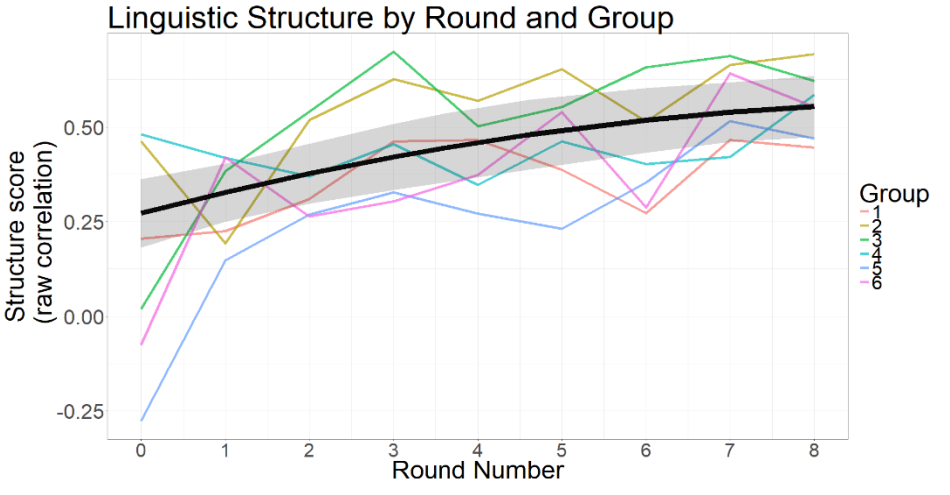


Figure 6: Summary statistics of the label-meaning correlations by Round Number. The different colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represent the model's standard error. Round number ranged from 0 (the group naming phase) to 8 (the final test phase).

Figure 7 illustrates one type of compositional structure that emerged in this experiment, using an example from the sixth group. For visualization purposes, we highlighted each meaningful sub-string in a different color, and added a "dictionary" to the language. This segmentation was statistically motivated by the mutual predictability segmentation algorithm (Stadler, under preparation), which looks at a given semantic dimension (e.g., shape) in the language of a given participant in the final test phase, and searches for non-overlapping sub-strings that co-occur with each of the different meanings. Then, it selects the sub-string that has the highest mutual predictability for each meaning, while merging different meanings if they are predicted by exactly the same string. This provides a new way to statistically confirm the existence of compositionality in artificial languages. Importantly, the segmentation algorithm identified all the sub-strings indicated in Figure 7[6].

[6] Since the label used to refer to Shape 2 had more variation in its final letters (i.e., "nena", "nenu"), the algorithm was able to recognize only part of the string as predictive (i.e,. "nen"). In addition, although the algorithm recognized all the relevant sub-strings for directions with a mutual predictability score of 1, this was not statistically significant for some directions (e.g., down; 270 degrees) due to the small number of scenes with this property.

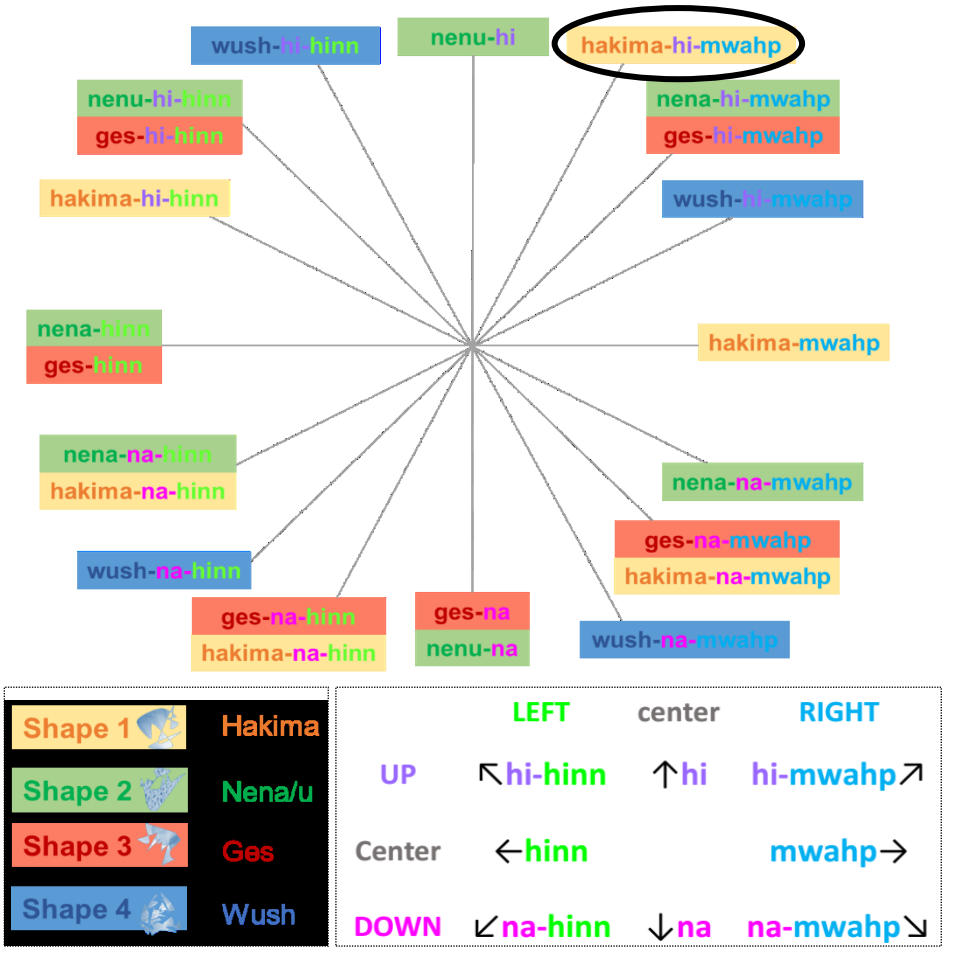


Figure 7: An example of a compositional language, produced in the final test phase by a participant in Group 6, along with a "dictionary". Different box colors represent the four different shapes which appeared in the scenes, and the grey axes indicate the direction in which the shape was moving on the screen. Different font colors represent different meaningful part-labels, as segmented by the authors for illustration purposes. For example, the label in the black circle ("hakima-hi-mwahp") was assigned to a scene in which shape 1 was moving in a 60º angle. It is comprised of several predictable parts: "hakima" indicates the type of shape which appeared in the scene, and the additional "hi-mwahp" indicates the type of motion (up-right). This latter part-label can also be decomposed to two meaningful parts: "hi" stands for "up" and "mwahp" stands for "right".

As can been seen in Figure 7, the language presented in this example distinguishes between the four shapes in a systematic way, with each shape represented by a unique prefix. For example, the segmentation

algorithm confirmed that the prefix "*wush*" was significantly associated with all labels for scenes with Shape 4, and with none of the other shapes (mutual predictability=1, $p<0.01$). Interestingly, some prefixes for shape (e.g., "*nenu*" and "*hakima*") originated from labels given during the naming phase to a specific scene with that shape. Over time, these strings spread to the rest of the group and were generalized to refer to all scenes containing that shape. Similar trajectories were observed in all groups. This process resembles the processes of Grammaticalization and semantic extension in natural languages, where specific lexical items can become functional markers over time, representing an entire class of items or events. Direction of motion was also systematically coded, with participants categorizing this continuous dimension into two orthogonal dimensions, horizontal and vertical: participants used one affix to encode right ("*mwahp*") vs. left ("*hinn*"), and another affix to encode up ("*hi*") vs. down ("*na*"). Participants combined these affixes in compositional ways to represent motion. For example, scenes that included a shape moving down-right (in 300, 315, or 330 degrees) were all given the suffix "*na-mwahp*" (mutual predictability=1, $p<0.01$).

Importantly, not all groups categorized angles in this way, and other types of categorization of the meanings space emerged, associated with different compositional structures. For example, Group 1 categorized scenes into seven prototypical directions which were each associated with a unique single-character suffix, and Group 4 used different orders and doubling of affixes to differentiate between directions. Interestingly, there were also cases in which motion affixes originated from a label given to a specific scene, which had a similar direction of movement.

Result Summary

The results of Study 1 show that groups became more accurate over the course of interactions, and developed languages that became increasingly stable, shared and structured over time. Importantly, as predicted, compositional structure emerged in closed groups even without generation turnover.

In the Introduction, we highlighted two mechanisms that may drive compressibility pressures in real language use and could lead to the emergence of compositionality during communication: (a) the need to interact with multiple people, and (b) the need to refer to and discriminate between more and more meanings over time. Since we wanted to

maximize the likelihood of compositional structure emerging, we included both pressures in our communication paradigm. Study 2 tries to tease apart these two pressures, and tests their individual role using a meta-analysis that included data from the six groups reported above, as well as data from 18 additional groups that were tested using an extended version of the same paradigm[7].

Study 2: Meta-analysis

In order to examine the unique contribution of our two communicative pressures, namely, interacting with multiple people and an expanding meaning space, we conducted a meta-analysis over data from 24 groups: the six groups reported in Study 1 above, and 18 additional groups that were tested using the same paradigm. All 18 additional groups played an extended version of the communication game, including eight additional rounds (seven more communication rounds + an additional test round). Of these 18 additional groups, six were small groups of four participants, and 12 were larger groups of eight participants. Below we report the details for these 18 additional groups.

Method

Participants

The meta-analysis includes data from a total of 144 adults: the 24 participants who took part in Study 1 (mean age: 23.2; 18 women), comprising of six small groups of four participants; and 120 additional participants who took part in the extended version (mean age: 24.9; 88 women), comprising a total of 6 small groups of four participants, and 12 larger groups of eight participants. Participants in Study 1 were paid between 20 and 26 Euros for their participation, depending on the amount of time they spent in the lab (ranging between 2:00 to 2:45 hours). Participants in the extended version were paid between 40 and 46 Euros for their participation, depending on the amount of time they spent in the lab (ranging between 4:30 to 5:15 hours, including a lunch break). All participants were native Dutch speakers and were recruited using the

[7] These 18 additional groups were run using the same paradigm to test other hypotheses (see Discussion) and are reported in Chapter 3 (Raviv, Meyer & Lev-Ari, 2019b). Importantly, this specific analysis is not reported anywhere else.

participant database of the Max Planck Institute for Psycholinguistics. Ethical approval was granted by the Faculty of Social Sciences of the Radboud University Nijmegen.

Stimuli

Identical to the stimuli used in Study 1.

Procedure

The additional participants played an extended version of the communication game reported in Study 1, in which the communication phase and the test phase were repeated for a second time. Importantly, this extended version had the same procedure, same settings and same instructions as in Study 1, and the first eight rounds were identical. Note that in the big groups, due to their larger size, implementing the same procedure led to each participant naming only one item in the naming phase, and for each pair interacting only half as many times as each pair in the small groups. The additional eight rounds also followed the same procedure as in the first eight rounds of Study 1, except for one difference: no new items were introduced after the first eight rounds. That is, the meaning space did not expand further in the additional rounds, and included all 23 scenes from Study 1 and only them.

After completing the first eight rounds, participants in the extended version had a lunch break (in which they were not allowed to talk about the experiment) and then reconvened to complete seven additional communication rounds (rounds 9-15) and an additional test round (round 16) in the same settings. Therefore, the extended version included 16 rounds in total, in three unique phases: a group naming phase (round 0), a communication phase (rounds 1-7, rounds 9-15) and a test phase (round 8, round 16).

Meta-analysis Results

Our meta-analysis was based on data from 24 groups: six small groups that played the short version (the original data reported in Study 1), six small groups that played the extended version, and 12 big groups that played the same extended version.

First, we replicated our findings that compositionality emerges during communication by running the same model employed in Study 1 over data from all 24 groups (see Appendix B). We found that, as predicted, there was a significant linear increase in linguistic structure over rounds whether we examined only the first eight rounds (β=4.65, SE=0.3, t=15.4, $p<0.001$), only the additional eight rounds (β=0.77, SE=0.2, t=3.7, $p<0.005$), or all 16 rounds together (β=5.6, SE=0.4, t=13.6, $p<0.001$). Notably, this increase in structure leveled off in later rounds: the quadratic term for ROUND NUMBER was significant during the first eight rounds (β=-0.74, SE=0.2, t=-3.6, $p<0.005$), and also when all rounds were taken into account (β=-2.4, SE=0.2, t=-11.1, $p<0.001$). Moreover, the effect of ROUND NUMBER was larger during the first eight rounds, as indicated by the effect sizes (i.e., the models' coefficients: 4.65 vs. 0.77). That is, most of the increase in structure happened in the first eight rounds, when the meaning space was still expanding and when participants experienced an increase in the number of partners. Together, these results consist a direct replication of the results we reported above for the six original groups in Study 1, and strengthen our conclusion that compositionality can indeed emerge in a closed group, without generation turnover. Moreover, they imply that our communicative pressures played a role.

Next, we examined the separate contribution of our two communicative pressures – multiple partners and an expanding meaning space – to the emergence of structure over time. To this end, we used mixed effects models similar to the one reported above to predict the structure scores at each round, with additional predictors for the NUMBER OF PARTNERS, the NUMBER OF SCENES, or both. First, we ran separate models that added to the model with ROUND NUMBER only one of the additional factors as a predictor. Then, we ran a full model that added both new predictors to the model, and compared the separate reduced models to the full model using model comparisons (likelihood ratio tests). This allowed us to examine the contribution of each additional predictor.

All models included a centered fixed effect for ROUND NUMBER (ranging from 0 to 16 before centering, linear and quadratic terms), and had the same random effects structure which included random intercepts and random slopes for the effect of ROUND NUMBER with respect to different participants nested in different groups. In the separate models, we included either a fixed effect for the NUMBER OF SCENES participants were exposed to so far (ranging between 8 to 23 scenes before centering), or a fixed effect for the NUMBER OF PARTNERS participants interacted with so far (ranging between 1 and 3 for the small groups and between 1 and 7

for the larger groups before centering). In the full model, all predictors were included. Even though these predictors are closely related, the maximal Variance Inflation Factor (VIF) for all predictors in all models was <6, indicating that the collinearity of these models was acceptable (see Kennedy, 1992; Hair, Anderson, Tatham & Black, 1995).

All models showed both linear and quadratic effects of ROUND NUMBER, indicating an increase in structure over rounds that leveled off in later rounds. Moreover, the separate models showed that both factors were significant positive predictors of linguistic structure on their own (see Appendix B): NUMBER OF PARTNERS, had a strong effect on structure (β=0.04, SE=0.005, t=7.05, p<0.001), and the NUMBER OF SCENES did too albeit with a smaller effect size (β=0.007, SE=0.002, t=2.9, p<0.01). Importantly, the full model was favored compared to the model that included only the NUMBER OF SCENES (ΔAIC = 26, p<0.0001), but was similar to a model that included only the NUMBER OF PARTNERS (ΔAIC = 2, p=0.96). Thus, this model comparison showed that NUMBER OF PARTNERS improved the model, while NUMBER OF SCENES did not add a unique contribution. In support of this finding, the full model showed that interacting with multiple people had a strong positive effect on structure scores, while the expanding meaning space did not (Table 6; Figure 8): When all factors were included in the model, structure scores significantly increased with the NUMBER OF PARTNERS (β=0.04, SE=0.006, t=6.37, p<0.001) but not with the NUMBER OF SCENES (β=0.0001, SE=0.002, t=0.04, p=0.96). Together, these results suggest that interacting with multiple people introduces a stronger pressure for compositionality than an expanding meaning space, and was the main driver for the emergence of compositionality in our design.

Table 6: Meta-Analysis Model

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.535	0.018	28.891	< 0.001 ***
No. of Scenes	0.0001	0.002	0.0443	0.9646
No. of Partners	**0.0407**	**0.006**	**6.3739**	**< 0.001 *** **
Round Number (linear)	**3.1029**	**0.642**	**4.8317**	**< 0.001 *** **
Round Number (quadratic)	**-0.8866**	**0.447**	**-1.9824**	**0.0481 * **

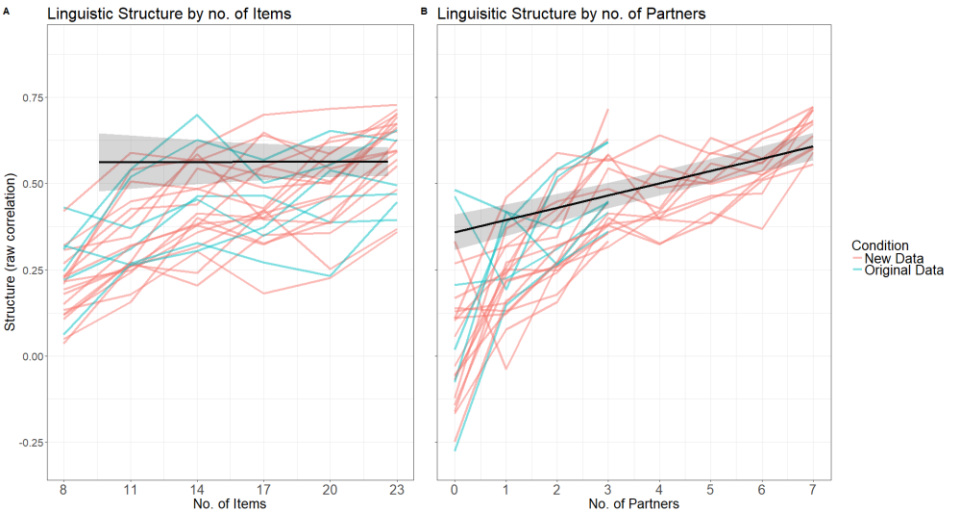


Figure 8: Summary statistics of structure score by the number of items (A) and the number of partners (B) to which participants were exposed. The colored lines represent the different groups in the meta-analysis. The black line represents the models' estimate, and its shading represent the models' standard error. The number of items ranged from 8 (during the group naming phase and round 1) to 23 (from round 6 onwards). The number of partners ranged from 0 (during the group naming phase) to 3 (for small groups) or 7 (for big groups).

General Discussion

In this paper we tested whether compositionality, one of the hallmarks of natural language, can emerge during communication given compressibility pressures other than learning by new generations. Kirby et al. (2015) argued that cross-generational transmission is crucial for the emergence of compositionality. Here, we hypothesized that properties of real-world communication, namely, interacting with multiple people on an expanding meaning space, could impose compressibility pressures that would lead to the emergence of compositional languages already in a single generation. We predicted that the need to converge with different partners and the need to refer to more and more meanings over time would give rise to structured, compositional languages during communication in closed groups.

To examine this claim, we tested six micro-societies of four participants each, who communicated in alternating pairs using an artificial language to refer to an expanding meaning space. We found that the languages developed in our micro-societies became significantly more structured over rounds of interaction, and developed compositional structure despite the absence of generational transmission. In particular, the micro-societies in our experiment developed languages in which different affixes were systematically combined to express different meanings. Additionally, those languages became more shared, more consistent, and more communicatively successful across rounds. Participants converged on stable and structured lexicons that allowed them to refer to new meanings with increasing efficiency: as languages became more structured, labels for new scenes became more predictable and stabilized faster. Our findings show that compositionality reliably emerges during communication without generational turnover, and advances our understanding of how communal interaction shapes grammatical structure in the process of language evolution and language change. We also conducted a meta-analysis with data from 18 additional micro-societies of four or eight participants, which replicated our main finding and extended it to groups of varying sizes: the additional groups also showed a significant increase in linguistic structure during multiple communication rounds, and developed compositionality without transmission to new learners. Thus, we have expanded on the theory brought forth in Kirby et al. (2015) by showing that natural properties of language use other than learning by new members can give rise to strong compressibility pressures during communication and therefore to compositional structure within a single generation.

One immediate implication of these findings is that compositionality can emerge in a linguistic signal system within the first generation, with no new learners needed. At first glance, these claims seem to be in conflict with the conclusions drawn from studies on developing sign languages and creoles, which stress the role of new learners in the formation of linguistic structure in the real world (Aronoff, Meir, Padden & Sandler, 2008; Bickerton, 1984; Senghas & Coppola, 2001; Senghas, Kita & Özyürek, 2004). However, developing sign languages and pidgins clearly show evidence of sentence-level compositionality in the first generation, as speakers re-use small units (i.e., words or gestures) to create sentences and refer to complex events. For example, over a fifth of the signers in the first cohort of the developing Nicaraguan Sign Language showed compositionality in representing manner and path of motion, and all first

cohort signers were able to recombine different signs to form sentences (Senghas et al., 2004). Moreover, compositionality at the sentence-level is present already in home-sign (Goldin-Meadow & Mylander, 1990), as well as in pidgin languages (Arends & Bruyn, 1994). What seems to change in languages over the course of generations is not the presence of compositionality per se, but rather the degree of its regularity (e.g., word order) and the degree of more fine-grained compositionality at the word-level (e.g., morphology). In our miniature language, there is no meaningful difference between sentence-level and word-level compositionality: descriptions in our paradigm could be interpreted as single words with different affixes, or alternatively as different words combined to a form a sentence (e.g., with a noun describing shape and a verb describing motion). Thus, our conclusions are in line with findings from developing sign languages, which also show that compositionality exists from very early stages.

A possible limitation of our study is that it is based on the behavior of adult participants rather than children, who may differ from adults in their biases and general cognitive skills. However, this limitation is relatively weak for several reasons. First, while children are indeed the prototypical majority of languages *learners* in real-world settings, they are not the prototypical majority of language *users*. As such, adults have been argued to play a larger role in the process of language innovation and change compared to children, given that they typically have a stronger social influence in the society (Labov, 2007; Nettle, 1999; Roberts & Winters, 2012). Second, the same cognitive principles outlined here (i.e., memory limitations; the need to communicate successfully) are likely to generalize to children as well. For example, children as young as four already adapt to their interlocutors by taking over structural and lexical forms used by their dialogue partners (e.g., Nilsenová & Nolting, 2010). Moreover, younger children are theoretically faced with an even stronger pressure for compressibility given their inferior working memory (e.g., Gathercole, Pickering, Ambridge & Wearing, 2004). Finally, a recent study compared children and adults' performance on an iterated language learning paradigm (similar to that used in Kirby et al., 2008), and found that children, like adults, can create linguistic structure in artificial languages (Raviv & Arnon, 2018). While adults significantly outperformed children in all experiments, children were able to create languages with simple systematic structures similar to those created by adults and in Kirby et al. (2008). Even though children did not introduce compositionality in that paradigm, Raviv & Arnon (2018) argue that children do not have

qualitatively different structural biases compared to the adults, and show that this difference can be attributed to children's worse learning overall. This study therefore suggests that our findings could be generalized to children (or more naturally, to mixed groups of children and adults).

Importantly, our meta-analysis tested the relative contribution of the two communicative pressures in our design, and revealed that having multiple interaction partners introduced a stronger compressibility pressure than the expanding meaning space. While both factors were significant predictors of structure individually, the expanding meaning space did not introduce an additional compressibility pressure beyond the pressure introduced by the number of interaction partners. In other words, while the need to discriminate between more and more items can lead to the emergence of more systematic structure (see also Nölle et al., 2018), it seems to be less crucial when another strong pressure for compressibility (i.e., interacting with multiple partners) already exists. Together, this meta-analysis showed that interacting with multiple people played a central role in shaping this pattern of results, and could be considered as the main driver for the emergence of compositionality in this paradigm. It is possible that a more extreme manipulation of the expanding meaning space would yield a stronger compressibility pressure. Future work could experimentally examine the emergence of compositionality when only one of these pressures is present, or use a computational model similar to the one used in Kirby et al. (2015) to examine the lower bound of each pressure and tweak the extent to which new meanings and new partners are introduced.

One possible implication that can be drawn from these findings is that cross-cultural differences in interaction patterns (e.g., group size) can affect the formation of linguistic structure: given the strong effect of having multiple communication partners, we predict that increasing the number of communication partners (and therefore the degree of input variability) will impose a stronger pressure for systemization and generalization, and should therefore result in languages with more linguistic structure. This prediction resonates with models of language evolution and language change: an increase in community size is argued to be one of the main drivers for the evolution of natural language (Dunbar, 1993), and interaction with multiple people is argued to promote the simplification of morphological structure (e.g., Wray & Grace, 2007). Moreover, this idea is supported by typological studies showing that languages spoken by more people have more transparent and more regular structures (e.g., Lupyan & Dale, 2010), and by computational models that

predict community size to have dramatic effects on linguistic structure (e.g., Dale & Lupyan, 2012; Reali & Griffiths, 2009). Our paradigm provides an efficient way to test the emergence of compositional languages with larger groups of interlocutors in laboratory settings, allowing for the manipulation of features such as group size and community structure. In the following chapters we experimentally examined how differences in population size and network configuration may affect the emergence of compositionality.

Conclusion

The results of the experiment and the meta-analysis show that languages can develop compositional structure over the course of communication, even in the absence of generational transmission to new learners. In particular, we found that when groups of participants interacted with multiple partners, their languages became more compositionally structured, more stable and more communicatively successful over time. This is the first demonstration that compositionality can reliably emerge in an artificial language in a closed-group setting and supports the idea that compressibility pressures can be imposed during communication.

Acknowledgments

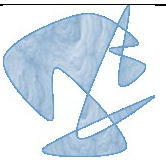
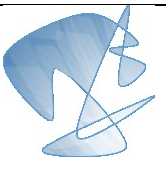
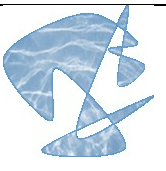
I wish to thank Caitlin Decuyper for programming the experiment, and Gary Lupyan, Sean Roberts, and Kevin Stadler for discussions and helpful input.

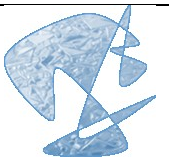
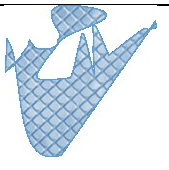
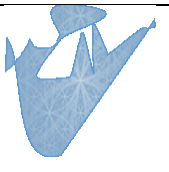
Appendix A: Materials

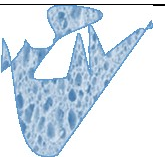
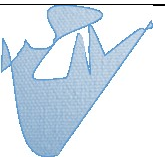
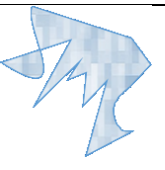
For each version of the stimuli, 23 scenes were created in the following way:

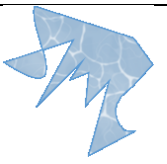
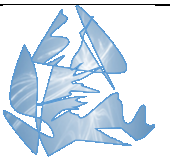
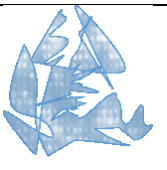
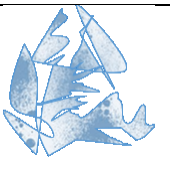
First, we created 28 static items (see *initial set*). The *initial set* contained exactly seven tokens of each of the four novel shapes, and each item was associated with a unique blue-hued fill pattern. For each version, 23 items were randomly drawn from the 28 static items in the *initial set,* with the constraint that each type of shape should appear between four to seven times. Then, each of these 23 static items was associated with an angle in order to create a moving scene. Angles were randomly selected from a set of 16 salient angles within the 360-degree-range (0º, 30º, 45º, 60º, 90º, 120º, 135º, 150º, 180º, 210º, 225º, 240º, 270º, 300º, 315º, 330º), following the constraint that in each version, each type of shape had to be associated with at least one angle from each of the four quadrants. The rest of the items' angles were randomly drawn from this set of angles.

Initial Set:

Item Number	Shape	Stimuli
1	1	
2	1	
3	1	
4	1	

5	1	
6	1	
7	1	
8	2	
9	2	
10	2	
11	2	
12	2	

13	2	
14	2	
15	3	
16	3	
17	3	
18	3	
19	3	
20	3	

21	3	
22	4	
23	4	
24	4	
25	4	
26	4	
27	4	
28	4	

Version 3:

Item Number	Angle
1	150
2	
3	360
4	210
5	60
6	240
7	315
8	270
9	90
10	210
11	45
12	135
13	330
14	180
15	45
16	135
17	
18	180
19	240
20	315
21	270
22	
23	300
24	
25	120
26	225
27	
28	30

Version 2:

Item Number	Angle
1	150
2	30
3	210
4	90
5	300
6	360
7	240
8	120
9	45
10	225
11	300
12	
13	180
14	330
15	60
16	180
17	315
18	
19	
20	
21	240
22	90
23	135
24	210
25	45
26	270
27	
28	330

Version 1:

Item Number	Angle
1	210
2	270
3	90
4	330
5	30
6	300
7	135
8	
9	270
10	150
11	330
12	30
13	
14	120
15	
16	45
17	315
18	
19	135
20	240
21	360
22	60
23	360
24	180
25	315
26	
27	120
28	225

Appendix B: Additional models reported in the Meta-Analysis

Models predicting Structure by round for all 24 groups (similar to the model of linguistic structure used in Study 1 and detailed in Table 5):

First 8 rounds

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.438	0.0178	24.555	< 0.001 ***
Round Number (linear)	**4.6564**	**0.3011**	**15.462**	**< 0.001 *****
Round Number (quadratic)	**-0.7435**	**0.2056**	**-3.615**	**0.0013 ** **

Last 8 rounds

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.6194	0.0263	23.544	< 0.001 ***
Round Number (linear)	**0.7719**	**0.2078**	**3.713**	**0.0015 ** **
Round Number (quadratic)	-0.0308	0.1589	-0.194	0.847

All 16 rounds

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.512	0.0191	26.719	< 0.001 ***
Round Number (linear)	**5.6481**	**0.4141**	**13.636**	**< 0.001 *****
Round Number (quadratic)	**-2.4717**	**0.2217**	**-11.147**	**< 0.001 *****

Reduced separate models predicting linguistic structure by only one of the communicative pressures:

Model for the effect of the number of different partners

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.5358	0.0184	29.01	< 0.001 ***
Round Number (linear)	**3.1189**	**0.4961**	**6.286**	**< 0.001 *** **
Round Number (quadratic)	**-0.8992**	**0.3302**	**-2.722**	**0.0078 ** **
Number of Partners	**0.0409**	**0.0058**	**7.052**	**< 0.001 *** **

Model for the effect of the number of different scenes:

	Estimate	**Std. Error**	**t-value**	***p*-value**
(Intercept)	0.5107	0.0192	26.522	< 0.001 ***
Round Number (linear)	**4.0289**	**0.6829**	**5.899**	**< 0.001 *** **
Round Number (quadratic)	**-1.383**	**0.4261**	**-3.245**	**0.0013 ** **
Number of Scenes	**0.0076**	**0.0025**	**2.998**	**0.003 ** **

3 Larger communities create more systematic languages

Abstract[8]

Understanding world-wide patterns of language diversity has long been a goal for evolutionary scientists, linguists and philosophers. Research over the past decade suggested that linguistic diversity may result from differences in the social environments in which languages evolve. Specifically, recent work found that languages spoken in larger communities typically have more systematic grammatical structures. However, in the real world, community size is confounded with other social factors such as network structure and the number of second languages learners in the community, and it is often assumed that linguistic simplification is driven by these factors instead. Here we show that in contrast to previous assumptions, community size has a unique and important influence on linguistic structure. We experimentally examine the live formation of new languages created in the lab by small and larger groups, and find that larger groups of interacting participants develop more systematic languages over time, and do so faster and more consistently than small groups. Small groups also vary more in their linguistic behaviors, suggesting that small communities are more vulnerable to drift. These results show that community size predicts patterns of language diversity, and suggest that an increase in community size might have contributed to language evolution.

[8] This chapter is based on *Raviv, L., Meyer, A. S., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. Proceedings of the Royal Society B: Biological Sciences, 286(1907): 20191262. doi:10.1098/rspb.2019.1262.*

Introduction

Almost 7,000 languages are spoken around the world (Dryer & Haspelmath, 2017; Lewis, Simons & Fennig, 2017), and the remarkable range of linguistic diversity has been studied extensively (Evans & Levinson, 2009; Maffi, 2005). Current research focuses on understanding the sources for this diversity, and attempts to understand whether differences between languages can be predicted by differences in their environments (Bentz & Winter, 2013; Everett, 2013; Everett, Blasi & Roberts, 2015; 2016; Lupyan & Dale, 2010; 2016; Nettle, 2012). If languages evolved as a means for social coordination (Beckner et al., 2009; Fusaroli & Tylén, 2012), they are bound to be shaped by their social environment and the properties of the cultures in which they evolved. Indeed, cross-linguistic and historical studies have suggested that different linguistic structures emerge in different societies depending on their size, network structure, and the identity of their members (Lupyan & Dale, 2010; Meir, Israel, Sandler, Padden & Aronoff, 2012; Milroy & Milroy, 1985; Nettle, 1999; Trudgill, 2002; Wray & Grace, 2007).

One social property, community size, might play a particularly important role in explaining grammatical differences between languages. First, an increase in human group size was argued to be one of the drivers for the evolution of natural language (Dunbar, 1993). Second, cross-linguistic work that examined thousands of languages found that languages spoken in larger communities tend to be less complex (Lupyan & Dale, 2010). Specifically, these languages have fewer and less elaborate morphological structures, fewer irregulars, and overall simpler grammars (Lupyan & Dale, 2010). In addition to shaping grammar, community size could affect trends of convergence and stability during language change (Meir et al., 2012; Milroy & Milroy, 1985; Nettle, 1999; Trudgill, 2002; Wray & Grace, 2007).

While there is correlational evidence for the relation between community size and grammatical complexity, cross-linguistic studies cannot establish a causal link between them. Furthermore, the relationship between bigger communities and linguistic simplification can be attributed to other social factors that are confounded with community size in the real world. In particular, bigger communities tend to be more sparsely connected, more geographically spread out, have more contact with outsiders, and have a higher proportion of adult second language learners (Meir et al., 2012; Trudgill, 2002; Wray & Grace, 2007). Each of these factors may contribute to the pattern of reduced complexity, and thus

provide an alternative explanation for the correlation between community size and linguistic structure (Bentz & Winter, 2013; Dale & Lupyan, 2012; Lou-Magnuson & Onnis, 2018; Lupyan & Dale, 2010; 2016; Nettle, 2012). In fact, many researchers assume that this correlation is accounted for by the proportion of second language learners in the community (Bentz & Winter, 2013; Dale & Lupyan, 2012; Lupyan & Dale, 2010; 2016) or by differences in network connectivity (Milroy & Milroy, 1985; Trudgill, 2002; Wray & Grace, 2007; Lou-Magnuson & Onnis, 2018; See discussion).

Here we argue that community size has a unique and casual role in explaining linguistic diversity, and show that it influences the formation of different linguistic structures in the evolution of new languages. Interacting with more people reduces shared history and introduces more input variability (i.e., more variants), which individuals need to overcome before the community can reach mutual understanding. Therefore, interacting with more people can favor systematization by introducing a stronger pressure for generalizations and transparency. That is, larger communities may be more likely to favor linguistic variants that are simple, predictable, and structured, which can in turn ease the challenge of convergence and communicative success. Supporting this idea, language learning studies show that an increase in input variability (i.e., exposure to multiple speakers) boosts categorization, generalization, and pattern detection in infants and adults (Bradlow & Bent, 2008; Gómez, 2002; Lev-Ari, 2016; 2018; Lively, Logan & Pisoni, 1993; Perry, Samuelson, Malloy & Schiffer, 2010; Rost & McMurray, 2009; 2010).

While existing studies cannot establish a causal link between community structure and linguistic structure or isolate the role of community size, teasing apart these different social factors has important implications for our understanding of linguistic diversity and its origins (Scott-Phillips & Kirby, 2010). Some computational models attempted to isolate the effect of community size on emerging languages using populations of interacting agents, but their results show a mixed pattern: while some models suggest that population size plays little to no role in explaining cross-linguistic patterns (Gong, Baronchelli, Puglisi & Loreto, 2012; Lou-Magnuson & Onnis, 2018; Wichmann & Holman, 2009), others report strong associations between population size and linguistic features (Reali, Chater & Christiansen, 2018; Spike, 2017; Vogt, 2009).

To date, no experimental work has examined the effect of community size on the emergence of language structure with human participants,

although it was suggested several times (Galantucci & Garrod, 2011; Gong, Shuai & Zhang, 2014; Roberts & Winters, 2012). We fill this gap by conducting a behavioral study that examines the live formation of new communicative systems created in the lab by small or larger groups. A couple of previous studies investigated the role of input variability, one of our hypothesized mechanisms, using an individual learning task, yet found no effect of learning from different models (Atkinson, Kirby & Smith, 2015; Atkinson, Smith & Kirby, 2018). Another related study compared the complexity of English descriptions produced for novel icons by two or three people, but reported no differences between the final descriptions of dyads and triads (Atkinson, Mills & Smith, 2018). These studies, however, did not test the emergence of systematic linguistic structure. Here we examine how group size influences the emergence of compositionality in a new language, and assess the role of input variability in driving this effect. In addition to examining changes in linguistic structure over time, we track other important aspects of the emerging systems (e.g., communicative success and the degree to which languages are shared across participants), shedding light on how community size affects the nature of emerging languages.

The Current Study

We used a group communication paradigm inspired by (Fay, Garrod, Roberts & Swoboda, 2010; Kirby, Cornish & Smith, 2008; Kirby, Tamariz, Cornish & Smith, 2015; Roberts, 2010; Roberts & Galantucci, 2012; Raviv, Meyer & Lev-Ari, 2019a) to examine the performance of small and larger micro-societies (See Figure 1; Appendix A). Participants interacted in alternating pairs with the goal of communicating successfully using only an artificial language they invented during the experiment. In each communication round, paired partners took turns in describing novel scenes of moving shapes, such that one participant produced a label to describe a target scene, and their partner guessed which scene they meant from a larger set of scenes. Participants in small and larger groups had the same amount of interaction overall, but members of larger groups had less shared history with each other by the end of the experiment. All other group properties (e.g., network structure) were kept constant across conditions.

We examined the emerging languages over the course of the experiment using several measurements (see Measures):

(1) Communicative Success; (2) Convergence, reflecting the degree of alignment in the group (3) Stability, reflecting the degree of change over time; and (4) Linguistic Structure, reflecting the degree of systematic mappings in the language. With these measures, we can characterize the emerging communication systems and understand how different linguistic properties change over time depending on community size.

Our main prediction was that larger groups would create more structured languages, given that they are under a stronger pressure for generalization due to increased input variability and reduced shared history. We also predicted that larger groups would show slower rates of stabilization and convergence compared to smaller groups. Furthermore, we ran analyses to test our proposed mechanism, namely, that larger groups create more structured languages because of greater input variability and reduced shared history.

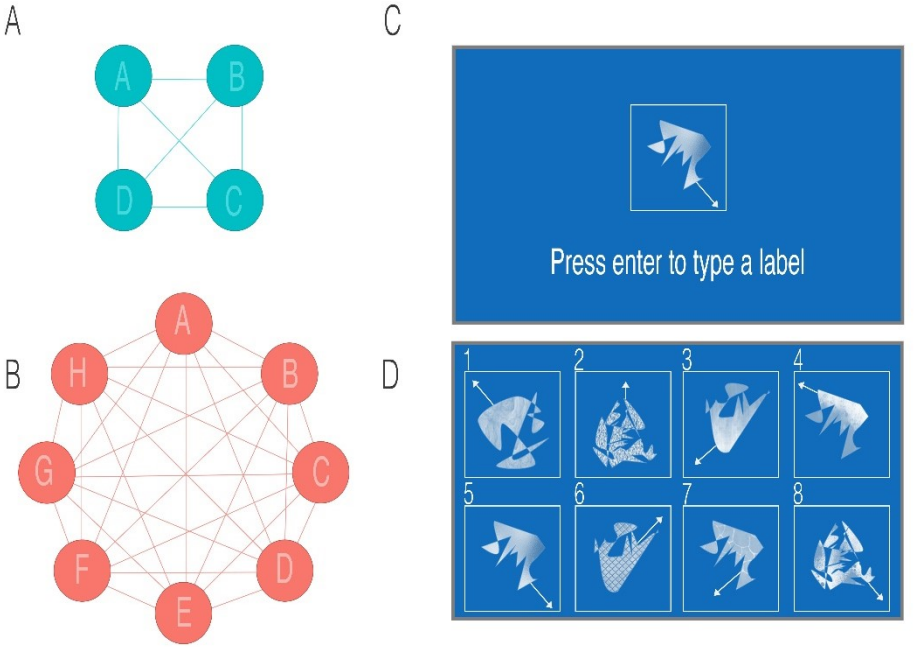


Figure 1. Group communication paradigm. We tested fully-connected groups of either four (A) or eight (B) participants. Panels (C) and (D) show the producer's and guesser's screens, respectively.

Methods

Participants

Data from 144 adults (mean age=24.9y, SD=8.9y; 103 women) was collected over the period of one year in several batches, comprising 12 small groups of four members and 12 larger groups of eight members. Participants were paid 40€ or more depending on the time they spent in the lab (between 270 to 315 minutes, including a 30-minutes break). Six additional small groups took part in a shorter version of the experiment (Raviv, Meyer & Lev-Ari, 2019a), which included only eight rounds. These additional groups showed similar patterns of results when compared to the larger groups. Their results are reported in Appendix B. All participants were native Dutch speakers. Ethical approval was granted by the Faculty of Social Sciences of the Radboud University Nijmegen.

Materials

We created visual scenes that varied along three semantic dimensions: shape, angle of motion, and fill pattern (see also Kirby, Cornish & Smith, 2008; Kirby, Tamariz, Cornish & Smith, 2015; Raviv, Meyer & Lev-Ari, 2019a). Each scene included one of four novel shapes, moving repeatedly in a straight line from the center of the frame in an angle chosen from a range of possible angles. The four shapes were unfamiliar and ambiguous in order to discourage labeling with existing words. Angle of motion was a continuous feature, which participants could have parsed and categorized in various ways. Additionally, the shape in each scene had a unique blue-hued fill pattern, giving scenes an idiosyncratic feature. Therefore, the meaning space promoted categorization and structure along the dimensions of shape and motion, but also allowed participants to adopt a holistic, unstructured strategy where scenes are individualized according to their fill pattern. There were three versions of the stimuli, which differed in the distribution of shapes and their associated angles (see Appendix A in Chapter 2). Each version contained 23 scenes and was presented to two groups in each condition. The experiment was programmed using Presentation.

Procedure

Participants were asked to create a fantasy language and use it in order to communicate about different novel scenes. Participants were not allowed

to communicate in any other way besides typing, and their letter inventory was restricted: it included a hyphen, five vowel characters (a,e,i,o,u) and ten consonants (w,t,p,s,f,g,h,k,n,m), which participants could combine freely.

The experiment had 16 rounds, comprising three phases: group naming (round 0), communication (rounds 1-7; rounds 9-15), and test (round 8; round 16).

In the naming phase (round 0), participants generated novel nonsense words to describe eight initial scenes, so that each group had a few shared descriptions to start with. Eight scenes were randomly drawn from the set of 23 scenes (see Materials) under the constraint that each shape and quadrant were represented at least once. During this phase, participants sat together and took turns in describing the scenes, which appeared on a computer screen one by one in a random order. Participants in larger groups named one scene each, and participants in small groups naming two scenes each. Importantly, no use of Dutch or any other language was allowed. An experimenter was present in the room throughout the experiment to ensure participants did not include known words. Once a participant had typed a description for a scene, it was presented to all group members for several seconds. This procedure was repeated until all scenes had been named and presented once. In order to establish shared knowledge, these scene-description pairings were presented to the group twice more in a random order.

Following the naming phase, participants played a communication game (the communication phase): the goal was to earn as many points as possible as a group, with a point awarded for every successful interaction. The experimenter stressed that this was not a memory game, and that participants were free to use the labels produced during the group naming phase, or create new ones. Paired participants sat on opposite sides of a table facing each other and personal laptop screens (see Appendix A). During this phase, group members exchanged partners at the start of every round, such that by end of the experiment, each pair in the small group has interacted at least four times and each pair in the large group has interacted exactly twice.

In each communication round, paired participants interacted 23 times, alternating between the roles of producer and guesser. In each interaction, the producer saw the target scene on their screen (see Fig. 1C) and typed a description using their keyboard. The guesser saw a grid of eight scenes on their screen (the target and seven distractors), and had to press the

number associated with the scene they thought their partner referred to. Participants then received feedback on their performance.

The number of target scenes increased gradually over the first six rounds, such that participants referred to more scenes in later rounds. While round 1 included only the eight initial scenes selected for the group naming phase, three new scenes were added in each following round until there were 23 different scenes in round 6. No more scenes were introduced afterwards, allowing participants to interact about all scenes for the following rounds. This method was implemented in order to introduce a pressure for developing structured and predictable languages (47), and resembles the real world with its unconstrained meaning space.

After the seventh communication round, participants completed an individual test phase (round 8), in which they typed their descriptions for all scenes one by one in a random order. After the test, participants had seven additional communication rounds (rounds 9-15) and the additional test round (round 16). These two individual test rounds allowed us to get a full representation of participants' entire lexicon at the middle and end of the experiment. Finally, participants filled out a questionnaire about their performance and were debriefed by the experimenter.

Due to a technical error, one large group played only six additional communication rounds instead of seven. Additionally, data from one participant in a large group was lost. The existing data from these groups was included in the analyses.

Measures

Communicative Success

Measured as binary response accuracy in a given interaction during the communication phase, reflecting comprehension.

Convergence

Measured as the similarities between all the labels produced by participants in the same group for the same scene in a given round: for each scene in round *n*, convergence was calculated by averaging over the normalized Levenshtein distances between all labels produced for that scene in that round. The normalized Levenshtein distance between two strings is the minimal number of insertions, substitutions, and deletions of

a single character that is required for turning one string into the other, divided by the number of characters in the longer string. This distance was subtracted from 1 to represent string similarity, reflecting the degree of shared lexicon and alignment in the group.

Stability

Measured as the similarities between the labels created by participants for the same scene on two consecutive rounds: for each scene in round *n*, stability was calculated by averaging over the normalized Levenshtein distances between all labels produced for that scene in round *n* and round *n+1*. This distance was subtracted from 1 to represent string similarity, reflecting the degree of consistency in the groups' languages.

Linguistic Structure

Measured as the correlations between string distances and semantic distances in each participant's language in a given round, reflecting the degree to which similar meanings are expressed using similar strings (Kirby, Cornish & Smith, 2008; Kirby, Tamariz, Cornish & Smith, 2015; Raviv, Meyer & Lev-Ari, 2019a). First, scenes had a semantic difference score of 1 if they differed in shape, and 0 otherwise. Second, we calculated the absolute difference between scenes' angles, and divided it by the maximal distance between angles (180 degrees) to yield a continuous normalized score between 0 and 1. Then, the difference scores for shape and angle were added, yielding a range of semantic distances between 0.18 and 2. Finally, labels' string distances were calculated using the normalized Levenshtein distances between all possible pairs of labels produced by participant *p* for all scenes in round *n*. For each participant, the two sets of pair-wise distances (i.e., string distances and meaning distances) were correlated using the Pearson product-moment correlation. While most iterated learning studies use the z-scores provided by the Mantel test for the correlation described above 43,44), z-scores were inappropriate for our design since they increase with the number of observations, and our meaning space expanded over rounds. Therefore, we used the raw correlations between meanings and strings as a more accurate measure of systematic structure (Raviv, Meyer & Lev-Ari, 2019a; Spike, 2016).

Input Variability

Measured as the minimal sum of differences between all the labels produced for the same scene in a given round. For each scene in round *n*, we made a list of all label variants for that scene. For each label variant, we summed over the normalized Levenshtein distances between that variant and all other variants in the list. We then selected the variant that was associated with the lowest sum of differences (i.e., the 'typical' label), and used that sum as the input variability score for that scene, capturing the number of different variants and their relative difference from each other. Finally, we averaged over the input variability scores of different scenes to yield the mean variability in that round.

Shared History

Measured as the number of times each pair in the group interacted so far, reflecting the fact that members of small groups interacted more often with each other. In small groups, pairs interacted once by round 3, twice by round 6, three times by round 10, four times by round 14, and started to interact for the fifth time in round 15. In larger groups, pairs only interacted once by round 7, and twice by round 15.

Analyses

We used mixed-effects regression models to test the effect of community size on all measuresusing the lme4 (Bates, Maechler Bolker & Walker, 2016) and pbkrtest (Halekoh & Højsgaard, 2014) packages in R (R Core Team, 2016). All models had the maximal random effects structure justified by the data that would converge. The reported p-values were generated using the Kenward-Roger Approximation, which gives more conservative p-values for models based on small numbers of observations. The full models are included in Appendix C. All the data and the scripts for generating all models can be openly found at https://osf.io/y7d6m/.

Changes in communicative success, stability, convergence and linguistic structure were examined using three types of models: (I) Models that analyze changes in the dependent variable over time; (II) Models that compare the final levels of the dependent variable at the end of the experiment; (III) Models that examine differences in the levels of variance in the dependent variable over time.

Models of type (I) predicted changes in the dependent variable as a function of time and community size. Models for communicative success included data from communication rounds only (excluding the two test rounds). In models for communicative success, convergence, and stability, the fixed effects were CONDITION (dummy-coded with small group as the reference level), ROUND NUMBER (centered), ITEM CURRENT AGE (centered), and the interaction terms CONDITION X ITEM CURRENT AGE and CONDITION X ROUND NUMBER. ITEM CURRENT AGE codes the number of rounds each scene was presented until that point in time, and measures the effect of familiarity with a specific scene on performance. ROUND NUMBER measures the effect of time passed in the experiment and overall language proficiency. The random effects structure of models for communicative success, convergence, and stability included by-scenes and by-groups random intercepts, as well as by-groups random slopes for the effect of ROUND NUMBER. Models for stability and communicative success also included by-scenes random slopes for the effect of ROUND NUMBER. As structure score was calculated for each producer over all scenes in a given round, the model for linguistic structure did not include ITEM CURRENT age as a fixed effect, and included fixed effects for ROUND NUMBER (quadratic, centered), CONDITION (dummy-coded with small group as the reference level), and the interaction term CONDITION X ROUND NUMBER. Following Beckner, Pierrehumbert & Hay (2017), who found that linguistic structure tends to increase nonlinearly, we included both the linear and the quadratic terms for the effect of ROUND NUMBER (using the poly() function in R to avoid collinearity). The model for linguistic structure included random intercepts and random slopes for the effect of ROUND NUMBER with respect to different producers who were nested in different groups.

Models of type (II) compared the mean values of the final languages created by small and larger groups in rounds 15-16. The fixed effect in these models was a two-level categorical variable for CONDITION (i.e., small groups vs. larger groups), dummy-coded with small groups as the reference level. In models for communicative success, stability and structure, the random effects structure included random intercepts for different groups and different scenes. In models for linguistic structure, the random effect structure included random intercepts for different producers nested in different groups.

Models of type (III) predicted the degree of variance in the dependent variable across groups and time. For linguistic structure, variance was calculated as the square standard deviation in participants' average

structure scores across all groups in a given round. For communicative success, convergence and stability, variance was calculated as the square standard deviation in the dependent variable on each scene across all groups in a given round. These models included by-scenes random intercepts and slopes for the effect of ROUND NUMBER. All models included fixed effects for ROUND NUMBER (centered), CONDITION (dummy-coded with small group as the reference level), and the interaction term CONDITION X ROUND NUMBER.

We also examined changes in input variability as a function of time and community size. This model included fixed effects for ROUND NUMBER (centered), CONDITION (dummy-coded with small group as the reference level), and the interaction between them. There were by-group random intercepts and by-group random slopes for the effect of ROUND NUMBER. Finally, we examined changes in linguistic structure scores over consecutive rounds as a function of (a) input variability, (b) shared history, or (c) both. In all three models, the dependent variable was the difference in structure score between round *n* and *n+1*, and there were random intercepts for different producers nested in different groups. In model (a), the fixed effect was MEAN INPUT VARIABILITY at round *n* (centered). In model (b), the fixed effect was SHARED HISTORY at round *n* (centered). Model (c) was a combination of models (a) and (b).

Results

We report the results for each of the four linguistic measures separately, using three types of analyses (see Methods). Figure 2 summarizes the average differences in the performance of small and larger groups over the course of all 16 rounds. Note that all analyses were carried over all data points and not over averages. All analyses are reported in full in Appendix C using numbered models, which we refer to here.

1. Communicative Success

Communicative Success increased over time (Model 1: β=0.08, SE=0.02, t=4, p<0.0001; Fig. 2A), with participants becoming more accurate as rounds progressed. This increase was not significantly modulated by group size (Model 1: β=0.04, SE=0.03, t=1.76, p=0.078), with small and larger groups reaching similar accuracy scores in the final communication round (Model 2: β=0.14, SE=0.08, t=1.8, p=0.083). Small and larger

groups differed in variance: while all groups became increasingly more varied over time (Model 3: β=0.002, SE=0.0004, t=5.18, p<0.0001), larger groups showed a slower increase in variance (Model 3: β=-0.002, SE=0.0005, t=-4.2, p<0.0001) and lower variance overall (Model 3: β=-0.007, SE=0.002, t=-3.48, p<0.001). These results indicate that while small groups varied in their achieved accuracy scores, and even more so as the experiment progressed, larger groups tended to behave more similarly to one another throughout the experiment.

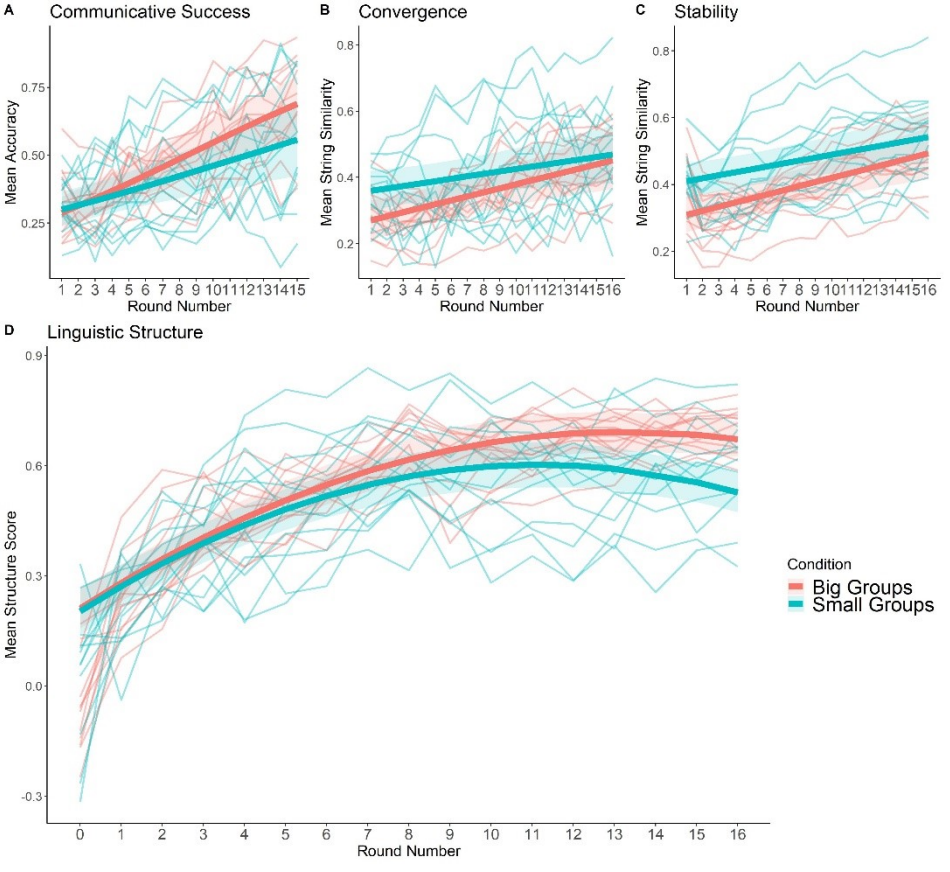


Figure 2. Changes in (A) Communicative Success, (B) Convergence, (C) Stability, and (D) Linguistic Structure over time as a function of community size. Thin lines represent average values for each group in a given round. Data from small and larger groups is plotted in blue and red, respectively. Thick lines represent the models' estimates, and their shadings represent the models' standard errors.

2. Convergence

Convergence increased significantly across rounds (Model 4: β=0.007, SE=0.003, t=2.31, p=0.029; Fig. 2B), with participants aligning and using more similar labels over time. Convergence was also better on more familiar scenes (Model 4: β=0.004, SE=0.001, t=2.62, p=0.014). Group size had no effect on convergence (Model 4: β=-0.06, SE=0.04, t=-1.37, p=0.18), so that small and larger groups showed similar levels of convergence by the end of the experiment (Model 5: β=-0.03, SE=0.05, t=-0.63, p=0.54). Interestingly, larger groups were not less converged than small groups, despite the fact that members of larger groups had double the amount of people to converge with and only half the amount of shared history with each of them. Variance increased over rounds (Model 6: β=0.001, SE=0.003 t=4.32, p<0.0001), but there was significantly less variance in the convergence levels of larger groups than across small groups throughout the experiment (Model 6: β=-0.04, SE=0.002 t=-23.68, p<0.0001). That is, larger groups behaved similarly to each other, showing a slow yet steady increase in convergence over rounds, while small groups varied more in their behavior: some small groups reached high levels of convergence, but others maintained a high level of divergence throughout the experiment, with different participants using their own unique labels.

3. Stability

Stability significantly increased over time, with participants using labels more consistently as rounds progressed (Model 7: β=0.009, SE=0.003, t=3.26, p=0.003; Fig. 2C). Labels for more familiar scenes were also more stable (Model 7: β=0.004, SE=0.001, t=3.68, p=0.001). Group size affected stability (Model 7: β=-0.08, SE=0.04, t=-2.08, p=0.047), with larger groups' languages being less stable (i.e., showing more changes). However, by the end of the experiment, the languages of small and larger groups did not differ in their stability (Model 8: β=-0.06 SE=0.05, t=-1.21, p=0.24). As in the case of convergence, larger groups showed significantly less variance in their levels of stability compared to small groups throughout the experiment (Model 9: β=-0.018, SE=0.001, t=-16.99, p<0.0001), reflecting the fact that smaller groups differed more from each other in their stabilization trends.

4. Linguistic Structure

Linguistic Structure significantly increased over rounds (Model 10: β=4.55, SE=0.48, t=9.46, $p<0.0001$; Fig 2D), with participants' languages becoming more systematic over time. This increase was non-linear and slowed down in later rounds (Model 10: β=-3, SE=0.38, t=-7.98, $p<0.0001$). As predicted, the increase in structure was significantly modulated by group size (Model 10: β=1.92, SE=0.63, t=3.06, p=0.004), so that participants in larger groups developed structured languages faster compared to participants in small groups. Indeed, the final languages developed in larger groups were significantly more structured than the final languages developed in small groups (Model 11: β=0.11, SE=0.04, t=2.93, p=0.006). Variance did not significantly decrease over time (Model 12: β=-0.0009, SE=0.0005, t=-1.73, p=0.094), yet larger groups varied significantly less overall in how structured their languages were (Model 12: β=-0.015, SE=0.004, t=-4.28, p=0.0002). That is, while small groups differed in their achieved levels of structure throughout the experiment, different larger groups showed similar trends and reached similar structure scores.

Although all groups started out with different random holistic labels, compositional languages emerged in many groups during the experiment. Many groups developed languages with systematic and predictable grammars (see Figure 3 for one example, and Appendix D for more examples), in which scenes were described using complex labels: one part indicating the shape, and another part indicating motion[9]. Interestingly, groups differed not only in their lexicons, but also in the grammatical structures they used to categorize scenes according to motion. While many groups categorized angles based on a two axes system (with part-labels combined to indicate up/down and right/left), other groups parsed angles in a clock-like system, using unique part-labels to describe different directions. Importantly, while no two languages were identical, the level of systematicity in the achieved structure depended on group size.

[9] Complex descriptions in the artificial languages could be interpreted as single words with different affixes, or alternatively as different words combined to a sentence (e.g., with a noun describing shape and a verb describing motion). Therefore, in the current paradigm, there is no meaningful distinction between syntactic and morphological compositionality.

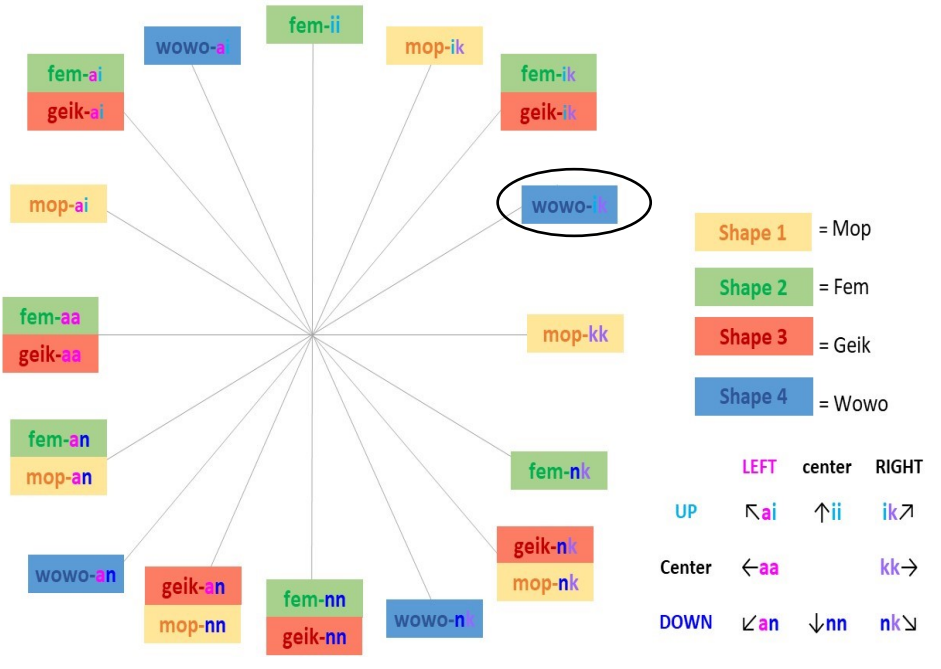


Figure 3. An example of the final language produced by a participant in a large group, along with a "dictionary" for interpreting it on the right. Box colors represent the four shapes, and the grey axes indicate the direction in which the shape moved. Font colors represent different meaningful part-labels, as segmented by the authors for illustration purposes only. For example, the label in the black circle ("*wowo-ik*") described a scene in which shape 4 moved in a 30º angle. It is comprised of several parts: "*wowo*" (indicating the shape) and "*ik*" (indicating the direction, comprised of two meaningful parts: "*i*" for "up" and "*k*" for "right").

We also tested our hypothesis that group size effects are driven by differences in input variability and shared history. First, we quantified the degree of input variability in each group at a given time point by measuring the differences in the variants produced for different scenes in different rounds. Then we examined changes in input variability over time across conditions. We found that input variability significantly decreased over rounds (Model 13: β=-0.1, SE=0.01, t=-8, $p<0.0001$), with a stronger decrease in the larger groups (Model 13: β=-0.08, SE=0.2, t=-4.42, p=0.0001). Importantly, this analysis also confirmed that larger groups were indeed associated with greater input variability overall (Model 13: β=1.45, SE=0.09, t=15.99, $p<0.0001$) – a critical assumption in the literature (Atkinson, Kirby & Smith, 2015; Meir et al., 2012; Nettle, 2012; Wray & Grace, 2007) and a premise for our hypothesis. We also quantified the degree of shared history between participants. Then, we

examined the role of input variability and shared history in promoting changes in linguistic structure by using these measures to predict differences in structure scores over consecutive rounds. We found that more input variability at round *n* induced a greater increase in structure at the following round (Model 14: β=0.015, SE=0.003, t=4.8, p<0.0001). Similarly, less shared history at round *n* induced a greater increase in structure at the following round (Model 15: β=-0.017, SE=0.004, t=-4.18, p=0.0004). When both predictors were combined in a single model, only input variability was significantly associated with structure differences (Model 16: β=0.011, SE=0.004, t=2.76, p=0.012), while the effect of shared history did not reach significance (Model 16: β=-0.008, SE=0.005, t=-1.42, p=0.17) – suggesting that input variability was the main driver for the increase in structure scores.

Discussion

We used a group communication paradigm to test the effect of community size on linguistic structure. We argued that larger groups were under stronger pressure to develop shared languages to overcome their greater communicative challenge, and therefore created more systematic languages. We found that while all larger groups consistently showed similar trends of increasing structure over time, some small groups never developed systematic grammars and relied on holistic, unstructured labels to describe the scenes. Importantly, linguistic structure increased faster in the larger groups, so that by the end of the experiment, their final languages were significantly more systematic than those of small groups. Our results further showed that the increase in structure was driven by the greater input variability in the larger groups. Remarkably, the languages developed in larger groups were eventually as globally shared across members, even though members of larger groups had fewer opportunities to interact with each other, and had more people they needed to converge with compared to members of small groups. Finally, the languages of small groups changed less over time, though larger groups reached an equal level of stability by the end of the experiment. Together, these results suggest that group size can affect the live formation of new languages.

The groups in our experiment were smaller than real-world communities. The results, however, should scale to real-world populations since the meaning space and speakers' life span scale up

proportionally. Concordantly, our results are consistent with findings from real developing sign languages, which show that given the same amount of time, a larger community of signers developed a more uniform and more systematic language compared to a small community of signers (Meir et al., 2012). It also resonates with psycholinguistic findings that show how input variability can affect generalization (Gómez, 2002): participants typically don't generalize over variants when they are able to memorize all of them individually, but do generalize when there are too many variants to remember. Similarly, greater input variability in larger groups promoted generalizations of the linguistic stimuli in our experiment, consistent with language change theories that argue for more systematicity in big communities of speakers for the same reasons (Milroy & Milroy, 1985; Nettle. 2012; Trudgill, 2002; Wray & Grace, 2007).

The proposed mechanisms assumes a close relationship between our linguistic measures, and is based on the hypothesis that linguistic structure can facilitate convergence and comprehension. We assumed that larger groups compensated for their greater communicative challenge by developing more systematic languages, which enabled them to reach similar levels of convergence and accuracy by the end of the experiment. Therefore, one may wonder whether more structure indeed facilitated convergence and communicative success in our experiment. To this end, we examined the relation between our measures of communicative success, convergence and linguistic structure after controlling for the effect of round (see Appendix C). One model predicted convergence as a function of time and linguistic structure. The model included ROUND NUMBER (centered), STRUCTURE SCORE (centered), and the interaction between them as fixed effects. Another model predicted communicative success as a function of time, convergence, and linguistic structure scores, with fixed effects for ROUND NUMBER (centered), STRUCTURE SCORE (centered), MEAN CONVERGENCE (centered), and the interaction terms STRUCTURE SCORE X ROUND NUMBER and MEAN CONVERGENCE X ROUND NUMBER. Both models included by-group random intercepts and by-group random slopes for all fixed effects. Indeed, we found that more linguistic structure predicted better convergence across different rounds (Model 17: β=0.018, SE=0.008, t=2.32, p=0.027). Additionally, communicative success was predicted by structure (Model 18: β=0.436, SE=0.06, t=7.48, $p<0.0001$) and convergence (Model 18: β=0.189, SE=0.06, t=2.95, p=0.008), so that better group alignment and more systematic structure predicted higher accuracy scores across rounds. Moreover, the relationship between structure and accuracy became stronger over rounds

(Model 18: β=0.051, SE=0.008, t=6.38, $p<0.0001$). These additional analyses provide important empirical evidence in support of the underlying mechanisms we proposed, and shed light on the nature of the group size effects reported in this paper.

Another important aspect of our results concerns the effect of group size on variance in behaviour. We found significantly more variance in the behaviors of small groups across all measures: some small groups reached high levels of communicative success, convergence, stability, and linguistic structure, while others did not show much improvement in these measures over time. By contrast, larger groups all showed similar levels of communicative success, stability, convergence, and linguistic structure by the end of the experiment. These results support the idea that small groups are more vulnerable to drift (Nettle, 1999; Spike, 2017): random changes are more likely to occur in smaller populations, while larger populations are more resilient to such random events and often show more consistent behaviors. This result may be underpinned by basic probability statistics: small samples are typically less reliable and vary more from each other, while larger samples show more normally distributed patterns and are more representative of general trends in the population ("the law of large numbers"; Blume & Royall, 2003).

Our findings support the proposal that community size can drive the cross-linguistic and historical findings that larger societies have more simplified grammars (Lupyan & Dale, 2010; Meir et al., 2012; Milroy & Milroy, 1985; Nettle, 2012; Trudgill, 2002; Wray & Grace, 2007), and suggest that differences in community size can help explain and predict patterns and trajectories in language formation and change. Our results show that the mere presence of more people to interact with introduces a stronger pressure for systemization and for creating more linguistic structure, suggesting that an increase in community size can cause languages to lose complex holistic constructions in favor of more transparent and simplified grammars. As such, our results are in line with the idea that increasing community size could have been one of the drivers for the evolution of natural language (Dunbar, 1993).

Our findings also stress the role of the social environment in shaping the grammatical structure of languages, and highlight the importance of examining other relevant social properties alongside community size. Particularly, network structure and connectivity are typically confounded with community size, and have been argued to play an important role in explaining cross-cultural differences in linguistic complexity.

Specifically, theories of language change suggest that differences in network density may be the true underling mechanism behind language simplification (Milroy & Milroy, 1985; Trudgill, 2002; Wray & Grace, 2007). This idea is supported by computational work showing that networks' structural properties, such as their degree of clustering and hierarchy, can influence linguistic complexity and modulate the effect of population size (Lou-Magnuson & Onnis, 2018; but see Spike, 2017). Chapter 4 examines the individual role of network structure on the formation of languages.

Acknowledgments

We wish to thank Caitlin Decuyper for programming the experiment, and Mark Atkinson, Rona Feldman, Gary Lupyan, Carolyn McGettigan, Alan Nielsen, Andrea Ravignani, Sean Roberts, Timo Roettger, Kevin Stadler, and Anastasia Ulicheva for discussions and helpful input.

Appendix A: Settings

The experimental settings between paired participants during the communication round:

The experimental settings across a group during the communication round:

Appendix B: Comparisons with the Short Version

In addition to the 24 groups reported in the paper, we also collected data from six small groups of four participants who played a shorter version of the experiment, which included only eight rounds instead of 16. These groups were tested in the first batch of data collection. The report of the individual performance of these six groups can also be found in Chapter 2 (Raviv, Meyer & Lev-Ari, 2019a).

In this appendix, we report the results of the comparison between these small "short-version" groups and the twelve larger groups reported in the paper. We compared the performance of these groups twice: first during the first eight rounds, when groups have the same amount of exposure but differ in their shared history; and then again at the end of the experiment (i.e., the seventh and eighth round for the small groups vs. the 15th and 16th round for the larger groups). At that point, the amount of shared history is equated - members of both types of groups have interacted with each other twice by that time point – but the amount of exposure differs.

Participants

A total of 24 adults participated in the short version (mean age=23.2y, SD=4.53; 18 women). All participants were native Dutch speakers and were recruited using the participant database of the Max Planck Institute for Psycholinguistics. Participants in the short version were paid 20€ or more depending on the time they spent in the lab (between 120 to 150 minutes).

Stimuli and Procedure

The stimuli and procedure of the short version were identical to those reported in Chapter 3, except for the fact that the experiment ended after the first eight rounds (i.e., participants completed the test round at round 8, filled the debriefing form and then left, without having lunch and without reconvening to continue the second half). The full description of the procedure can be found in Chapter 2.

Analyses

We used mixed-effects regression models to test the effect of community size on all measures, using two types of models: (1) Models that analyze changes over the course of the first eight rounds; (2) Models that compare the final languages created by small and larger groups before and after the additional rounds, that is, at rounds 7 and 8, and at rounds 15 and 16. All models were generated using the lme4 and pbkrtest packages in R (see references in the main chapter). All the data and the scripts for generating the models can be found online at https://osf.io/y7d6m/.

Models of type (1) were used to predict changes in the dependent variable as a function of time and community size, and included all data from the six small groups who played the short version and data the first eight rounds of the 12 larger groups who played the full version and were collected in the same batch. Models for communicative success included data from communication rounds only (excluding the eighth test round). In models for communicative success, convergence, and stability, the fixed effects were CONDITION (dummy-coded with small group as the reference level), ROUND NUMBER (centered), ITEM CURRENT AGE (centered), and the interaction terms CONDITION X ITEM CURRENT AGE and CONDITION X ROUND NUMBER. The random effects structure of models for communicative success, convergence, and stability always included by-scenes and by-groups random intercepts, as well as by-scenes and by-groups random slopes with respect to the effect of ROUND NUMBER. Because the structure score was calculated for each producer over all scenes in a given round, the model for linguistic structure did not include ITEM CURRENT AGE as a fixed effect. The model for linguistic structure therefore included fixed effects for ROUND NUMBER (linear and quadratic terms, centered), CONDITION (dummy-coded with small group as the reference level), and the interaction term CONDITION X ROUND NUMBER (linear and quadratic terms). The model for linguistic structure included random intercepts and random slopes for the effect of ROUND NUMBER (linear and quadratic terms) with respect to different producers who were nested in different groups.

Models of type (2) were used to compare the final languages created by small and larger groups in rounds 7-8 and in rounds 15-16, whenever we found evidence for group size influences in type (1) models. Since group size did not influence communicative success, we did not run models of type (2) for this measure. The fixed effect was a three-level categorical variable for CONDITION (i.e., small groups at rounds 7-8, larger

groups at rounds 7-8, larger groups at rounds 15-16). This variable was dummy-coded with small groups at rounds 7-8 as the reference level. In models for convergence and stability, the random effects structure included random intercepts for different groups and different scenes. In models for linguistic structure, the random effect structure included random intercepts for different producers nested in different groups.

Results

Communicative Success

Communicative Success increased during the first eight rounds (Model 1: β=0.20, SE=0.05, t=4.5, p<0.0001). Participants became more accurate as rounds progressed, and this increase was not affected by community size (Model 1: β=-0.27, SE=0.2, t=-1.3, p=0.19). While small groups were more accurate than larger groups at the seventh round (Model 2: β=-0.11, SE=0.03, t=-3.86, p=0.0009), larger groups reached more accuracy when given additional rounds (Model 2: β=0.11, SE=0.03, t=3.87, p=0.009).

(1) Type (I) Model: Accuracy during the first 8 rounds

Accuracy ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 + centered.Round | ItemID) + (1 + centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	-0.2710	0.1710	-1.5845	0.1131
Round Number	0.2044	0.0450	4.5445	0.0000
Condition (Big vs. Small)	-0.2688	0.2061	-1.3040	0.1922
Item Current Age	-0.0118	0.0289	-0.4077	0.6835
Round Number X Condition	-0.0838	0.0520	-1.6130	0.1067
Item Current Age X Condition	0.0208	0.0316	0.6578	0.5107

(2) Type (II) Model: Final Accuracy comparison at round 7 and round 15

MeanAccuracy ~ Condition + (1 | Group) + (1| ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.5800	0.0361	16.0783	0e+00
Condition: Big Groups at Round 15 vs. Small Groups at round 7	0.1084	0.0280	3.8744	9e-04
Condition: Big Groups at Round 7 vs. Small Groups at rounds 7	-0.1081	0.0280	-3.8636	9e-04

Convergence

Convergence increased significantly during the first eight rounds (Model 3: β=0.03, SE=0.01, t=3.2, p=0.004), with participants in the same community aligning over time and using more similar labels. Larger groups were significantly less converged than small groups during the first eight rounds (Model 3: β=-0.1, SE=0.03, t=-2.79, p=0.01). A comparison of convergence levels before and after the additional rounds confirmed that larger groups were significantly less converged by the end of the eighth round (Model 4: β=-0.08, SE=0.02, t=-4.11, p=0.0005). However, this disadvantage disappeared once larger groups completed all 16 rounds and had the same shared history (Model 4: β=0.02, SE=0.02, t=1.2, p=0.245). This result suggests that larger groups needed more time in order to develop globally shared languages, but eventually reach similar levels of convergence as small groups.

(3) Type (I) Model: Convergence during the first 8 rounds

Convergence ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 + centered.Round | ItemID) + (1 + centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.3873	0.0291	13.3205	0.0000
Round Number	0.0276	0.0086	3.1855	0.0044
Condition (Big vs. Small)	-0.0963	0.0345	-2.7916	0.0108
Item Current Age	0.0003	0.0030	0.0946	0.9255
Round Number X Condition	-0.0149	0.0105	-1.4214	0.1697
Item Current Age X Condition	0.0002	0.0035	0.0634	0.9501

(4) Type (II) Model: Final Convergence comparison at rounds 7-8 and rounds 15-16

MeanConvergence ~ Condition + (1 | Group) + (1| ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4190	0.0271	15.4482	0.0000
Condition: Big Groups at Rounds 15-16 vs. Small Groups at rounds 7-8	0.0234	0.0196	1.1954	0.2453
Condition: Big Groups at Rounds 7-8 vs. Small Groups at rounds 7-8	-0.0806	0.0196	-4.1132	0.0005

Stability

Stability significantly increased during the first eight rounds, with participants using labels more consistently over time (Model 5: β=0.03, SE=0.01, t=3.76, p=0.0011). Larger groups were significantly less stable than small groups during the first eight rounds (Model 5: β=-0.09, SE=0.03, t=-2.98, p=0.0069), and a comparison of the stability levels before and after the additional rounds confirmed that by the end of the eighth round, larger groups showed less stability compared to small groups (Model 6: β=-0.08, SE=0.02, t=-5.3, p<0.0001). Yet again, this pattern disappeared once larger groups were given the additional rounds (Model 6: β=0.025, SE=0.02, t=1.55, p=0.134). That is, while larger groups needed more time to develop consistent languages, they eventually reached the same level of stability as small groups.

(5) Type (I) Model: Stability during the first 8 rounds

Stability ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 + centered.Round | ItemID) + (1 + centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4214	0.0272	15.4865	0.0000
Round Number	0.0299	0.0079	3.7617	0.0011
Condition (Big vs. Small)	-0.0948	0.0318	-2.9800	0.0069
Item Current Age	0.0011	0.0028	0.3918	0.6990
Round Number X Condition	-0.0163	0.0095	-1.7098	0.1015
Item Current Age X Condition	-0.0019	0.0032	-0.5823	0.5663

(6) Type (II) Model: Final Stability comparison at rounds 7-8 and rounds 15-16

MeanStability ~ Condition + (1 | Group) + (1| ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4639	0.0225	20.6421	0.0000
Condition: Big Groups at Rounds 15-16 vs. Small Groups at rounds 7-8	0.0248	0.0160	1.5540	0.1345
Condition: Big Groups at Rounds 7-8 vs. Small Groups at rounds 7-8	-0.0848	0.0160	-5.3054	0.0000

Linguistic Structure

Linguistic Structure significantly increased over the first eight rounds in a linear way (Model 7: β=2.58, SE=0.52, t=4.98, p=0.0001), with participants' languages becoming more systematic over time. This increase in structure was modulated by group size (Model 7: β=0.02, SE=0.04, t=2.95, p=0.0077), with participants in larger groups developing structured languages faster compared to participants in small groups. Although the languages of small and larger groups were equally structured after eight rounds (Model 8: β=0.0005, SE=0.01, t=0.36, p=0.97), members of larger groups developed languages with significantly more linguistic structure after given additional rounds (Model 8: β=0.066, SE=0.01, t=5.3, p<0.0001).

(7) Type (I) Model: Linguistic Structure over time

Linguistic Structure ~ poly(centered.Round,2) * Condition + (1 + poly(centeredRound ,2) | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4409	0.0321	13.7289	0.0000
Round Number (Linear)	2.5766	0.5173	4.9811	0.0001
Round Number (Quadratic)	-0.3875	0.3984	-0.9726	0.3418
Condition (Big vs. Small)	0.0194	0.0378	0.5133	0.6131
Round Number (Linear) X Condition	1.7576	0.5960	2.9488	0.0077
Round Number (Quadratic) X Condition	-0.1959	0.4501	-0.4353	0.6678

(8) Type (II) Model: Final Linguistic Structure comparison

MeanStructure ~ Condition + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.6305	0.0147	42.7549	0.0000
Condition: Big Groups at Rounds 15-16 vs. Small Groups at rounds 7-8	0.0663	0.0125	5.2978	0.0000
Condition: Big Groups at Rounds 7-8 vs. Small Groups at rounds 7-8	0.0005	0.0128	0.0358	0.9717

Appendix C: Models

Communicative Success

(1) Type (I) Model: Accuracy over time

Accuracy ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 | ItemID) + (1 +centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	-0.31050	0.17486	-1.77572	0.07578
Round Number	0.07713	0.01925	4.00637	0.00006
Condition (Big vs. Small)	0.25736	0.24271	1.06037	0.28898
Item Current Age	0.00590	0.01278	0.46191	0.64415
Round Number X Condition	0.04443	0.02523	1.76106	0.07823
Item Current Age X Condition	0.00114	0.01479	0.07712	0.93853

(2) Type (II) Model: Final Accuracy comparison

MeanAccuracy ~ Condition + (1 | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.5489	0.0548	10.0130	0.0000
Big vs. Small Groups at Round 15	0.1395	0.0775	1.7993	0.0835

(3) Type (III) Model: Accuracy variance

SD_Accuracy ~ centered.Round * Condition + (1 + centered.Round | ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.23960	0.00150	160.17920	0.00000
Round Number	0.00231	0.00045	5.17516	0.00000
Condition (Big vs. Small)	-0.00696	0.00200	-3.48421	0.00075
Round Number X Condition	-0.00193	0.00046	-4.20020	0.00006

Convergence

(4) Type (I) Model: Convergence over time

Convergence ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 + centered.Round | ItemID) + (1 + centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4110	0.0291	14.1310	0.0000
Round Number	0.0073	0.0032	2.3125	0.0287
Condition (Big vs. Small)	-0.0558	0.0408	-1.3671	0.1830
Item Current Age	0.0039	0.0015	2.6192	0.0143
Round Number X Condition	0.0047	0.0044	1.0827	0.2886
Item Current Age X Condition	-0.0008	0.0020	-0.4232	0.6755

(5) Type (II) Model: Final Convergence Comparison

MeanConvergence ~ Condition + (1 | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4745	0.0368	12.9100	0.0000
Big vs. Small Groups at Rounds 15-16	-0.0326	0.0520	-0.6268	0.5362

(6) Type (III) Model: Convergence variance

SD_Convergence ~ centered.Round * Condition + (1 + centered.Round | ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0669	0.0016	40.6846	0.0000
Round Number	0.0013	0.0003	4.3256	0.0000
Condition (Big vs. Small)	-0.0414	0.0017	-23.6760	0.0000
Round Number X Condition	0.0001	0.0004	0.2328	0.8161

Stability

(7) Type (I) Model: Stability over time

Stability ~ centered.Round * Condition + centered,ItemCurrentAge * Condition + (1 | ItemID) + (1 +centered,Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4701	0.0262	17.9676	0.0000
Round Number	0.0088	0.0027	3.2577	0.0031
Condition (Big vs. Small)	-0.0763	0.0367	-2.0806	0.0472
Item Current Age	0.0043	0.0012	3.6838	0.0010
Round Number X Condition	0.0035	0.0038	0.9226	0.3645
Item Current Age X Condition	-0.0008	0.0015	-0.5318	0.5993

(8) Type (II) Model: Final Stability comparison

MeanStability ~ Condition + (1 | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.5427	0.0322	16.8386	0.0000
Big vs. Small Groups at Rounds 15-16	-0.0553	0.0456	-1.2122	0.2363

(9) Type (III) Model: Stability variance

SD_Stability ~ centered.Round * Condition+ (1 + centered.Round | ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0382	0.0013	30.4467	0.0000
Round Number	0.0004	0.0002	1.4867	0.1403
Condition (Big vs. Small)	-0.0183	0.0011	-16.9857	0.0000
Round Number X Condition	0.0002	0.0002	0.7191	0.4738

Linguistic Structure

(10) Type (I) Model: Linguistic Structure over time

Linguistic Structure ~ poly(centered.Round,2) * Condition + (1 + poly(centeredRound ,2) | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.5015	0.0268	18.7272	0.0000
Round Number (Linear)	4.5508	0.4808	9.4645	0.0000
Round Number (Quadratic)	-3.0024	0.3762	-7.9804	0.0000
Condition (Big vs. Small)	0.0562	0.0368	1.5281	0.1352
Round Number (Linear) X Condition	1.9213	0.6284	3.0574	0.0042
Round Number (Quadratic) X Condition	0.4474	0.4848	0.9228	0.3623

(11) Type (II) Model: Final Linguistic Structure comparison

MeanStructure ~ Condition + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.5884	0.0282	20.8361	0.0000
Big vs. Small Groups at Rounds 15-16	0.1119	0.0382	2.9307	0.0063

(12) Type (III) Model: Linguistic Structure variance

SD_Structure ~ centered.Round * Condition

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0446	0.0025	17.7510	0.0000
Round Number	-0.0009	0.0005	-1.7318	0.0936
Condition (Big vs. Small)	-0.0152	0.0036	-4.2823	0.0002
Round Number X Condition	-0.0008	0.0007	-1.1229	0.2704

Input Variability

(13) Input Variability over time

MeanInputVariability ~ centered.Round * Condition + (1 + centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.64720632	0.06066371	10.668757	0.0000000000
Round Number	-0.09768587	0.01221066	-8.000047	0.0000000031
Condition (Big vs. Small)	1.44830290	0.09054961	15.994579	0.0000000000
Round Number X Condition	-0.07938256	0.01796243	-4.419366	0.0001003847

Changes in Linguistic Structure by Input Variability and Shared History

(14) Differences in linguistic structure by input variability

StructureDiff ~ centered.MeanInputVariability + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.03261559	0.004439334	7.346957	0.0000002669
Mean Input variability	0.01511464	0.003143137	4.808777	0.0000882001

(15) Differences in linguistic structure by shared history

StructureDiff ~ centered.History + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0341	0.0044	7.6844	0.0000
Shared History	-0.0172	0.0041	-4.1809	0.0004

(16) Differences in structure by input variability and shared history

StructureDiff ~ centered.MeanInputVariability + centered.History + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0330	0.0044	7.4226	0.0000
Shared History	-0.0076	0.0054	-1.4189	0.1707
Input variability	0.0113	0.0041	2.7594	0.0118

Relationship between measures

(17) Convergence by linguistic structure and round

MeanConvergence ~ centered.Structure * centered.Round + (1 + centered.Structure * centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.3932966	0.0226627	17.354342	0.0000000
Structure	0.0184677	0.0079378	2.326542	0.0266781
Round Number	0.0147805	0.0028131	5.254104	0.0000103
Round Number X Structure	0.0030277	0.0025288	1.197279	0.2402469

(18) Communicative Success by convergence, linguistic structure and round

MeanaAccuracy ~ centered.Structure * centered.Round + centered.Structure * centered.Convergence + (1 + centered.Structure * centered.Round | Group)

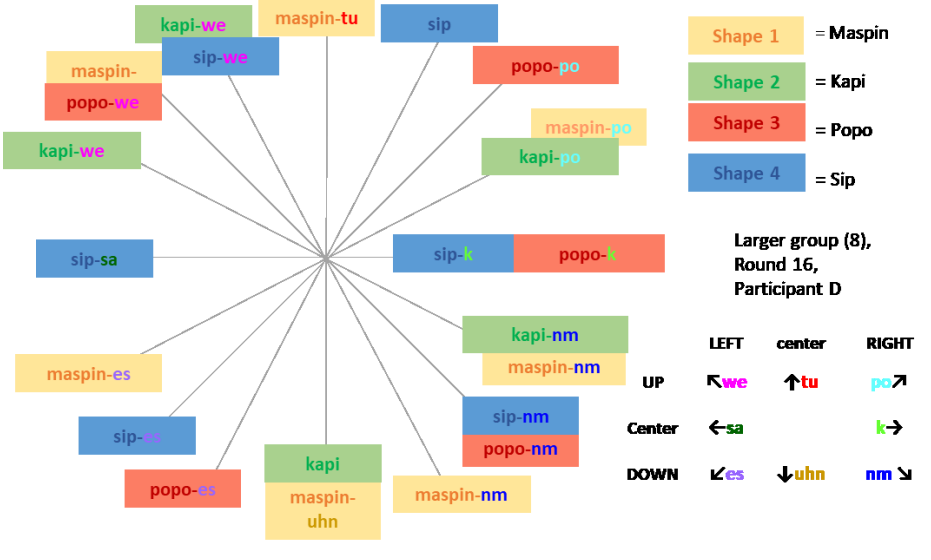
	Estimate	Std.Error	t-value	p-value
(Intercept)	0.42829439	0.018622207	22.999121	0.000000000
Structure	0.43612062	0.058279532	7.483255	0.000000493
Round Number	0.01158424	0.001826365	6.342788	0.000004712
Convergence	0.18890233	0.064083119	2.947771	0.008362724
Round Number X Structure	0.05153705	0.008070811	6.385609	0.000004315
Round Number X Convergence	-0.01658259	0.013990141	-1.185306	0.250743903

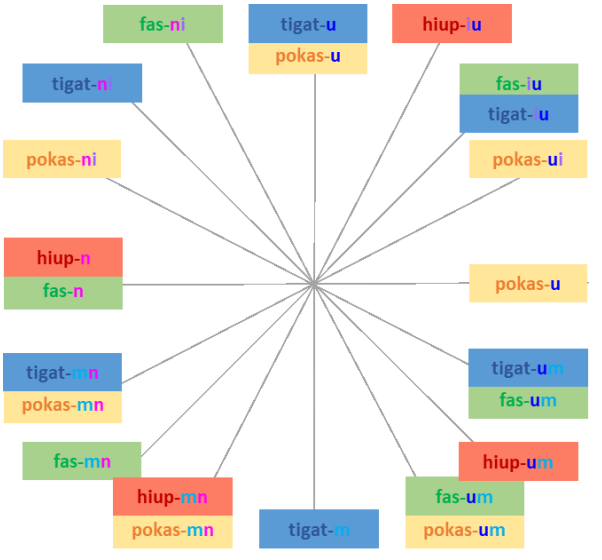
Appendix D: Examples of Structured Languages

Below we include four additional examples of structured languages produced by participants in small and larger groups at the final test round (round 16).

Each language is accompanied by a "dictionary" for interpreting the language on the right. Different box colors represent the four different shapes which appeared in the scenes, and the grey axes indicate the direction in which the shape was moving on the screen. Different font colors represent different meaningful part-labels, as segmented by the authors.

The dictionary and colors are solely for the purpose of illustration and were not used for any of the statistical analyses.

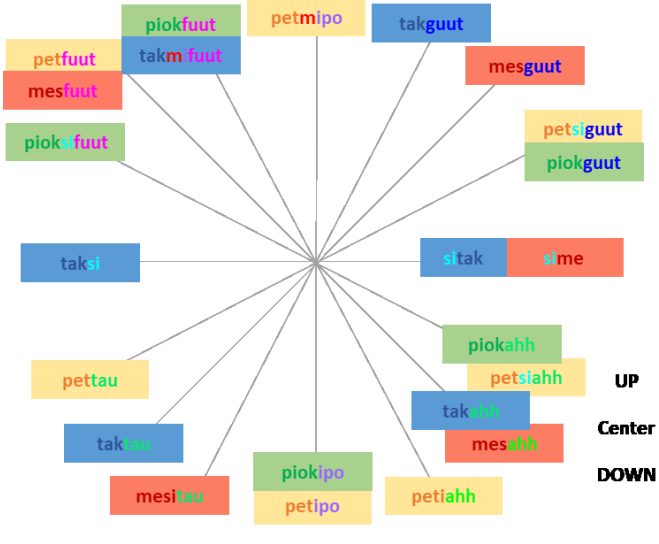




Shape 1 = Pokas
Shape 2 = Fas
Shape 3 = Hiup
Shape 4 = Tigat

Larger group (9),
Round 16,
Participant A

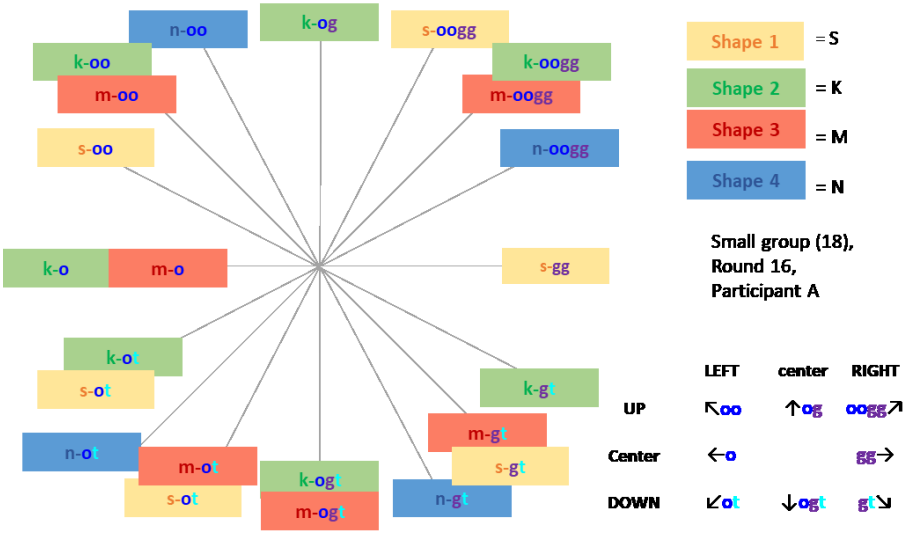
	LEFT	center	RIGHT
UP	↖ni	↑u	iu↗
Center	←n		u→
DOWN	↙mn	↓m	um↘



Shape 1 = Pet
Shape 2 = Piok
Shape 3 = Mes
Shape 4 = Tak

Small group (14),
Round 16,
Participant C

	LEFT	center	RIGHT
UP	↖fuut	↑mipo	guut↗
Center	←_si		si_→
DOWN	↙tau	↓ipo	ahh↘


n-oo
k-og
s-oogg
k-oo
k-oogg
m-oo
m-oogg
s-oo
n-oogg
k-o
m-o
s-gg
k-ot
s-ot
k-gt
m-gt
n-ot
m-ot
s-gt
k-ogt
s-ot
n-gt
m-ogt
Shape 1 = S
Shape 2 = K
Shape 3 = M
Shape 4 = N
Small group (18), Round 16, Participant A
LEFT
center
RIGHT
UP
Center
DOWN

4 The role of social network structure in the emergence of linguistic structure

Abstract

Social network structure has been argued to shape the structure of languages, as well as affect the spread of innovations and the formation of conventions in the community. Specifically, theoretical and computational models of language change predict that sparsely connected communities develop more systematic languages, while tightly knit communities can maintain high levels of linguistic complexity and variability. However, the role of social network structure in the cultural evolution of languages has never been tested experimentally. Here, we present results from a behavioral group communication study, in which we examined the formation of new languages created in the lab by micro-societies that varied in their network structure. We contrasted three types of social networks: fully connected, small-world, and scale-free. We examined the artificial languages created by these different networks with respect to their linguistic structure, communicative success, stability, and convergence. Results did not reveal any effect of network structure for any measure, with all languages becoming similarly more systematic, more accurate, more stable, and more shared over time. At the same time, small-world networks showed the greatest variation in their convergence, stabilization and emerging structure patterns, indicating that network structure can influence the community's susceptibility to random linguistic changes (i.e., drift).

Introduction

Why are languages so different from each other? One possible explanation is that selective pressures associated with social dynamics and language use can influence the emergence and distribution of different linguistic properties – making language typology a mirror of the social environment (Lupyan & Dale, 2016). According to this hypothesis, often referred to as the *Linguistic Niche Hypothesis*, the structure of languages is shaped by the structure of the community in which they evolved. Research in the past decades supports this theory by showing that different types of languages tend to develop in different types societies (Bentz & Winter, 2013; Lupyan & Dale, 2010; Meir, Israel, Sandler, Padden, & Aronoff, 2012; Nettle, 1999, 2012; Raviv, Meyer, & Lev-Ari, 2019b; Reali, Chater, & Christiansen, 2018).

Esoteric vs. Exoteric Languages

Models of language change typically draw a distinction between two types of social environments – *esoteric* communities and *exoteric* communities – and argue that there are substantial differences in the grammatical structure and overall uniformity of the languages used in such environments (Milroy & Milroy, 1985; Roberts & Winters, 2012; Trudgill, 1992, 2002, 2009; Wray & Grace, 2007). Specifically, esoteric communities are generally small and tightly-knit societies with little contact with outsiders, and therefore few if any non-native speakers. In contrast, exoteric communities tend to be much bigger and sparser societies, in which there is a higher degree of language contact and more interaction with strangers, and consequently also a higher proportion of non-native speakers.

Importantly, computational models, typological studies, and empirical work on the formation of new sign languages all suggest that esoteric and exoteric settings promote the emergence of different linguistic structures. For example, languages spoken in esoteric environments are claimed to be more morphologically complex, and have higher chances of developing rich and non-transparent systems of case marking and grammatical categories (Lupyan & Dale, 2010). Exoteric languages, on the other hand, tend to have fewer and less elaborate morphological paradigms and are more likely to express various grammatical relations (e.g., negation, future tense) by using lexical forms (individual words) rather than inflections (affixes). That is, there seems to be a greater

pressure for creating simpler and more systematic languages in exoteric compared to esoteric settings (Nettle, 2012; Trudgill, 2009; Wray & Grace, 2007). This is presumably because (a) members of exoteric communities are more likely to interact with strangers, resulting in communicative pressure in favor of generalization and transparency; and (b) there is a relatively high proportion of adult second-language learners in exoteric communities, who often struggle with learning complex and opaque morphologies.

Exoteric and esoteric languages are also claimed to show different rates of convergence. Members of esoteric communities are highly familiar with each other and share much common ground, which often entails more alignment and a stronger conservation of existing linguistic norms (Milroy & Milroy, 1985; Trudgill, 2002). Yet this high degree of familiarity between members of esoteric communities can preserve variation and reduce the pressure to establish *new* norms in the early stages of language development, as was found in the case of emerging sign languages (Meir et al., 2012). Specifically, new sign languages that developed in esoteric contexts tend to exhibit more variability across speakers, more irregularities, and overall greater context-dependence in comparison to sign languages developed in an exoteric context. In other words, because members of exoteric communities are far less connected to one another and typically share less common grounds with each other, such settings can increase the need for conventions and conformity in the early stages of language emergence, but hinder its preservation later on.

Teasing apart conflating social factors

The distinction between exoteric and esoteric communities relies on several parameters, namely, community size (small vs. big), network structure (highly connected vs. sparsely connected), and the proportion of adult non-native speakers in the community (low vs. high). These social parameters are naturally confounded in real-world environments (e.g., smaller groups also tend to be highly connected), making it hard to evaluate the unique contribution of each of these factors to the observed pattern of results (i.e., that languages used in exoteric contexts have simpler and more systematic morphologies; Lupyan & Dale, 2010). That is, we currently know very little about how each of these properties affects the structure of languages independently, and whether all features are equally influential in shaping linguistic patterns. Disentangling these social features from one another is important for understanding how

exactly languages adapt to fit their social environment, and for assessing the individual role of each factor.

Several computational models have attempted to isolate specific parameters associated with the difference between esoteric and exoteric communities, and to manipulate it separately from the others in order to examine its effects on various linguistic outcomes (Dale & Lupyan, 2012; Fagyal, Swarup, Escobar, Gasser, & Lakkaraju, 2010; Gong, Baronchelli, Puglisi, & Loreto, 2012; Lou-Magnuson & Onnis, 2018; Spike, 2017; Vogt, 2007, 2009; Wichmann, Stauffer, Schulze, & Holman, 2008). Such models generally suggest that different properties of esoteric and exoteric societies are associated with different pressures, yet often report conflicting results due to differences in model setup and parameter selection. For example, similar computational simulations examining the effect of community size can yield opposite results if agents' learning strategies are defined differently (Wichmann et al., 2008): when agents are assumed to copy globally (i.e., from all other agents in their network), larger groups seem to show slower rates of language change, yet when agents are assumed to copy more locally (i.e., from their closest neighbors), community size has no effect. Therefore, while computational models are valuable for teasing apart different social feature, they should be tested against experimental data.

Recently, a behavioral study focused on the role of community size, one of the features differentiating between esoteric and exoteric communities, and tested its individual effect on language emergence by contrasting languages created in the lab by big and small communities, while keeping all other social properties equal (Raviv et al., 2019b). Results showed that groups of eight interacting participants created more systematic languages, and did so faster and more consistently than groups of four interacting participants. The languages developed in the larger groups were more structured (i.e., more compositional) compared to those developed in smaller groups – a finding that was explained by the fact that larger groups faced a greater communicative challenge (due to more input variability). These results are in line with the cross-linguistic observations and the theoretical models reported above, and suggest that at least some of the typological and theoretical differences between exoteric and esoteric languages can indeed be attributed to differences in community size. As such, the study provided the first experimental evidence that community size has a unique and causal role in shaping linguistic patterns.

The postulated role of network structure

What about the other social features that differentiate between esoteric and exoteric communities? Does network structure also have a unique effect, above and beyond community size? An important feature of esoteric societies is their dense nature, in which members are typically connected via strong ties (i.e., family, close friends), and most if not all members of the community are familiar with one another. In contrast, exoteric societies are much sparser, and typically include many weak ties (i.e., acquaintances) and many members that never interact (i.e., strangers). This difference in network connectivity means that members of exoteric societies generally have fewer opportunities to develop common ground and globally align with each other (given that many of them will rarely or never meet), potentially resulting in more variability in the entire network.

Indeed, recent work on the cultural evolution of technology found that an increase in sparse connections from a state of high density (perhaps due to more geographical spread) leads to more innovations and more diversity in the community (Derex & Boyd, 2016). In this study, well-connected populations were less likely to produce complex technological solutions because of the ability to learn from all members and quickly converge on a local optimum, reducing exploration and cultural diversity in the population. In contrast, individuals in partially connected groups were more likely to progress along different paths of technological accumulation, leading to larger and more diverse technological repertoires and eventually to more complex solutions. These findings complement a long line of work on the prevalence and spread of innovations in social networks, which suggest that sparser ties generally promote more innovations and more variability. Specifically, work on social network structure shows that weak ties in sparser networks provide individuals with access to information, beliefs and behaviors beyond their own social circle, making the presence and prevalence of weak ties important for cultural innovation, technological accumulation, and the transmission and spread of ideas, behaviors and norms (Bahlmann, 2014; Granovetter, 1983; Liu, Madhavan, & Sudharshan, 2005).

Additionally, weak ties between members of sparser communities can affect the process of conventionalization, as they may entail less language stability, more variability, and more potential for changes. In contrast, strong ties between members of dense communities can inhibit language change and increase linguistic conformity: tight-knit connections often

function as a conservative force, preserving and amplifying existing norms and resisting external pressures to change (Granovetter, 1983; Milroy & Milroy, 1985; Trudgill, 2002, 2009). That is, denser communities may exhibit stricter maintenance of group conventions and therefore more preservation of linguistic norms, even when these norms are relatively complex and irregular (Trudgill, 2002, 2009). However, even though dense networks are postulated to show more stability, once a change does occur it is more likely to quickly spread to the entire community. This is because individuals are more likely to copy the behavior of strong than weak ties (Centola, 2010) and the propagation of variants is typically faster in dense networks than in sparser networks (Centola, 2010; Milroy & Milroy, 1985; Trudgill, 2009). Importantly, sparser networks' difficulty in convergence can trigger a stronger need for generalizations and regularizations, which may eventually lead to the creation of more systematic languages (Raviv et al., 2019b; Wray & Grace, 2007).

Although network structure is postulated to have an important effect in shaping linguistic patterns, to date there is no experimental evidence demonstrating its causal role in language complexity. As such, the theoretical claims described above remain hypothetical or anecdotal, and it is still unclear whether and how languages actually change in different types of network structures. The goal of the current study is to fill in this gap in the literature, and experimentally test the effect of social network structure on the emergence of new languages using a similar paradigm to that used in Raviv et al. (2019a) for demonstrating community size effects.

Computational evidence for network structure effects in language change

While experimental data is currently lacking, several computational models have examined the effect of social network structure using agent-based simulations. These models typically examine populations of communicating agents in three different types of networks: (1) dense, fully connected networks, in which all agent are connected to each other; (2) small-world networks, which are sparser in comparison to fully connected networks (i.e., there are fewer connections between agents), but in which most "strangers" are indirectly linked by a short chain of shared connections (Watts & Strogatz, 1998); and (3) scale-free networks, which are also characterized by sparsity and short paths but their distribution of connections follows a power law (i.e., most agents have few connections, yet some agents, the "hubs", have many; (Barabási & Albert, 1999).

A typical interaction in such models consists of two agents, who are randomly selected depending on the networks' available connections and their likelihood. Then, one agent (the producer) produces a linguistic variant (e.g., a vowel, word, or phrase) based on their inventory at the time of the interaction, and the other agent (the receiver) updates their own inventory based on that production and whether it is novel or familiar. This simple type of communication and learning (i.e., updating agent's representations) is then repeated for many iterations, allowing researchers to observe how variants spread and change over time in a given network. Importantly, the vast majority of these models do not examine the complexity or the systematicity of communication systems themselves, but rather focus only on the formation of linguistic conventions. This is done either by examining the time it takes for a population of agents to converge on a single linguistic variant or a shared lexicon, or by examining the degree of global alignment in the population after a fixed amount of time.

In most cases, computational models support the claim that differences in the structural properties of networks can lead to differences in convergence rates and in the spread of variants in the population. Specifically, multiple models report that denser networks show more successful diffusion of innovations compared to sparser networks, and that extra-dense networks (e.g., fully connected) typically converge most rapidly (Fagyal et al., 2010; Gong et al., 2012; Ke, Gong, & Wang, 2008). In addition, the existence of "hubs" (i.e., highly connected agents) in scale-free networks was shown to improve convergence and uniformity by advancing the spread of innovations to all agents in the community (Fagyal et al., 2010; Zubek et al., 2017). Nevertheless, one model suggested that, as long as networks have small-world properties (i.e., as long as "strangers" are indirectly linked by a short chain of shared connections), the network's specific configuration plays a minor role in the formation of conventions (Spike, 2017).

Interestingly, two models did examine the structure of the languages themselves, and they both report that network structure affected linguistic structure in some way. One model looked at the origin and the evolution of linguistic categorization of color terms, and found that scale-free networks were the fastest to develop color categories, and that those categories were more structured and more efficient compared to those developed in other types of networks (Gong et al., 2012). The second model introduced comprehensive, real-world mechanisms of social learning and language change, and looked at the creation and maintenance

of complex morphology (Lou-Magnuson & Onnis, 2018). The results of this model showed that more transitive networks (i.e., with a higher degree of "intimate" connections) were more likely to develop languages with complex morphological structures. Moreover, fully connected networks showed the highest levels of complexity, regardless of community size.

Together, computational models generally support the hypothesis that network structure can affect linguistic outcomes. They show that sparser networks tend to exhibit more structured languages but overall less convergence compared to dense networks, and suggest that the existence of "hubs" can further promote systemization and alignment. However, such computational models need to be further tested against empirical data obtained from human participants, seeing as they often lack ecological validity in terms of agents' cognitive capacities (e.g., agents have an unlimited memory capacity) or their behavior (e.g., agents update their inventories after every interaction by overriding all previous variants). As such, the causal role of network structure warrants further experimental validation.

The Current Study

Here, we experimentally tested the individual effect of network structure using a group communication paradigm (Raviv, Meyer, Lev-Ari, 2019a; 2019b). We examine the formation of new languages that develop in different micro-societies that varied in their network structure. Community size was kept constant across conditions, such that all networks were comprised of eight participants, yet differed in their degree of connectivity (i.e., how many people each participant interacts with) and homogeneity (i.e., whether all participants are equally connected). Specifically, we contrasted three different types of networks, which are typically used in computational models (Figure 1; see Network Properties for more details):

(1) Fully connected network (Figure 1A): This network is maximally dense, such that all possible connections are realized (i.e., all participants in the group get to interact with each other). It is also homogenous, as every participant has the same number of connections (i.e., seven people). This type of network resembles early human societies, hunter-gatherer communities and some villages, yet it is overall rare nowadays (Coward, 2010; Johnson & Earle, 2000).

(2) Small-world network (Figure 1B): This network is also relatively homogenous such that everyone has approximately the same number of connections (i.e., either three or four other participants), yet it is much sparser than the fully connected network and realizes only half of the possible connections. Importantly, this network has the small-world property where "strangers" are indirectly linked by a short chain of individuals (Watts & Strogatz, 1998). For example, participants G and H never interact, but they are connected via participants F, D and B, so innovations can still spread across the group and conventions can be formed.

(3) Scale-free network (Figure 1C): This network is equally sparse as the small-world network, and has the same number of possible connections overall. However, it is not homogenous: not everyone has the same number of connections. While some agents are highly connected, others are more isolated. The distribution of connections in this network roughly follows a power-law distribution (Barabási & Albert, 1999), with few participants having many connections, and a tail of participants with very few connections. For example, participant A is the "hub" and interacts with almost everyone in the group, while participants E and D are more isolated.

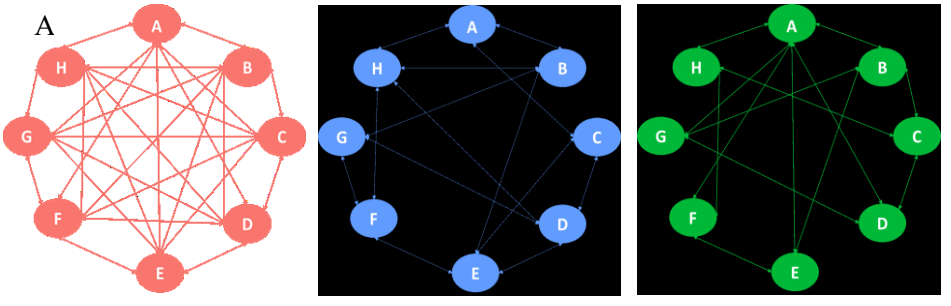


Figure 1. Network structure conditions. We tested groups of eight participants who were connected to each other in three different setups: a fully connected network (A), a small-world network (B), and a scale-free network (C).

Across conditions, participants' goal was to communicate successfully with each other using only an artificial language they created during the experiment. Participants in the same group interacted in alternating pairs according to the structural properties on their allocated network condition

(see Network Properties). In each communication round, paired partners took turns in describing novel scenes of moving shapes, such that one participant produced a label to describe a target scene, and their partner guessed which scene they meant from a larger set of scenes (see Procedure; Figure 2).

Over the course of the experiment, we analyzed the emerging languages using several measurements (see Measures): (1) Communicative Success, reflecting guessing accuracy; (2) Convergence, reflecting the degree of global alignment in the network (3) Stability, reflecting the degree of change over time; and (4) Linguistic Structure, reflecting the degree of systematic label-to-meaning mappings in participants' languages.

These measures are all related to real-world properties of natural languages: our measure of convergence reflects language uniformity (i.e., the number of dialects in the community and how much people's lexicons differ from one another); our measure of communicative success is related to mutual understanding; our measure of stability can be taken to reflect languages' rate of change (i.e., how fast innovations spread in the network); and our measure of linguistic structure can capture various grammatical properties, such as the systematicity of inflectional paradigms and the number of irregulars in a given language. Looking at these four measures enabled us to characterize the emerging languages and to consider how various linguistic properties change over time depending on network structure.

Our predictions are summarized in Table 1. Our main prediction was that sparser networks would develop more compositional languages, as a result of higher levels of input variability and diversity in such networks, which increase the pressure for generalization and systematization (Lou-Magnuson & Onnis, 2018; Raviv et al., 2019b; Wray & Grace, 2007). We also predicted that scale-free networks would show higher compositionality levels compared to small-world networks, since the existence of "hubs" in scale-free networks can further increase the chances of a compositional innovation spreading to the entire population (Fagyal et al., 2010; Gong et al., 2012; Zubek et al., 2017). That is, we predicted that scale-free networks would show the highest degree of linguistic structure (thanks to the "hub"), followed by small-world networks, and then by fully connected networks. We also expected the difference in linguistic structure to be closely linked to the degree of input

variability in dense vs. sparse networks: scale-free and small-world networks should show higher levels of input variability compared to fully connected networks, though the hub in scale-free networks might reduce variability compared to small world networks by increasing convergence.

Table 1: Predictions for Each Measure in the Current Experiment

<table>
<tr><th>Network Type / Measure</th><th>Fully connected (FC)</th><th>Small-world (SW)</th><th>Scale-free (SF)</th><th>Predication</th></tr>
<tr><td>Input Variability</td><td colspan="3">More input variability in sparse networks</td><td>FC < SF < SW</td></tr>
<tr><td>Linguistic structure</td><td colspan="2">Sparse networks =
more variability
→ More pressure for generalization and systematicity</td><td>“Hubs” can further promote the spread of systematic languages</td><td>FC < SW < SF</td></tr>
<tr><td>Convergence</td><td colspan="3">Sparse networks = more variability,
more strangers → less convergence
BUT
Sparse networks =? more systematic languages → Similar levels of convergence</td><td>FC=SW=SF</td></tr>
<tr><td>Stability</td><td>Dense networks = less diversity
→ More/faster stability</td><td colspan="2">Sparse networks =
more variability,
more innovations
→ Less/slower stability</td><td>FC > SW=SF</td></tr>
<tr><td>Communicative success</td><td colspan="3">No difference between conditions</td><td>FC=SW=SF</td></tr>
<tr><td colspan="5">Note. The predictions in the table are for the final languages. As described in more detail in the text, there could be differences across conditions in the rate of achieving these final outcomes. For example, we predicted that languages in all conditions would eventually show convergence, but we predicted it to occur faster in fully connected (FC) networks.</td></tr>
</table>

Based on the results of Raviv et al. (2019a), we hypothesized that the emergence of more structured languages in sparser networks would promote convergence in such networks (i.e., it should be easier to converge on more systematic variants). That is, while computational models suggest that sparser networks show less convergence in comparison to fully connected networks (given that some participants never interact with each other), we hypothesized that the creation of more structured languages in such networks would facilitate global alignment and lead to similar levels of convergence across networks. Moreover, scale-free networks may exhibit even better global alignment thanks to the existence of a "hub". In other words, if our prediction about linguistic structure is correct and sparser networks create more systematic languages, then convergence levels should be the same across dense and sparse networks. Otherwise, there should be relatively less convergence in sparser networks.

As for stability, we predicted that sparser networks would be less stable than the dense networks, given that there is a higher chances of innovations occurring in sparser networks and more variability overall (Derex & Boyd, 2016), and that changes take longer to stabilize in sparser networks (Ke et al., 2008). As such, we expected to see a difference in the rates of stabilization across conditions, with fully connected networks showing faster stabilization (i.e., less changes over rounds) compared to small-world and scale-free networks. Nevertheless, we expected similar levels of communicative success across all conditions, with all interacting members being equally good at understanding each other.

Methods

Participants

We collected data from 168 adults (mean age=24.6 years, SD=8.1 years; 132 women), comprising 21 groups of eight members (seven groups in each of the three conditions). Participants were paid 40€ or more depending on the time they spent in the lab (between 270 to 315 minutes, including a 30-minutes break). All participants were native Dutch speakers. Ethical approval was granted by the Faculty of Social Sciences of the Radboud University Nijmegen.

Materials

The materials used in this experiment were identical to those used in Chapters 2 and 3 (Raviv et al., 2019a; 2019b). For the full list of stimuli, see Appendix A in Chapter 2. Below we summarize the most relevant details:

We created 23 visual scenes that varied along three semantic dimensions: shape, angle of motion, and fill pattern. Each scene included one of four novel unfamiliar shapes, which moved repeatedly in a straight line from the center of the frame in a given direction (i.e., in an angle chosen from a range of possible angles). The shapes were created to be novel and ambiguous in order to prevent easy labeling with existing words. While the dimension of shape included four distinct categories, angle of motion was a continuous feature that could have been parsed and categorized by participants in various ways. Additionally, the shape in each scene had a unique blue-hued fill pattern, giving scenes an idiosyncratic feature. Therefore, the meaning space promoted categorization and structure along the dimensions of shape and motion, but also allowed participants to adopt a holistic, unstructured strategy where scenes are individualized according to their fill pattern.

Procedure

The procedure employed in this experiment was identical to that of Chapter 3 (Raviv et al., 2019b), except for the fact that all groups were comprised of eight participants, and were split up into pairs at the beginning of each communication round depending on their allocated network structure (see Network Properties; Appendix A). For a comprehensive description of the procedure, see Chapters 2 and 3. Below we recap the most relevant details:

Participants were told they were about to create a new fantasy language ("Fantasietaal" in Dutch) in the lab and use it in order to communicate with each other about different novel scenes. Participants were not allowed to talk, gesture, point, or communicate in any other explicit way besides the fantasy language and their assigned laptop. Participants' letter inventory was restricted and included a hyphen, five vowel characters (a,e,i,o,u) and ten consonants characters (w,t,p,s,f,g,h,k,n,m) which participants could combine freely.

The experiment had 16 rounds, and included three phases: a group naming phase (round 0), a communication phase (rounds 1-7; rounds 9-15), and a test phase (round 8; round 16).

In the initial naming phase (round 0), participants came up with novel nonsense words to describe eight initial scenes, so that the group had a few shared descriptions to start with. For each of the eight initial scenes, one of the participants was asked to use their creativity and describe it using one or more nonsense words. Participants took turns in describing the scenes, so the first scene was described by participant A, the second scene was described by participant B, and so on. Importantly, no use of Dutch or any other language was allowed, and participants were instructed to come up with novel nonsense labels. In order to establish mutual knowledge, we presented the scene-description pairings to all participants three times in a random order.

Following the naming phase, participants played a communication game with each other (the communication phase; Figure 2): the goal was to be communicative and earn as many points as possible as a group, with a point awarded for every successful interaction. The experimenter stressed that this was not a memory game, and that participants were free to use the labels produced during the group naming phase, or choose to create new ones. In each communication round, paired participants interacted with each other 23 times, with participants alternating between the roles of producer and guesser. In a given interaction, the producer saw the target scene on their screen (Fig. 2A) and produced description for it. Then, they rotated their screen and showed the description (without the target scene) to their partner, the guesser. The guesser was presented with a grid of eight scenes on their screen (the target and seven distractors; Fig. 2B), and had to select the scene they thought their partner referred to. Both participants then received feedback on whether their interaction was successful or not, including the target scene and the selected scene. The number of different target scenes increased gradually over the first six rounds (from eight initial scenes to a total of 23 scenes, with three new scenes introduced at each round), such that participants needed to refer to more and more new scenes as rounds progressed (Raviv et al., 2019a).

At the end of the seventh communication round, participants completed an individual test phase (round 8), in which they were presented with all scenes one by one in a random order, and needed to type their descriptions for them using the fantasy language. After the test, participants received a 30-minutes break and then reconvened to complete seven additional

communication rounds (rounds 9-15) and an additional test round (round 16). At the end of the experiment, all participants filled out a questionnaire about their performance and were debriefed by the experimenter.

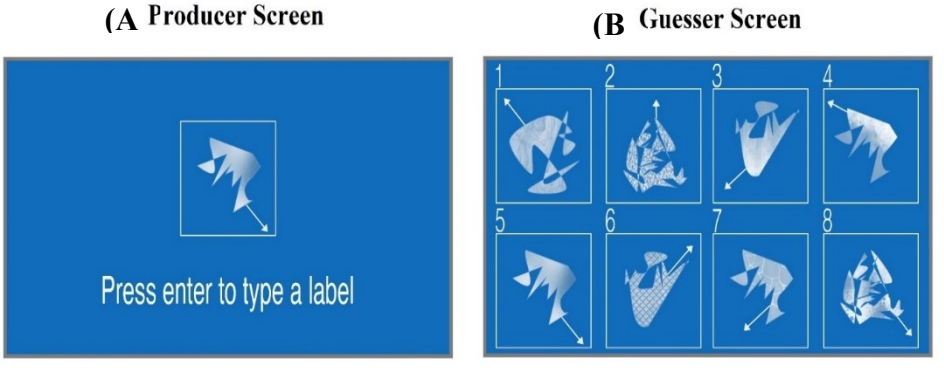


Figure 2. Example of the computer interfaces in a single interaction during the communication phase. The producer saw the target scene on their screen (A), while the guesser was presented with a grid of eight different scenes on their screen (the target and seven distractors; B). The producer typed a description for the target scene using the artificial language, and the guesser pressed the number associated with the scene they thought their partner was referring to. Paired participants alternated between the roles of producer and guesser. Note that scenes were dynamic events which included a moving shape. The arrows represent the direction of motion.

Network Properties

We created three different network structures: a fully connected network, a small-world network, and a scale-free network. Each network was comprised of eight individuals (also referred to as nodes or agents), but differed in how these individuals were connected to one another. Figure 1 shows the exact configuration of each network. Appendix A includes a detailed description of the order of interactions amongst pairs in each network condition. These networks can be described using formal measures that are typically used in graph theory. Below we characterize the three different networks used in this study in detail, and compare them based on the following three measures (see Tables 2 and 3).

Network density

This measure reflects the proportion of possible ties which are actualized among the members of a given network. It is measured as the ratio between the number of actual connections in the network and the number of all possible connections (Granovetter, 1976). A possible connection is one that could potentially exist between every two nodes. In a network with n individuals, the number of possible connections is $n*(n-1)/2$. By contrast, an actual connection is one that really exists in the given network. In a fully connected network where all possible connections are realized, density equals 1 (i.e., 100% connectivity). In a totally isolated network, in which there are no connections between nodes, density equals 0 (i.e., 0% connectivity). All other networks have density values between 0 and 1 (e.g., 0.5, or 50% connectivity, in our experiment).

Global clustering coefficient

This measure, also referred to as transitivity, reflects the degree to which nodes in the network tend to cluster together. In social networks, this measure indicates whether an individual's connections also tend to be connected to each other. In other words, it is the probability that two of one's friends are friends themselves. The global clustering coefficient equals 1 in a fully connected network where everyone knows everyone else, but has typical values in the range of 0.1 to 0.5 in many real-world networks (Girvan & Newman, 2002). For a given network, this measure is calculated in the following way: for a given node i, the local clustering coefficient is the ratio between the number of realized connections in the neighborhood of node i and the number of all possible connections in that neighborhood if it was fully connected. Then, the average of all nodes' local clustering coefficients yields the global clustering coefficient of the entire network (Watts & Strogatz, 1998).

Betweenness centrality

This measure reflects a given node's centrality, i.e., how necessary a specific node is for the communication between all the other nodes in the network. In social networks, this measure identifies the most important or influential individuals in the network. That is, having a high betweenness centrality value suggests that the node is necessary for mediating connections between otherwise unconnected nodes. It is calculated in the

following way: for a given node *i*, betweenness centrality is the number of times node *i* acts as a bridge along the shortest path between two other nodes (i.e., the number of shortest paths that pass through node *i*).

Table 2: Comparison of Networks' Density and Global Clustering Coefficients

	Fully connected	Small-world	Scale-free
Number of realized connections	28/28	14/28	14/28
Network density	100%	50%	50%
Global clustering coefficient	1	0.1667	0.4167

Table 3: Comparison of Nodes' Betweenness Centrality Across Network

	Fully connected		Small-world		Scale-free	
Node	connections	Betweenness	connections	Betweenness	connections	Betweenness
A	7	0	3	0.047619048	6	**0.325396825**
B	7	0	4	0.119047619	4	0.111111111
C	7	0	3	0.047619048	3	0.063492063
D	7	0	4	0.119047619	3	0.03968254
E	7	0	4	0.119047619	3	0.023809524
F	7	0	3	0.047619048	3	0.023809524
G	7	0	3	0.047619048	3	0.015873016
H	7	0	4	0.119047619	3	0.063492063

Condition 1: Fully connected network

In this condition, depicted in Figure 1A, all individuals in the network get to interact with one another. As such, all possible connections in the network are realized, and the network is maximally dense and maximally clustered (i.e., density and the clustering coefficient both equal 1). Since all individuals are directly connected to all others, the number of connections per node is identical (i.e., seven), and the betweenness centrality of each node equals 0 – no individual is necessary for the others to interact. In our experimental paradigm, it takes seven communication rounds for all pairs in the network to interact (see also Appendix A).

Condition 2: Small-world network

In this condition, depicted in Figure 1B, only half of the possible connections are realized. As such, this network is much sparser than the fully connected one, and its density is only 0.5 or 50%. In addition, every node in the network has a relatively similar number of connections, with each individuals connected to either three or four other individuals. An important feature of small-world networks, which is crucially present in our chosen network, is that the neighbors of any given node are also likely to be neighbors of each other (Watts & Strogatz, 1998). Therefore, unconnected nodes ("strangers") are still linked by a short chain of shared acquaintances. Indeed, every pair of individuals in our selected small-world network is linked by just one other individual, and typically there is more than one possible mediating individual (resulting in fairly similar and relatively low betweenness centrality values for all nodes, i.e., 0.047 and 0.119). For example, while participants G and H are not connected directly, they are nonetheless indirectly connected via participants F, D and B. In our experimental paradigm, it takes four communication rounds for all pairs in the network to interact (see also Appendix A).

Condition 3: Scale-free network

In this condition, depicted in Figure 1C, only half of the possible connections are realized, such that the network's density is identical to that of the small-world network in condition 2 (i.e., 50% connectivity). Scale-free networks are characterized by the same properties as small-world networks, with an additional important property: the distribution of node degree (i.e., the number of connections the node has to other nodes) follows a power-law (Barabási & Albert, 1999). That is, there are many low-degree nodes (individuals with fewer connections), and a few high-degree nodes (individuals with many connections). The less-connected individuals are often indirectly connected via the highly-connected agents, who are often referred to as "hubs". In our selected network, most participants (i.e., six out of eight) have only three connections, one participant has four connections, and one participant ("A", the hub) is connected to almost everyone else in the group. Accordingly, this participant has a very high betweenness centrality score compared to all other participants (i.e., 0.32 vs. 0.11, 0.06, 0.03, 0.02 and 0.01), indicating that they are central for the network's connectivity, and are necessary for connecting the other participants. In our experimental paradigm, it takes

six communication rounds for all pairs in the network to interact (see also Appendix A).

Measures

Communicative Success

Measured as binary response accuracy in a given interaction during the communication phase, reflecting comprehension.

Convergence

Measured as the similarities between all the labels produced by participants in the same group for the same scene in a given round: for each scene in round *n*, convergence was calculated by averaging over the normalized Levenshtein distances between all labels produced for that scene in that round. The normalized Levenshtein distance between two strings is the minimal number of insertions, substitutions, and deletions of a single character that is required for turning one string into the other, divided by the number of characters in the longer string. This distance was subtracted from 1 to represent string similarity, reflecting the degree of shared lexicon and alignment in the group.

Stability

Measured as the similarities between the labels created by participants for the same scenes on two consecutive rounds: for each scene in round *n*, stability was calculated by averaging over the normalized Levenshtein distances between all labels produced for that scene in round *n* and round *n+1*. This distance was subtracted from 1 to represent string similarity, reflecting the degree of consistency in the groups' languages.

Linguistic Structure

Measured as the correlations between string distances and semantic distances in each participant's language in a given round, reflecting the degree to which similar meanings are expressed using similar strings (Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015). First, scenes had a semantic difference score of 1 if they differed in shape, and 0 otherwise. Second, we calculated the absolute difference

between scenes' angles, and divided it by the maximal distance between angles (180 degrees) to yield a continuous normalized score between 0 and 1. Then, the difference scores for shape and angle were added, yielding a range of semantic distances between 0.18 and 2. Finally, the labels' string distances were calculated using the normalized Levenshtein distances between all possible pairs of labels produced by participant *p* for all scenes in round *n*. For each participant, the two sets of pair-wise distances (i.e., string distances and meaning distances) were correlated using the Pearson product-moment correlation, yielding a measure of systematic structure (Raviv et al., 2019a, 2019b).

Input Variability

Measured as the minimal sum of differences between all the labels produced for the same scene in a given round (Raviv et al., 2019b). For each scene in round *n*, we made a list of all label variants for that scene. For each label variant, we summed over the normalized Levenshtein distances between that variant and all other variants in the list. We then selected the variant that was associated with the lowest sum of differences (i.e., the 'typical' label) and used that sum as the input variability score for that scene, capturing the number of different variants and their relative difference from each other. Finally, we averaged over the input variability scores of different scenes to yield the mean variability in that round.

Analyses

We used mixed-effects regression models to test the effect of network condition on all measures using the lme4 package (Bates, Maechler Bolker & Walker, 2016) in R (R Core Team, 2016). The reported p-values were generated using the Kenward-Roger Approximation via the pbkrtest package (Halekoh & Højsgaard, 2014), which gives conservative p-values for models based on small numbers of observations. All models had the maximal random effects structure justified by the data that would converge, and are included in full in Appendix B. The data and the scripts for generating the models can be found at https://osf.io/utjsb/.

We examined communicative success, stability, convergence and linguistic structure using three types of models: (I) Models that predict changes in the dependent variable with respect to time and network condition; (II) Models that compare the different networks' final levels of

the dependent variable at the end of the experiment; (III) Models that predict the variance of the dependent variable with respect to time and network condition. In all models, NETWORK CONDITION was a three-level categorical factor that was simple-coded (i.e., similar to dummy-coding except that the intercepts correspond to the grand mean) with fully connected groups as the reference level. That is, we separately contrasted the small-world networks and the scale-free networks with the fully connected networks.

Models of type (I) predicted changes in the dependent variable over time as a function of network structure. Models for communicative success included data from communication rounds only (excluding the two test rounds). In models for communicative success, convergence, and stability, the fixed effects were NETWORK CONDITION, ROUND NUMBER (centered), ITEM CURRENT AGE (centered), and the interaction terms NETWORK CONDITION X ITEM CURRENT AGE and NETWORK CONDITION X ROUND NUMBER. ITEM CURRENT AGE codes the number of rounds each scene was presented until that point in time, and measures the effect of familiarity with a specific scene on performance. ROUND NUMBER measures the effect of time passed in the experiment and overall language proficiency. The random effects structure of models for communicative success, convergence, and stability included by-scenes and by-groups random intercepts and random slopes for the effect of ROUND NUMBER. As linguistic structure score was calculated for each producer over all scenes in a given round, the model for linguistic structure included fixed effects for NETWORK CONDITION, ROUND NUMBER (quadratic[10], centered), and the interaction term NETWORK CONDITION X ROUND NUMBER, as well as random intercepts and random slopes for the effect of ROUND NUMBER with respect to different producers nested in different groups.

Models of type (II) compared the mean values of the final languages in the last two relevant rounds of the experiment with respect to NETWORK CONDITION. The models for communicative success, stability and convergence included random intercepts for different groups, and the model for linguistic structure included random intercepts for different producers nested in different groups.

Models of type (III) predicted changes over time in the variance of each measure (i.e., the degree to which different groups differ from each other)

[10] We included both the linear and the quadratic terms using the poly() function in R to avoid collinearity.

as a function of NETWORK STRUCTURE. For linguistic structure, variance was calculated as the square standard deviation in participants' average structure scores across all groups in a given round. For communicative success, convergence and stability, variance was calculated as the square standard deviation in the dependent variable on each scene across all groups in a given round. All models included fixed effects for NETWORK CONDITION, ROUND NUMBER (centered), and the interaction between them. Models for communicative success, convergence and stability also included by-scenes random intercepts and random slopes for the effect of ROUND NUMBER.

Following (Raviv et al., 2019b), we also examined changes in input variability as a function of time and network structure. This model included fixed effects for NETWORK CONDITION, ROUND NUMBER (quadratic, centered), and the interaction between them, and by-group random intercepts and random slopes with respect to ROUND NUMBER. Finally, we examined changes in linguistic structure over consecutive rounds as a function of input variability. The dependent variable was the difference in structure scores between rounds n and $n+1$, the fixed effect was MEAN INPUT VARIABILITY at round n (centered), and there were random intercepts for different producers nested in different groups.

Results

Below we report the results for each of the four linguistic measures separately. All analyses are reported in full in Appendix B using numbered models, which we refer to here. Figure 3 summarizes the average performance of different network conditions over the course of the experiment, and Table 4 summarizes the main findings with respect to our predictions.

1. Communicative Success

Communicative Success increased over time (Model 1: β=0.1, SE=0.01, t=9.74, $p<0.0001$; Fig. 3A), indicating that participants became better at understanding each other as rounds progressed. All networks shows similar levels of accuracy overall (Model 1: Scale-free vs. fully connected: β=0.08, SE=0.27, t=0.3, p=0.76; Small-world vs. fully connected: β=-0.007, SE=0.27, t=-0.03, p=0.98), and the increase in accuracy over time was not significantly modulated by network structure (Model 1: Scale-

free vs. fully connected: β=0.01, SE=0.27, t=0.55, p=0.58; Small-world vs. fully connected: β=-0.01, SE=0.27, t=-0.49, p=0.62). Indeed, all networks reached similar levels of accuracy in the final communication rounds (Model 2: Scale-free vs. fully connected: β=0.26, SE=0.55, t=0.47, p=0.64; Small-world vs. fully connected: β=0.03, SE=0.55, t=0.05, p=0.96). No other effect was significant.

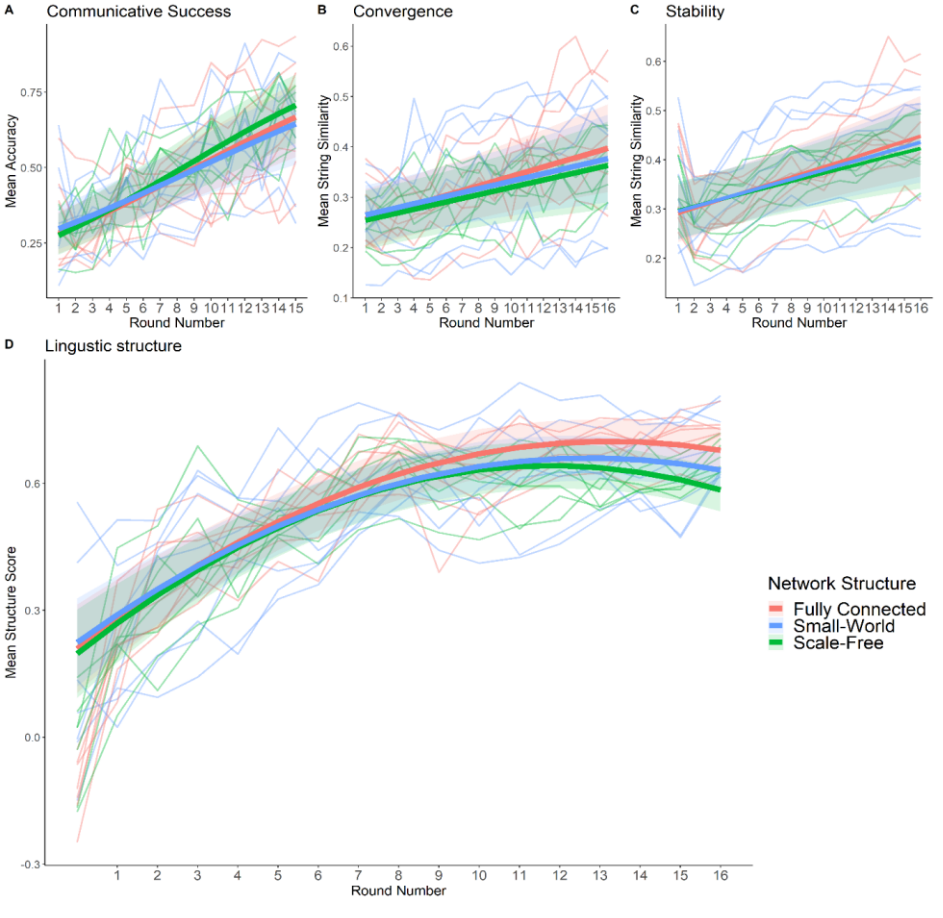


Figure 3. Changes in (A) Communicative Success, (B) Convergence, (C) Stability, and (D) Linguistic Structure over time as a function of network structure. Thin lines represent average values for each group in a given round. Thick lines represent the models' estimates, and their shadings represent the models' standard errors.

As for variance in communicative success, there was no significant difference across network structure conditions (Model 3: Scale-free vs. fully connected: β=0.001, SE=0.002, t=0.45, p=0.66; Small-world vs. fully connected: β=0.003, SE=0.002, t=1.06, p=0.3). Variance in accuracy generally increased over rounds (Model 3: β=0.001, SE=0.0004, t=2.97, p=0.006), but not in scale-free networks (Model 3: β=-0.001, SE=0.0006, t=-2.4, p=0.02). Together, these results indicate that while different groups differed from each other in their accuracy more and more as the experiment progressed (and especially those in the fully connected condition), the difference across groups in the scale-free condition did not change throughout the experiment.

2. Convergence

Convergence increased significantly over rounds (Model 4: β=0.008, SE=0.001, t=5.42, p<0.0001; Fig. 3B), with participants aligning, that is, using more similar labels over time. All networks shows similar levels of convergence overall (Model 4: Scale-free vs. fully connected: β=-0.018, SE=0.05, t=-0.4, p=0.7; Small-world vs. fully connected: β=-0.006, SE=0.05, t=-0.14, p=0.89), and the increase in convergence over time was not significantly modulated by NETWORK STRUCTURE (Model 4: Scale-free vs. fully connected: β=-0.002, SE=0.003, t=-0.61, p=0.55; Small-world vs. fully connected: β=-0.002, SE=0.003, t=-0.55, p=0.58). Indeed, all networks reached similar levels of convergence by the end of the experiment (Model 5: Scale-free vs. fully connected: β=-0.06, SE=0.06, t=-1.02, p=0.32; Small-world vs. fully connected: β=-0.05, SE=0.06, t=-0.92, p=0.37). Although there was no significant main effect of ITEM CURRENT AGE (Model 4: β=0.001, SE=0.0008, t=1.43, p=0.17), the interaction between NETWORK STRUCTURE and ITEM CURRENT AGE was significant, indicating that only fully connected networks showed greater convergence with item age compared to both sparse networks (Model 4: Scale-free vs. fully connected: β=-0.004, SE=0.002, t=-2.2, p=0.04; Small-world vs. fully connected: β=-0.005, SE=0.002, t=-2.83, p=0.01).

Network conditions significantly differed in their degree of variance overall, with scale-free networks showing the lowest variance, and small-world networks showing the highest variance (Model 6: Scale-free vs. fully connected: β=-0.007, SE=0.001, t=-5.91, p<0.0001; Small-world vs. fully connected: β=0.006, SE=0.001, t=5.07, p<0.0001). Variance in convergence increased over rounds (Model 6: β=0.0008, SE=0.0002, t=4.84, p<0.0001), but a significant interaction between ROUND NUMBER

and NETWORK CONDITION indicated that this was not the case for scale-free networks (Model 6: β=-0.001, SE=0.0003, t=-4.28, p=0.0001). Together, these results suggest that scale-free networks were most consistent in their convergence behavior, while small-world networks were least consistent and varied from each other in their convergence patterns. That is, while some small-world and fully-connected networks reached high levels of convergence, others maintained a high level of divergence throughout the experiment, with participants using their own unique labels. In contrast, scale-free networks behaved fairly similar to each other, and reached relatively similar convergence levels throughout the experiment.

3. Stability

Stability increased significantly over rounds (Model 7: β=0.009, SE=0.001, t=6.71, p<0.0001; Fig. 3C), with participants using labels more consistently as rounds progressed. All networks shows similar levels of convergence overall (Model 7: Scale-free vs. fully connected: β=-0.008, SE=0.04, t=-0.19, p=0.85; Small-world vs. fully connected: β=-0.002, SE=0.04, t=-0.06, p=0.96), and network structure did not modulate the increase in stability over rounds (Model 7: Scale-free vs. fully connected: β=-0.002, SE=0.003, t=-0.67, p=0.51; Small-world vs. fully connected: β=-0.001, SE=0.003, t=-0.36, p=0.73). Indeed, all networks reached similar levels of stability by the end of the experiment (Model 8: Scale-free vs. fully connected: β=-0.05, SE=0.06, t=-0.82, p=0.42; Small-world vs. fully connected: β=-0.05, SE=0.06, t=-0.85, p=0.4). Although there was no significant main effect of ITEM CURRENT AGE (Model 7: β=0.002, SE=0.0008, t=2.07, p=0.0516), the interaction between NETWORK STRUCTURE and ITEM CURRENT AGE was significant, indicating that only fully connected networks showed more stability with item age (Model 7: Small-world vs. fully connected: β=-0.005, SE=0.002, t=-3.06, p=0.006).

Additionally, and as in the case of convergence, network conditions significantly differed in their degree of variance overall, with scale-free networks showing the lowest variance, and small-world networks showing the highest variance (Model 9: Scale-free vs. fully connected: β=-0.006, SE=0.001, t=-6.35, p<0.0001; Small-world vs. fully connected: β=0.005, SE=0.001, t=5.65, p<0.0001). Even though there was no significant increase in variance in stability over rounds (Model 9: β=0.003, SE=0.0002, t=1.64, p=0.11), a significant interaction between ROUND NUMBER and NETWORK CONDITION indicated that variance

increased less over time in scale-free networks (Model 9: β=-0.0006, SE=0.0002, t=-2.77, p=0.009) In other words, while scale-free networks were most consistent in their behavior, and even more so as the experiment progressed, small-world networks varied most from each other in their stabilization patterns.

4. Linguistic Structure

Linguistic Structure significantly increased over rounds (Model 10: β=6.39, SE=0.36, t=17.51, p<0.0001; Fig 3D), with participants' languages becoming more systematic over time. The increase in structure over time was non-linear and leveled off in later rounds (Model 10: β=-2.92, SE=0.25, t=-11.76, p<0.0001). All networks shows similar levels of linguistic structure overall (Model 10: Scale-free vs. fully connected: β=-0.03, SE=0.04, t=-0.82, p=0.42; Small-world vs. fully connected: β=-0.02, SE=0.04, t=-0.48, p=0.64), and the increase in structure over time was not significantly modulated by network structure (Model 10: Scale-free vs. fully connected: β=-1.23, SE=0.89, t=-1.38, p=0.18; Small-world vs. fully connected: β=-0.93, SE=0.89, t=-1.04, p=0.31). Indeed, all networks reached similar levels of structure by the end of the experiment (Model 11: Scale-free vs. fully connected: β=-0.08, SE=0.04, t=-2.05, p=0.055; Small-world vs. fully connected: β=-0.05, SE=0.04, t=-1.27, p=0.22). These findings suggest that networks developed languages with systematic and compositional grammars, and did so to similar extents. To formally test this, we compared the level of structure in the final round of the experiment to chance using the Mantel test with respect to 1000 random permutations (for a similar procedure, see Kirby et al., 2008). Results indicated that the level of structure in all network conditions was significantly above chance (Fully connected networks: Mean structure score = 0.72, Mean *z*-score = 11.45, p<0.0001; Small-world networks: Mean structure score = 0.7, Mean *z*-score = 11.43, p<0.0001; Scale-free networks: Mean structure score = 0.67, Mean *z*-score = 10.98, p<0.0001). In these systematic languages, participants used complex labels for describing the scenes, with one part typically indicating the shape, and another part typically indicating motion (see Appendix C for multiple examples of final languages created by different groups).

Variance in structure significantly decreased over time (Model 12: β=-0.002, SE=0.003, t=-6.13, p<0.0001). Additionally, small-world networks were significantly more varied overall in terms of how structured their languages were (Model 12: Small-world vs. fully connected: β=0.02,

SE=0.003, t=5.29, p<0.0001). Given their greater variance to begin with, small-world networks also showed a faster decrease in variance over rounds (Model 12: Small-world vs. fully connected: β=-0.002, SE=0.0007, t=-2.86, p=0.006). These results suggest that even though small-world networks initially varied most in their level of structure, by the end of the experiment all networks showed similar and relatively little variability in their level of structure.

Following Raviv et al. (2019b), we also quantified the degree of input variability in each network at a given time point by measuring the differences in the variants produced for different scenes in different rounds. First, we tested whether input variability predicted changes in linguistic structure over consecutive rounds. Our results were in line with the findings of Raviv et al. (2019b), and confirmed that more input variability at round *n* induced a greater increase in structure at the following round (Model 13: β=0.02, SE=0.003, t=6.2, p<0.0001). We also found that input variability significantly decreased with time (Model 14: β=-23.69, SE=1.05, t=-22.58, p<0.0001), but the rate of the decrease slowed down in later rounds (Model 14: β=28.71, SE=0.99, t=28.95, p<0.0001). There was also a significant interaction between the linear term of ROUND NUMBER and NETWORK CONDITION (Model 14: Scale-free vs. fully connected: β=5.85, SE=2.57, t=2.27, p=0.028; Small-world vs. fully connected: β=7.54, SE=2.57, t=2.93, p=0.005), showing that input variability decreased more slowly in small-world and scale-free networks than in fully connected networks. Importantly, there was no significant main effect of NETWORK CONDITION (Model 14: Scale-free vs. fully connected: β=0.07 SE=0.18, t=0.37, p=0.71; Small-world vs. fully connected: β=0.05, SE=0.18, t=0.26, p=0.8). This result suggests that, in contrast to our original prediction (i.e., that sparse networks would show more variability), there was no effect of network structure on input variability, such that all networks had similar levels of input variability overall. Given the assumed causal relationship between the amount of input variability and the creation of more linguistic structure, the lack of difference in the degree of input variability across the different network conditions may explain why there was no effect of network structure on linguistic structure, as we further discuss below.

Discussion

The current study experimentally tested the effect of social network structure on the formation of new languages using a group communication paradigm. We compared the behaviors of groups that varied in their network architecture, contrasting three types of networks: (1) Fully connected networks, in which all members interact with each other; (2) Small-world networks, which are much sparser and have many members that never interact, although these "strangers" are nevertheless linked indirectly via a short chain of shared connections; And (3) Scale-free networks, which are as sparse as small-world networks, but whose members' distribution of connectivity roughly follows a power law such that one of the participants is highly connected to almost everyone in the network (a "hub") and others are much less connected.

Based on theoretical and computational models we generated several predictions (Table 1). First, we predicted that there would be more input variability in sparser networks, given that in such networks, some of the community members never directly interact (i.e., there are more strangers). We hypothesized that this greater input variability and difficulty in convergence would induce a stronger pressure for generalization and systemization, which would result in the sparser networks creating more systematic languages compared to fully connected networks. We further predicted that the emergence of more structured languages in sparser networks would facilitate convergence, allowing members of sparser networks to align on a shared language more easily and therefore resulting in similar convergence to that of fully-connected networks. Moreover, we predicted that scale-free networks would develop even more structured languages thanks to the existence of the hub, who can potentially promote the spread of conventions and systematic innovations. Furthermore, we predicted that sparser networks would stabilize to a lesser extent or slower compared to fully connected networks, given that changes take longer to stabilize in sparser networks. Finally, we predicted that all networks would reach similar levels of communicative success, such that across conditions, members that interacted with each other would understand each other equally well.

Table 4 summarizes our experimental results and compares them to our predictions. We found that over time, all groups developed languages that were highly systematic, communicatively efficient, stable, and shared across members. However, there were no significant differences between the three network conditions on any measure: All networks showed the

same behavioral patterns, had similar degrees of input variability, and reached similar levels of linguistic structure, stability, convergence and communicative success. While the results for communicative success and convergence are in line with our predictions (i.e., that all networks would show similar levels of communicative accuracy and global alignment), the remaining predictions were not met. Below we discuss potential reasons for this.

Table 4: Experimental Results vs. Predictions for Each Measure

Network Type / Measure	**Fully connected (FC)**	**Small-world (SW)**	**Scale-free (SF)**	**Prediction**	**Experimental Results**
Input Variability	More input variability in sparse networks			**FC < SF < SW**	**FC=SW=SF**
Linguistic structure	Sparse networks = more variability → More pressure for generalization and systematicity		Hubs can further promote the spread of systematic languages	**FC < SW < SF**	**FC=SW=SF**
Convergence	Sparse networks = more variability, more strangers → less convergence *BUT* Sparse networks =? more systematic languages → Similar levels of convergence			**FC=SW=SF**	**FC=SW=SF**
Stability	Dense networks = less diversity → More/faster stability	Sparse networks = more variability, more innovations → Less/slower stability		**FC > SW = SF**	**FC=SW=SF**
Communicative success	No difference between conditions			**FC=SW=SF**	**FC=SW=SF**

One consistent pattern that emerged from all our analyses, however, was that small-world networks showed the most variance in their observed behaviors, with different small-world networks behaving very differently from one another (not to be confused with the similar levels of input variability within each network). Fully connected networks and scale-free networks were generally similar to other fully connected networks and other scale-free networks respectively in terms of their convergence, stability and linguistic structure levels. However, small-world networks showed a great deal of variance, with different groups in the same condition showing very different levels of these three measures (also visually evident in Figure 3, which shows a high degree of dispersity for small-world networks). These results suggest that small-world networks may be more sensitive to random events (i.e., drift). Specifically, frequent interactions amongst small sub-groups can preserve random behaviors more easily, resulting in small-world networks being more likely to fixate on local (and possibly costly) strategies instead of converging on more optimal solutions (Bahlmann, 2014; Kurvers, Krause, Croft, Wilson, & Wolf, 2014). Our finding that small world networks show more variance in their linguistic behaviors also raises several predictions worth investigating. First, it suggests that changes in community structure across history that required greater geographical spread and reduced contact may have led to greater diversification in linguistic structure. Second, it might suggest that community structure can predict how likely communities are to exhibit common linguistic features compared to more rare ones (e.g., common vs. uncommon word order). Future research should investigate how community structure can influence the likelihood of a given language to follow or violate common trajectories of language change.

As mentioned earlier, the results of the study differed from those we had originally predicted. We predicted that different networks would show similar levels of convergence, but the rationale behind this prediction was not met. We hypothesized that the similar levels of convergence across networks would be the result of sparser networks initially showing greater input variability (hindering convergence in comparison to the fully connected networks), but that this greater variability would eventually lead sparser networks to create more systematic languages, which would in turn help them overcome this disadvantage. That is, our prediction was based on the idea that different networks would reach a sort of equilibrium between their difficulty to converge and their need to converge. Crucially, this was not the case: all networks showed similar levels of input variability and systematic

structure. This discrepancy fits our findings of equal convergence across conditions: different networks showed the same convergence patterns because their degree of input variability was the same.

While our results are surprising given the literature reviewed in the Introduction, they are in line with the computational model described in Spike (2017), who concluded that network structure plays a relatively small role in the development and maintenance of linguistic complexity and linguistic norms. This model simulated the process of conventionalization in populations of agents that varied in their community size, network structure, and learning biases (Spike, 2017). While the learning capacity of agents and the size of the population influenced the final outcomes of the model, results from multiple simulations showed that network structure had no apparent long-term effects on language change. Spike (2017) concluded that as long as populations exhibit a small-world property, i.e., that the average path-length between any two people is small (which is the case in all our three network conditions), the diffusion of variants across the network is sufficiently large to ensure similar linguistic trends. As in our experimental manipulation, real-world social networks are small-world in nature (Watts & Strogatz, 1998). That is, it is possible that network structure has little to no effect on the formation linguistic trends, at least in relatively natural networks.

However, we believe this interpretation is unlikely given the theoretical and computational models that argue in favor of network structure effects (Fagyal et al., 2010; Gong et al., 2012; Ke et al., 2008; Lou-Magnuson & Onnis, 2018). We believe it is more likely that the current study did not sufficiently capture the potential role of network structure. One possibility is that network structure interacts with group size in complex ways (as suggested by Lou-Magnuson & Onnis, 2018), and/or that network structure effects only manifest themselves once a certain group size threshold has been crossed. That is, it is possible that our eight-person networks were simply too small, and that running this experiment with bigger networks (e.g., of 200 people) would yield different results. Disentangling the relation between group size and network structure experimentally would require further investigation, potentially using online adaptations of this paradigm, which would allow testing much larger groups of interacting participants.

Another possibility is that, regardless of group size, our network structure manipulation was not strong enough to create meaningful

differences between network types, or was not representative of real-world differences between dense and sparse networks. For example, the sparse networks might not have been sparse enough, or the difference between the small-world and scale-free networks might have been too subtle. Notably, the nature of our experimental procedure restricted the specific architecture of sparser networks to a great extent. At any given communication round, each network had to be divided into pairs who play the game simultaneously, with no participant left out. Given this constraint, our choice of possible connections between group members was highly limited: many possible network configurations did not adhere to this constraint and were therefore inappropriate for our design. For the sake of illustration, imagine designing a four-person network that is sparsely connected, such that only four out of the six possible connections are realized. While there are 15 hypothetical network configuration that qualify this definition, only three of them satisfy the condition of being able to be divided into two unique pairs at a given time point and can therefore be used in our experimental paradigm. In the remaining 12 theoretical networks configurations, one participant would need to be included in two pairs at the same time, or would have no available communication partner. While it is relatively simple to find out which of the 15 hypothetical four-person networks could be suitable for our design, the problem was exponentially worse with the larger networks used in the current study: for sparser networks with eight individuals and 14 realized connections, there are over 40 million possibilities for network configurations, and only few on them are suitable for our design. As such, we cannot rule out the possibility that the networks that were selected for this experiment were not representative of real-world sparser networks, and/or had biased characteristics that made them too similar to one another. In other words, it is reasonable to assume that network structure had no effect in the current design because our selected networks did not differ sufficiently from each other. This possibility is supported by the lack of observed differences in input variability across conditions, which stands in sharp contrast with the general consensus that sparser networks should be more diversified (Bahlmann, 2014; Derex & Boyd, 2016; Liu et al., 2005).

The similar levels of input variability across network conditions may, in fact, explain the remaining results obtained in this study. Evidently, the prediction that sparse networks would show more input variability was a key component underlying the predictions for stability and linguistic structure. Since it turned out to be false, it is perhaps not surprising that

the predictions that were based on it also turned out to be false. In the case of stability, we hypothesized that more input variability in sparser networks would lead to slower or less stabilization in such networks. Given that there were no differences in input variability between the dense networks and sparse networks, it is not surprising that they also showed similar degrees of stability over time. In the case of linguistic structure, our prediction for structural differences between network conditions relied on the causal relation between input variability and systematic structure. This relation, i.e., that more input variability promotes more linguistic structure, was demonstrated in Raviv et al. (2019b) and further confirmed in the current study. We found that, across conditions and across experimental rounds, more input variability at time point *n* induced more structure at time point *n+1*. Therefore, if sparse networks indeed show greater input variability, they should consequently show more linguistic structure. However, if all networks show similar levels of input variability, they should also show similar levels of linguistic structure – which is what we found in the current study. Together, these results support the idea that network structure had no effect in our study because our selected networks did not differ substantially from each other. It is possible that a stronger manipulation of networks' sparsity would have yielded different results. Therefore, more research is required in order to confirm or to refute the influence of network structure on linguistic patterns.

We also predicted that scale-free networks would develop even more structured languages due to the existence of a highly-connected participant (a "hub"), who should potentially promote the spread of systematic variants to the entire community once they emerge (Fagyal et al., 2010; Zubek et al., 2017). This prediction was not met, and scale-free networks showed similar levels of linguistic structure to the other two network types. In retrospect, this discrepancy is very likely to be the result of the specific properties of our design: given that all networks in our experiment received the same amount of time for interaction (14 communication rounds in total, see Procedure) and given that each communication round included simultaneous communication between pairs, having more connections inevitably resulted in having less time to interact with each connection. Given these features, a highly connected participant would require more rounds to interact with all their possible connections, while a less-connected participant would in the meantime repeatedly interact with the same few connections. While such sub-groups can be seen as a relevant feature of sparser networks, this configuration

also resulted in the highly connected participants interacting less with each of their connections. That is, while the highly connected agent was indeed well-connected in the sense that they communicated with almost every person in the group, they were actually *less connected* to each person in terms of their frequency of interactions: the hub interacted approximately twice with each of their connections by the end of the experiment, while the less-connected participants interacted amongst themselves for approximately six times in the meanwhile. From the perspective of the less-connected participants, who repeatedly conversed with the same people and only rarely interacted with the hub, the hub could have effectively be seen as "an outsider", i.e., a person they rarely interacted with, and consequently a person who mattered less. That is, our design may have maintained the structural property of the hub but stripped it of their commonly associated social meaning, namely, having greater rather than lesser social importance. If true, it would again suggest that a different design or a different network selection would have revealed different results.

One possible way of dealing with the methodological issues described above is to move away from our current design and introduce more flexible communication conditions, while maintaining equal experience across all individuals in the group. For example, it is possible to include individual rounds or semi-communicative rounds, in which a participant is not assigned a partner, but nevertheless engages in some form of communicative behavior, for example with a computer-simulated agent. Alternatively, it is possible to introduce multi-player rounds, in which three participants are assigned to communicate together so that one participant produces a word and the other two participants guess the corresponding scene separately. Such modifications would dramatically improve the flexibility of our paradigm and expand the pool of suitable networks, while also introducing more varied conversational settings. Nevertheless, they introduce new challenges: the degree of input variability (and consequently, the difficulty of convergence) may be reduced if participants can interact with several group members at the same time, and it is not clear how to simulate a computerized participant in a way that mimics human participants' behavior and produces the same communicative challenges faced by people interacting with a real participants.

Finally, it is worth mentioning that our network structures were fixed, and did not change over time. Therefore, our sparse networks differed from real-world sparse networks in the sense that pairs of participants who

were not directly connected to each other would in fact *never* interact, and may soon figure this out. Some researchers have argued that an important feature of real-world sparse communities is the increased possibility of interacting with strangers, and treat interaction with strangers as a crucial mechanism driving morphological simplification (Wray & Grace, 2007). The idea behind this argument is that increasing the chances of interacting with unfamiliar people (with whom you have no shared history) introduces a stronger pressure for creating languages with simpler, transparent and regular structure (Granito, Tehrani, Kendal, & Scott-Phillips, 2019). In other words, the *potential* of encountering a new member of one's community may be relevant for explaining cross-linguistic differences. One way of testing this hypothesis is by introducing a more dynamic, open-ended network design to future studies, for example by assigning an unexpected connection every few rounds (so that individuals who are not directly connected may nevertheless encounter each other randomly from time to time).

Conclusions

The current study attempted to experimentally test the influence of social network structure on emerging languages using a group communication paradigm. We found no effect of network structure on any measure, with fully connected, small-world, and scale-free networks all showing similar patterns of communicative success, convergence, stability, and linguistic structure. We argue that these findings could be traced back to the lack of differences in input variability between network conditions in our current design, and that further research is needed in order to confirm or refute the postulated role of network structure on language evolution and language change. Nevertheless, our results show that network structure can significantly affect communities' susceptibility to drift, with small-world networks being more likely to vary from each other and fixate on local strategies.

Acknowledgments

We wish to thank Caitlin Decuyper for programming the experiment, and Phillip Alday, Mark Atkinson, Mattan S. Ben-Shachar, Rona Feldman, Gary Lupyan, Joe Rodd, William Schueller, and Jose Segovia-Martin for discussions and helpful input.

Appendix A: Network Details

Below we specify the exact configuration of each network condition in the experiment.

For each type of network, we include an adjacency matrix (also referred to as connection matrix) representing the network's structure using rows and columns for each node. Nodes (i.e., participants) range from A to H. Cell (*i*,*j*) in the adjacency matrix will have a value of 1 with dark shading if nodes *i* and *j* are connected, or a value of 0 with light shading otherwise. In undirected networks such as those used in this study, where communication goes in both ways, the adjacency matrix is symmetrical such that cell (*i*,*j*) and cell (*j*,*i*) have the same value.

For each type of network, we also include the order of communication between pairs over the course of 16 rounds.

Condition 1: Fully connected network

Table 1: Adjacency Matrix for the Fully Connected Network

Fully connected	A	B	C	D	E	F	G	H
A		1	1	1	1	1	1	1
B			1	1	1	1	1	1
C				1	1	1	1	1
D					1	1	1	1
E						1	1	1
F							1	1
G								1
H								

Table 2: Order of Pair-wise Interactions for the Fully Connected Network

Round	**Pair**			
1	A-B	C-D	E-F	G-H
2	H-A	B-G	F-C	D-E
3	A-D	E-B	C-H	G-F
4	C-A	E-G	B-D	F-H
5	D-H	G-C	B-F	A-E
6	G-D	F-A	C-B	H-E
7	D-F	E-C	F-B	A-G
8	TEST			
9	A-B	C-D	E-F	G-H
10	H-A	B-G	F-C	D-E
11	A-D	E-B	C-H	G-F
12	C-A	E-G	B-D	F-H
13	D-H	G-C	B-F	A-E
14	G-D	F-A	C-B	H-E
15	D-F	E-C	F-B	A-G
16	TEST			

Condition 2: Small-world network

Table 3: Adjacency Matrix for the Small-world Network

Small-world	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**
A		1	1	0	0	0	0	1
B			0	0	1	0	1	1
C				1	1	0	0	0
D					1	0	1	1
E						1	0	0
F							1	1
G								0
H								

Table 4: Order of Pair-wise Interactions for the Small-world Network

Round	**Pair**			
1	A-C	B-H	G-F	E-D
2	A-B	C-E	H-F	D-G
3	B-E	H-D	C-A	F-G
4	A-H	D-C	G-B	E-F
5	A-C	B-H	G-F	E-D
6	A-B	C-E	H-F	D-G
7	B-E	H-D	C-A	F-G
8	TEST			
9	A-H	D-C	G-B	E-F
10	A-C	B-H	G-F	E-D
11	A-B	C-E	H-F	D-G
12	B-E	H-D	C-A	F-G
13	A-H	D-C	G-B	E-F
14	A-C	B-H	G-F	E-D
15	A-B	C-E	H-F	D-G
16	TEST			

Condition 3: Scale-free network

Table 5: Adjacency Matrix for the Scale-free Network

Scale-free	A	B	C	D	E	F	G	H
A		1	0	1	1	1	1	1
B			1	0	1	0	1	0
C				1	0	0	0	1
D					0	0	1	0
E						1	0	0
F							0	1
G								0
H								

Table 6: Order of Pair-wise Interactions for the Scale-free Network

Round	**Pair**			
1	A-B	E-F	H-C	D-G
2	G-A	C-D	F-H	B-E
3	A-D	G-B	H-C	E-F
4	A-E	F-H	C-B	D-G
5	H-A	E-F	B-G	C-D
6	F-A	B-E	H-C	G-D
7	A-B	E-F	H-C	D-G
8			TEST	
9	G-A	C-D	F-H	B-E
10	A-D	G-B	H-C	E-F
11	A-E	F-H	C-B	D-G
12	H-A	E-F	B-G	C-D
13	F-A	B-E	H-C	G-D
14	A-B	E-F	H-C	D-G
15	G-A	C-D	F-H	B-E
16			TEST	

Appendix B: Models

All models use simple coding. Therefore the intercept and all effects are reported for the grand mean.

Communicative Success

(1) Type (I) Model: Accuracy over time

Accuracy ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 + centered.Round | ItemID) + (1 +centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	-0.1010	0.1135	-0.8898	0.3736
Round Number	0.1181	0.0121	9.7382	0.0000
Scale-free (SF) vs. Fully Connected (FC)	0.0807	0.2690	0.3002	0.7641
Small-world (SM) vs. Fully Connected (FC)	-0.0069	0.2688	-0.0255	0.9797
Item Current Age	-0.0007	0.0075	-0.0990	0.9212
Round Number X Condition (SF vs. FC)	0.0148	0.0268	0.5515	0.5813
Round Number X Condition (SM vs.FC)	-0.0132	0.0269	-0.4914	0.6232
Item Current Age X Condition (SF vs. FC)	-0.0193	0.0156	-1.2305	0.2185
Item Current Age X Condition (SM vs. FC)	-0.0049	0.0158	-0.3118	0.7552

(2) Type (II) Model: Final Accuracy comparison

MeanAccuracy ~ Condition + (1 | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.3923	0.2245	1.7473	0.0806
Scale-free (SF) vs. Fully Connected (FC) at rounds 14-15	0.2588	0.5492	0.4713	0.6374
Small-world (SM) vs. Fully Connected (FC) at rounds 14-15	0.0299	0.5510	0.0542	0.9567

(3) Type (III) Model: Accuracy variance

SD_Accuracy ~ centered.Round * Condition + (1 + centered.Round | ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.2356	0.0012	198.2406	0.0000
Round Number	0.0011	0.0004	2.9748	0.0061
Condition (SF vs. FC)	0.0011	0.0025	0.4462	0.6590
Condition (SM vs. FC)	0.0026	0.0025	1.0630	0.2973
Round Number X Condition (SF vs. FC)	-0.0014	0.0006	-2.4011	0.0236
Round Number X Condition (SM vs. FC)	-0.0008	0.0006	-1.3732	0.1811

Convergence

(4) Type (I) Model: Convergence over time

Convergence ~ centered.Round * Condition + centered.ItemCurrentAge * Condition + (1 + centered.Round | ItemID) + (1 + centered.Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.3191	0.0193	16.4973	0.0000
Round Number	0.0080	0.0015	5.4197	0.0000
Scale-free (SF) vs. Fully Connected (FC)	-0.0184	0.0466	-0.3952	0.6971
Small-world (SM) vs. Fully Connected (FC)	-0.0065	0.0466	-0.1392	0.8907
Item Current Age	0.0012	0.0008	1.4278	0.1694
Round Number X Condition (SF vs. FC)	-0.0021	0.0034	-0.6149	0.5458
Round Number X Condition (SM vs.FC)	-0.0019	0.0034	-0.5524	0.5871
Item Current Age X Condition (SF vs. FC)	-0.0039	0.0018	-2.2024	0.0401
Item Current Age X Condition (SM vs. FC)	-0.0050	0.0018	-2.8251	0.0107

(5) Type (II) Model: Final Convergence Comparison

MeanConvergence ~ Condition + (1 | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.3809	0.0233	16.3291	0.0000
Scale-free (SF) vs. Fully Connected (FC) at rounds 15-16	-0.0584	0.0571	-1.0221	0.3203
Small-world (SM) vs. Fully Connected (FC) at rounds 15-16	-0.0523	0.0571	-0.9161	0.3717

(6) Type (III) Model: Convergence variance

SD_Convergence ~ centered.Round * Condition + (1 + centered.Round | ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0227	0.0010	22.7241	0.0000
Round Number	0.0008	0.0002	4.8385	0.0000
Condition (SF vs. FC)	-0.0068	0.0012	-5.9114	0.0000
Condition (SM vs. FC)	0.0058	0.0012	5.0687	0.0000
Round Number X Condition (SF vs. FC)	-0.0011	0.0003	-4.2832	0.0001
Round Number X Condition (SM vs. FC)	0.0001	0.0003	0.4981	0.6217

Stability

(7) Type (I) Model: Stability over time

Stability ~ centered.Round * Condition + centered,ItemCurrentAge * Condition + (1 | ItemID) + (1 +centered,Round | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.3648	0.0187	19.5018	0.0000
Round Number	0.0094	0.0014	6.7163	0.0000
Scale-free (SF) vs. Fully Connected (FC)	-0.0084	0.0447	-0.1880	0.8528
Small-world (SM) vs. Fully Connected (FC)	-0.0025	0.0447	-0.0560	0.9559
Item Current Age	0.0016	0.0008	2.0710	0.0516
Round Number X Condition (SF vs. FC)	-0.0021	0.0031	-0.6704	0.5103
Round Number X Condition (SM vs.FC)	-0.0011	0.0031	-0.3563	0.7254
Item Current Age X Condition (SF vs. FC)	-0.0028	0.0016	-1.7533	0.0950
Item Current Age X Condition (SM vs. FC)	-0.0048	0.0016	-3.0625	0.0062

(8) Type (II) Model: Final Stability comparison

MeanStability ~ Condition + (1 | Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.4349	0.0225	19.3519	0.0000
Condition: Scale Free (SF) vs. Fully Connected (FC) at Rounds 15-16	-0.0452	0.0551	-0.8204	0.4227
Condition: Small World (SM) vs. Fully Connected (FC) at Rounds 15-16	-0.0470	0.0551	-0.8532	0.4048

(9) Type (III) Model: Stability variance

SD_Stability ~ centered.Round * Condition+ (1 + centered.Round | ItemID)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0183	1e-03	18.8913	0.0000
Round Number	0.0003	2e-04	1.6387	0.1109
Condition (SF vs. FC)	-0.0061	1e-03	-6.3515	0.0000
Condition (SM vs. FC)	0.0054	1e-03	5.6538	0.0000
Round Number X Condition (SF vs. FC)	-0.0006	2e-04	-2.7661	0.0093
Round Number X Condition (SM vs. FC)	0.0004	2e-04	1.7406	0.0912

Linguistic Structure

(10) Type (I) Model: Linguistic Structure over time

Linguistic Structure ~ poly(centered.Round,2) * Condition + (1 + poly(centeredRound ,2) | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.5437	0.0175	30.9805	0.0000
Round Number (Linear)	6.3878	0.3648	17.5087	0.0000
Round Number (Quadratic)	-2.9182	0.2482	-11.7561	0.0000
Scale Free (SF) vs. Fully Connected (FC)	-0.0351	0.0430	-0.8164	0.4246
Small World (SM) vs. Fully Connected (FC)	-0.0206	0.0430	-0.4804	0.6366
Round Number (Linear) X Condition (SF vs. FC)	-1.2341	0.8937	-1.3810	0.1837
Round Number (Quadratic) X Condition (SF vs. FC)	-0.4190	0.6080	-0.6890	0.4993
Round Number (Linear) X Condition (SM vs. FC)	-0.9311	0.8937	-1.0419	0.3108
Round Number (Quadratic) X Condition (SM vs. FC)	0.1089	0.6080	0.1792	0.8597

(11) Type (II) Model: Final Linguistic Structure comparison

MeanStructure ~ Condition + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.6722	0.0153	43.8621	0.0000
Condition: Scale Free (SF) vs. Fully Connected (FC) at Rounds 15-16	-0.0771	0.0375	-2.0532	0.0549
Condition: Small World (SM) vs. Fully Connected (FC) at Rounds 15-16	-0.0476	0.0375	-1.2677	0.2210

(12) Type (III) Model: Linguistic Structure variance

SD_Structure ~ centered.Round * Condition

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.0354	0.0014	25.5396	0.0000
Round Number	-0.0017	0.0003	-6.1305	0.0000
Condition (SF vs. FC)	0.0054	0.0034	1.5990	0.1168
Condition (SM vs. FC)	0.0180	0.0034	5.2931	0.0000
Round Number X Condition (SF vs. FC)	0.0000	0.0007	-0.0581	0.9539
Round Number X Condition (SM vs. FC)	-0.0020	0.0007	-2.8595	0.0064

Input Variability

(13) Changes in linguistic structure by input variability

StructureDiff ~ centered.MeanInputVariability + (1 | Group/Producer)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.03484474	0.003928631	8.869434	3.8000e-09
Mean Input variability	0.01711272	0.002760437	6.199277	1.8496e-06

(14) Input Variability over time

MeanInputVariability ~ poly(centeredRound ,2) * Condition + (1 + poly(centeredRound ,2)| Group)

	Estimate	Std.Error	t-value	p-value
(Intercept)	2.45956026	0.0750661	32.7652609	0.000000000
Round Number (Linear)	-23.69336281	1.0492407	-22.5814375	0.000000000
Round Number (Quadratic)	28.71256189	0.9918768	28.9477104	0.000000000
Scale Free (SF) vs. Fully Connected (FC)	0.06792667	0.1838736	0.3694204	0.713587191
Small World (SM) vs. Fully Connected (FC)	0.04752040	0.1838736	0.2584405	0.797272601
Round Number (Linear) X Condition (SF vs. FC)	5.84619297	2.5701043	2.2746910	0.027855087
Round Number (Quadratic) X Condition (SF vs. FC)	3.18386240	2.4295920	1.3104515	0.196845851
Round Number (Linear) X Condition (SM vs. FC)	7.53641263	2.5701043	2.9323373	0.005323155
Round Number (Quadratic) X Condition (SM vs. FC)	4.03674360	2.4295920	1.6614903	0.103728170

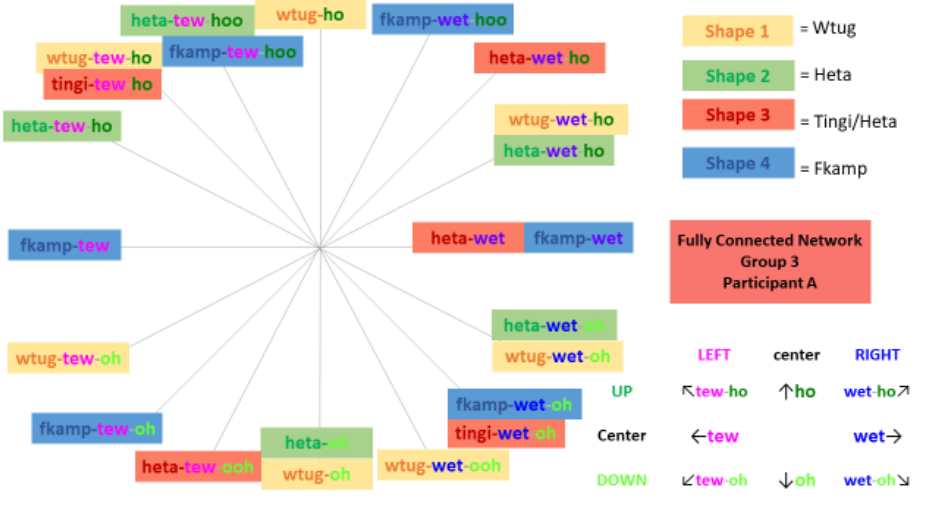
Appendix C: Examples of Structured Languages

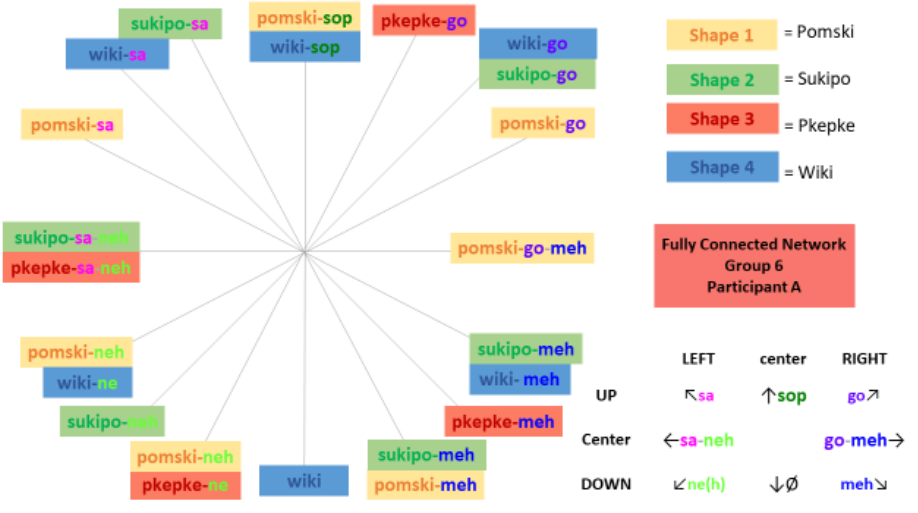
Below we include 15 examples of structured languages produced by participants in the final test round (round 16), with five examples from each network condition.

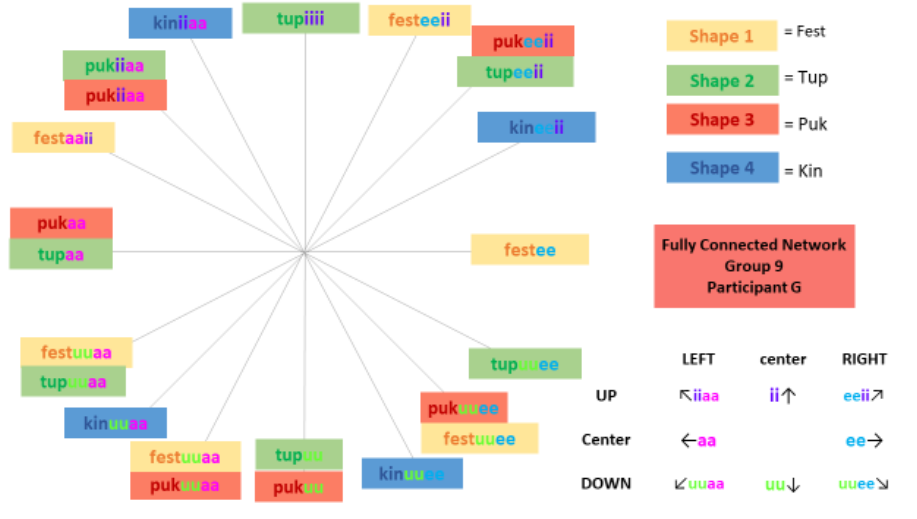
Each language is accompanied by a "dictionary" for interpreting the language on the right.

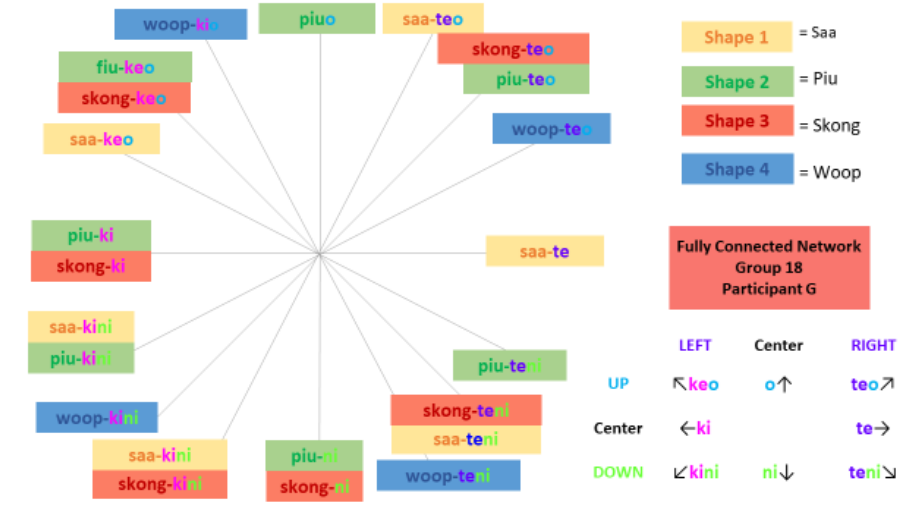
Different box colors represent the four different shapes which appeared in the scenes, and the grey axes indicate the direction in which the shape was moving on the screen. Different font colors represent different meaningful part-labels, as segmented by the authors.

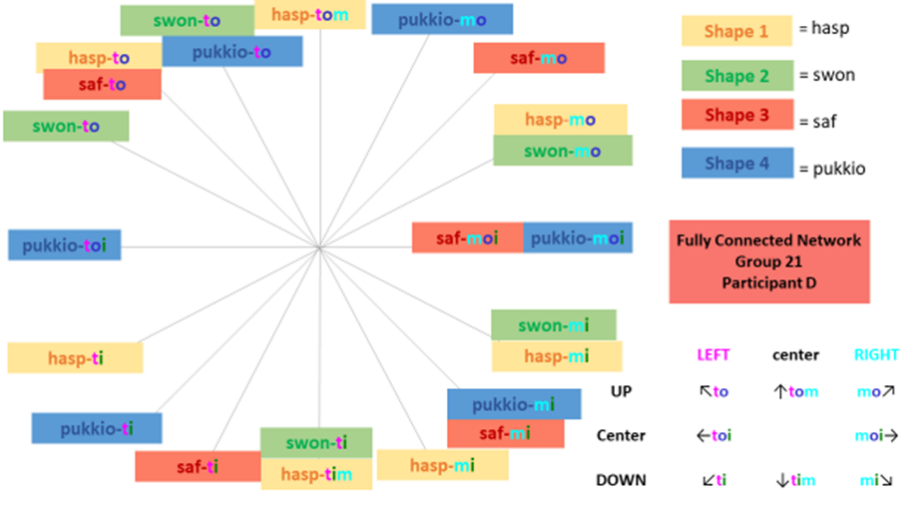
The dictionary and colors are solely for the purpose of illustration and were not used for any of the statistical analyses.

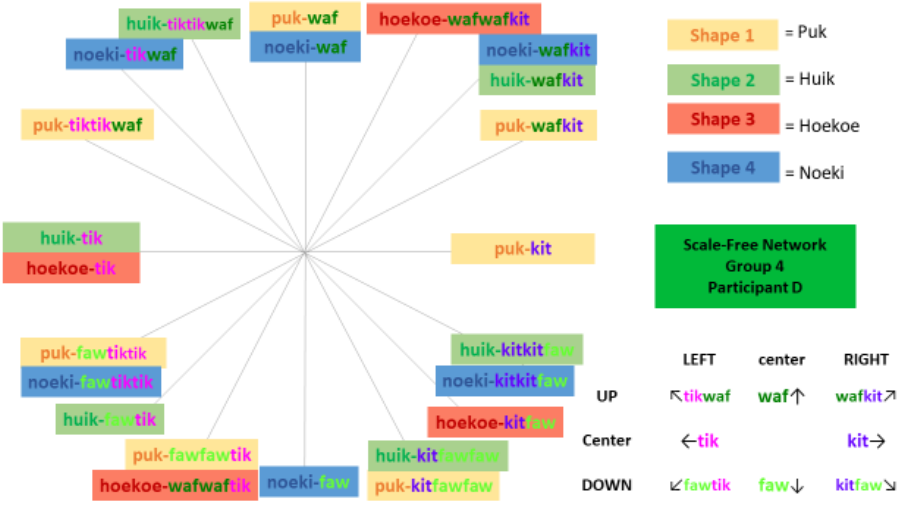


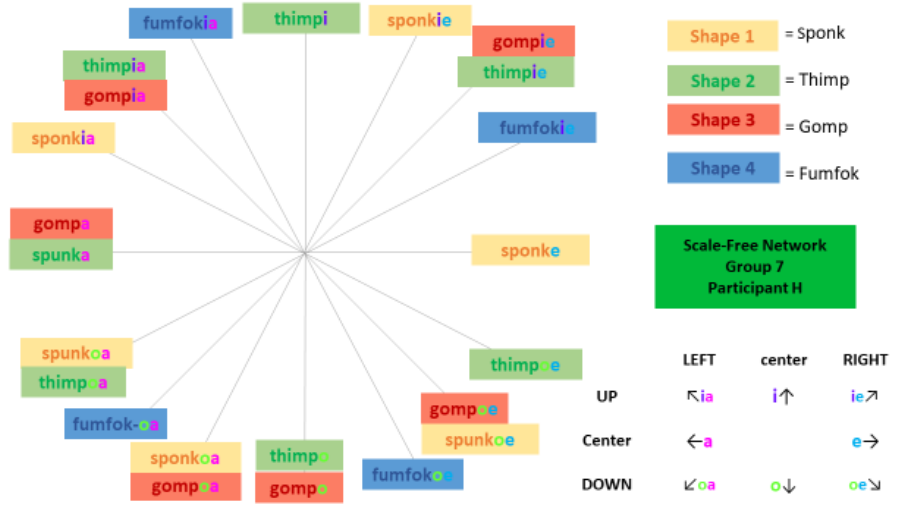

sukipo-sa
wiki-sa
pomski-sop
wiki-sop
pkepke-go
wiki-go
sukipo-go
pomski-sa
pomski-go
sukipo-sa-neh
pkepke-sa-neh
pomski-go-meh
pomski-neh
wiki-ne
sukipo-meh
wiki- meh
sukipo-neh
pkepke-meh
pomski-neh
pkepke-ne
wiki
sukipo-meh
pomski-meh
Shape 1 = Pomski
Shape 2 = Sukipo
Shape 3 = Pkepke
Shape 4 = Wiki
Fully Connected Network
Group 6
Participant A
LEFT center RIGHT
UP ↖sa ↑sop go↗
Center ←sa-neh go-meh→
DOWN ↙ne(h) ↓Ø meh↘

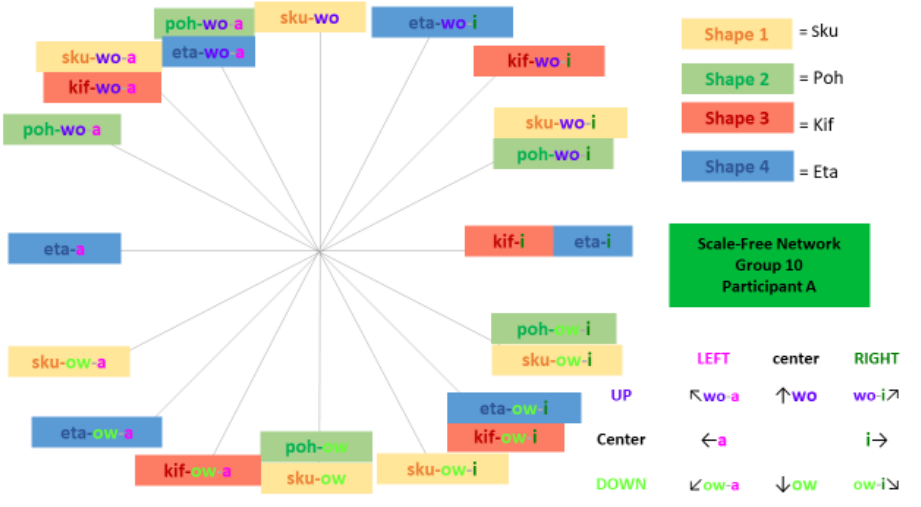

kiniiaa
tupiiii
festeeii
pukeeii
pukiiaa
pukiiaa
tupeeii
kineeii
festaaii
pukaa
tupaa
festee
festuuaa
tupuuaa
tupuuee
pukuuee
kinuuaa
festuuee
festuuaa
pukuuaa
tupuu
pukuu
kinuuee
Shape 1 = Fest
Shape 2 = Tup
Shape 3 = Puk
Shape 4 = Kin
Fully Connected Network
Group 9
Participant G
LEFT center RIGHT
UP ↖iiaa ii↑ eeii↗
Center ←aa ee→
DOWN ↙uuaa uu↓ uuee↘

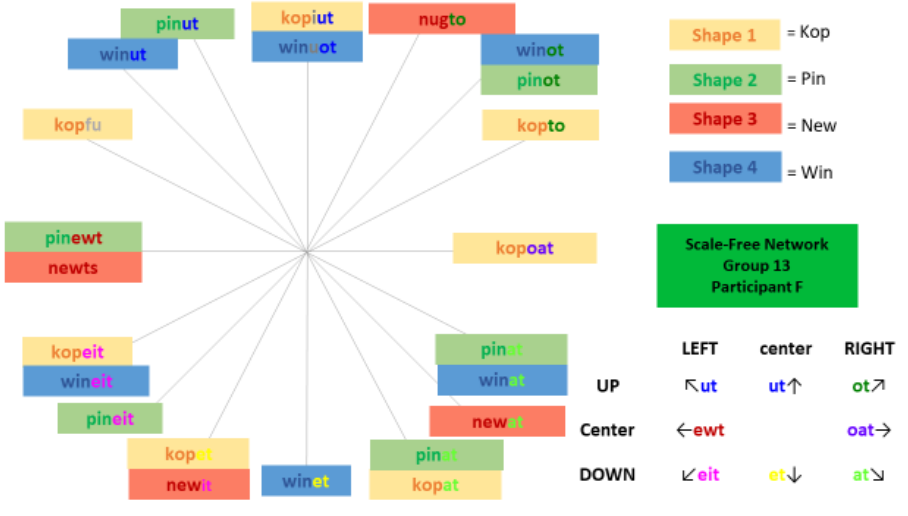

woop-kio
piuo
saa-teo
skong-teo
fiu-keo
piu-teo
skong-keo
woop-teo
saa-keo
piu-ki
saa-te
skong-ki
saa-kini
piu-kini
piu-teni
woop-kini
skong-teni
saa-teni
saa-kini
piu-ni
skong-kini
skong-ni
woop-teni
Shape 1 = Saa
Shape 2 = Piu
Shape 3 = Skong
Shape 4 = Woop
Fully Connected Network
Group 18
Participant G
LEFT Center RIGHT
UP ↖keo o↑ teo↗
Center ←ki te→
DOWN ↙kini ni↓ teni↘

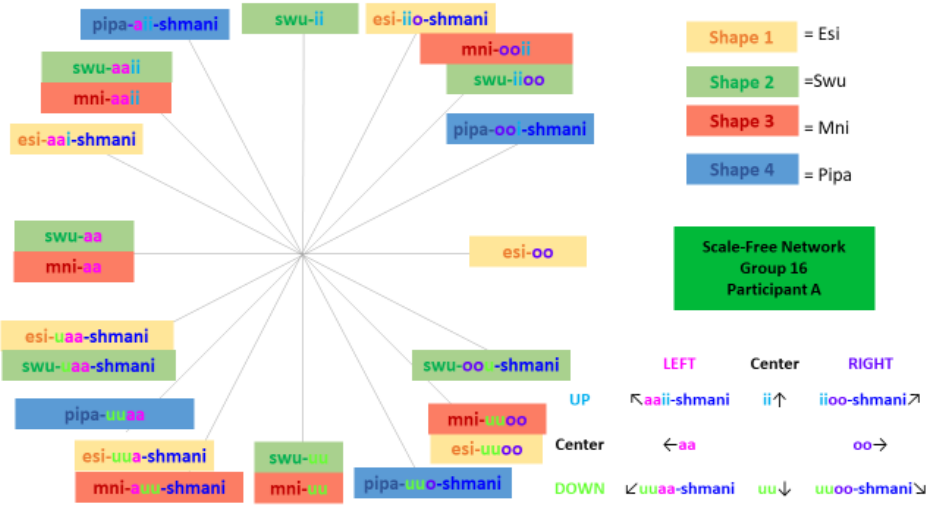

swon-to
hasp-tom
pukkio-mo
hasp-to
pukkio-to
saf-mo
saf-to
swon-to
hasp-mo
swon-mo
pukkio-toi
saf-moi
pukkio-moi
swon-mi
hasp-ti
hasp-mi
pukkio-mi
pukkio-ti
saf-mi
swon-ti
saf-ti
hasp-tim
hasp-mi
Shape 1 = hasp
Shape 2 = swon
Shape 3 = saf
Shape 4 = pukkio
Fully Connected Network
Group 21
Participant D
LEFT center RIGHT
UP ↖to ↑tom mo↗
Center ←toi moi→
DOWN ↙ti ↓tim mi↘

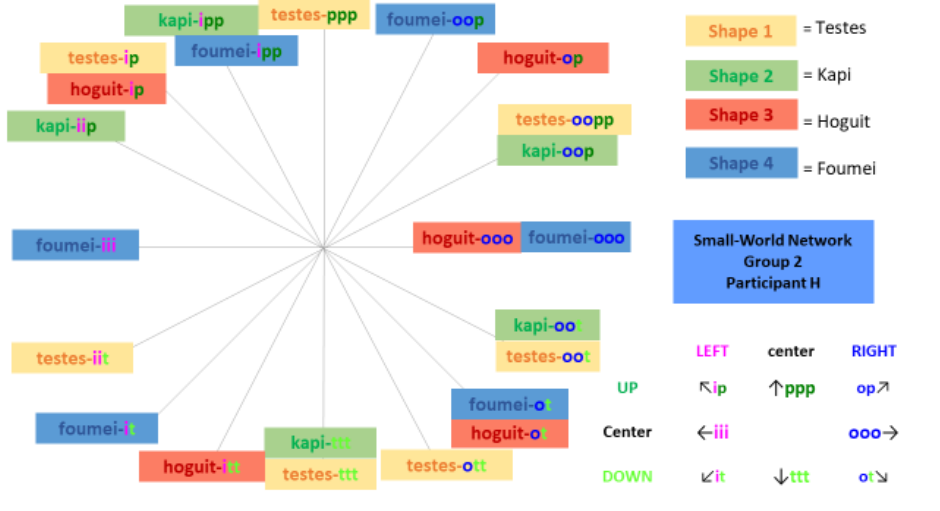

huik-tiktikwaf
puk-waf
hoekoe-wafwafkit
noeki-tikwaf
noeki-waf
noeki-wafkit
huik-wafkit
puk-tiktikwaf
puk-wafkit
huik-tik
hoekoe-tik
puk-kit
puk-fawtiktik
huik-kitkitfaw
noeki-fawtiktik
noeki-kitkitfaw
huik-fawtik
hoekoe-kitfaw
puk-fawfawtik
huik-kitfawfaw
hoekoe-wafwaftik
noeki-faw
puk-kitfawfaw
Shape 1 = Puk
Shape 2 = Huik
Shape 3 = Hoekoe
Shape 4 = Noeki
Scale-Free Network
Group 4
Participant D
LEFT center RIGHT
UP ↖tikwaf waf↑ wafkit↗
Center ←tik kit→
DOWN ↙fawtik faw↓ kitfaw↘

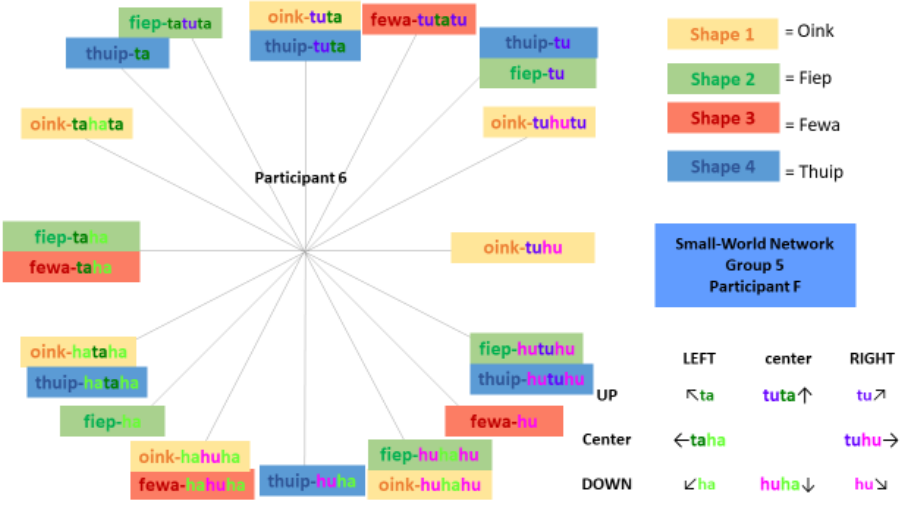

fumfokia
thimpi
sponkie
gompie
thimpia
thimpie
gompia
fumfokie
sponkia
gompa
spunka
sponke
spunkoa
thimpoe
thimpoa
gompoe
spunkoe
fumfok-oa
sponkoa
thimpo
fumfokoe
gompoa
gompo
Shape 1 = Sponk
Shape 2 = Thimp
Shape 3 = Gomp
Shape 4 = Fumfok
Scale-Free Network
Group 7
Participant H
LEFT center RIGHT
UP ↖ia i↑ ie↗
Center ←a e→
DOWN ↙oa o↓ oe↘

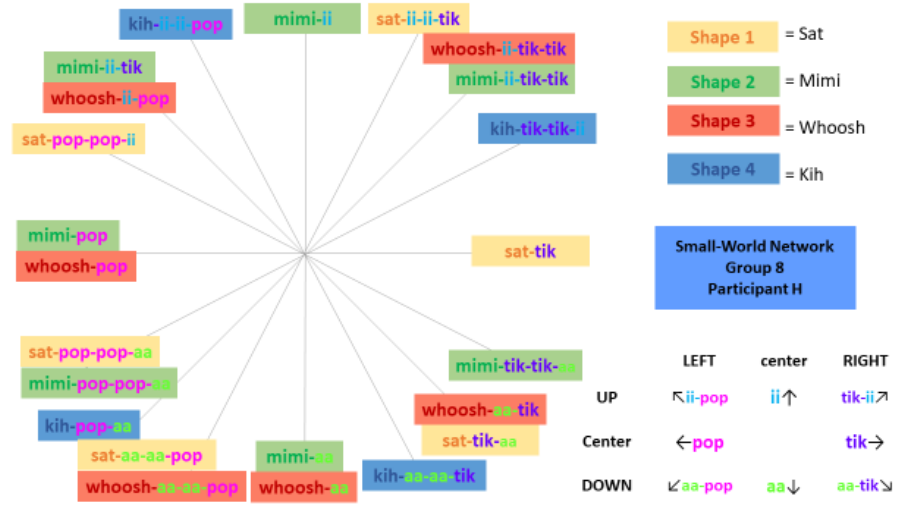

poh-wo a
sku-wo
eta-wo i
sku-wo-a
eta-wo a
kif-wo i
kif-wo a
sku-wo-i
poh-wo a
poh-wo i
eta-a
kif-i
eta-i
poh-ow-i
sku-ow-a
sku-ow-i
eta-ow-i
eta-ow-a
kif-ow-i
poh-ow
kif-ow-a
sku-ow
sku-ow-i
Shape 1 = sku
Shape 2 = Poh
Shape 3 = Kif
Shape 4 = Eta
Scale-Free Network
Group 10
Participant A
LEFT center RIGHT
UP ↖wo-a ↑wo wo-i↗
Center ←a i→
DOWN ↙ow-a ↓ow ow-i↘

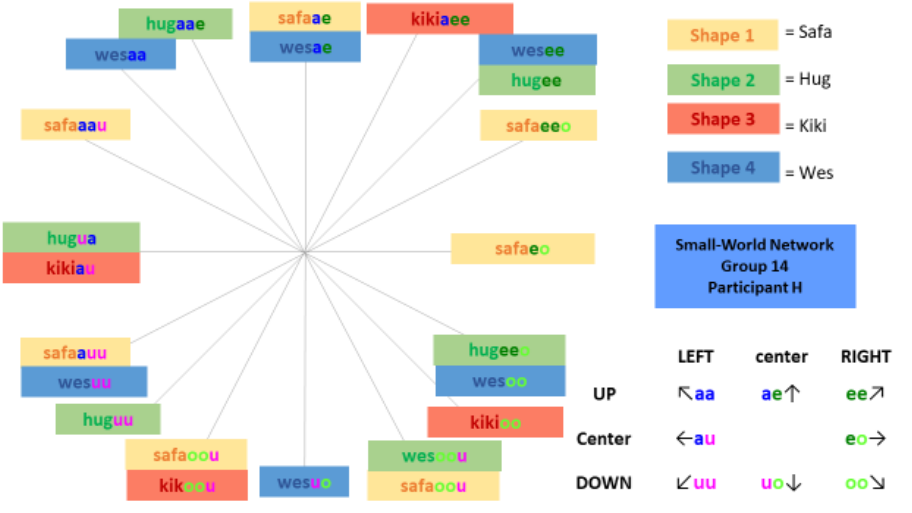

pinut
kopiut
nugto
winut
winuot
winot
pinot
kopfu
kopto
pinewt
newts
kopoat
kopeit
winelt
pineit
kopat
newit
winet
pinat
kopat
Shape 1 = Kop
Shape 2 = Pin
Shape 3 = New
Shape 4 = Win
Scale-Free Network
Group 13
Participant F
LEFT center RIGHT
UP ↖ut ut↑ ot↗
Center ←ewt oat→
DOWN ↙eit et↓ at↘

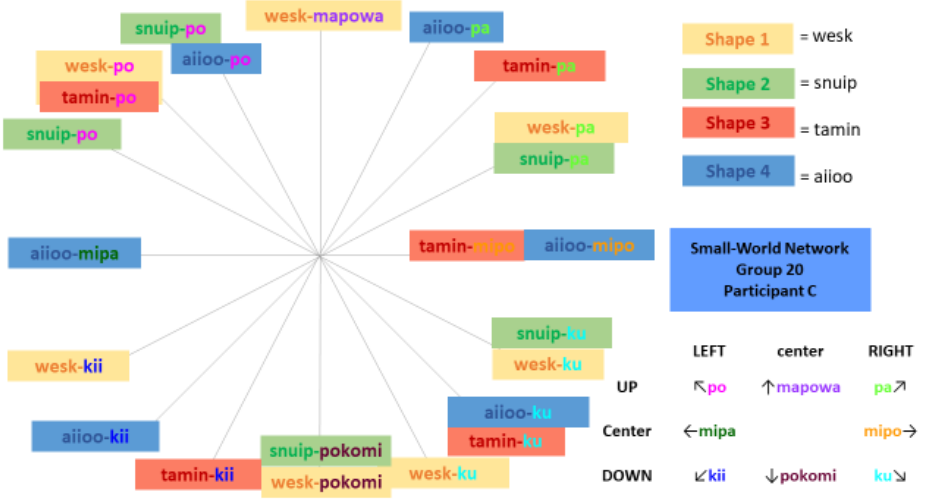

pipa-aii-shmani
swu-ii
esi-iio-shmani
mni-ooii
swu-aaii
mni-aaii
swu-iioo
esi-aai-shmani
pipa-oo -shmani
swu-aa
mni-aa
esi-oo
esi-uaa-shmani
swu- aa-shmani
swu-oo -shmani
pipa-uuaa
mni-uuoo
esi-uuoo
esi-uua-shmani
swu-uu
mni-uu
mni-auu-shmani
pipa-uuo-shmani
Shape 1 = Esi
Shape 2 =Swu
Shape 3 = Mni
Shape 4 = Pipa
Scale-Free Network
Group 16
Participant A
LEFT
Center
RIGHT
UP
↖aaii-shmani
ii↑
iioo-shmani↗
Center
←aa
oo→
DOWN
↙uuaa-shmani
uu↓
uuoo-shmani↘


kapi-ipp
testes-ppp
foumei-oop
testes-ip
foumei-ipp
hoguit-op
hoguit-ip
kapi-iip
testes-oopp
kapi-oop
foumei-iii
hoguit-ooo
foumei-ooo
kapi-oo
testes-iit
testes-oot
foumei-ot
foumei-it
hoguit-ot
kapi-
hoguit-it
testes-ttt
testes-ott
Shape 1 = Testes
Shape 2 = Kapi
Shape 3 = Hoguit
Shape 4 = Foumei
Small-World Network
Group 2
Participant H
LEFT
center
RIGHT
UP
↖ip
↑ppp
op↗
Center
←iii
ooo→
DOWN
↙it
↓ttt
ot↘


fiep-tatuta
oink-tuta
fewa-tutatu
thuip-ta
thuip-tuta
thuip-tu
fiep-tu
oink-tahata
oink-tuhutu
Participant 6
fiep-taha
fewa-taha
oink-tuhu
oink-hataha
thuip-hataha
fiep-hutuhu
thuip-hutuhu
fiep-ha
fewa-hu
oink-hahuha
fewa-hahuha
thuip-huha
fiep-huhahu
oink-huhahu
Shape 1 = Oink
Shape 2 = Fiep
Shape 3 = Fewa
Shape 4 = Thuip
Small-World Network
Group 5
Participant F
LEFT center RIGHT
UP ↖ta tuta↑ tu↗
Center ←taha tuhu→
DOWN ↙ha huha↓ hu↘


kih-ii-ii-pop
mimi-ii
sat-ii-ii-tik
whoosh-ii-tik-tik
mimi-ii-tik
whoosh-ii-pop
mimi-ii-tik-tik
kih-tik-tik-ii
sat-pop-pop-ii
mimi-pop
whoosh-pop
sat-tik
sat-pop-pop-aa
mimi-pop-pop-aa
mimi-tik-tik-aa
whoosh-aa-tik
kih-pop-aa
sat-tik-aa
sat-aa-aa-pop
mimi-aa
whoosh-aa-aa-pop
whoosh-aa
kih-aa-aa-tik
Shape 1 = Sat
Shape 2 = Mimi
Shape 3 = Whoosh
Shape 4 = Kih
Small-World Network
Group 8
Participant H
LEFT center RIGHT
UP ↖ii-pop ii↑ tik-ii↗
Center ←pop tik→
DOWN ↙aa-pop aa↓ aa-tik↘


hugaae
safaae
kikiaee
wesaa
wesae
wesee
hugee
safaaau
safaeeo
huguа
kikiau
safaeo
safaauu
wesuu
huguu
hugeeo
wesoo
kikioo
safaoou
kikoou
wesuo
wesoou
safaoou
Shape 1 = Safa
Shape 2 = Hug
Shape 3 = Kiki
Shape 4 = Wes
Small-World Network
Group 14
Participant H
LEFT center RIGHT
UP ↖aa ae↑ ee↗
Center ←au eo→
DOWN ↙uu uo↓ oo↘


snuip-po
wesk-mapowa
aiioo-pa
aiioo-po
wesk-po
tamin-pa
tamin-po
wesk-pa
snuip-po
snuip-pa
aiioo-mipa
tamin-mipa
aiioo-mipo
snuip-ku
wesk-kii
wesk-ku
aiioo-ku
aiioo-kii
tamin-ku
snuip-pokomi
tamin-kii
wesk-pokomi
wesk-ku
Shape 1 = wesk
Shape 2 = snuip
Shape 3 = tamin
Shape 4 = aiioo
Small-World Network
Group 20
Participant C
LEFT center RIGHT
UP ↖po ↑mapowa pa↗
Center ←mipa mipo→
DOWN ↙kii ↓pokomi ku↘

5 What makes a language easy to learn? A preregistered study on how systematic structure and community size affect language learnability

Abstract

Cross-linguistic differences in morphological complexity and social structure could have important consequences for language learning. Specifically, it is assumed that languages with more regular, compositional, and transparent grammars are easier to learn by both children and adults. It has also been shown that such grammars tend to evolve in bigger communities. Together, this suggests that some languages are acquired faster than others, and that this advantage can be traced back to community size and to the degree of systematicity in the language. However, the causal relationship between systematic linguistic structure and language learnability has not been formally tested, despite its importance for theories on language evolution, second language learning, and the origin of linguistic diversity. In this pre-registered study, we experimentally tested the effects of community size and systematic structure on language learning. We compare the acquisition of different yet comparable artificial languages that were created by either big or small groups in a previous communication experiment, and varied on their degree of systematic linguistic structure and their group size origin. We ask (a) whether more structured languages are easier to learn; and (b) whether languages created by bigger groups are easier to learn. Our results confirm that structured languages are advantageous for learning by adults, with highly systematic languages being learned faster and more accurately. We also found that the relationship between language learnability and linguistic structure is typically non-linear, so that high systematicity was indeed advantageous for learning, but learners did not seem to benefit from partly or semi-structured languages, i.e., languages that have some systematic rules but multiple irregulars and inconsistencies. Community size did not affect learnability: languages that evolved in big and small groups were equally learnable. Crucially, our results show that an important advantage of systematic structure is its productivity: with increasing structure, participants were better at generalizing the language they learned to new, unfamiliar meanings, and different participants were more likely to produce similar labels. That is, systematic structure may allow speakers to converge effortlessly, so that strangers can immediately understand each other.

Introduction

Languages differ greatly in how they map different meanings into morpho-syntactic structures (Dryer & Haspelmath, 2013; Evans & Levinson, 2009). Some languages appear to be relatively simple in terms of their morphology, while other languages are viewed as highly complex. For example, English makes minimal use of verb inflection to express grammatical relations: most English verbs have only one basic inflection paradigm to express time, such as adding [*-ed*] to express past tense, and this inflection is consistent across grammatical persons (i.e., *she* and *they* receive the same inflected form). Even verbs that are considered irregular in English (e.g., *sing, ring, buy, seek*) often follow a systematic inflectional rule (i.e., *sang, rang, bought, sought*). In contrast, Georgian has an elaborate set of verb inflection paradigms based on time, grammatical person, grammatical case, mood and more (Hewitt, 1995; Imedadze & Tuite, 1992). Verbs in Georgian can take an astonishing number of different forms (estimated at around 200), and many verbs are truly irregular and follow unique rules, requiring speakers to learn the inflections of these verbs independently. Beyond such anecdotal examples, cross-linguistic studies have confirmed that languages differ in their degree of morphological complexity (Ackerman & Malouf, 2013; Bentz & Berdicevskis, 2016; Hengeveld & Leufkens, 2018; Lewis & Frank, 2016; Lupyan & Dale, 2010; McCauley & Christiansen, 2019).

This cross-linguistic difference in morphological complexity may have important consequences for learning: some languages may be easier to learn than others. This idea goes against a wide-spread axiom in the field of linguistics, which is that all languages are equally difficult to learn and take the same effort to acquire (Sweet, 1899). Recent work has challenged this axiom, and provided initial support for the premise that languages differ in their degree of learnability. In particular, corpus studies report that the trajectory of children's first language acquisition (L1) can vary across languages (Armon-Lotem et al., 2016; Bleses, Basbøll, & Vach, 2011; Bleses et al., 2008; Dressler, 2003; Xanthos et al., 2011), and work on second language learning (L2) has shown that adults are better at learning some languages than others (Kempe & Brooks, 2008; Kempe & MacWhinney, 1998). These differences in learning outcomes and proficiency are often assumed to relate to several factors, amongst which is differences in languages' morphological complexity, i.e., the degree to which inflectional morphemes are informative, productive, and clearly marked. Specifically, languages with more regular, compositional, and

transparent structures are generally considered to be easier to learn by both children and adults when compared to languages with opaque and irregular structures (DeKeyser, 2005; Dressler, 2003, 2010; Hengeveld & Leufkens, 2018; Slobin, 1985).

While there is no widely accepted way to measure morphological complexity, various metrics have been used – from counting the number of inflected word forms per lemma (Xanthos et al., 2011), to conditional entropy (Ackerman & Malouf, 2013; Winters, Kirby, & Smith, 2015), type/token ratio (McCauley & Christiansen, 2019), and the degree of regularity in the mapping between forms and meaning (Cornish, Tamariz, & Kirby, 2009; Tamariz, Brown, & Murray, 2010; Tamariz & Smith, 2008). Although the quantitative definition of morphological complexity varies across researchers, its descriptive notion is relatively stable. Generally speaking, a language is considered to be simpler if it is compositional, regular, and transparent, i.e., if there are systematic one-to-one relations between units of meanings and units of form (DeKeyser, 2005; Hay & Baayen, 2005; Hengeveld & Leufkens, 2018). For example, the word [*walked*] consists of two parts: the verbal stem [*walk*] and the past tense morpheme [*ed*], which are combined in a transparent way to express the act of walking in the past. In comparison, the irregular past form [*bought*] cannot be as easily divided into separate bits, making it more holistic and opaque. Similarly, a language is considered to be more complex if the meanings of words are not directly predictable from their constituents. Such opacity can stem from multiple sources, such as having redundant or optional marking, syncretism, and/or a high prevalence of inconsistencies and irregularities. In this sense, more complexity is seen as the result of having less transparency. Complexity can also stem from having multiple inflectional paradigms and many obligatory grammatical rules. As such, the relation between complexity and transparency is not always straight-forward (Kempe & Brooks, 2018; Kempe & MacWhinney, 1998). For example, languages such as Russian have complex and elaborate inflectional paradigms with multiple grammatical cases, which are nevertheless transparent and informative; in contrast, languages such as German have considerably simpler paradigms with fewer grammatical cases, but high levels of syncretism that render the system fairly opaque and uninformative. While it is important to consider this potential discrepancy, the main theoretical notion of linguistic complexity incorporates the idea that more regularity, more transparency, and more compositionality are simpler and therefore beneficial for learning.

Intuitively, it seems reasonable that languages with more regular and compositional morphology will be easier to learn, given that they allow learners to derive a set of productive rules rather than memorizing individual forms (Kirby, 2002; Zuidema, 2003). This intuition is supported by information theory, as data with systematically recurring elements can be compressed into fewer bits. However, the causal relationship between linguistic structure and language learnability is currently untested. Very few studies have attempted to examine this link by investigating learning difficulty as a function of linguistic complexity. Only a handful of correlational and experimental studies have examined learning outcomes and learning trajectories in natural languages that differ in their morphological complexity. These studies exhibit a mixed patterns of results: some studies report slower acquisition and worse overall proficiency in languages that are more morphologically opaque (Kempe & Brooks, 2008; Kempe & MacWhinney, 1998; Slobin, 1985), while others suggest similar learning rates across languages (Armon-Lotem et al., 2016; Braginsky, Yurovsky, Marchman, & Frank, 2019), or the opposite pattern altogether, i.e., that morphologically complex languages are acquired faster (Dressler, 2003; Xanthos et al., 2011). These conflicting findings can be related to different complexity metrics used across studies, different variables of interest (e.g., acquisition of vocabulary, passive constructions, verb inflections, case marking, etc.), and/or the presence of multiple confounding factors in natural languages such as phonological complexity, inconsistent word order, and more.

Crucially, no study to date has systematically compared the acquisition of a broad yet controlled range of morphological structures using an experimental paradigm. As such, it is not clear whether languages with more regular and transparent structures are indeed easier to learn. While direct empirical evidence for this argument is lacking, two studies provide it with some initial support. Brooks et al. (1993) and Monaghan, Christiansen, and Fitneva (2011) both conducted artificial language learning experiments to test the acquisition of artificial languages that differed in their degree of sound-systematicity, i.e., the mapping between forms and categories (Brooks, Braine, Catalano, Brody, & Sudhalter, 1993; Monaghan, Christiansen, & Fitneva, 2011). In the studies, participants were trained on a miniature vocabulary containing two word classes, corresponding to grammatical gender (Brooks et al., 1993) or to actions/objects (Monaghan et al., 2011). In the arbitrary condition, there were no similarities between the words' phonological forms and their grammatical class, such that different sounds were distributed randomly

between the two classes. This condition was contrasted with a fully systematic condition, in which words from the same grammatical category contained distinct sounds (e.g., words for objects contained fricatives while words for actions contained plosives)Monaghan et al., 2011), and with a partially systematic condition in which members of each noun class shared a subset of phonological features (Brooks et al., 1993; Monaghan et al., 2011). Results showed that participants were better at learning the distinction between the two categories when there was full or partial systematicity in the mapping between forms and meanings, i.e., when there was a phonological cue indicating the nouns' grammatical category. These findings provide initial support for the idea that learning outcomes can be affected by the degree of systematic structure, at least in terms of grammatical categories being systematically mapped to specific sounds. But since these studies did not directly test the effect of morphological complexity or compositionality, they are not sufficient for concluding that compositional, transparent, and regular languages are indeed easier to learn.

Nonetheless, the causal relationship between systematic linguistic structure and language learnability serves a crucial component in two strands of influential literature: (a) language evolution simulations on the emergence of linguistic structure, and (b) the social origin of linguistic diversity. The assumption that transparent and regular grammars are more easily learned is essential for the theoretical reasoning in both fields. Therefore, it is important to validate the postulated effect of linguistic structure on language learning. The current study aimed to fill in this gap and experimentally test the learnability of artificial languages that vary in their degree of systematic structure (i.e., in how transparent, compositional, and predictable the mapping between meanings and labels is).

In the first line of research, language evolution models explicitly argue that compositional structure emerges as a consequence of learnability pressures combined with expressivity pressures, and that compositional structure facilitates accurate transmission of languages over multiple generations of learners, who would struggle to learn a holistic and unstructured lexicon (e.g., Cornish et al., 2009; Kirby et al., 2015, 2008; Smith, 2011). Using iterated learning and diffusion chain paradigms, multiple studies have reported that, over time, artificial languages become more compositional (as reflected by greater form-meaning mapping) and more faithfully reproduced (as reflected by fewer transmission errors). The observed increase in language learnability over time is argued to be

the direct result of the increase in linguistic structure, given that systematic languages are supposed to be easier to learn (i.e., there are fewer unique forms to remember, and thus it is easier to predict the forms of unseen meanings given a limited subset of examples). The emergence of compositional languages is therefore attributed to learning pressures: because of cognitive limitations on learners' memory, more compressed and predictable signals are favored, and these in turn ease the learning process and allow for more accurate reproduction (Cornish, 2010). Moreover, compositional languages are argued to be advantageous for generalizations, and allow learners to overcome the "poverty of stimulus" (Kirby, 2002; Kirby, Smith, & Brighton, 2004; Zuidema, 2003): because learners must acquire their linguistic competence from finite and partial input, languages with more regular and transparent structures are favored since they allow learners to easily refer to new, unfamiliar meanings using the same system. In other words, iterated learning studies assume a close and causal relationship between linguistic structure and learnability, and the hypothesized mechanism behind the emergence of structure strongly relies on the intuition that more systematic languages are easier to learn and are more generalizable.

Notably, iterated languages learning studies typically report a simultaneous increase over time in both systematic structure and learnability, which is taken as evidence that more structured languages are easier to learn (Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015; Saldana, Fagot, Kirby, Smith, & Claidière, 2019). Yet crucially, these studies typically do not examine this relation directly (e.g., using statistical analyses to demonstrate a significant correlation between structure and learning), and do not test the *causal* relation between linguistic structure and learnability beyond the mediating variable of generation number. As such, iterated language learning paradigms have not directly confirmed the causal role of linguistic structure on learning. Nevertheless, there is some evidence in support of the correlation between accuracy and systematicity in such paradigms. For example, Tamariz and Smith (2008) found that participants who learned languages with more regular form-to-meaning mappings also produced languages with more regular form-to-meaning mappings, but participants' accuracy in learning the input language was not reported. Another study reported a significant correlation between learning accuracy and producing systematic structure, albeit in the opposite direction of causality: Raviv and Arnon (2018) reported that transmission error was a significant negative predictor of linguistic structure across all generations of learners, so that participants

who showed better learning of the input language also introduced more linguistic structure when reproducing the language. Interestingly, the results of one iterated learning study suggest that linguistic structure and learnability are not always related: Berdicevskis (2012) found that even though artificial languages became more compositional over generations of learners, they did not become more learnable: there was no significant increase in reproduction fidelity over time despite the increase in systematic structure, and there was no correlation between how compositional languages were and how accurately they were learned. That is, the increase in linguistic structure did not facilitate learning. Together, these findings strengthen the need for conducting a careful examination of the causal relation between language learnability and systematicity.

As for the second line of research, on the social origin of linguistic diversity, cross-linguistic work has found that languages spoken by big communities are typically less morphologically complex than languages spoken by small communities – a finding that has been attributed to learnability pressures caused by the presence of a higher proportion of second-language learners in larger communities (Bentz, Verkerk, Kiela, Hill, & Buttery, 2015; Bentz & Winter, 2013; Lupyan & Dale, 2010). Specifically, the inverse correlation between morphological complexity and population size (which is taken a proxy for the proportion of non-native speakers) is argued to be driven by the difficulty of adult L2 learners in acquiring morphologically complex and opaque languages (Bentz & Berdicevskis, 2016; Dale & Lupyan, 2012; Lupyan & Dale, 2010, 2016; McWhorter, 2007; Trudgill, 1992, 2002, 2009). In other words, the reduced morphological complexity observed in languages of larger communities is argued to be the direct result of the postulated relationship between linguistic structure and learnability, which is presumably amplified in adults.

According to this theory, languages adapt to fit their social niches: if adult non-native speakers constitute a substantial part of the community (as is typically the case in larger societies), then adults' difficulty in learning complex morphological structures may push languages towards simplification and regularization. This line of reasoning includes three explicit assumptions: (1) that morphologically simpler languages are advantageous for learning; (2) that adults struggle more with learning complex morphological structures; and (3) that imperfect learning by non-native speakers can lead to morphological simplification over time due to cross-generational transmission and/or accommodation by native speakers. The third assumption has been confirmed experimentally in

three studies, which report that given insufficient exposure, adults tend to simplify a morphologically complex artificial language (Atkinson, Smith, & Kirby, 2018; Bentz & Berdicevskis, 2016), and that native speakers tend to adapt more to the syntactic choices of non-native confederates, even when they produce ungrammatical sentences (Loy & Smith, 2019). The second assumption receives support from the literature on second language learning, which suggests that adults generally struggle with learning and using morphology in a second language compared to children: adults L2 speakers typically show optional or variable use of verbal and nominal inflections related to case marking, tense, agreement, aspect, and gender marking (DeKeyser, 2005; Haznedar, 2006; Parodi, Schwartz, & Clahsen, 2004), and learn faster when languages exhibit more reliable morphological cues (Kempe & MacWhinney, 1998). Importantly, the first assumption, which is essential for the theory's main argument, has yet to be tested.

An interesting alternative explanation for the documented correlation between morphological complexity and community size is that, instead of being mediated by the proportion of adult non-native speakers and their difficulty in language learning, it is directly derived from differences in community size itself (Nettle, 2012; Raviv, Meyer, & Lev-Ari, 2019b; Wray & Grace, 2007). According to this hypothesis, the total number of speakers in the community can affect language structure in relevant ways, and there is no need to assume the prevalence of second language learning as a mediating factor: big communities might develop simpler and more transparent languages simply because they are big. In a big community with many individuals there is more input variability, and people have less common ground with one another (some people are strangers who rarely or never interact). Given that each person in the community may have unique and possibly unfamiliar morpho-lexical variations, interacting with more people without establishing common ground can be increasingly taxing for individuals' memory, making it harder to maintain a holistic, irregular or unstructured lexicon as community size grows. Therefore, big communities are under a stronger pressure to develop systematic and transparent languages that can in turn facilitate convergence and mutual understanding. As such, members of larger communities may favor simpler and more compositional linguistic variants that are easier to remember. In contrast, small communities have fewer individuals (and therefore, less variability overall), and its members typically have more shared history with each other and more common knowledge as a result of frequent interactions. As such, members of small

communities should be able to rely on common ground and language-external knowledge when communicating, and consequently, may develop and more easily maintain languages with more complex, arbitrary and/or idiosyncratic variants.

The hypothesis that community size can shape the structure of languages was recently confirmed using a group communication experiment, in which groups of four or eight interacting participants needed to create a new artificial language to communicate with each other about different novel scenes (Raviv et al., 2019b). Results showed that larger groups developed more systematic and compositionally structured languages over time, and did so faster and more consistently than small groups. Furthermore, the increase in linguistic structure was driven by the greater input variability in the larger groups, and facilitated better convergence and accuracy. In other words, the emergence of more systematic and compositional languages in larger groups was indeed advantageous for communication, and allowed larger groups to converge and understand each other equally well as small groups despite being faced with a greater communicative challenge (i.e., interacting with more people while having less shared history with each person). Together, the findings of Raviv et al. (2019b) show that community size has a unique and causal role in shaping linguistic structure beyond learning constraints, and that having languages with systematic structure can facilitate convergence between more individuals. Importantly, this line of reasoning still implies that there is a processing advantage for compositional and predictable variants: systematic languages should be more efficient for learning and use because they ease individuals' memory load and allow them to communicate more successfully and more productively.

Notably, an important question inevitably arises: are the languages of big communities easier to learn? That is, did larger groups create languages that would be better acquired by new members of the community? This question draws a direct link between the two literatures discussed above, i.e., iterated language learning and the social origin of linguistic diversity: if larger communities have more systematic languages, and if more systematic languages facilitate learning by the next generation of learners, the languages of larger groups should be better acquired by naïve individuals. In other words, languages used in big communities may be more learnable because typically they are also more systematically structured. If true, then languages created in larger communities may be easier to learn for adult second language learners not because of the presence of such L2 learners to begin with, but simply

because of the size of the community. This idea shifts the explanatory load offered in previous work (e.g., Lupyan & Dale, 2010) from constraints of language learning to constraints on efficient communication between strangers (e.g., Wray & Grace, 2007).

There may be additional advantages to signals developed in bigger groups, above and beyond their degree of systematicity. Considering the fact that languages developed in larger communities have passed the processing filter of more people and were used by more different individuals, it is possible that they are better adapted to humans' general preferences. Specifically, computational models of iterated learning have shown that languages adapt to fit individuals' cognitive biases over time, and that weak individual tendencies can become greatly amplified as languages are transmitted by more and more individuals (Kirby, Dowman, & Griffiths, 2007; Kirby et al., 2004; Reali & Griffiths, 2009; Smith, 2011). As such, it is possible that signals that evolved in larger communities are somehow better fitted to our cognitive and learning biases, and are therefore more efficient for processing, learning and use. If so, we may expect that languages of larger communities would be easier to learn for reasons other than their structure, i.e., even when they have similar degrees of systematic structure as languages of small communities.

While there is no direct evidence that languages of larger communities are easier to learn, one study has attempted to test the effect of group size on the complexity and transparency of linguistic conventions that were created by two vs. three individuals (Atkinson, Mills, & Smith, 2018). In that study, dyads or triads used English to describe novel icons to each other, and their final descriptions were transmitted to naïve learners who had to match them to their referents. Atkinson et al. (2018) found that matching accuracy did not differ significantly across conditions (i.e., the descriptions created by two vs. three people were guessed equally well), providing no evidence that larger communities create more transparent form-to-meaning mappings. However, we cannot draw strong conclusions from this null result. It is possible that the group size manipulation used in that study was not sufficiently strong (i.e., contrasting productions by two vs. three people may not be enough to detect community size differences), and/or that examining descriptions in a pre-established language such as English does not give rise to transparency differences. Therefore, it is possible that novel communication systems developed in big groups are easier to learn after all. Importantly, studies on visual signals (drawings) suggests that this is indeed the case in the non-

linguistic domain. When groups of eight people and groups of two people played multiple rounds of Pictionary with each other, the final drawings of the big groups were superior to those of the small groups in terms of their learnability and processing by new individuals, despite being comparable in terms of their visual complexity (Fay, Garrod, & Roberts, 2008; Fay & Ellison, 2013). Naïve learners were more accurate in guessing the meanings of drawings that evolved in larger groups, and were able to learn them faster, recognize them faster, recall them faster, and reproduce them with better fidelity compared to drawings that evolved in small groups. This advantage was attributed to large groups' drawings being more iconic, i.e., having more transparent form-to-meaning mappings. Fay et al. (2008) conclude that the better "fitness" or quality of signs developed by big communities was derived from the increased diversity of potential signs: larger groups have a greater pool of variants to draw on, allowing for the selection of simpler and optimized signs. If such reasoning extends to language, then the greater input variability reported in the big groups in Raviv et al. (2019b) may actually benefit learners in the long run.

In sum, it is of high interest to test the causal relationship between language complexity and learnability, as well the role of community size in shaping such patterns. Confirming that more systematic languages are easier to learn is crucial for theories on language evolution and linguistic diversity, which assume this link as an essential underlying mechanism. Moreover, discovering an overall learning advantage for languages created by larger communities would suggest that social structure shapes cross-linguistic patterns. Together, such findings would have important implications for language learning and language typology by suggesting that some languages are acquired more slowly or more quickly compared to others because of their grammatical structure and/or the size of the community in which they are spoken.

The Current Study

The goal of the current study was to experimentally test the causal effect of group size and linguistic structure on language learnability. To this end, we used an artificial language learning paradigm in which individuals needed to learn a new miniature language with labels for describing different types of novel scenes (see Procedure). After training, participants were tested on their knowledge of the input language in two ways: (a) a

memory test, testing participants' reproduction accuracy on the scene-label pairings; and (b) a generalization test, testing participants' ability to label new, unseen scenes.

Importantly, participants were trained on different input languages, all of which had been created in a previous experiment by real groups of either four or eight interacting participants playing a communication game (Raviv et al., 2019b). We contrasted learning across several conditions by selecting 10 different input languages, which varied in their degree of systematic structure and in their origin group size, while being relatively similar in their average word length and internal confusability (see Materials). For example, one participant could learn a high-structured language created by a big group, and another participant could learn a medium-structured language created by a small group.

In order to promote open-science and increase the transparency and credibility of our research, the entire study (e.g., design, procedure, predictions, analyses plans, etc.) was pre-registered on OSF and is available online: https://osf.io/9vw86/. Additionally, all the data collected in this experiment and the scripts for generating all analyses can be openly found at https://osf.io/d5ty7/.

For our confirmatory analyses, the main prediction was that linguistic structure would significantly affect language learnability, so that more compositional languages that display systematic form-to-meaning mappings would be easier to learn (i.e., more accurately learned). Therefore, we expected that participants who learned more structured languages would show higher reproduction accuracy. We also hypothesized that group size would have an additional effect on language learnability, beyond the effect of linguistic structure: languages created by bigger groups are postulated to be easier to learn compared to languages created by small groups, above and beyond their degree of systematic structure. Therefore, we expected that across all structure levels, participants who learned languages that were created by big groups would show higher reproduction accuracy. We also planned to carry out exploratory analyses to examine the speed of learning across conditions, and to test the effect of linguistic structure and community size on participants' ability to generalize the language to a new set of meanings.

Methods

Participants

We analyzed data from 100 adults (79 women) between the ages of 18 and 35 (mean age=22.9y). This sample size was determined in advance using a power analysis based on pilot data and power simulations for a range of possible effect sizes (see Appendix A). We tested two additional participants who did not complete the experiment, and so their data was not included in the analyses. Each participant was paid 10€. All participants were native Dutch speakers, aged between 18 and 35 years old, and had no reported visual or reading difficulties. Ethical approval was granted by the Faculty of Social Sciences of Radboud University Nijmegen.

Materials

We selected ten languages from a bigger database of artificial languages, which were created in a previous experiment (Raviv et al., 2019b). The database contained 144 languages that were created by individual participants in either small or larger groups after completing a group communication game. Each language consisted of 23 scene-label pairings. i.e., 23 written labels that corresponded to 23 dynamic visual scenes. The scenes varied along three semantic dimensions: shape, angle of motion, and fill pattern. Each scene consisted of one out of four possible shapes, moving repeatedly in a straight line from the center of the screen in a given direction. Additionally, each scene had a unique blue-hued fill pattern. There were three versions of the stimuli, which differed in the distribution of shapes and their associated angles.

Each language in the database had a *structure score*, which reflected the degree of systematic mapping between labels and meanings in the language (Kirby, Cornish, & Smith, 2008; Kirby, Tamariz, Cornish, & Smith, 2015; Raviv, Meyer, & Lev-Ari, 2019a). The structure score for each language was calculated as the correlation between the pair-wise semantic distances between scenes' features and the pair-wise string distances between their labels. First, we calculated the semantic differences between different scenes, resulting in a similarity matrix for all pairs of scenes in the language. This was done using Hamming distances, in the following way: First, two scenes had a semantic difference of 1 if they differed in shape, and a semantic difference of 0 if they included the same shape. Second, the difference between two scenes'

angles was calculated and divided by the maximal distance between angles (180 degrees) to yield a continuous normalized score between 0 and 1. Then, the difference scores for shape and angle were added, yielding a possible semantic distance between 0.18 and 2 for each pair of scenes in the language. Next, we calculated the string differences between all pairs of labels in the languages using normalized Levenshtein distances, which is the minimum number of character changes (insertions, deletions or substitutions) needed in order to transform one label into the other, divided by the length of the longest label. This resulted in a similarity matrix for all pairs of labels in the language. Finally, the two sets of pair-wise distances (i.e., string distances and meaning distances) were correlated using the Pearson product-moment correlation, yielding a measure of systematic structure.

This continuous measure was divided into five equally sized bins of possible structure scores[11]: low structure (0.0-0.2), low-medium structure (0.2-0.4), medium structure (0.4-0.6), medium-high structure (0.6-0.8), and high structure (0.8-1.0). Figure 1 gives a general description of the structural properties of languages in each structural bin, along with an illustration. Low structure scores reflect the absence of systematic mapping between labels in the language and their corresponding scenes, resulting in a holistic lexicon where labels seem to be randomly assigned to the scenes regardless of their sematic features (see Figure 1 for an illustration). In low structured languages, each scene has an opaque label that cannot be decomposed into small components based on scenes' shape or direction of motion. In contrast, high structure scores reflect the existence of systematic mappings between meanings and labels, resulting in compositional languages in which similar semantic features are expressed using similar part-words (see Figure 1 for an illustration). Specifically, a highly systematic language would include a consistent part-word for describing each of the four shapes (e.g., *"tup"* for Shape 1 and *"fest"* for Shape 2), and a consistent part-word for describing the direction of motion (e.g., *"o"* for up, *"i"* for right, and *"oi"* for up-right). In addition to the structure score, we characterized each language using two other measures: *average word length*, i.e., the average number of

[11] Although correlations can potentially range from -1 to 1, there were no languages with a correlation below 0 (i.e., a languages with "anti-systematic" or "counter-systematic" mapping between labels and scenes). The structure scores of the languages in the data set ranged from 0.07 (i.e., an unstructured, holistic language) to 0.9 (i.e., a fully systematic, compositional language).

characters in the language's labels; and *confusability*, i.e., the average normalized Levenshtein distance between all possible pairs of labels in a language, capturing the phonological similarity across all labels in a given language.

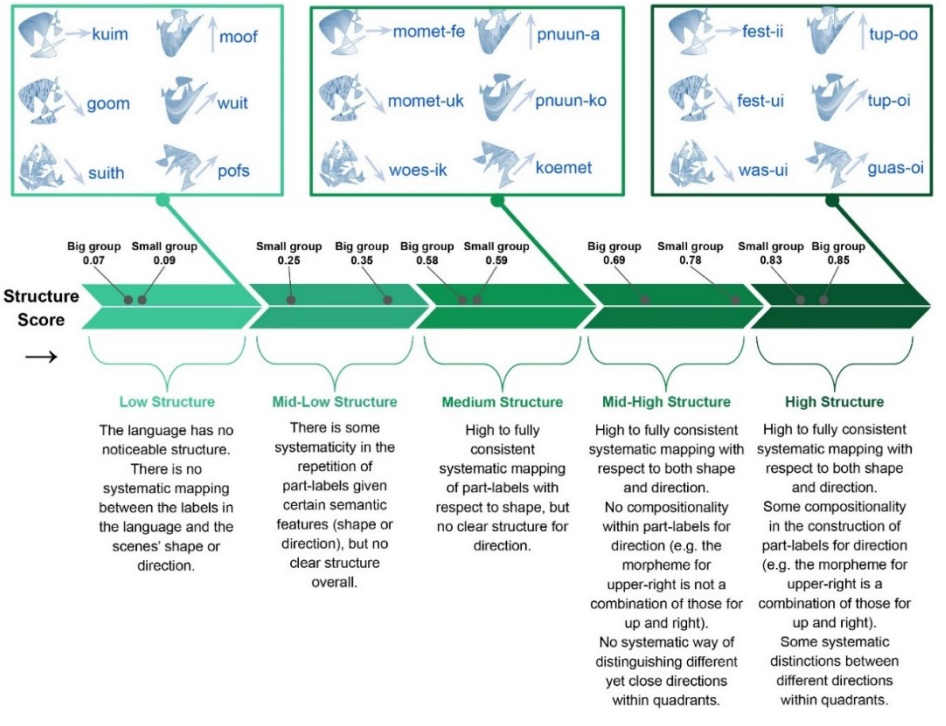


Figure 1. An illustration of the structure levels of input languages learned by participants in the experiment. The axis represents languages' structure scores, ranging from 0 to 1. For descriptive purposes, this continuous measure can be divided into five equally sized bins: low structure (0.0-0.2), low-medium structure (0.2-0.4), medium structure (0.4-0.6), medium-high structure (0.6-0.8), and high structure (0.8-1.0). Each bin can be characterized by a different degree of systematicity, which is described verbally below it. We included illustrations of three miniature lexicons: a language with low structure in light green, a language with medium structure in green, and a language with high structure in dark green. For example, in the low structured language, there is no similarity between the labels for scenes with similar shapes (e.g., *moof, wuit*) or for scenes with similar directions (e.g., *wuit, pofs*). In the high structured language, part-labels are consistently associated with a given shape (e.g., *fest, tup*) or with a given direction (e.g., *ui, oi*). The direction morphemes are also compositional, and are comprised of two meaningful parts: for example, the morpheme for the direction up-right (oi) is a combination of the morpheme for up (o) and the morpheme for right (i). The grey dots on the axis point to the structure scores of ten specific languages originally created in a group communication game (Raviv et al., 2019b), which were selected as the input languages for this experiment. From each of structure bin, we selected one language that was created by a small group and one language that was created by a big group.

We then selected ten languages from the database to be used as input languages for the current study (Figure 1; see Appendix B for the full list of stimuli). Specifically, we picked two languages from each of the five structure score bins described above: one language that was created by a small group, and one language that was created by a big group. This resulted in a 2 X 5 factorial design, with the two factors being group size (with two levels: big vs. small) and structure score (with five levels of structure degrees). Note that although we used these descriptive bins to select our input languages, structure score was treated as a continuous variable in our analyses (ranging from 0.07 in the low structure bin, to 0.84 in the high structure bin). Ten participants were assigned to learn each of the ten input languages using a pre-made randomization list.

Since we wanted to ensure that differences in language learnability can indeed be attributed to their structural properties and/or group size origin, we picked languages that were comparable in several ways. First, all languages fell within a reasonably similar range of average word length and confusability scores. Under the assumption that longer and more confusable words are harder to learn (Laufer, 2009; Willis & Ohashi, 2012), we chose languages from the lower half of the distributions of these two measures, i.e., languages with relatively short words (i.e., between 4 and 7 characters) and relatively low confusability (i.e., between 0.14 and 0.37). Second, languages in the same structure bin were comparable in terms of their descriptive grammatical properties and had similar types of consisted mappings (as judged by the authors; see Fig. 1 and Appendix B). Third, languages within the same structure bin had similar numbers of irregulars, as counted by the authors. Fourth, across the different structure bins, differences in structure scores were balanced with group size, so that it was not the case that one group size condition was consistently higher/lower in structure compared to the other. The structure scores of the selected languages can be seen in Figure 1. For the full set of input languages and their detailed descriptions, see Appendix B.

Finally, we created 13 new scenes for each stimuli version. These additional scenes were not included in the learning phase and the memory test, and were presented to participants for the first time during the generalization test. The new scenes varied along the same semantic dimensions as the 23 original scenes in the input languages, and were comprised of one of the four possible shapes moving in one of the possible directions. Each new scene included a new combination of shape and angle of motion (with each of the four shapes appearing in at least two new scenes), and a completely different blue-hued fill pattern. That is, the

new scenes matched the general meaning space of the language, but included new combinations of shape, direction, and fill pattern which were unfamiliar to the participants and not present beforehand.

Procedure

Participants were told they were about to learn a new fantasy language ("Fantasietaal" in Dutch) to describe different scenes of moving shapes, and that their goal was to learn the language as best as they could in order to succeed in a subsequent test. The experiment consisted of two phases: (1) a learning phase, which was comprised of three leaning blocks with a similar procedure; (2) a test phase, which was comprised of two parts, i.e., memory test and generalization test. For example screenshots of each phase of the experiment, along with detailed descriptions of the accompanying instructions and procedure, see https://osf.io/mkv5r/.

The learning phase consisted of three blocks. The first learning block included half of the language (12 scene-label pairings), the second block included the other half of the language (11 scene-label pairings), and the third block included the entire language (23 scene-label pairings). Each input language was randomly divided into two halves in advance, so that the set of target scenes in each block was identical for all participants in a given condition. The order of appearance of target scenes within a given block and during the test phase was randomized separately per participant at the beginning of the experiment.

Each learning block comprised of three tasks: passive exposure, guessing, and production. During passive exposure, participants were exposed to scene-label pairings one by one in a random order, with each target label appearing on the screen together with its corresponding scene for the duration of 10 seconds. In this task, participants only had to look at their screen and try to remember the scene-label pairings. In the guessing task, participants were presented with target labels one by one in a random order, and needed to select the scene to which that label referred to from a set of possible scenes. In the first two blocks, this set included four scenes (i.e., the target scene and three distractors), while in the third block this set included eight scenes (i.e., the target scene and seven distractors). The distractors were randomly selected for each participant and for each trial from the set of possible scenes in that block. Participants received feedback after each guess indicating whether they were right or wrong, along with the target label, the correct scene, and the scene they

selected in case it was different. In the production task, participants were presented with target scenes without labels one by one in a random order, and needed to type the correct label for it using their keyboard. Participants' letter inventory was restricted, and matched the letter inventory of the original input languages from Raviv et al. (2019b): it included a hyphen, five vowel characters (a,e,i,o,u), and ten consonants characters (w,t,p,s,f,g,h,k,n,m), which participants could combine freely. Participants received feedback after each production, along with the target scene, the correct label, and the label they typed in case it was different.

In the first two learning blocks, which included only half the language, each of the three tasks (i.e., passive exposure, guessing, production) was repeated twice with all the available target scenes-label pairings for that block, so that each scene-label pairing appeared twice in each task and six times in total. In the third learning block, which included the entire language, each task was repeated once, so that all scene-label pairings appeared once in each task and three times in total. This resulted in a total of nine exposures per scene-label pairing during the learning phase: three times during the passive exposure task, three times during the guessing task, and three times during the production task.

Following learning, participants completed a test phase. The first part of the test phase was a memory test, in which participants demonstrated how well they had learned the input language. During the memory test, participants were presented with each of the 23 target scenes without labels one by one in a random order, and typed in a label for them. The second part of the test was a generalization test, in which participants were asked to use the language they had just learned to label new scenes that they had not seen before. Participants were presented with 13 unfamiliar scenes (see Materials) without labels one by one in a random order, and typed in a label for each of them based on their acquired knowledge of the Fantasy language. Participants were asked to label the new scenes as if they were communicating to another person, who had learned the same language as they did but knew no other language (i.e., no use of Dutch, English, or any other language was allowed). No feedback was provided during the memory and generalization tests, and participants' letter inventory was restricted in the same manner as in the production phase.

After the test phase, participants filled out a questionnaire about their performance in the experiment, including questions such as "How hard was it to learn the fantasy language?", and "Did you notice anything about

the structure of the fantasy language during the experiment?". Finally, all participants were debriefed by the experimenter.

Measures

Binary Accuracy

This measure reflects whether participants were correct or incorrect on a given trial during the learning phase or the memory test, and is calculated as binary response accuracy. If participants produced/guessed the target label correctly, accuracy equaled 1; otherwise, it equals 0[12].

Production Similarity

This continuous measure reflects how closely participants reproduced their input language by measuring the similarity between a target label (i.e., an original label as it appeared in the input language) and the corresponding label produced by a participant in production trials (during the learning phase and during the memory test). For each production trial, we calculated the normalized Levenshtein distance between the label produced by the participant and the original input label. The normalized Levenshtein distance is the minimum number of character changes (insertions, deletions or substitutions) needed in order to transform one label into the other, divided by the length of the longest label. This distance was subtracted from 1 to represent string similarity, i.e., how much the labels participants produced resembled the labels they had learned. High production similarity indicates good reproduction fidelity, with participants producing labels that are similar to those they learned (i.e., a score of 1 indicates that the produced label matched the target label exactly). Low production similarity indicates poor reproduction fidelity, with participants producing labels that are very different from those they learned.

[12] In cases where the target label described more than one scene (i.e., homonym), participants' accuracy in guessing trials (during the learning phase) would equal 1 if they had guessed any one of the possible scenes associated with that target label.

Guessing Similarity

This continuous measure reflects how well participants learned the label-scene mapping in the input language by measuring the similarity between the target scene (i.e., the correct scene given a specific label) and the scene selected by the participant during guessing trials. We used Hamming distances to quantify the semantic differences between the selected scene and the target scene based on the differences in scenes' shapes and directions of motion. This measure was calculated in a similar way to the semantic distances used for calculating the structure score (see Materials). First, two scenes had a semantic difference of 1 if they differed in shape, and 0 otherwise. Second, the difference between the two scenes' angles was calculated and divided by the maximal distance between angles (180 degrees) to yield a continuous normalized score between 0 and 1. Then, the difference scores for shape and angle were added, yielding a range of semantic distances between 0 and 2. This distance was then subtracted from 2 to represent guessing similarity, i.e., how much the scene participants guessed resembled the correct scene. High guessing similarity indicate that, given a target label, participants guessed a scene which was similar to the target scene in terms of its features (i.e., a similarity score of 2 indicates that the selected scene matched the target scene exactly). Low guessing similarity indicates that, given a target label, the participant's guess was very different from the target scene (i.e., a similarity score of 0 indicates that the participant selected a maximally different scene with a different shape going to the opposite direction).

Generalization Score

This continuous measure reflects the degree of similarity between the labels participants produced for each new scene during the generalization test, and the labels they produced for familiar scenes during the memory test. A high generalization score reflects the fact that, given an unfamiliar scene, participants produced a label which was as similar as possible to the labels they produced for familiar scenes with similar features (e.g., the same shape and/or the same direction). That is, their labels during the memory and the generalization test followed the same principles. A low generalization score reflects the fact that, given an unfamiliar scene, participants produced a label which was different from the labels they produced for familiar scenes with similar features. That is, the labels they produced for unfamiliar scenes did not resemble those they produced in the memory test. For each participant, the generalization score is the

normalized correlation between (a) the pair-wise semantic distances between each new scene and all familiar scenes, and (b) the pair-wise string distances between each new label produced in the generalization test and all labels produced for familiar scenes during the memory test. This correlation was normalized to account for the fact that high-structure languages offer more possibilities to generalize to begin with. The generalization score is calculated in the following way: For each new scene in the generalization test, we first calculated the semantic differences between that new scene and all familiar scenes using Hamming distances, in the same way as described above for structure score and for guessing similarity. Second, we calculated the string differences between the new label produced for this scene and the labels produced for familiar scenes during the memory test using normalized Levenshtein distances, in the same way as described above for structure score and for production similarity. We repeated this calculation for all new scenes and their corresponding labels. Then, these two sets of pair-wise distances (i.e., string distances and meaning distances between new and familiar scenes/labels) were correlated using the Pearson product-moment correlation. Finally, this correlation was scaled using a procedure inspired by the min-max normalization procedure (also called unity-based normalization and feature-scaling), yielding the final generalization score per participant. This normalization procedure was implemented in order to ensure that all conditions show similar ranges of generalization scores, and that we do not bias against low structured languages, which by default would show lower generalization scores given that participants' productions for familiar items are likely to be less structured in such languages. Specifically, we linearly transformed the correlation scores to fit in the range [0,1], and scaled across different conditions so that the final generalization score was proportionate to the range of achieved values in that condition: low generalization scores relative to the range of possible scores are mapped to values closer to 0, and high generalization scores relative to the range of achieved scores are mapped to values closer to 1. This was done using the formula $x' = (x-\min(x))/(\max(x)-\min(x))$, where $\min(x)$ in the lowest value for x achieved by a participant across all conditions (-0.069), and $\max(x)$ is the highest value for x achieved by a participant in a specific condition (i.e., $\max(x)$ varied for different input languages, with each input language having a different maximal value). For example, the highest value achieved by a participant in a low-structure language was 0.5, while the highest value achieved by a participant in a high-structure language was 0.88.

Generalization Convergence

This continuous measure reflects the degree of similarity between the labels produced during the generalization test by different participants who learned the same input language. For each of the new scenes in the ten input languages, we calculated the normalized Levenshtein distances between all pairs of labels produced by different participants for the same new scenes. The average distance between all pairs of labels was subtracted from 1 to represent string similarity, i.e., how much the labels of different participants resembled each other. A high convergence score indicates that participants who learned the same language also produced similar labels for the unfamiliar scenes during the generalization test. A low convergence score indicates that participants who learned the same language produced different labels for unfamiliar scenes during the generalization test.

Analyses and Results

We analyzed the data using mixed effects regression models generated by the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2016; R Core Team, 2016). All reported p-values were generated using the pbkrtest package (Halekoh & Højsgaard, 2014), which uses the Kenward-Roger Approximation to calculate conservative p-values for models based a relatively small number of observations. All analyses are reported in Appendix C using numbered models, which we refer to here. The data and the R code to generate all analyses can be openly accessed at https://osf.io/d5ty7/.

Confirmatory analysis: Final Binary Accuracy (Figure 2a)

As declared in the preregistration (under "Analysis Plan"), our main model had final binary accuracy (i.e., whether participants were right or wrong in the memory test) as the dependent variable, and included fixed effects for GROUP SIZE ORIGIN (dummy-coded, with small groups as reference level) and STRUCTURE SCORE (continuous, centered), as well as random intercepts for participants and scenes. Since we suspected that the effect of structure score would be non-linear (Beckner, Pierrehumbert, & Hay, 2017; Raviv et al., 2019b), we used Likelihood ratio tests to compare models with 1- and 2-degree polynomials (generated using the poly() function in R to avoid collinearity). These model comparisons revealed

that the best fitting model (Model 1) included both a linear and a quadratic term for the effect of STRUCTURE SCORE (see Appendix C).

Results from this model showed that STRUCTURE SCORE was a positive significant predictor of participants' binary accuracy during the memory test (Model 1: β=31.47, SE=6.93, z=4.54, p=0.00001), and that this effect was non-linear (Model 1: β=31, SE=6.87, z=4.51, p=0.00001). Specifically, the effect of STRUCTURE SCORE on accuracy followed a U-shape: participants' binary accuracy was poorer when trained on medium structure languages than when trained on low structured languages, but the highest when trained on high structured languages (Fig, 2A).

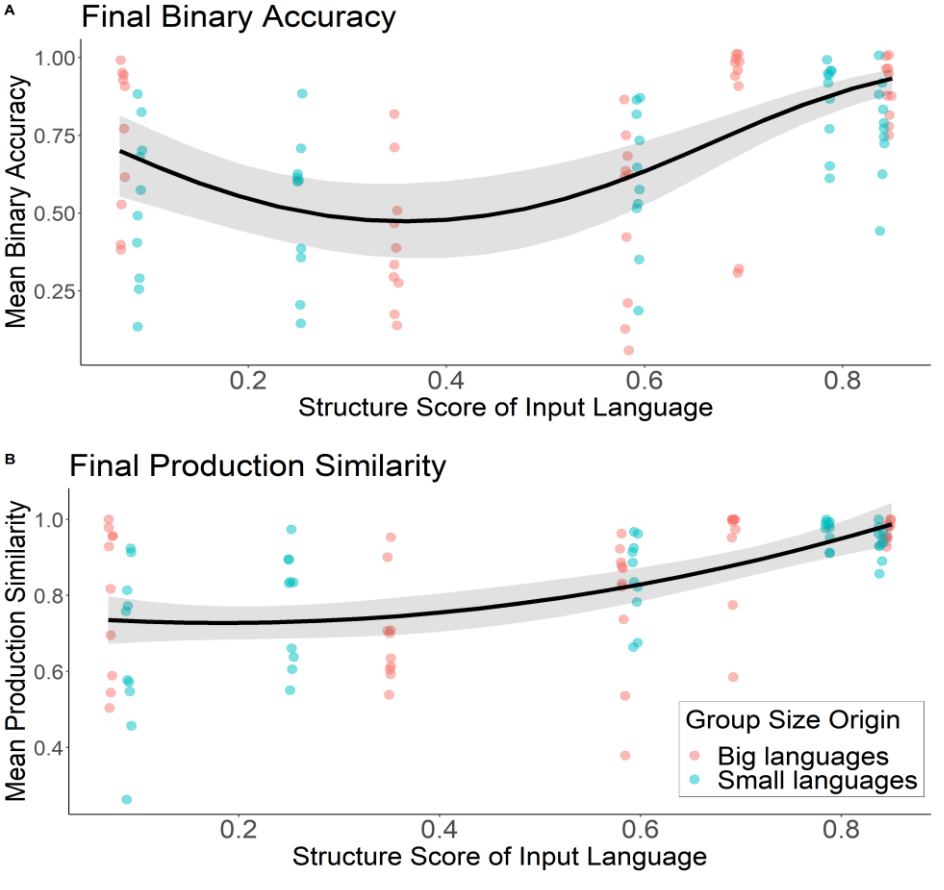


Figure 2. (A) Binary Accuracy and (B) Production Similarity at the final memory test, as a function of learned languages' structure score and group size origin**.** Each point represents the average accuracy of a single participant. The thick line represent the model's estimate, and its shadings represent the model's standard error.

The U-shape pattern is evident in the global minimum of the polynomial function predicted by the model, which can be directly calculated when running the same model without the orthogonal polynomials and comparing its derivative to 0. After re-centering, we found that the minimum value for binary accuracy was obtained when structure equals 0.36, which is within the medium structure bin. In other words, participants' performance was worst when learning semi-structured languages, and the increase in structure only benefited accuracy as languages became highly systematic. The effect of GROUP SIZE ORIGIN was not significant, with languages originating from big and small groups eliciting similar levels of accuracy (Model 1: β=0.48, SE=0.29, z=1.67, p=0.096).

Exploratory analysis: Final Production Similarity (Figure 2b)

Originally, we believed that binary accuracy was a good measure to examine learning, considering an "all or nothing" approach. However, during data collection we observed that this measure was too crude, and did not reliably reflect how well participants learned the languages. Specifically, many participants were able to reproduce the language with relatively high fidelity but not perfectly, which the binary accuracy measure did not capture. For example, if a participant correctly typed five letters out of a six-letter label, the binary accuracy measure would treat this one-letter error as if the entire label was incorrect. This led to an overestimation of errors, with some participants receiving low scores despite making very minor mistakes (e.g., one letter difference between the label they learned and the label they reproduced). As such, we decided to use a more subtle proxy of participants' learning accuracy, namely, production similarity (see Measures). This continuous measure reflects the degree of reproduction accuracy more reliably by quantifying the similarity between participants' input and output languages, and is broadly used in iterated language learning paradigms (Kirby et al., 2008, 2015).

We therefore ran an identical model to that described in the confirmatory analysis section, but used production similarity during test as the dependent variable instead of binary accuracy during test. Importantly, the predictions for this measure were identical to those of binary accuracy: more structured languages should be reproduced more accurately, i.e., show more production similarity. Accordingly, the model for production similarity had the same effect structure as the binary

accuracy model reported above, and included fixed effects for GROUP SIZE ORIGIN (dummy-coded, with small groups as reference level) and STRUCTURE SCORE (continuous, centered), and random intercepts for participants and scenes. As in the confirmatory analysis, Likelihood ratio tests favored the 2-degree polynomial model (Model 2) with both a linear and a quadratic term for the effect of STRUCTURE SCORE (see Appendix C).

Results from this model showed that STRUCTURE SCORE was a positive significant predictor of production similarity during the memory test (Model 2: β=4.41, SE=0.68, t=6.49, p<0.0001). This effect was also non-linear (Model 2: β=1.6, SE=0.68, t=2.34, p=0.02), yet in an exponential way: participants produced labels that were increasingly more similar to those they learned as structure increased, so that the advantage of structure was stronger in highly structured languages (Fig. 2B). That is, the increase in structure benefited accuracy most as languages became more systematic. As for binary accuracy, we calculated the global minimum of the polynomial function predicted by the model for production similarity, and found that the minimum value for reproduction fidelity was obtained when structure equaled 0.18, which is within the low structure bin. That is, participants' performance was worst when learning unstructured languages. The effect of GROUP SIZE ORIGIN was not significant, with languages originating from big and small groups eliciting similar levels of production accuracy (Model 2: β=0.007, SE=0.03, t=0.26, p=0.8).

Exploratory analyses: Learning Trajectory (Figure 3)

As declared in the preregistration (under "Analysis Plan"), we also planned to perform an exploratory analysis to examine participants' learning trajectory during the three blocks of the learning phase. Specifically, we were interested in seeing whether improvement in performance during the first three blocks was modulated by structure score and/or group size (e.g., are highly structured languages learned faster?). To this end, we generated three models in which the dependent variable was either binary accuracy, production similarity or guessing similarity (see Measures). All three models had the same effects structure, and included fixed effects for BLOCK NUMBER (continuous, centered), GROUP SIZE ORIGIN (dummy-coded, with small groups as reference level), STRUCTURE SCORE (continuous, centered), and the interaction terms BLOCK NUMBER X GROUP SIZE ORIGIN and BLOCK NUMBER X STRUCTURE SCORE. All models included by-participant and by-scene random

intercepts, as well as random by-participant slopes with respect to the effect of BLOCK NUMBER. We used Likelihood ratio tests to compare 1- and 2-degree polynomial models with respect to the effect of STRUCTURE SCORE (see Appendix C), and found that models with a quadratic term were favored in the case of binary accuracy (Model 3) and guessing similarity (Model 5), but not for production similarity (Model 4).

All three models yielded similar results (Fig. 3), and showed that performance significantly improved over learning blocks, with participants showing higher binary accuracy (Model 3: β=0.29, SE=0.05, z=5.99, p<0.0001), higher production similarity (Model 4: β=0.04, SE=0.007, t=5.57, p<0.0001) and higher guessing similarity (Model 5: β=0.04, SE=0.01, t=3.74, p=0.0003) over time. There was also a significant effect of STRUCTURE SCORE for all measures, indicating that, across blocks, performance was overall better on more structured languages (Model 3: β=66.49, SE=10.67, z=6.23, p<0.0001; Model 4: β=0.34, SE=0.05, t=7.35, p<0.0001; Model 5: β=14.65, SE=2.13, t=6.88, p<0.0001). This effect was non-linear for binary accuracy and guessing similarity, suggesting that the advantage of structure for these two measures was increasingly higher as structure increased (Model 3: β=51.83, SE=10.74, z=4.83, p<0.0001; Model 5: β=8.07, SE=2.14, t=3.76, p=0.0003). Additionally, there was a significant interaction between STRUCTURE SCORE and BLOCK NUMBER for binary accuracy (Model 3: β=25.88, SE=4.32, z=6, p<0.0001; β=14.46, SE=4.76, z=3.04, p=0.0024) and production similarity (Model 4: β=0.05, SE=0.02, t=2.83, p=0.00564), indicating that the improvement in participants' performance over time in these two measures was even faster in more structured language, i.e., the learning slope was steeper for highly structured languages. This interaction was not significant for guessing similarity (Model 5: β=1.75, SE=1.19, t=1.48, p=0.14), suggesting that the slope of improvement in participants' guessing performance over time was similar across all structural levels.

Finally, GROUP SIZE ORIGIN did not significantly affect performance on any of our three measures (Model 3: β=0.23, SE=0.19, z=1.19, p=0.23; Model 4: β=0.001, SE=0.03, t=0.04, p=0.97; Model 5: β=0.01, SE=0.03, t=0.41, p=0.69) or participants' learning trajectories (Model 3: β=0.04, SE=0.07, z=0.54, p=0.59; Model 4: β=-0.01, SE=0.01, t=-1.37, p=0.17; Model 5: β=-0.01, SE=0.02, t=-0.86, p=0.39).

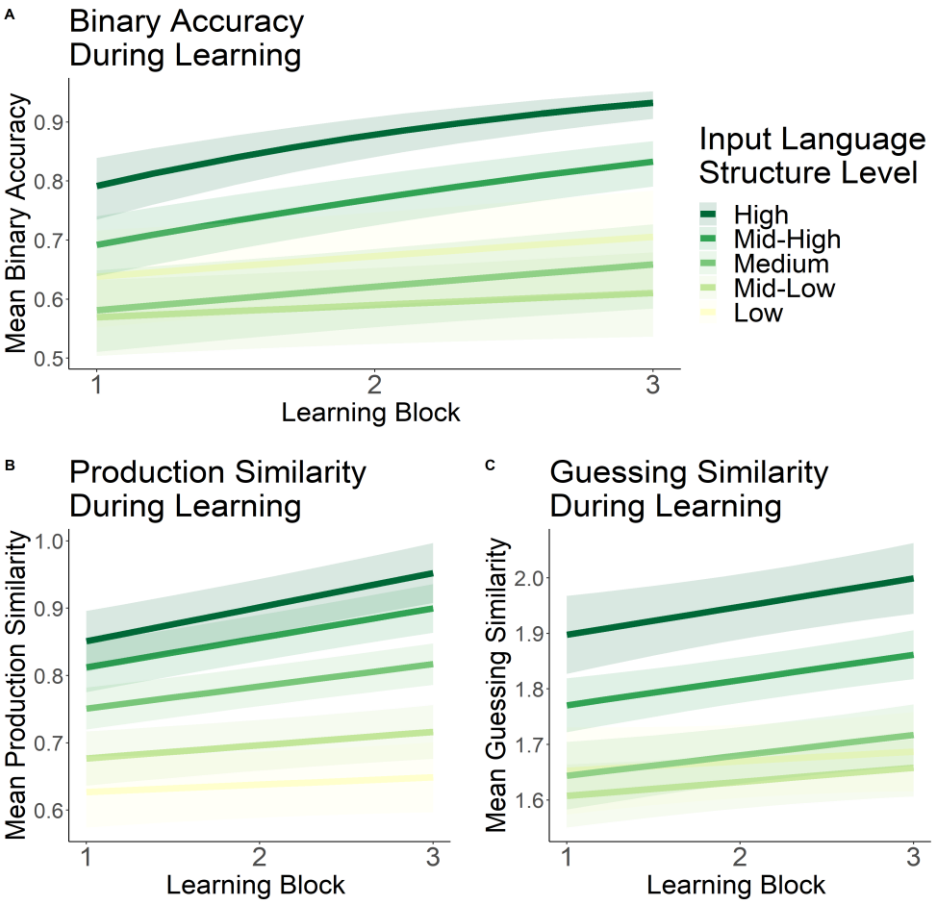


Figure 3. Changes in Mean (A) Binary Accuracy, (B) Production Similarity, and (C) Guessing Similarity over time as a function of learned languages' structure score. The colored lines and their shadings represent the models' estimates and standard errors, averaged over the five descriptive structure levels (i.e., collapsed over big and small groups' languages).

Exploratory analyses: Generalization Behavior (Figure 4)

As declared in the preregistration (under "Analysis Plan"), we also planned to examine participants' behavior during the generalization phase. In particular, we wanted to see whether participants would generalize the linguistic patterns of their input language to new, unseen scenes. If participants learned a systematic language and learned its underlying structure, generalizations could potentially take place in the form of reusing the learned structural patterns (i.e., part-words) when

producing new labels (e.g., combining existing morphemes for shape and motion to describe a new scene with a new combination of shape and motion). If participants learned an unstructured language, generalizations could potentially take place in the form of reusing existing full words to describe scenes with similar elements (i.e., creating homonyms), or combining exiting words. In both cases, if the participants' generalized their input language and maintained its patterns, then their productions for each new scene during the generalization test should be similar to their productions of the input language during the memory test, resulting in a high generalization score (see Measures). If participants did not generalize and instead produced random, unrelated labels, then their generalization score should be lower. This score was also adjusted to take into account the fact that low-structured languages allow for less generalizations to begin with.

To test participants' generalization behavior, we used a general linear regression model with normalized generalization score as the dependent variable, and fixed effects for GROUP SIZE ORIGIN (dummy-coded, with small groups as reference level) and STRUCTURE SCORE (continuous, centered). We used Likelihood ratio tests to compare 1- and 2-degree polynomial models with respect to the effect of STRUCTURE SCORE, and found that the model with only a linear term (Model 6) was favored (Appendix C).

Results from this model showed that STRUCTURE SCORE was a significant predictor of generalization score: participants who had acquired more structured languages also generalized more (Model 6: β=0.51, SE=0.07, t=7.22, $p<0.00001$; Fig. 4). There was no significant effect of GROUP SIZE ORIGIN (Model 6: β=0.01, SE=0.04, t=0.31, p=0.76), suggesting that generalization behavior was similar for languages originating from big and small groups.

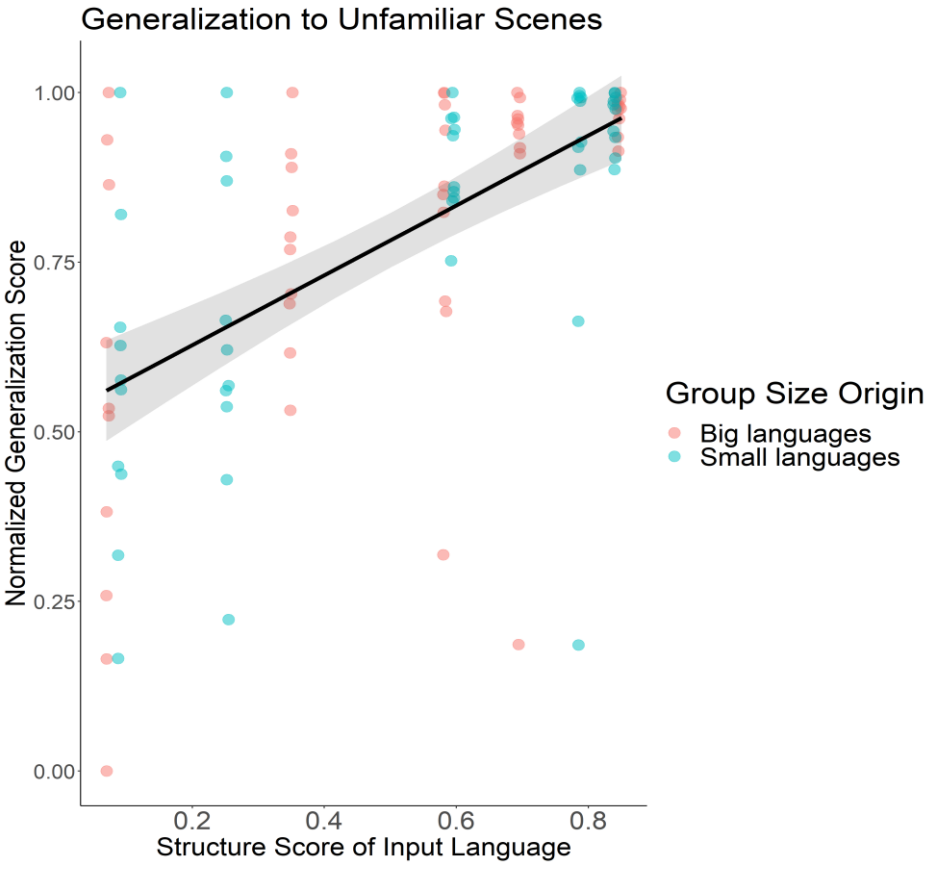


Figure 4. Generalization as function of learned languages' structure score and group size origin. Each point represents the normalized generalization score of a single participant. The thick line represent the model's estimate, and its shadings represent the model's standard error.

Exploratory analyses: Generalization Convergence (Figure 5)

Finally, we looked for similarities in participants' generalizations: do participants in the same condition make similar generalizations, i.e., produce similar labels for unseen scenes? We assumed that when languages are highly systematic and rule-governed, they allow for transparent and productive labeling – resulting in different participants producing the same labels, i.e., generalizing in the same way. By contrast, when languages are unstructured or inconsistent in their mapping of labels to meanings, it may be less clear what or how to generalize (e.g., which features of the scenes are relevant?) and therefore less clear how to label new scenes. This may result in participants producing new labels more randomly, or attempting to make generalizations based on the

idiosyncratic features of scenes (i.e., fill-pattern). In other words, we assumed that highly structured languages would facilitate convergence amongst participants, potentially enabling them to understand each other even without previously interacting.

Results from this model showed that STRUCTURE SCORE was a significant predictor of generalization score, so that participants who learned more structured languages also produced labels that were more similar to one another (Model 7: β=0.74, SE=0.03, t=21.63, p<0.00001; Fig. 5). There was no significant effect of GROUP SIZE ORIGIN (Model 7: β=-0.03, SE=0.02, t=-1.71, p=0.09), suggesting that languages originating from big and small groups did not differ in their convergence.

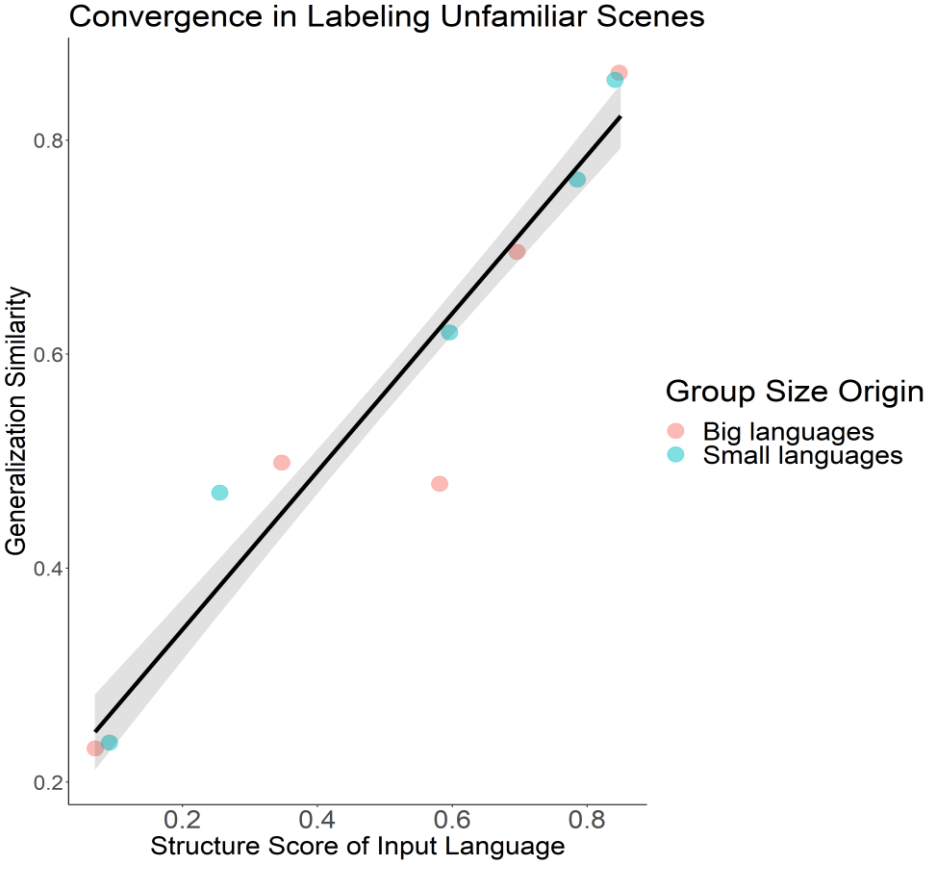


Figure 5. Generalization convergence across participants as function of learned languages' structure score and group size origin. Each point represents the average convergence of ten participants on each of the ten input languages. The thick line represent the model's estimate, and its shadings represent the model's standard error.

Discussion

In this pre-registered study, we tested the effects of systematic structure and community size on language learnability using an artificial language learning paradigm. We compared participants' acquisition of a broad yet controlled set of input languages for describing novel dynamic events (see Fig. 1). These input languages varied in their degree of linguistic structure (ranging from low to high systematicity) and in their group size origin (created by either big or small groups in a previous communication experiment, Raviv et al., 2019b). Language learnability was assessed by examining participants' final reproduction accuracy, their learning trajectories over time, and their ability to generalize the language they learned to a new set of unseen events.

Our main prediction was that participants would show better learning of languages with more systematic structures. This prediction was motivated by previous literature reviewed in the Introduction (e.g., second language learning, iterated learning), which argued for a causal link between the grammatical structure of languages and their relative ease of learning. Specifically, more regular and transparent languages with more systematic form-to-meaning mappings are considered to be easier to learn. As such, we hypothesized that linguistic structure would positively affect learnability, so that languages with more systematic grammars would be better learned. We expected that this learning advantage would be reflected first and foremost in higher reproduction accuracy during the memory test, and potentially also in a faster improvement in performance over time during the learning phase. Additionally, we reasoned that systematic and rule-governed languages would facilitate clear and productive labeling (Ackerman & Malouf, 2013; Kirby, 2002). As such, we predicted that more structured languages would be more easily generalizable to new meanings. To this end, we tested participants' ability to generalize the language they learned in order to produce new labels for unseen events. Finally, we hypothesized that community size may have an additional effect of learning. Specifically, we considered the possibility that, even when equating for the degree of structure, languages that evolved in bigger groups may be better fitted to learners' individual biases, and would potentially be easier to learn across all structural levels. This hypothesis was motivated by studies showing that bigger groups generated visual signs (i.e., drawings) that, despite being equally efficient, were processed and learned faster by new individuals, and were overall

better suited for production and comprehension (Fay et al., 2008; Fay & Ellison, 2013).

Results from the confirmatory analysis showed that the relationship between language learnability and linguistic structure followed a U-shape (Fig. 2A): although participants' mean accuracy was, as predicted, highest when learning highly structured languages, it was poorest when learning medium structured languages, and not when learning low structured languages (as one would expect if the relationship between structure and learning was simply linear). That is, learners struggled most with learning languages that were partly or semi-structured, i.e., languages that contained some patterns but also multiple irregulars and inconsistencies. This pattern, however, was not fully replicated in a similar exploratory model, where we examined participants' learning by using a more subtle measure of reproduction fidelity (i.e., production similarity) that reflected the degree of similarity between the labels participants learned and the labels they eventually reproduced. Results from this model also supported a non-linear relationship between structure and learnability, albeit an exponential relation and not U-shaped: participants produced more similar labels to those they learned as linguistic structure increased, and especially so for highly compositional languages (Fig. 2B). In other words, the benefit of linguistic structure for learning was proportionate to the level of structure in the language, and increased as structure increased. Similar findings were obtained from a set of exploratory analyses that investigated participants' learning trajectories over time: participants' performance was better on more structured languages across all learning blocks, and gradually improved over time. Moreover, the reproduction accuracy of participants who learned highly structured languages improved more quickly.

Together, our results confirm that a higher degree of linguistic structure is advantageous for language learning, and that languages with highly structured grammars are learned faster and more accurately. These findings are in line with our main prediction, and corroborate the postulated link between the degree of systematicity in the language and its relative learnability. This link is important for theories of language evolution and language diversity, which rely on it as an explanatory mechanism. Although the non-linear nature of the relationship between language structure and language learnability warrants further explanation, our results do support a causal relationship between them: highly regular and systematic morphologies indeed seem easier to learn. This conclusion has broader implications for theories on second language learning and

language acquisition, and strengthen the premise that not all grammatical systems are equally easy to acquire. As such, our study supports the claim that cross-linguistic differences in structural complexity and morphological opacity may potentially result in different learning trajectories and in different proficiency levels for adult L2 learners learning different languages.

Our results also show that systematic structure is advantageous for making generalizations: in an exploratory analysis, we found that participants generalized significantly more as linguistic structure in their input language increased. Specifically, participants who learned more systematic languages created new labels that matched the patterns of their input language more closely. This finding shows that, in addition to being beneficial for learning, an important advantage of linguistic structure is its productivity. That is, learners can exploit transparent, systematic and regular patterns found in their language to make informed guesses about unknown forms of words based on exposure to known forms, allowing them to effectively produce new labels for unfamiliar meanings. Indeed, the advantage of highly structured languages for generalization was also evident when looking at participants' self-reported behavior in the final questionnaire: all participants learning languages with systematic structure indicated that they "knew" how to label the new scenes in the generalization test, and some of them did not even notice that these scenes were not seen before. However, given that these results were based on a preliminary, exploratory measure, they should be taken with caution and require further experimental validation.

Notably, formally quantifying participants' generalization behavior was not a trivial task. In particular, it was not clear what counts as a generalization in low structured languages, which had no obvious structure. For example, if there is no systematic label for scenes with the same shape (e.g., different scenes with Shape 1 can be called *mipo*, *lex*, or *fuit)*, then a label for a new scene with Shape 1 can be potentially generalizable and referred to using three different homonyms. But are all these homonyms equally good generalizations, or is the best generalization achieved when the chosen homonym is also the label for the closest scene in terms of direction? Since there was no prior measure of generalizations we could rely on, it was not clear what the right way to measure it would be. In a first attempt to explore the complex realm of generalizations in artificial languages, we chose to use a metric that quantifies generalization behavior as the similarity between participants' new labels and their own productions during the final memory test, and

normalized based on the best and worst observed generalization behavior in a given input language condition. This normalization procedure was implement in order to account for the fact that different input languages allowed for different degrees of generalizations: a language with little to no structure naturally permits fewer generalizations (as there are no clear rules) compared to a highly structured language. We chose to compare participants' productions in the generalization test to their own productions in the memory test (and not to the original input language) since we wanted to avoid a confound between learning accuracy and generalization behavior, i.e., not to bias against people who learned the input poorly: if a participant learned the input language only partially but nevertheless generalized based on what they did learn, comparing their productions to the input language would have yielded a low score despite their ability to generalize. Therefore, participants' generalization behavior was based on their own final productions of the input language.

While we believe this measure reflects participants' generalization behavior, it is important to acknowledge that it suffers from several issues that may render it biased or problematic. For example, it is not clear what the overall distribution of possible generalization scores is, and whether is it similar across different input language conditions. Moreover, the scores were normalized with respect to minimal and maximal values achieved by participants in our experiment, rather than by the absolute minimal and maximal generalization values, since it was not clear what these values would be or how to calculate them. Nevertheless, we believe the achieved maximal and minimal values obtained by participants were close to these theoretical absolute values: the best performing participant in the high-structure condition generalized the compositional system perfectly, and the worst performing participant in the low-structure condition seemed to have produced completely random labels. However, it is possible that the maximal generalization value achieved by one participant in a given condition was actually not within the range of generalization scores available to other participants in that condition, that is, that the maximal possible value for participant X was not the maximal possible value for participant Y. This is quite likely given that participants' generalization scores were based on their own productions in the final test, which differed across participants. One way to address these problems in future work would be to develop a new, unbiased measure of generalization behavior, for example, one that is based on simulations of random labels or on ratings by naïve participants.

In addition to being beneficial for individuals' generalization behavior, high structure languages were advantageous for communication between individuals. When we examined the new labels produced for unseen events by different participants who learned the same language, we found that participants who learned more structured languages produced labels that were significantly more similar. That is, systematic structure led different participants to produce similar labels for new meanings without previously interacting with each other. This finding suggests that systematicity allows strangers to converge effortlessly: strangers who never interacted before could potentially communicate successfully about new events – and immediately be understood. This finding supports the postulated mechanism behind larger communities' tendency to develop more systematic languages (Raviv et al., 2019b). Small communities typically have tightly connected networks of individuals who are highly familiar with each other, and can rely on common ground and shared history when communicating about novel events. In contrast, bigger communities have more strangers (i.e., individuals who don't communicate regularly or never interact), who cannot rely on shared history to support mutual understanding. Nevertheless, they need to be able to understand each other when interacting for the first time. As such, it was argued that members of bigger communities are under a stronger pressure to develop transparent, predictable, and systematic structures that aid convergence and allow strangers to successfully communicate (Wray & Grace, 2007). Our findings suggest that the benefits of systematic linguistic structure go beyond learnability and may aid communication and productivity in general language use.

As for the possible contribution of group size beyond linguistic structure, we found no evidence that languages that developed in bigger groups differed from languages that developed in small groups. Across all measures and all analyses, we found no significant effect of group size on learnability or generalization behavior. Although we cannot draw strong conclusions from this null result, it suggests that once the level of linguistic structure is controlled for, there may be no additional benefits to learning languages created by big groups. In other words, the most relevant difference between big and small communities could, in fact, be their tendency to develop different degrees of systematicity (Raviv et al., 2019b).

However, the lack of significant group size effects in our study might not necessarily reflect the state of affairs in natural languages: it is possible that community size does affect language learnability, but that

we did not capture this difference. One possibility is that big groups' languages only show a learning advantage once all members of the group have fully converged on one single language, but that individuals' variations do not possess the same advantage. Specifically, the selected input languages used in this experiment originated from individual members within groups tested by Raviv et al., (2019b). While the languages of members of the same group were similar to each other, they were not identical. It is possible that if groups were fully converged on one single language, or if we had selected only labels that were shared across all group members, these languages might have encompass some other features that would have make them easier to learn.

Another reason why we cannot draw strong conclusions from these null results is that we intentionally chose input languages that were similar in terms of their structural properties. Specifically, we made sure that in the same structure bin, languages from big and small groups would be comparable in terms of their descriptive grammatical properties, such as having a similar type of form-to-meaning mapping and a similar number of irregulars. It is possible that by doing so, we selected languages that were more similar in terms of their structure than the average big/small group languages, and were therefore not representative of their group size origin. This is rather likely when considering the skewed distribution of big vs. small group languages across structure bins in the full set of 144 final languages. For example, there were very few small-group languages in the highest structural bin (3 out of 23 languages), meaning that (a) our choice of a small-group language in that bin was highly limited, and (b) the selected big-group language from that bin had to match the properties of this "rare" small-group language, and may therefore not be the most "representative" big-group language. Similarly, there were relatively few big-group languages in the mid-low structure bin (2 out of 9 languages). It is possible that a random selection of input languages from the full set of final languages (or, alternatively, using the full set of final languages) would have yielded a different result.

Moreover, when selecting the input languages, we controlled for other linguistic features that may make languages more or less learnable, above and beyond their structural properties. Specifically, we chose languages that were similar in terms of their average word length (i.e., the average number of characters in the language's labels) and in their confusability scores (i.e., the average similarity between all labels in a given language). Given that longer and more confusable words are assumed to be harder to learn (Laufer, 2009; Papagno & Vallar, 1992; Willis & Ohashi, 2012), we

chose our input languages from the lower half of the distributions of these two measures, i.e., languages with relatively short words (i.e., between 4 and 7 characters) and with relatively low confusability (i.e., between 0.14 and 0.37). In addition to restricting our possible pool of languages to select from, these selection criteria may have incidentally washed away relevant differences between the two group size conditions, which could have subsequently affected the languages' processing difficulty and overall learnability. In other words, it is possible that word length and confusability are some of the features that differentiate the languages created by big and small groups, and that by controlling for them we eliminated relevant variation. If this is indeed the case, then we may find group size effects when these two measures are varied systematically according to their distributions across small and larger groups, and/or when they serve as predictors of accuracy rather than controls.

To look into this possibility, we examined the distributions of average word length (Figure 6) and average confusability (Figure 7) in the full set of 144 final languages document by Raviv et al., (2019b). We found that languages created by small groups generally had shorter and less confusable words compared to languages created by bigger groups, except for when they were highly structured: although the observation for this bin relies on just a handful of small-group languages, it suggests that highly structured languages created by small groups actually tended to have longer and more confusable words. This pattern may suggest that (a) small groups' languages are easier to learn overall (counter to our predictions) given their shorter and less confusable words, but that (b) highly structured languages of big groups are more learnable.

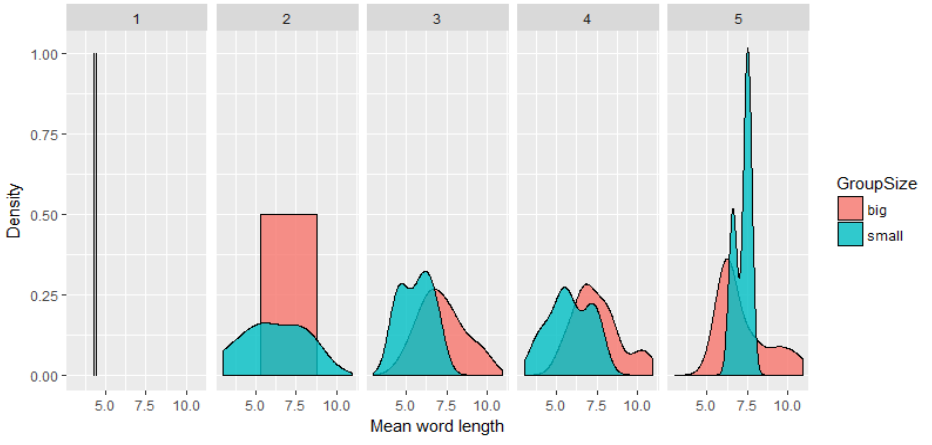


Figure 6. Density of average word length values in the full set of 144 final languages created by big and small groups in Chapter 3, faceted by structure bin (1=low structure, 5=high structure). Since there were exactly two languages in the lowest structure bin (one big group language and one small group language), the density for each group size condition in this bin is exactly 1, with no spread.

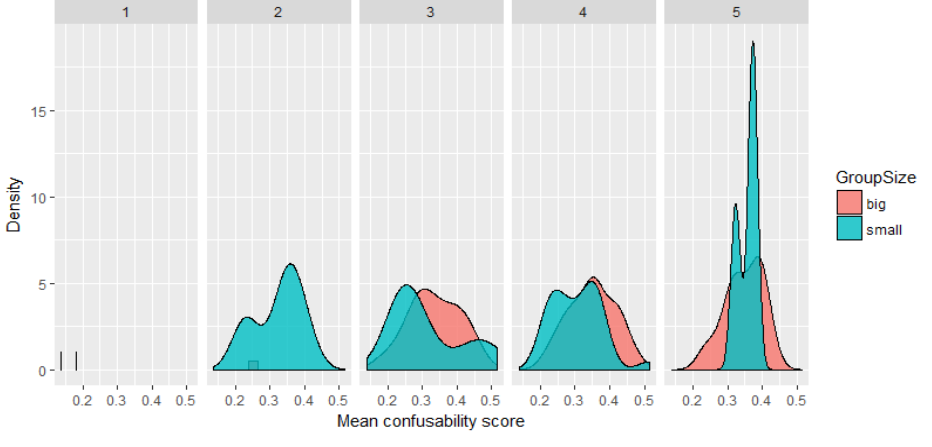


Figure 7. Density of average confusability scores in the full set of 144 final languages created by big and small groups in Chapter 3, faceted by structure bin (1=low structure, 5=high structure). Since there were exactly two languages in the lowest structure bin (one big group language and one small group language), the density for each group size condition in this bin is exactly 1, with no spread.

Interestingly, it is also possible that the effects of word length and confusability on language learnability are modulated by the degree of linguistic structure in the language. Specifically, the effect of word length (i.e., that longer words are harder to learn) may be reduced or even eliminated in highly structured languages. If compositional structure reduces the memory load by making languages more compressible (i.e., given that each word is created by combining repeating morphemes), it may matter less how long the words are in total, or how long the morphemes are. That is, structured languages with short and long words may be just as easy to learn. Similarly, the effect of confusability (i.e., that phonologically similar words are harder to learn) may be modulated by structure. On one hand, compositional languages are potentially more confusable than holistic languages given the repetitions of morphemes, but this increase in confusability may nevertheless be advantageous for learning given the increase in systematic structure. On the other hand, the difficulty in learning a confusable language may actually be amplified in highly structured languages: If a compositional language has highly similar words, the morphemes corresponding to different meanings are relatively similar in form (e.g., the prefix for Shape 1 is very similar to the prefix of Shape 2). If this is the case, such similarity could cause severe problems for learning the mappings between words and meanings in the language.

Importantly, the relationship between linguistic structure and language learnability was not a straight-forward, linear relationship. Although we did predict that this relationship may be non-linear (e.g., that it would be stronger or weaker as structure increases), we were not expecting a U-shape pattern where completely unstructured languages are easier to learn than medium structured languages. Rather, we hypothesized that holistic languages with no systematic structure whatsoever would be harder to learn than languages that exhibit *some* systematic structure, i.e., that any increase in structure would be advantageous for learning. Counterintuitively, participants' final binary accuracy suggested that the hardest languages to learn were those that exhibit some structure, as opposed to none. Even when looking only at participants' final production fidelity, it was not the case that completely holistic and unstructured systems were harder to learn. Rather, low-structured languages and medium-structured languages showed similar production fidelity. One way to account for these unexpected findings is that the nonlinear pattern does not actually hold in natural languages, and does not faithfully represent speakers' true learning biases. Notably, real-world natural

languages are never truly holistic or structure-free: there are no known languages which are fully suppletive or consist only of unpredictable inflections (Ackerman & Malouf, 2013). Instead, languages are inherently quasi-regular, and typically consist of some regular and transparent patterns alongside pockets of opacity and exceptions to the rule (Kempe & Brooks, 2008). Since the low-structure languages in our experiment do not really resemble natural languages, it is possible that the non-linear relationship we observed between language learnability and linguistic structure was merely a quirk caused by our artificial choice of stimuli. If natural languages realistically range only from medium-structure to high-structure, then the actual relationship between systematicity and learning in the real-world may indeed be linear.

While it is possible that a non-linear relationship between language learnability and grammatical structure is less relevant for natural language environments, the nonlinear result (i.e., that partly structured languages are not easier to learn than unstructured languages) is still puzzling. In particular, our original expectation was based on findings from the two artificial language learning studies that examined the benefit of systematic sound-mapping for learning (Brooks et al., 1993; Monaghan et al., 2011). In those studies, languages with partially consistent mapping between phonological features and noun classes were learned better than completely arbitrary languages. Importantly, the stimuli used in those studies can also be seen as unrepresentative of natural languages, given that all natural languages have some degree of iconicity and are never completely arbitrary (Perlman, Little, Thompson, & Thompson, 2018). Yet despite the equally artificial nature of their stimuli, those studies suggested that partial systematicity did aid learning. As such, the unnaturalness of fully unstructured languages does not exempt us from explaining this unpredicted pattern and the discrepancy from previous studies.

A reasonable explanation for the nonlinear relationship we found between learnability and systematicity is that, although partial structure can provide some regularity in the form of statistical cues for meaning, it might also result in more uncertainty and a high cognitive load for learners. Specifically, the inconsistent patterns in medium structured languages may be similarly or even more confusing to learn than a set of unrelated words given (a) participants' learning strategies, and (b) cue validity. First, let us consider that learners are trying (explicitly or implicitly) to build hypotheses about potential linguistic rules (MacWhinney, 1978). This idea is supported by studies showing that

speakers automatically attempt to decompose pseudo-words and non-words into smaller components in a lexical similarity task (Post, Marslen-Wilson, Randall, & Tyler, 2008): any stimulus that can be potentially interpreted as ending in an inflection, whether real or not, is responded to more slowly than an unambiguous stimulus. Moreover, adults tend to assume that unpredictable variation is, in fact, meaningful, and tend to treat random patterns as if they rely on factors not yet discovered (Perfors, 2016). Such findings suggest that speakers try to figure out the underlying structure of word forms, and that morphotactic ambiguity can therefore elicit processing costs and learning difficulties when these hypotheses are not met. Furthermore, it implies that participants' learning strategy may differ across conditions. Learners of highly systematic languages might start out with an item-based learning strategy and initially memorize individual words, but, over time, could detect consistent patterns in the language that regularly associate part-words with semantic features, and consequently switch to forming rules and abstractions (Kempe & Brooks, 2008). In contrast, learners of completely unstructured languages may soon realize that word forms appear to be random, and that there are no meaningful or useful patterns in the language. By hypothesis, they then may simply "give up" looking for rules, and focus on memorizing the holistic lexicon in an item-based manner (i.e., rote learning). But since medium structured languages contain some partial patterns (e.g., shapes have consistent markings, but angles don't) and/or some inconsistent patterns (e.g., some morphemes appear with a given angle only sometimes), learners may be motivated to keep looking for systematic cues and abstractions, even when these do not exist. The fact that their input actually does not contain clear and systematic governing rules may lead to confusion and even frustration, and could require increasing effort. Even if one abandons a rule-based learning model in favor of associative learning, i.e., learning as gradual strengthening of the association between co-occurring elements of the language, the absence of valid and reliable cues would still hinder learning (Kempe & MacWhinney, 1998).

In any case, the finding that the relation between linguistic structure and learnability is not linear (i.e., so that holistic languages are not necessarily more difficult to learn) poses a potential problem for iterated language learning models, which rely on a learning advantage of *some* structure compared to none (Kirby et al., 2008, 2015; Smith, 2011). Specifically, studies on the cultural evolution of compositionality via iterated learning have shown that compositional linguistic structure gradually emerges over time from a state of a holistic lexicon. Crucially,

this slow accumulation in structure is typically attributed to learnability pressures, i.e., to the difficulty in memorizing a completely unstructured lexicon. Accordingly, such models assume that the learning advantage provided by linguistic structure is already present in the early stages of language evolution, and facilitates the emergence of linguistic structure to begin with. One way to reconcile these claims with our findings is to argue that creating linguistic structure has additional benefits to language users, above and beyond the benefits to learning. Indeed, our study suggests that this is the case: highly systematic languages are favored not only because they are more learnable, but also because they are predictable and allow for clear generalizations and quick convergence. This idea resonates with early iterated learning models (Kirby, 2002; Smith, Brighton, & Kirby, 2003), which stress the benefit of linguistic structure for generalizations: although agents are usually not exposed to the entire repertoire of the language, they must be able to produce labels to new events despite their partial exposure.

Finally, the relation between morphological structure and learning difficulty may differ in strength across different populations of learners. In particular, the current study was based on adult participants, who may differ from children in their learning biases (Dale & Lupyan, 2012; Hudson Kam & Newport, 2005; Lupyan & Dale, 2015; Nettle, 2012). The possible differences between children and adults' language learning preferences are especially relevant given the postulated role of adult second-language learners in simplifying morphologically complex languages (Dale & Lupyan, 2012; Lupyan & Dale, 2015). As discussed in the Introduction, the tendency of big communities to have simpler languages is often attributed to the higher proportion of adult second-language learners in bigger communities. This argument is based on the assumption that adults indeed differ from children in their learning biases, which may lead to different learnability pressures across age groups. For example, it has been suggested that children, but not adults, benefit from the existence of redundant cues, even though such redundancy is typically considered to increase linguistic complexity. If this suggestion is true, more complex languages may indeed be harder to learn for adult learners, but may be equally learnable (or even more learnable) for children. In other words, it is possible that the advantage of systematic structure demonstrated in this study does not hold for child learners.

However, the idea that children's learning biases are radically different from those of adults, so much so that they would not benefit from systematic linguistic structure, seems unlikely for several reasons. First,

the postulated advantage of learning a more systematic language is based on general memory constraints and cognitive principles of compressibility, which should be present in learners across all ages. If anything, the benefit of systematic structure may even be greater in children given their lower working memory capacity (Gathercole, Pickering, Ambridge, & Wearing, 2004). That is, children should benefit from systematicity just as much as adults, if not more. Supporting this idea, children's acquisition of more morphologically complex languages is typically argued to be slower than languages with simpler morphologies, suggesting that children's language learning is indeed affected by the morpho-syntactic properties of their language (Hengeveld & Leufkens, 2018; Slobin, 1985). Second, although adults are typically viewed as being inferior in learning a second language compared to children (DeKeyser, 2005), the differences between children and adults with respect to language learning outcomes do not necessarily reflect fundamental differences in their learning biases. Rather, adults' learning difficulty is often attributed to language-external factors such as their meta-linguistic awareness and prior knowledge, their learning strategies (implicit vs. explicit), the type or quantity of input they are exposed to, their motivation and social immersion, etc. (Birdsong, 2006; DeKeyser, 2013). Supporting this point, children and adults were shown to be equally affected by the degree of systematic mapping between phonological forms and grammatical categories in an artificial language learning task, with both age groups significantly benefitting from having systematic cues to indicate noun classes (Brooks et al., 1993). Additionally, the only study that compared children and adults' performance on iterated language learning reported similar learning patterns across age groups, despite children's overall inferior performance (Raviv & Arnon, 2018). Although there is anecdotal evidence that children can benefit from rich inflectional environments (Xanthos et al., 2011) and from redundant cues (Tal & Arnon, 2019), as long as there is no direct empirical evidence to support the claim that *only* children benefit in this way, there is no reason to assume that the learning advantage of more systematic languages does not hold across the lifespan.

Conclusions

The current study tested the acquisition of different artificial languages that varied in their degree of systematic structure and in their community size origin. We found that more linguistic structure generally benefited

language learning, with highly structured languages were learned fastest and most accurately. Interestingly, the relationship between language learnability and linguistic structure was not straight forward: high systematicity was indeed advantageous for learning, but learners did not seem to benefit from partly or semi-structured languages (i.e., languages that contained some patterns but also multiple irregulars and inconsistencies). We also found Community size did not affect learnability: languages that evolved in big and small groups were equally learnable. Crucially, our results suggest that systematic structure is not only beneficial for learning, but also for generalizations and convergence. Participants who learned more structured languages were better at generalizing the language they learned to new, unfamiliar meanings. Moreover, different participants tended to create similar new labels as structure increased. That is, systematicity facilitated convergence and mutual understanding between strangers.

Appendix A: Power Analysis

The sample size for our study was determined by conducting a power analysis for both a linear and a polynomial effect of Structure Score on binary accuracy (1 for accurate, 0 for inaccurate). We based our power simulations for this analysis on scaled down effect sizes that were estimated on data collected in a pilot study. Additionally, we computed power for two hypothesized effects of Group Size on binary accuracy.

Here we explain the rationale and procedure of the power analysis, as well as the resulting plots that we used to determine our study's sample size. The simulations and analyses was performed in R (v. 3.5.0; R Core Team, 2016) using the *simr* package (v. 1.0.5; Green & McLeod, 2016). The full script for running the simulations and power analyses is available at https://osf.io/abvcg/, and includes detailed comments describing the procedure. The full output of our simulation run is available at https://osf.io/htvqy/.

Effect of structure

Our hypothesized effects of structure on binary accuracy were based on results from a pilot study, in which we tested three out of our ten input languages (S1, S3 and S5), with two participants per language. We expected that the effect of Structure on binary accuracy would either be a linear or a 2-degree polynomial effect, so two separate generalized linear mixed effect models were fitted to the pilot data accordingly.

Figure 1 visualizes the data obtained in the pilot experiment and the estimates of the two models for the fixed effects of Structure. We used these models to simulate the data for the rest of the analyses after scaling down the effect sizes by factors of 0.1, 0.15 and 0.2. Table 1 lists the effects as estimated by the pilot model, as well as the scaled down effect sizes for each scaling factor.

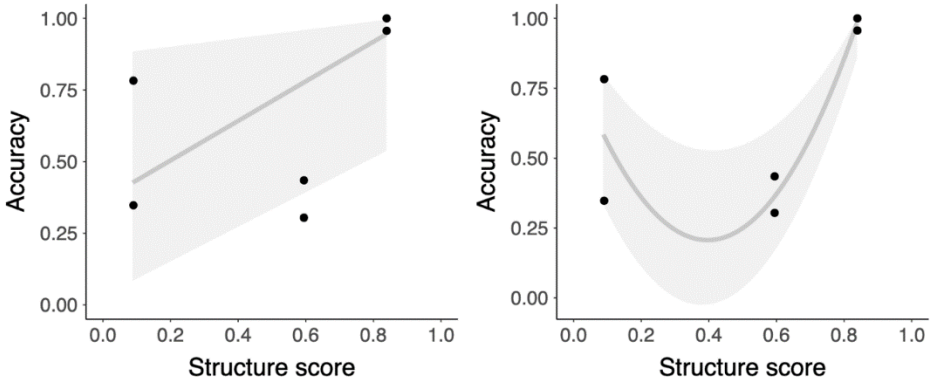


Figure 1. Linear and 2-degree polynomial effects of Structure on binary accuracy, as predicted by the linear mixed-effect models fitted on the pilot dat**a.** Mean accuracy scores per participant are visualized as black dots**.**

Table 1: Model estimates for the fixed effect of Structure on binary accuracy as fitted on the pilot data, as well as the scaled down effects used for our power simulations, by the three different scaling factors (0.1, 0.15 and 0.2).

	Pilot model estimate	**0.1-scaled effect**	**0.15-scaled effect**	**0.2-scaled effect**
Linear model	13.4	1.34	2.01	2.68
Polynomial model, linear term	14.6	1.46	2.19	2.92
Polynomial model, quadratic term	18.5	1.85	2.78	3.70

Effect of group size

We expected the potential effect of group size to be positive, i.e., that participants learning languages created by bigger groups would obtain higher accuracies on the memory test. To simulate an effect of Group Size, we scaled the predicted mean accuracy of the big group languages by a factor of either 1.05 or 1.10, simulating an effect of either 5% or 10% increased accuracy respectively. We chose these effect sizes as we estimated them to be the smallest possible effects on accuracy that would still be reasonably measurable with our planned experimental setup.

Simulation procedure

Power for the effect of group size was calculated for sample sizes ranging from 2 participants to 15 participants per each input language condition (i.e. from 20 to 150 participants in total), and included all combinations of the scaled effects of structure and group size. For each of these 84 possible settings, the simulation was run 1000 times to calculate the rates of correctly detecting each specified effect using linear mixed effect models that were equivalent to the confirmatory analysis for binary accuracy used for our experimental data.

In the case of structure, power was estimated for both detecting an effect and preferring it over the other effect type in model comparison (e.g. correctly detecting a polynomial effect on data simulated from a polynomial model, and preferring the 2-degree polynomial effect over the linear effect). To estimate power for the effect of group size, we calculated the average rate of correctly detecting an effect of group size across all effect types and sizes of structure.

Power simulation results

Our obtained estimates of statistical power varied per simulation setting (i.e., sample size, effect size, and effect type). Below we visualize the power curves for the two different effect types of Structure, as well as the effect of Group Size, for different effect sizes and sample sizes. The script for reproducing these graphs is available at https://osf.io/ywat5/, and the full results of these simulations are available at https://osf.io/htvqy/. Overall, the results show that power varied according to effect size, but rapidly increased when the effect size was higher than our smallest simulated effect size (scaling of 0.1).

Figure 2 visualizes the statistical power for finding a significant effect of Structure on binary accuracy and for preferring the model with the linear fixed effect over the model with the polynomial fixed effect (in both cases with $\alpha = 0.05$) given that the data was generated by a linear model. Figure 3 visualizes the statistical power for finding a significant effect of Structure on binary accuracy and for preferring the model with the polynomial fixed effect over the model with the linear fixed effect (in both cases with $\alpha = 0.05$) given that the data was generated by a polynomial model. Figure 4 visualizes the statistical power for finding a significant effect of Group Size on binary accuracy ($\alpha = 0.05$) given that the data was

generated with an either 5% or 10% increase in accuracy for bigger groups.

Based on these results, we decided on a sample size of 10 participants per input language condition, corresponding to 100 participants in total. This sample size provided us with reasonable statistical power (>60%) even for very small effect sizes. Importantly, the effect sizes of Structure on binary accuracy estimated by the models fitted on our *actual* experimental data were higher than our largest simulated effects (scaling of 0.2). As such, we are confident that the statistical power for our confirmatory analysis was over 80%.

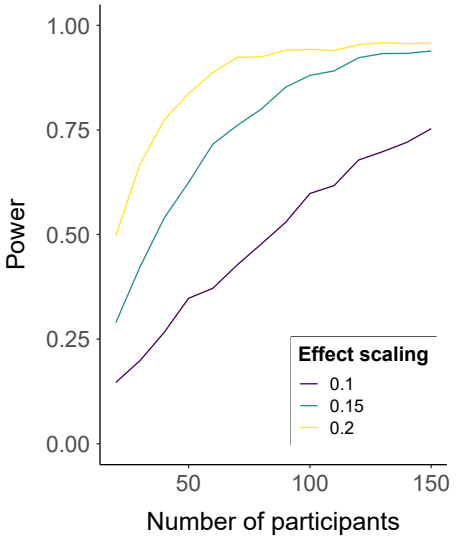


Figure 2. Power curves as estimated based on our simulations for a linear effect of Structure on binary accuracy. Shadings indicate the 95% binomial confidence intervals. Depending on effect size the power for our chosen sample size of 100 participants is approximately between 60% and 90%.

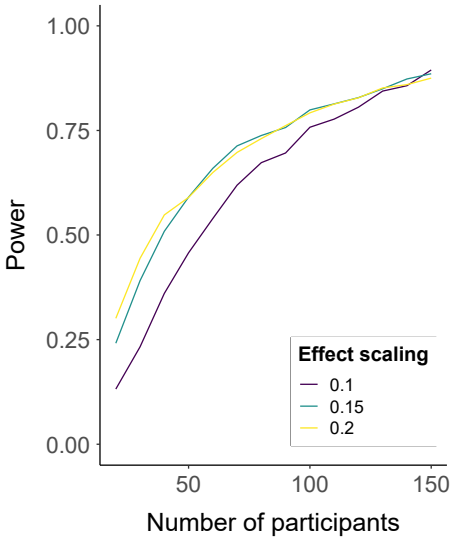


Figure 3. Power curves as estimated based on our simulations for a second-degree polynomial effect of Structure on binary accuracy. Shadings indicate the 95% binomial confidence intervals. Depending on effect size the power for our chosen sample size of 100 participants is approximately between 70% and 80%.

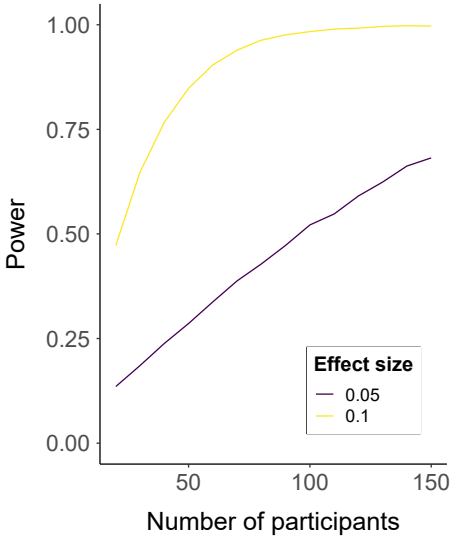


Figure 4. Power curves as estimated based on our simulations for 5% vs. 10% effect of Group Size. Shadings indicate the 95% binomial confidence intervals. Depending on effect size the power for our chosen sample size of 100 participants is over 60%.

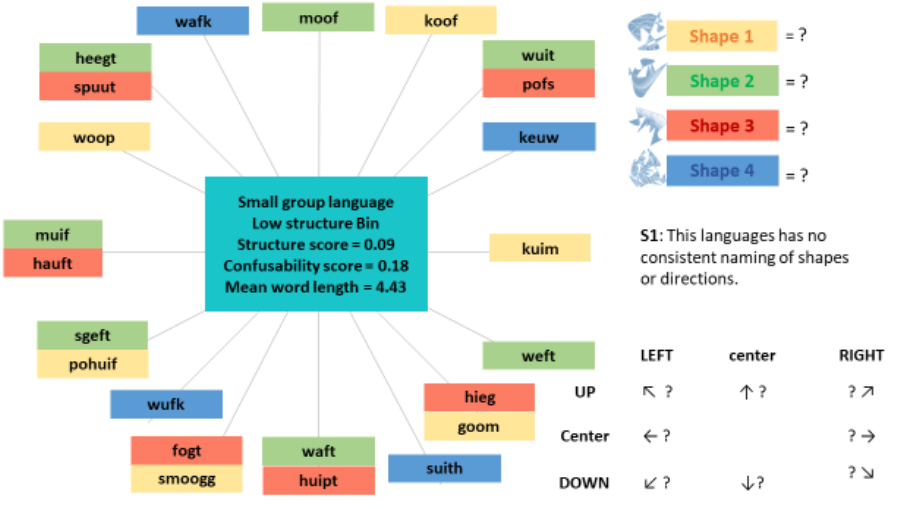
Appendix B: Input Languages

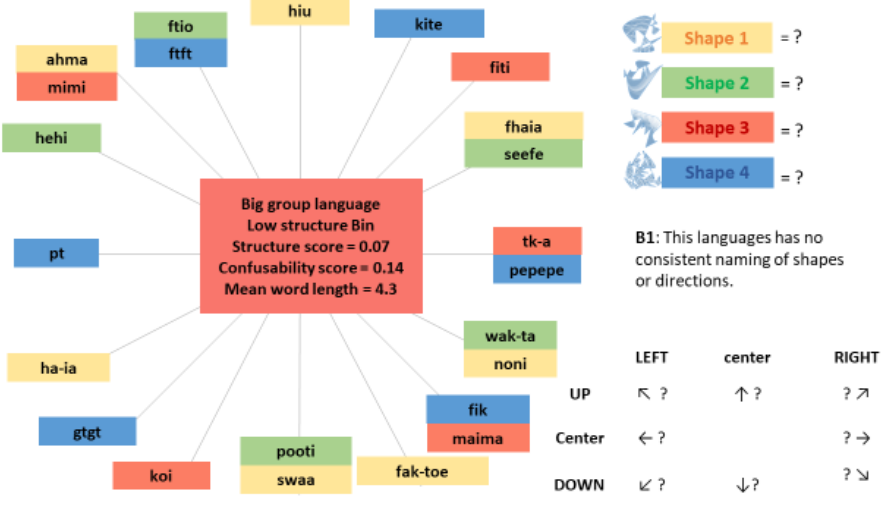
This appendix was adapted from a similar pre-registered file (https://osf.io/ya2ps/) and includes a detailed description of each of the 10 input languages used in the experiment.

Each language is characterized by a short description, as well as its structure score, confusability score, and average word length.

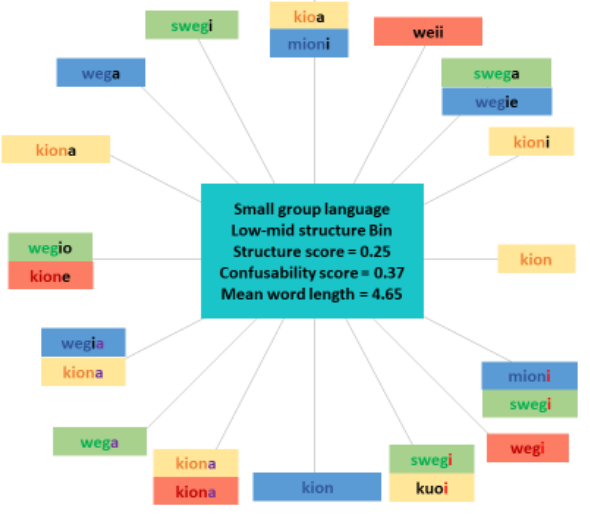
Each language is accompanied by a "dictionary" for interpreting the language on the right. Different box colors represent the four different shapes which appeared in the scenes, and the grey axes indicate the direction in which the shape was moving on the screen. Different font colors represent different meaningful part-labels, as segmented by the authors.

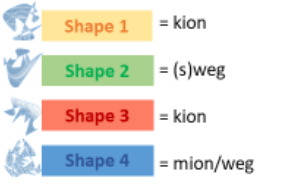
Low structure bin (1)





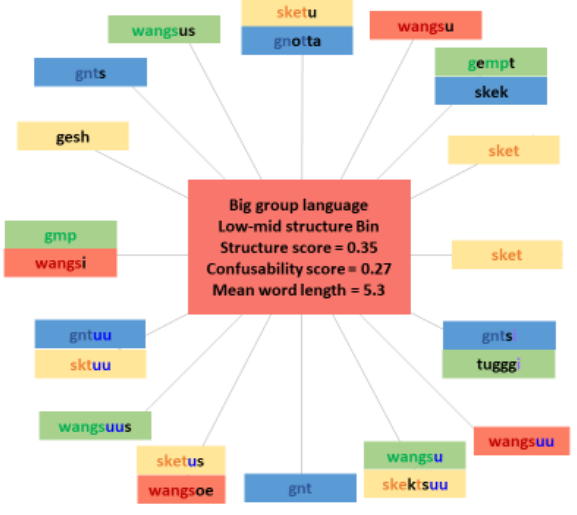
Low-Mid structure bin (2)

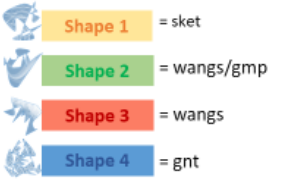




S2: Some part-labels repeat more with certain shapes than with others, but there is still no clear labeling for shapes (e.g., *kion* and *weg* are used for multiple shapes). The directions down-left and down-right are consistently labelled, but there is no consistent labeling for any of the other directions, and the same endings repeat without clear rules.

	LEFT	center	RIGHT
UP	↖ ?	↑ ?	? ↗
Center	← ?		? →
DOWN	↙ a	↓ ?	i ↘

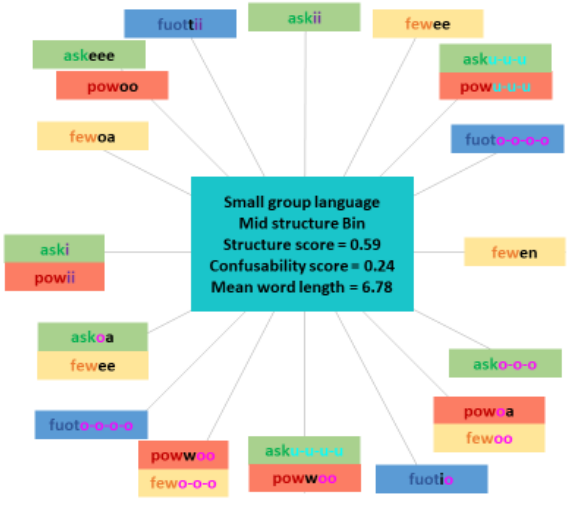


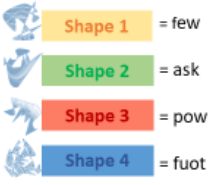


B2: Some part-labels repeat more with certain shapes than with others, but there is still no clear labeling for shapes (e.g., *wangs* is used for multiple shapes). There is little structure in the labelling of directions (some repetitions of *uu* for down-left, and *i* for one direction in the down-right quadrant, but no clear structure for motion otherwise.

	LEFT	center	RIGHT
UP	↖ ?	↑ ?	? ↗
Center	← ?		? →
DOWN	↙ uu	↓ ?	uu/i ↘

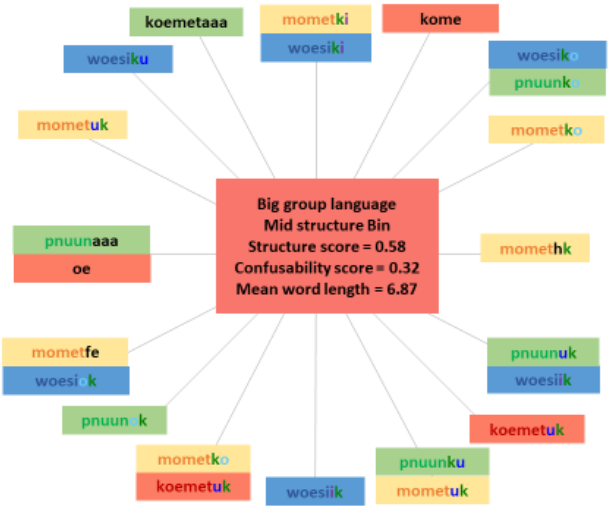
Medium structure bin (3)

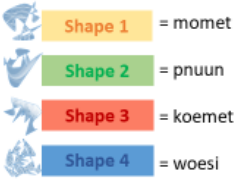




S3: All shapes are consistently labeled with the same part-words. However, there is no clear way to label directions. The same vowels are always added to the end of words, but with only partially consistent structure (e.g., *o-o* is used for up-right, down-right and down-left).

	LEFT	center	RIGHT
UP	↖ ?	↑ ?	u-u-u ↗
Center	← i		? →
DOWN	↙ o-o	↓ ?	o ↘

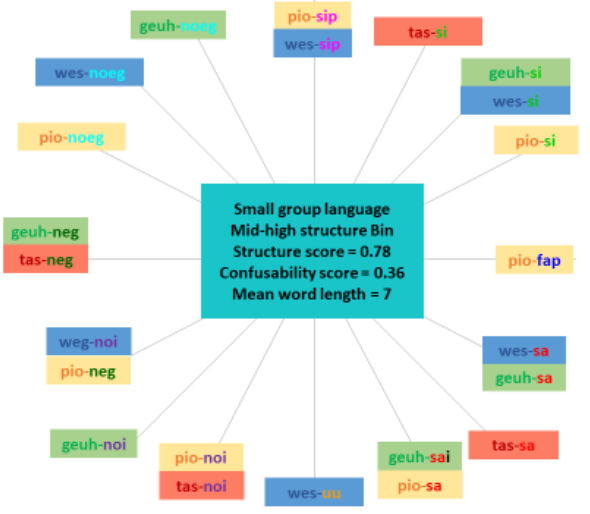


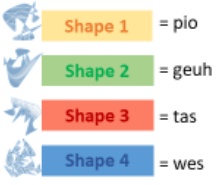


B3: Labeling of shapes 1 and 4 is fully consistent, but only partly consistent for shapes 2 and 3. There is some structure in the labeling of directions, with mirroring the order of part-labels for opposite directions (*ik, ki, ko, ok*), yet it is not fully consistent.

	LEFT	center	RIGHT
UP	↖ ?	↑ ki	ko ↗
Center	← ?		? →
DOWN	↙ ok	↓ ik	uk ↘

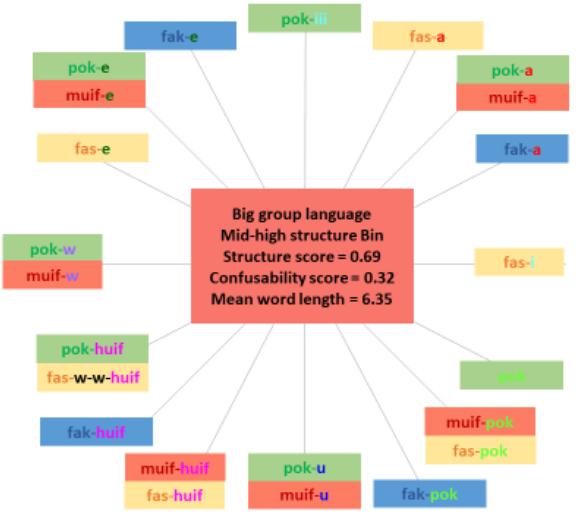
Mid-High structure bin (4)

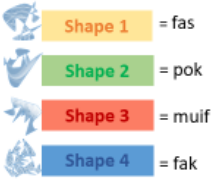




S4: Labels for shapes and for directions are consistently combined across the whole language, with only two small exceptions in direction morphemes. The direction morphemes themselves are holistic and not constructed compositionally (e.g., the morpheme for down-left is not a combination of the morphemes for down and left).

	LEFT	center	RIGHT
UP	↖noeg	sip↑	si↗
Center	←neg		fap→
DOWN	↙noi	uu↓	sa↘

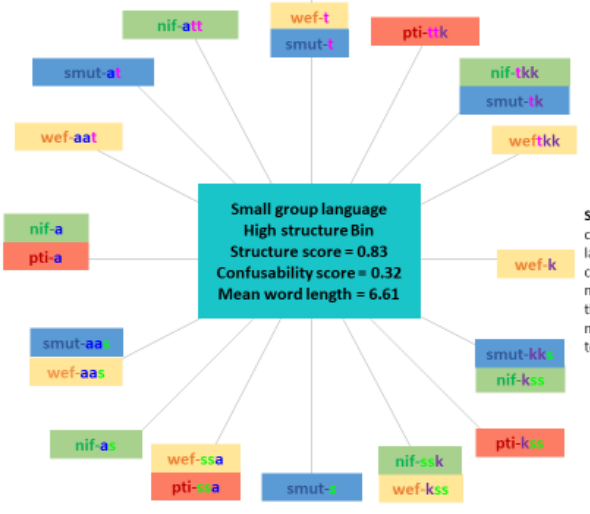


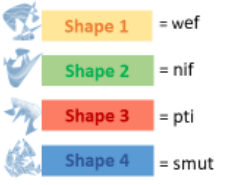


B4: Labels for shapes and for directions are consistently combined across the whole language, with only two small exceptions in direction morphemes. The direction morphemes themselves are holistic and not constructed compositionally (e.g., the morpheme for down-left is not a combination of the morphemes for down and left).

	LEFT	center	RIGHT
UP	↖e	iii↑	a↗
Center	←w		i→
DOWN	↙huif	u↓	pok↘

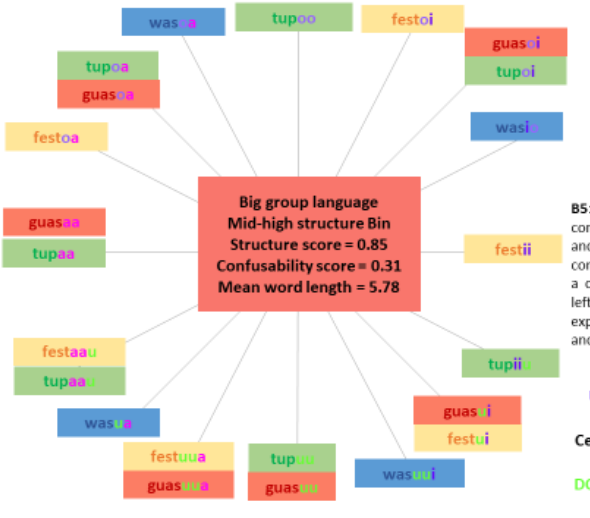
High structure bin (5)

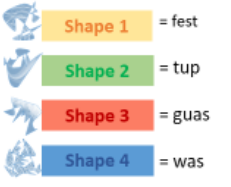




S5: Labels for shapes and for directions are consistently combined across the whole language, and the motion morphemes are compositionally constructed (e.g., the morpheme for down-left is a combination of the morphemes for down and left). Doubling of motion morphemes is used to express closeness to a given axis, albeit with some exceptions.

	LEFT	center	RIGHT
UP	at↖	t↑	tk↗
Center	a←		k→
DOWN	as↙	s↓	ks↘





B5: Labels for shapes and for directions are consistently combined across the whole language, and the motion morphemes are compositionally constructed (e.g., the morpheme for down-left is a combination of the morphemes for down and left). Doubling of motion morphemes is used to express closeness to a given axis in the down-left and down-right quadrants.

	LEFT	center	RIGHT
UP	↖oa	oo↑	oi/io↗
Center	←aa		ii→
DOWN	↙ua/au	uu↓	ui/iu↘

Appendix C: Models

Final Memory Test

(1) Binary Accuracy (confirmatory)

Accuracy ~ poly(centered.Structure,2) + Condition + (1 | Item) + (1 | Participant)

(Model with 2-degree polynomial favored: ΔAIC = 16.4, p<0.0001)

	Estimate	Std.Error	z-value	p-value
(Intercept)	0.82547	0.21370	3.86280	0.00011
Group Size Origin (Big vs. Small)	0.48359	0.29018	1.66649	0.09562
Structure Score (Linear)	31.47067	6.92838	4.54228	0.00001
Structure Score (Quadratic)	30.99636	6.86715	4.51372	0.00001

(2) Production Similarity (exploratory)

ProdSimilatity ~ poly(centered.Structure,2) + Condition + (1 | Item) + (1 | Participant)

(Model with 2-degree polynomial favored: ΔAIC = 3.52, p=0.01877)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.83338	0.02088	39.91962	0.00000
Group Size Origin (Big vs. Small)	0.00737	0.02851	0.25851	0.79657
Structure Score (Linear)	4.40701	0.67901	6.49030	0.00000
Structure Score (Quadratic)	1.59779	0.68384	2.33651	0.02154

Learning Trajectory

(3) Binary Accuracy Over Time (exploratory)

Accuracy ~ centered.Block * Condition + centered.Block * poly(centered.Structure,2) + (1 | Item) + (1 + centered.Block | Participant)

(Model with 2-degree polynomial favored: ΔAIC = 13, p=0.0002)

	Estimate	Std.Error	z-value	p-value
(Intercept)	0.87848	0.14082	6.23849	0.00000
Group Size Origin (Big vs. Small)	0.22843	0.19145	1.19314	0.23281
Block Number	0.28930	0.04827	5.99326	0.00000
Structure Score (Linear)	66.49299	10.67087	6.23126	0.00000
Structure Score (Quadratic)	51.82917	10.73681	4.82724	0.00000
Block X Group Size	0.03748	0.06992	0.53606	0.59192
Block X Structure Score (Linear)	25.88165	4.31615	5.99646	0.00000
Block X Structure Score (Quadratic)	14.45578	4.75638	3.03924	0.00237

(4) Production Similarity Over Time (exploratory)

ProdSimilatity ~ centered.Block * Condition + entered.Block * centered.Structure + (1 | Item) + (1 + centered.Block | Participant)

(Model with 2-degree polynomial not favored: ΔAIC = 1.3, p=0.26)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.78620	0.01953	40.25844	0.00000
Group Size Origin (Big vs. Small)	0.00103	0.02609	0.03956	0.96852
Block Number	0.04038	0.00724	5.57434	0.00000
Structure Score	0.33807	0.04602	7.34551	0.00000
Block X Group Size	-0.01402	0.01023	-1.37147	0.17311
Block X Structure Score	0.05101	0.01806	2.82546	0.00564

(5) Guessing Similarity Over Time (exploratory)

GuessSimilatity ~ centered.Block * Condition + centered.Block * centered.Structure + (1 | Item) + (1 + centered.Block | Participant)

(Model with 2-degree polynomial favored: ΔAIC = 9.9, p=0.0009)

	Estimate	Std.Error	t-value	p-value
(Intercept)	1.77965	0.02198	80.96479	0.00000
Group Size Origin (Big vs. Small)	0.01219	0.02981	0.40889	0.68353
Block Number	0.04375	0.01171	3.73519	0.00032
Structure Score (Linear)	14.65422	2.12964	6.88108	0.00000
Structure Score (Quadratic)	8.06983	2.14480	3.76251	0.00029
Block X Group Size	-0.01421	0.01660	-0.85623	0.39400
Block X Structure Score (Linear)	1.75145	1.18646	1.47620	0.14316
Block X Structure Score (Quadratic)	-0.05540	1.19550	-0.04634	0.96314

Generalization Behavior

(6) Generalization Score (exploratory)

Normalized.Generalization ~ Condition + centered.Structure

(Model with 2-degree polynomial not favored: ΔRSS = 0.0351. p=0.34)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.78112	0.02858	27.32870	0.00000
Group Size Origin (Big vs. Small)	0.01259	0.04042	0.31151	0.75608
Structure Score	0.51498	0.07129	7.22319	0.00000

(7) Generalization Convergence (exploratory)

Mean.Convergence ~ Condition + centered.Structure + (1 | Item)

(Model with 2-degree polynomial not favored: ΔAIC = 0.61. p=0.23)

	Estimate	Std.Error	t-value	p-value
(Intercept)	0.58823	0.01368	42.98378	0.00000
Group Size Origin (Big vs. Small)	-0.03308	0.01935	-1.70924	0.08989
Structure Score	0.73820	0.03414	21.62583	0.00000

6 Summary and General Discussion

Why are there so many different languages in the world? How much do languages differ from each other in terms of their linguistic structure and their learnability? And how do such differences come about?

This doctoral thesis attempted to shed light on the *social* origin of language diversity by experimentally examining the live-formation and acquisition of new languages that were created under different social conditions. Specifically, it looked at how real-world communicative pressures can give rise to systematic, compositional languages, and tested how this process is shaped by the fact that languages evolve in different communities, with different population sizes and different types of social networks. It also examined whether languages that evolved under different social conditions differ from each other in how easily they are learned and used.

This chapter summarizes the main findings of the preceding experimental chapters, and reflects on their main themes and broader implications. It also discusses methodological issues and suggests possible directions for future research.

1 Summary of main findings

Chapter 2 tested the prediction that systematic compositional structure can emerge during communication as a result of communicative needs. Previous work suggested that communication alone is insufficient for compositionality to emerge (Kirby et al., 2015), and that languages only develop systematic structures when they are also subjected to a learning pressure (i.e., when they are transmitted across multiple generations of learners). However, we hypothesized that natural properties of language use can give rise to similar pressures and to the creation of structured languages even without generational transmission. Specifically, we predicted that two aspects of real-world communication, namely, interaction with multiple people and interaction about an expanding meaning space, would lead to the emergence of compositional languages in closed groups. To test this prediction, we introduced a novel group communication paradigm in which different micro-societies comprised of four participants interacted using an artificial language they created on-the-go. Participants interacted with all other members of their group in

alternating pairs, and needed to refer to and discriminate between more and more novel meanings (i.e., dynamic scenes) over time. We tested whether groups developed compositional languages with consistent form-to-meaning mappings, and examined the relative individual contribution of each of the two communicative pressures in the design (i.e., interaction with multiple people and an expanding meaning space) using a meta-analysis. We also characterized the emerging languages in terms of convergence, stability, and communicative success.

The results of this chapter showed that the languages that evolved in micro-societies became significantly more structured over the course of multiple interactions, and developed compositionality despite the absence of generation turnover. In particular, the groups in this experiment developed languages in which different affixes were systematically combined to express different meanings (e.g., scenes' shape or direction of motion). Additionally, the emerging languages became more shared across different members of the group, more stable, and more communicatively successful over time. These findings show that systematic languages can evolve under communicative pressures, and that new learners are not necessary for the formation of grammatical structure within a community. Results also showed that having multiple people to interact with was the main driver for the emergence of compositionality in this paradigm. This result implied that differences in the number of interaction partners (i.e., group size) can affect the formation of linguistic structure and the degree of systematicity in the evolving languages. Specifically, it suggested that larger communities may be under a stronger pressure for systemization and generalization, and may therefore develop more compositional languages.

Chapter 3 directly tested the role of community size in the formation of languages by contrasting the performance of small and big groups using the same group communication paradigm described in Chapter 2. Specifically, it compared the languages that emerged in micro-societies comprised of either four or eight members in order to evaluate the prediction that bigger groups would create more structured languages. This prediction was motivated by cross-linguistic correlational studies and by theories of language change, which suggested that languages spoken in big communities tend to have more systematic and transparent grammars (Lupyan & Dale, 2010; Trudgill, 2002, 2009; Wray & Grace, 2007). While these findings are often attributed to factors that are naturally confounded with a large community size (i.e., a high proportion of non-native learners and/or more interaction with strangers), we

hypothesized that the sheer number of people in the community could already affect the formation of languages in meaningful ways, and help explain the observed patterns of linguistic diversity. This chapter also tested two potential factors that may underlie group size effects, namely differences in input variability and differences in shared history.

The results of this study showed that larger groups indeed developed more systematic languages over time, and did so faster and more consistently than small groups. The results further suggested that this increase in linguistic structure was driven by the greater input variability in larger groups. Specifically, more input variability introduces a greater communicative challenge, which members of larger groups needed to overcome in order to communicate successfully. Consequently, larger groups were under a stronger pressure to generalize their languages and favor systematic variants, which could in turn ease mutual understanding and facilitate convergence and effectively reduce input variability. Moreover, results showed that small groups varied more in their linguistic behaviors, while larger groups behaved relatively similarly to each other. We therefore suggested that smaller communities were more susceptible to random events (i.e., drift), and may therefore exhibit less consistent and rarer behaviors (Spike, 2017). Together, the results of this chapter showed that community size had a unique and causal influence on the formation of new languages, and provided the first experimental evidence that larger communities create more systematic languages. They also supported the claim that a growth in early humans' population size may have been one of the main drivers for the evolution of compositional and systematic grammars from a state of an unstructured protolanguage (Dunbar, 2017).

Chapter 4 investigated the role of social network structure in the process of language emergence, using the same paradigm as in Chapters 2 and 3. Here, we compared equally sized groups of eight members in three different network conditions: fully connected, small-world, and scale-free. These network configurations differed in their degree of connectivity (i.e., how many people each participant interacted with) and in their homogeneity (i.e., whether all participants were equally connected). Our main prediction was that sparsely connected networks (small-world and scale-free networks in this experiment) would develop more systematic languages compared to a dense fully connected network. This prediction was motivated by work on social network structure and theories on language typology, which suggest that weak ties in sparser networks promote diversity and consequently the creation of complex innovations (Derex & Boyd, 2016; Granovetter, 1983; Lou-Magnuson &

Onnis, 2018; Trudgill, 2002, 2009). We hypothesized that sparser networks would create more structured languages as a result of greater input variability, which should increase the pressure for generalization and systematization. We also predicted that the presence of a highly connected agent (i.e., "hub") in scale-free networks would further advance convergence and the spread of compositional languages (Fagyal et al., 2010; Zubek et al., 2017).

In contrast to our predictions, results showed no significant effect of network structure for any measure. Groups in all network conditions developed languages that were highly compositional and systematic, and did so to similar extents. Similarly, there were no significant differences in the levels of communicative success, stability, and convergence achieved by different groups. We argued that these null findings could be traced back to the absence of significant differences in input variability across network conditions, which were a prerequisite for our predictions. More research is therefore needed in order to test the role of network structure in explaining patterns of language diversity. At the same time, a consistent and significant finding across all linguistic measures was that small-world networks showed the greatest variation in their behaviors: while different fully connected and scale-free groups behaved relatively similarly to other groups in the same condition (i.e., reaching similar levels of structure, convergence, stability, and accuracy), small-world groups differed from each other more in their behaviors (i.e., reaching varying levels of structure, convergence, stability, and accuracy). This pattern suggested that the frequent interactions amongst small sub-groups in small-world networks could preserve random behaviors more easily and could result in small-world groups being more likely to fixate on local (and possibly costly) strategies instead of converging on more optimal solutions. These findings indicated that network structure can nevertheless affect the community's vulnerability to drift, and resonated with the findings of Chapter 3 that showed a similar vulnerability to drift in small groups.

Chapter 5 addressed a crucial assumption underlying the work presented in the previous chapters of this dissertation – the idea that more systematic and compositional languages are easier to learn. The postulated link between language learnability and language structure is also a crucial component in related theories of language evolution, language diversity, and language acquisition (Cornish, 2010; Cornish et al., 2009; Dale & Lupyan, 2012; Kirby, 2002; Kirby et al., 2008; Zuidema, 2003), but so far had not been confirmed experimentally. Additionally, it has been shown

that big groups tend to develop signal systems that are superior in terms of their learning and processing, above and beyond their complexity (Fay et al., 2008; Fay & Ellison, 2013), which suggests that language learnability may also be affected by community size. That is, languages that evolved in bigger groups may be easier to learn for reasons unrelated to their degree of linguistic structure, for example by being better adapted to individuals' cognitive biases and/or general linguistic preferences. This chapter directly probed the postulated causal links between language learnability, linguistic structure and community size by experimentally testing whether languages that evolved in different-sized groups and had different degrees of systematic linguistic structure differed in how easily they were learned by new individuals. Specifically, we compared the acquisition of a range of artificial languages that were created by participants in Chapter 3 and had different levels of systematicity in their form-to-meaning mappings. We also tested how well learners could generalize the languages they learned to describe new, unfamiliar meanings, and whether different participants generalized the languages in similar ways.

Results showed that more linguistic structure was advantageous for language learning, such that languages with highly systematic grammars were learned faster and more accurately. However, the relationship between language learnability and linguistic structure was non-linear: while highly structured languages were easier to learn, learners did not seem to benefit from partly or semi-structured languages. Results also showed that community size did not affect learnability, such that languages that evolved in big and small groups were equally learnable. Finally, participants who learned highly structured languages were better at generalizing them to new, unfamiliar meanings, with different participants being more likely to produce similar labels. Together, these results showed that linguistic structure is advantageous not only for language learning, but also for language use: systematic languages allow for productive labeling, which in turn promotes quick and effortless convergence between strangers (Wray & Grace, 2007). This result is directly related to the mechanism suggested to underlie the results of the previous chapters (i.e., that compositional structure helps to relieve participants' memory load and can facilitate convergence), and showed that the creation of more systematic languages can indeed help communities overcome communicative challenges such as interacting with multiple people.

2 Discussion

The goal of this doctoral thesis was to explore how communicative pressures and different aspects of societies shape the formation and distribution of linguistic properties in an artificial language game. Specifically, I tried to experimentally tease apart different social features that are confounded in the real-world, and to examine whether different degrees of linguistic structure emerge in different types of communities. This was done using a novel group communication paradigm in which different micro-societies created new languages over time (Chapters 2-4), as well as an individual learning experiment in which these emerging languages were assessed in terms of their learnability and productivity (Chapter 5).

Taken together, the results presented in this doctoral thesis show that:

(1) Individuals' communicative needs (e.g., the need to successfully interact with multiple people) can lead to the creation of linguistic structure.

(2) The process of language evolution and change is affected by community size, but perhaps not by network structure, at least not in the current design.

(3) The emergence of linguistic structure in a community can in turn serve its members' communicative needs by benefiting language use, and can also serve its future members in by benefiting language learning.

The implications for our understanding of how language diversity and complexity are influenced by the social environment are discussed in detail below.

2.1 Languages are shaped by their social environment

The findings presented in this dissertation show that pressures associated with language usage and with social dynamics influence the formation of languages, and affect the emergence and distribution of different grammatical structures. Specifically, some grammatical constructions may be favored over their competitors in a given community's language because they are better fitted to some of that community's needs and pressures. As such, language typology can be effectively seen as a potential mirror for communities' socio-demographic properties (Gibson

et al., 2019; Lupyan & Dale, 2016; Nettle, 2012). In other words, looking at the structural features of *languages* can shed light on the structural features of *communities*. While there is no doubt that many patterns of language diversity arise by chance without an obvious causal explanation (i.e., drift,), in this work I focused on pinpointing patterns that could be driven and predicted by social properties. This idea implies that at least some cross-linguistic differences could reflect relevant cross-cultural differences, and could potentially explain why certain properties of language evolved in certain communities but not in others.

This idea does not entail a deterministic or comprehensive explanation of languages' origin: human languages clearly evolve (and continue to evolve) in complex landscapes, and are subjected to multiple pressures, external influences, historical events and random changes, which all shape languages in tandem. As such, identifying social factors that can affect languages only provides partial and probabilistic explanation for why languages look the way they do. Nonetheless, the findings presented in this thesis do suggest that at least some of the differences between languages' grammatical structures may be traced back to the social environment in which they evolved, and specifically, to the size of the community. Moreover, community size is not only a relevant factor in explaining typological patterns of language diversity, but it can also be relevant for understanding the process of language evolution in our species. In particular, the findings of Chapter 3 lend support to the idea that a growth in the average size of social groups was one of the drivers for the evolution of modern human languages (Dunbar, 2017).

The results of Chapters 3 and 4 also show that social factors such as community size and network structure can affect languages' vulnerability to stochastic changes, analogous to the concept of genetic drift. Specifically, the results of these chapters suggest that small communities and small-world networks are more severely affected by random events. Similar claims have been previously made in the literature, specifically with respect to community size. For example, Henrich (2004) demonstrated that small populations are more likely to lose cultural practices (e.g., technology) by chance. More related to the issue of language complexity, Nettle (1999; 2012) argued that small communities are more likely to drift and fixate on communicatively suboptimal grammatical strategies, and Trudgill (2005) suggested that small communities are more likely to develop disfavored sizes of phonological inventories (i.e., too big or too small). Importantly, the finding that small-world networks are also more susceptible to drift is in line with these

claims. This is because a small-world network structure is likely to be the structure of real-world small communities. Typical small communities in natural settings are comprised of thousands of individuals (which is still very small compared to the millions of individuals in larger communities). Given their size, such small communities are unlikely to be fully connected. Instead, they are likely to exhibit small-world characteristics, where strangers are indirectly linked by short chains of shared acquaintances, and where one's friends are also likely to be friends with each other. This is in contrast with real-world larger communities, which have been claimed to be scale-free (e.g., comprised of multiple sub-communities with small-world properties that are linked by few highly connected agents). Taken together, these empirical results imply that natural languages spoken in real-world small communities may be more likely to display rarer linguistic properties, such as uncommon word orders (e.g., OSV), rare sounds (e.g., Bilabial trills), or unique morpho-syntactic alignments (e.g., tripartite alignment).

Crucially, even though the results obtained in this dissertation were based on very small groups in comparison to real-world communities, I believe the conclusions scale up to larger scenarios if one takes into account that all relevant aspects of the experiment scale up accordingly. Specifically, the experiments presented in Chapters 2-4 involved a "miniature world": people needed to refer to a small set of meanings that vary only along two semantic dimensions; they were part of a small micro-society with relatively few participants; they interacted with each other for only a several hours; and they never met their partners more than a handful of times. In the real world, everything is scaled-up: people interact about immensely more things that vary along many different semantic dimensions; they are a part of a society that is, even when considered to be small, comprised of at least a few hundreds of individuals (and in big societies, even millions); but they also have years and years to interact with each other, and interact with their peers regularly. In other words, there is proportional scaling of the relevant aspects of the group communication paradigm to real-world scenarios, such as the amount of experience with the language, the population size, and the familiarity with other members of the community. As such, I believe that the conclusions drawn from the findings presented above would generalize to much larger communities in natural settings.

Similarly, even though the results obtained in this dissertation were based on written language (selected for pragmatic reasons), I believe the conclusions hold for the auditory and manual modality. That is,

conducting these experiments with vocalizations and/or gestures instead of written labels should yield similar results. This assumption is based on the fact that iterated learning studies generally yield similar results across modalities: signal systems tend to become more compositional and less iconic over time in the written, spoken and signed modality (Jones et al., 2014; Motamedi et al., 2019; Perlman et al., 2015). While it is possible that holistic signals may be sustained for longer in the gestural modality given its greater affordance for iconicity, there is evidence that compositionality arises nonetheless due to the general existence of compressibility and expressivity pressures (Bohn et al., 2019; Motamedi et al., 2019; Senghas et al., 2004). Importantly, such pressures are shaped by the social environment regardless of modality, and are therefore expected to act on languages in general. Compelling evidence for this is also found in the work on emerging sign languages, which inspired my PhD project to begin with, and showed that larger and sparser communities of signers tended to create languages with less variability and fewer irregulars compared to those languages created by small and tightly knit communities (Meir et al., 2012).

2.2 The relation between language complexity and learnability

The findings of this dissertation also suggest that not all languages are equally complex and equally learnable, at least not in terms of their morphologies. Specifically, different-sized communities were found to develop languages with different degrees of systematic form-to-meaning mapping (Chapter 3), and more regular and systematic languages were found to be acquired faster by adult learners (Chapter 5). Assuming that these results reflect real-world tendencies, they suggest that there are meaningful differences in languages' learnability and complexity, such that some languages can be seen as simpler than others, and can consequently be learned faster and more easily. Importantly, this by no means suggests that some languages are *better* than others (Gil, 2001). All languages are equally good at expressing messages, as reflected by similar communicative success rates across different conditions in Chapters 3-4.

Alternatively, it is possible that all languages are equally complex when taking into account all levels of linguistic analysis, not just morphology. That is, languages may "balance-out" different degrees of complexity across multiple domains such as phonology, word order, and pragmatics (Crystal, 1987; Joseph & Newmeyer, 2012). For example, a common cross-linguistic observation is that languages with extensive case

marking systems tend to have flexible or free word order, showing an efficient trade-off between complex morphology and simple syntax (Gibson et al., 2019; McFadden, 2003). Similarly, languages may have highly elaborate inflectional paradigms, but fairly simple sound systems, or the other way around. Although different languages show different levels of complexity in some domains, their global complexity may be relatively similar. Since there is no agreed metric for formally quantifying language complexity (especially not combined across different levels of linguistic analysis), this idea remains untested. Notably, looking at language learnability as a proxy for language complexity may be a promising venue for assessing linguistic complexity: learning outcomes and language acquisition trajectories can potentially serve as a window into the language's degree of complexity. Relatively slow acquisition rates of linguistic feature X in language Y may indicate high complexity of language Y with respect to feature X.

However, it is important to keep in mind that prior knowledge may affect what is perceived by language learners as complex: it is easier for adult second-language learners to learn a new language if it has similar grammatical structures to their native language (Baptista et al., 2016; Barking, 2016). Nevertheless, examining children's language acquisition trajectories across different languages and across different domains (e.g., vocabulary, phonology, word order, verb inflections, etc.) may yield interesting results. So far, cross-linguistic studies that compared child language acquisition rates have mostly focused on only one feature (e.g., the passive construction, word learning), rendering the question of global complexity unanswered. Yet some work has attempted to relate children's learning difficulty in one linguistic domain to language complexity in another domain. For example, Bleses, Basbøll and Vach (2011) showed the children's acquisition of inflectional past-tense morphology is considerably slower in Danish compared to similar Nordic languages, and that this inferior learning may be due to the phonetic structure of Danish, which makes it difficult to segment words and identify their endings due to heavy reductions. In contrast, Icelandic has a much richer morphology compared to Danish, yet it is more easily acquired by children given that Icelandic phonology makes different suffixes relatively easy to perceive (e.g., using sonority to saliently mark word boundaries). By highlighting the link between language learnability and morphological complexity, our results encourage a more global approach to testing and quantifying cross-linguistic differences, in which multiple linguistic features are compared across multiple languages.

2.3 Input variability as an underlying force

A crucial part of the predictions of this dissertation relied on the postulated relation between input variability and regularization, which was suggested as a possible force behind cross-linguistic differences in language complexity (Nettle, 2012). Specifically, it was suggested that larger and sparser communities tend to have simpler and more regular morphologies because members of such communities are exposed to more variation, which acts against complex and irregular morphological structures in terms of acquisition, emergence, preservation, and use. Following this line of reasoning, I hypothesized that larger and sparser groups would develop more systematic languages as a means of overcoming the increase in input variability and the communicative challenge it entails. In other words, I predicted that compositional variants would be more likely to emerge and more likely to be adopted and spread in larger groups, whereas complex, non-transparent, and/or irregular variants would have more chances to survive in small groups.

This hypothesis was drawn from several different literatures, and relied on two crucial assumptions: **(1)** that members of bigger and sparser communities are exposed to more variation in their linguistic input compared to members of small and dense communities; and **(2)** that exposure to more input variability promotes regularizations. Below I provide detailed evidence for each of these two assumptions. I then lay out my original hypotheses in light of these assumptions and describe how the experimental results obtained in Chapters 2-4 relate to them. Finally, I discuss the role of heterogeneity as a possible source of input variability, and highlight its implications with respect to assumption (1).

2.3.1 Assumption 1: Larger and sparser communities exhibit more input variability

Broadly speaking, the term *input variability* refers to the extent to which available data points differ from each other. This broad definition indicates that input variability in language can arise at multiple different levels of linguistic analysis, from morpho-lexical variability (i.e., different words/morphemes for describing the same meaning) to syntactic variability (i.e., different word orders) to phonetic-acoustic variability (i.e., different pronunciations of sounds). The first assumption behind the idea that cross-linguistic differences are related to differences in social structure is that larger and sparser communities are more likely to exhibit

more variability across all these linguistic levels (e.g., more dialectal variation, more acoustic variation, more lexical variation). In the case of phonetic variation, this assumption is intuitively reasonable: given that different individuals always differ in their pronunciations from one another, and given that the same target phonemes are typically uttered slightly differently every time even when produced by the same individual, it is highly likely that larger communities with more individuals would also feature more phonetic variability. In the case of morpho-lexical and syntactic variability, the prediction is similar (albeit less straight forward), and is motivated by several arguments, which I discuss below.

First, the available linguistic input in small and dense communities is predicted to be more restricted and homogeneous (i.e., less variable) given that individuals in such communities are expected to have more constrained social networks. This is because the pool of potential individuals they can interact with is typically smaller to begin with, and because different members of the community are typically highly familiar with one another (i.e., interact often) and/or are closely related (i.e., even if they don't directly interact, they have multiple shared connections). Specifically, it has been empirically shown that individuals in small communities exhibit greater network closure (i.e., interconnectedness), even when holding the number of direct connections constant: since the pool of potential connections is limited in smaller communities, the connections of every two individuals are likely to overlap (Allcott et al., 2007). In such cases, there is a higher chance that individuals will receive similar information from different connections since different connections are also connected to each other (Liu et al., 2005). As such, members of smaller and denser communities are likely to be exposed to less variable input compared to members of larger and sparser communities, whose connections are less likely to be related.

Moreover, it has been suggested that more innovations are likely to take place in larger communities, so that members of such communities are likely to have access to more diverse input (Fay et al., 2019; Henrich, 2004). The idea is that in big populations, there are more individual models from whom knowledge can be copied and additively combined, leading to more potential solutions (and by chance, to more successful and/or better adapted solutions). As a result, larger communities are argued to display more variable and complex repertories (e.g., for technological tools).

Finally, there are more chances for preserving diversity in larger and sparser communities (Bahlmann, 2014; Derex & Boyd, 2016; Liu et al., 2005). Computational models have shown that the rate of conventionalization is proportional to population size, so that information takes longer to spread when there are more agents in the community (Baronchelli et al., 2006; Gong et al., 2014). Given that global alignment is considerably faster and easier to achieve in a small community with fewer individuals, variations can be more quickly and more efficiently eliminated (i.e., replaced with similar variants that are shared across all community members). As a result, small communities are likely to show less variability overall. Even when the number of agents in the community is kept constant (and therefore, the number of potential variations is kept constant), the spread of information to the entire community is typically reduced in sparser networks compared to dense networks, given that some members of sparse networks rarely or never interact (Fagyal et al., 2010; Gong et al., 2012; Martín et al., 2019; Zubek et al., 2017). Taken together, larger and sparser communities are expected to show less convergence in the same amount of time, and consequently, enable variations to be maintained for longer.

2.3.2 Assumption 2: More input variability promotes regularization

The second assumption behind the idea that cross-linguistic differences are related to differences in social structure is that communication in high variability conditions (i.e., when members of the community are exposed to many different variants) can lead to systematization, which can in turn reduce variability. That is, I hypothesized that greater input variability would be the driving force behind the emergence of systematic linguistic structure and the creation of more compositional languages in larger and sparser groups. Supporting this idea, multiple studies suggest that more input variability plays an important role in learning, generalization, categorization, and pattern detection in both infants and adults (e.g., Bradlow & Bent, 2008; Estes & Burke, 1953; Gómez, 2002; Lev-Ari, 2016, 2018; Lively, Logan, & Pisoni, 1993; Munsinger & Kessen, 1966; Perry, Samuelson, Malloy, & Schiffer, 2010; Rost & McMurray, 2009; Seidl, Onishi, & Cristia, 2014; for a nuanced review, see Van Heugten, Bergmann, & Cristia, 2015). Interestingly, different types of input variability have different effects on language learning and categorization, depending on the linguistic target behavior (e.g., speech perception, word learning, morphology acquisition), the task (e.g., production,

categorization, recall), and learners' prior knowledge (e.g., familiar vs. new categories, L1 vs. L2).

Below I draw on past literature to disentangle the complex relationship between variation, memory, and regularization across different linguistic domains. The main conclusion is that exposure to more input variability is initially taxing for individuals' memory, yet over time can boost long-term performance by favoring the formation of robust and abstract representations. Simply put, it seems that the increased processing costs associated with greater variability eventually benefit learning and categorization by promoting generalizations.

2.3.2.1 Lessons from phonetic variability studies

Language learning studies have shown that exposure to multiple speakers (and consequently, to more phonetic variability) enhances the learning of phonological features: it leads to better speech perception in noise (Lev-Ari, 2018), better discrimination between minimal pairs (Rost & McMurray, 2009, 2010), better learning of phonotactic rules (Seidl et al., 2014), better adaption to foreign accented speech (Bradlow & Bent, 2008), better perceptual categorization of regional dialects (Clopper & Pisoni, 2004), and better identification of non-native phonetic contrasts, including better ability to generalize these contrasts to unfamiliar speakers (Lively et al., 1993). Importantly, exposure to different tokens coming from the same speaker also leads to better adaptation to non-native speech (Sumner, 2011) and to better discrimination between minimal pairs (Galle et al., 2015). Together, these results highlight the benefits of learning from acoustically variable input (i.e., different pronunciations): what seems to improve learning in the studies reported above is the fact that learners were exposed to high variability in the phonetic realizations of sounds, rather than to multiple speakers per se (but see Lively et al., 1993, for evidence that generalizing to new speakers does require across-speaker variation). Other studies have shown that phonetic variability can facilitate higher-order aspects of language learning, leading to better L2 vocabulary learning in adults (Barcroft & Sommers, 2005; Sinkeviciute et al., 2019; Sommers & Barcroft, 2007), better novel word production in children (Richtsmeier et al., 2009), and better word segmentation in infants (Estes & Lew-Williams, 2015).

In contrast, some studies report that speaker input variability does not affect infants' acquisition of native vowel categories (Bergmann &

Cristia, 2018) or children's and adults' ability to discriminate between non-native vowel contrasts (Giannakopoulou et al., 2017), and that it can even hinder sound categorization (Bergmann et al., 2016) and phonetic classification (Green et al., 1997). Phonetic variability has also been shown to negatively affect adults' speech processing and perception of their native language in tasks such as word recognition, word recall, and word naming (Martin et al., 1989; Mullennix et al., 1989; Sommers & Barcroft, 2006). Such findings are taken to show that talker-specific acoustic variation can also mask linguistically relevant information, and make learning and processing overall more difficult and taxing for memory (Rost & McMurray, 2010; Van Heugten et al., 2015).

How come phonetic variability benefits some aspects of speech perception and learning, but hinders others? One way of reconciling these findings was offered by Lev-Ari (2018), who suggested that the effect of input variability depends on (a) whether the variation occurs in relevant or irrelevant features of the target behavior, and (b) whether the target behavior is familiar or new (Lev-Ari, 2018). Using computational simulations, Lev-Ari (2018) showed that input variability along relevant phonetic features of vowel categories improved performance for established phonological categories, but not when learners were still in the process of learning new categories. These results suggest that input variability benefits adaptations and category robustness for known behaviors, but that its positive effect may be absent (or even reversed) at the earliest stages of learning. When learners are in early stages of acquiring a target behavior, high input variability along relevant dimensions can actually make learning more difficult, as it makes it harder for learners to figure out how many categories there are and how these categories differ from one another.

This account can help explain why certain aspects of speech perception benefit from phonetic variability while others do not, and helps reconcile conflicting results obtained from different age groups and from first vs. second language learning. Specifically, it suggests that varying critical aspects of the input may hinder learning at first, but have benefits later on. For example, infants were shown to successfully differentiate between similar sounding words that differed in the voicing of one phoneme (i.e., minimal pairs such as *buk* and *puk*) only when exposed to variability in *irrelevant* aspects of the words' pronunciation, such a prosody and vowel quality, which actually do not help differentiate between these words (Rost & McMurray, 2010). When there was variation along aspects of pronunciations that were directly relevant for differentiating between

voiced and unvoiced consonants (e.g., voice onset time), infants failed to discriminate between the words. These results showed that when infants are still in the process of establishing categorical distinctions based on voice onset time, variation along this relevant feature hinders the learning of these relevant categories. In contrast, variability of irrelevant phonetic features, which has typically been thought of as noise, can boost learning by indicating which aspects of the input are, in fact, critical to the target behavior, and which are not. That is, high variation in feature X signals to learners that feature X is not directly relevant (and can therefore potentially be ignored). At the same time, high variation in feature X can highlight the existence of other features that, in contrast to feature X, exhibit little to no variance (signaling to learners that these features may be crucial). As such, exposure to variation along irrelevant dimensions can help infants' formation of robust and generalized representations that include only phonetically relevant cues while excluding irrelevant ones. On the other hand, other studies have shown that in later stages of life, in which learners already have well-established phonemic categories, variability along *relevant* phonetic aspects can also facilitate learning: Sumner (2011) reported that adults listening to non-native accented speech showed better adaptation when the words varied along the relevant dimension (i.e., voice onset time). That is, the same type of variability that hindered infants' learning in Rost & McMurray's study was nevertheless beneficial for proficient language users, who only needed to tune their existing knowledge in order to successfully comprehend an unfamiliar non-native speaker. The account proposed by Lev-Ari (2018) offers a unified explanation for these results by suggesting that the advantage of variation along relevant vs. irrelevant dimensions is modulated by the stage of learners' language acquisition.

2.3.2.2 Can phonetic variability explain cross-linguistic differences?

Atkinson and colleagues asked whether phonetic variability can explain why larger communities tend to have simple morphological structure (Atkinson et al., 2015). In an artificial language learning task, Atkinson et al. (2015) tested whether learning a morphologically complex artificial language was harder for participants when exposed to greater phonetic variability, i.e., when learning from three speakers as opposed to only one. In that study, adult participants needed to learn a miniature language with three case markers that changed their form depending on vowel harmony, following exposure to the exact same sentences produced by single or

multiple speakers. Their working hypothesis was that speaker input variability would hinder learning, and would therefore lead to more learners producing more simplified versions on their input. Results showed no significant effect of phonetic input variability on participants' production accuracy – speakers learned the language equally well when exposed to one or many speakers. While it is not possible to draw strong conclusions from such null results, this study suggests that talker-specific acoustic variability may not affect morphology acquisition. Atkinson et al. (2015) therefore concluded that phonetic input variability is unlikely to be an explanatory mechanism for how group size determines languages' morphological complexity.

Importantly, the findings and conclusions of Atkinson et al. (2015) are in line with the account presented above, given that the study manipulated input variability along *irrelevant* features of the target behavior to be learned (i.e., different pronunciations of the same morphological inflections). As such, the fact that phonetic variability did not affect learning is unsurprising. Furthermore, Atkinson et al. (2015) reported that exposure to multiple speakers did not benefit adults' speech segmentation in a classic statistical learning task. Given that infants do show positive effects of phonetic variation in a similar paradigm (Estes & Lew-Williams, 2015), these results are again in line with the idea that input variability can have different effects in different stages of language learning (Lev-Ari, 2018).

Based on the literature on phonetic variation, greater input variability is only expected to hinder the learning of new categories (and consequently, lead to more simplifications, see Assumption (2)) when variation is along *relevant* features (i.e., morpho-lexical variation rather than phonetic variation). Accordingly, in the current dissertation I argued that the relevant force for understanding group size effects on linguistic structure relies on input variability that is directly relevant to the target behavior i.e., to *morpho-lexical variability*. In fact, the group communication paradigm employed in this dissertation did not include phonetic speaker variability at all, given that participants interacted by typing labels using uniform keyboards. Instead, participants in the group communication paradigm were exposed to variability in the actual form of the labels. Importantly, I hypothesized that *this* type of variability would be more taxing for participants' memory, and would create an increased pressure for the creation and adoption of more systematic and generalizable variants.

2.3.2.3 Morpho-lexical variability and the generalization of grammatical patterns

Several studies have shown that *morpho-lexical* variability (i.e., variation in the actual word form, such as different labels and/or different morphemes) can affect the learning of grammatical patterns and lead to better generalizations. For example, in an artificial language learning task, Gómez (2002) showed that nonadjacent dependencies in three-element strings (e.g., *pel kicey jic,* where *pel* and *jic* are depended) were learned better by both infants and adults when they were exposed to more varied exemplars, i.e., when the middle element of the string (e.g., *kicey*) was drawn from a larger pool of different words. Crucially, when the middle element did not vary, participants were unable to learn the grammatical dependencies in the artificial language. Gómez (2002) reasoned that the positive effect of morpho-lexical variability on grammar learning and generalization was the result of memory constraints: in the presence of high variability along relevant features of the language[13], memorizing each three-element string and storing it separately was hard, and this difficulty encouraged learners to generalize over items and detect the underlying pattern. In other words, the results of this study suggest that generalization over variants occurs only when there are too many variants to remember, but not when learners are able to memorize them all individually.

The account described above suggests that although high variation is harder to process, it can have positive effects on learning in the long run. This idea is in line with exemplar-based frameworks and associative learning theories, where all information from encountered examples (including both relevant and irrelevant aspects) is stored in memory in early stages of learning, and is later used to form robust representations that allow for abstractions and generalizations (Apfelbaum & McMurray, 2011; Barcroft & Sommers, 2005; Kleinschmidt & Jaeger, 2015; Posner & Keele, 1968). This idea also resonates with the concept of *desirable difficulties* (Bjork & Bjork, 2011; Bjork, 1994), which suggests that varying the input and the learning conditions may impair immediate performance, but subsequently triggers better encoding and retrieval

[13] Although variation in the middle word of the string could be mistaken for an irrelevant aspect of the artificial language, it is in fact crucial for understanding the underlying rule, i.e., that the dependency in the string is non adjacent rather than adjacent.

processes that support retention and generalization. In other words, while high variation is more costly for processing and more taxing for memory, it also enhance long-term performance. In support of these theories, studies have shown that exposure to multiple exemplars (i.e., to more variable input) helps the formation of categories and is crucial for generalizing these categories to novel input. For example, infants who were taught labels for different object categories were able to generalize those labels to novel objects only when exposed to variable exemplars of each category (Perry et al., 2010). Similarly, infants who were taught arbitrary animal-sound pairings were shown to acquire these pairing and generalize them to novel items only when they were familiarized with multiple exemplars of each category (Vukatana et al., 2015).

Complementary findings on the relation between variability and generalizations have been found in the literature on regularizing unpredictable variation, which showed that people tend to introduce more regularities to the language when faced with morpho-lexical variation that is taxing for memory (Fehér et al., 2016; Ferdinand et al., 2019; Hudson Kam & Newport, 2009; Samara et al., 2017). In particular, multiple studies have shown that when children and adults are exposed to an artificial language containing an inconsistent, difficult-to-learn grammatical pattern (e.g., variable plural marking of nouns, variable labels for the same stimuli, variable use of different determiners for marking gender, etc.), they tend to regularize their input in various ways, for example, by systematically producing only one variation, by making the variation lexically conditioned (and therefore predictable), or by eliminating the inconsistent variation altogether (Fehér et al., 2014; Hudson Kam & Newport, 2005, 2009; Samara et al., 2017; Wonnacott & Newport, 2005). Together, these studies suggest that high morpho-lexical variation in the input hinders the acquisition of morpho-lexical target behaviors, but promotes more regularizations in learners' output. Moreover, increasing the memory load in such paradigms leads to more regularizations: learners regularize more when they are given less exposure (Samara et al., 2017), and when they are required to learn more items (Ferdinand et al., 2019). Similarly, learners tend to regularize less when retrieval is facilitated and memory load is reduced (Hudson Kam & Chang, 2009). In line with the argument presented in the previous paragraph, this line of work demonstrates that exposure to morpho-lexical input variability increases processing costs, but consequently results in the creation of more systematic linguistic patterns.

2.3.3 Summary of my hypotheses and relevant results

Based on the evidence presented above, I predicted that exposure to more morpho-lexical input variability would promote the emergence of structured languages. Considering that participants in a given group needed to create their own languages in order to communicate with each other and had no established or shared language to rely on, they were likely to be faced with multiple variations of words/morphemes created by different participants in the group for the same stimuli. For example, participants were likely to encounter different labels and morphemes for describing the same meanings (e.g., *pok* vs. *muif*), and/or different variations of existing words (e.g., spelling differences such as *muif* vs. *mif*). Since their goal was to interact with each other using these labels/morphemes, I hypothesized that such variability would impose a communicative challenge that group members would need to overcome in order to successfully communicate with each other. In other words, since high morpho-lexical variability should be taxing for participants' memory, it should favor the creation and adoption of more systematic and generalizable variants, which are assumed to be easier to remember (an assumption directly confirmed in Chapter 5). In turn, I hypothesized that the creation of systematic variants would relieve participants' memory load, facilitate successful interaction and aid the process of convergence, effectively reducing input variability.

The results of Chapter 2 supported this prediction, and showed that structured, compositional languages can emerge within a single generation when there is a pressure for developing systematicity, i.e., when people needed to refer to an expanding meaning space in which there are more and more meanings to communicate about, and when the interaction included multiple people in a group setting (in contrast to interaction with just one other person in a dyadic scenario, as in Kirby et al. 2015). Together, the results of Chapter 2 showed that more variable communicative contexts can indeed boost the emergence of systematic and productive languages.

Next, I hypothesized that small and tightly knit groups would exhibit less morpho-lexical input variability compared to larger and sparser groups, at least in early stages of language formation. Consequently, I hypothesized that small and tightly knit groups would be under weaker pressure to develop systematic languages. Specifically, I reasoned that in the early stages of language formation, community members may employ two potential strategies in order to successfully

interact with each other: memorize each other's unique morpho-lexical variants, or try to align on a shared language. Importantly, the efficacy of these strategies and the ease with which they can be employed should differ depending on group size and network structure. Members of small and dense groups may be better able to remember each other's unique variants thanks to generally lower variability (i.e., having fewer and more similar variants to remember overall). As such, they may be better able to cope with the existing variability, leading to a reduced pressure to develop a shared language. In other words, convergence on similar labels is potentially less needed in a small community. In contrast, such a strategy would be much harder for members of larger and sparser groups, where there should be many more different variants to keep track of. Since members of larger and sparser groups should be faced with a greater memory load and a greater communicative challenge, they should therefore be more likely to develop and favor more transparent and simplified variants that can facilitate memory and aid convergence. This idea was motivated by evidence from emerging sign languages, which show that there is less convergence and more lexical variability in a small community of signers, while a larger community of signers was more uniform (Meir et al., 2012). This observed pattern has been attributed to a weaker pressure to conventionalize in the smaller community, given that members of that community are highly familiar with each other and can therefore maintain a surprisingly large amount of lexical variability. In contrast, members of the larger community were not able to cope with the high degree of variability, and were more prone to develop shared linguistic regularities. In sum, I hypothesized that the increased morpho-lexical variability postulated to occur in larger and sparser communities would result in a stronger pressure for conventionalization and systemization in such communities, eventually leading to the creation of more structured languages.

The results of Chapter 3 confirmed this hypothesis in several ways. First, a direct examination of the levels of morpho-lexical input variability across conditions and time confirmed assumption (1) with respect to population size, and showed that members of larger groups were indeed faced with more variability. Second, higher levels of morpho-lexical variability were shown to induce a greater increase in structure over time, confirming the rationale behind assumption (2). Importantly, although linguistic structure increased in both conditions, this increase was faster in larger groups. As such, languages created by members of larger groups were indeed more systematic by the end of the experiment. This result

confirmed the main hypothesis, i.e., that larger communities would develop more regular languages, and supported the claim that this difference could be traced back to differences in input variability. Third, the results of Chapter 3 showed that input variability decreased over time, and that this decrease was faster in larger groups. This was presumably because members of larger groups developed more structured languages over the course of communication, which enabled them to align on a shared lexicon (reducing variability) more easily. Specifically, our results showed that, on average, larger groups reached similar levels of convergence as small groups, despite convergence being much harder to achieve in larger groups. That is, developing more structured languages indeed allowed larger groups to overcome their greater communicative challenge and greater input variability. This advantage was further supported using an analysis showing that more linguistic structure indeed predicted better convergence. Finally, small groups in Chapter 3 varied more in behavior – while some small groups showed high levels of convergence and high levels of structure (in line with the findings from Chapter 2, i.e., that groups of four participants *can* develop compositionality), other groups did not – some small groups showed much less alignment and did not develop linguistic structure. This result is in line with the prediction that members of small groups may be able to memorize each other's unique variants, and as such are under a weaker pressure to establish a shared lexicon and to favor more systematic variants. Such results are also in line with observations from emerging sign languages, which suggest that smaller, tightly-knit communities are less converged (Meir et al., 2012). Taken together, the results of Chapter 3 suggest that an increase in morpho-lexical input variability is one of the driving forces behind group size effects, and can therefore serve as a possible explanation for cross-linguistic differences in language complexity.

Notably, I had very similar predictions for the effect of network structure: I hypothesized that sparser networks would develop more structured languages, and that this difference would be the result of more input variability in such networks. However, neither prediction was borne out: the results of Chapter 4 showed that dense and sparser networks reached similar levels of input variability throughout the experiment, as well as similar levels of linguistic structure and convergence. I reasoned that these results can be explained in light of the relationship between input variability and systematicity, that is, that the lack of differences between networks' levels of linguistic structure in Chapter 4 can be traced

back to the lack of differences in morpho-lexical variability across network conditions.

One possible explanation for the lack of input variability differences across network conditions in Chapter 4 is that in the current design, members of the sparser small-world and scale-free networks interacted with *fewer* individuals compared to members of the fully-connected network, i.e., their personal social network size was smaller. Although the size of the community was the same across conditions (N=8), members of fully connected networks interacted with all seven other members of their group, while members of sparser networks interacted with only two to four other people. Importantly, the existence of sub-groups (i.e., few members who frequently interacted with each other) in sparser networks may have facilitated convergence within these smaller sub-groups, leading to less global (i.e., communal) variability in sparser networks than originally expected. That is, the existence of small and dense social networks within the sparser networks could have eliminated the difference between network conditions. In other words, although the global level of input variability was supposed to be higher in the sparser conditions, it is possible that the local (i.e., individual) level of input variability experienced by single members in sparser networks in early rounds was, in fact, reduced – leading to more convergence and less variability in the entire network. Even though the measure of input variability used in Chapters 3 and 4 was a global measure (i.e., quantifying how much variation is present in the entire network), it is theoretically possible to measure the changes in the local levels of input variability experienced by individuals across different networks over time by quantifying the degree of variability present only in their personal social network (i.e., only in the limited circle of people they were directly connected to). If this hypothesis holds, local variability would differ from global variability in early communication rounds. It is possible to measure convergence within different sub-groups of connected participants. If convergence in local sub-groups was indeed faster/better, this would explain why sparser networks did not show more input variability, and could potentially account for the null results obtained in Chapter 4. Although it was not possible for me to test these hypotheses before submitting this dissertation, I hope to carry out these additional analyses in the future.

2.3.4 Heterogeneity: The source of input variability

Importantly, the speculations discussed above raise an additional issue: what are the sources of input variability? In the current dissertation, more input variability was the result of experiencing more morpho-lexical variants due to being in larger or sparser communities, and due to having less chances to converge with all members of the community (i.e., sustaining more diversity for longer). The idea that more input variability can be equated with an increase in population size (or in the number of individuals people can learn from and/or interact with) is evident in many psycholinguistic and cultural evolution studies (e.g., Caldwell & Millen, 2010; Kempe & Mesoudi, 2014; Fay, Kleine, Walker & Caldwell, 2019). In many studies, group size differences are simulated by manipulating the number of models people are exposed to, which is treated as a proxy for the amount of variability in the input. While this is a reasonable assumption (as I explain above, and as the results from Chapter 3 confirm), it is not always the case that more people per se entail more variable input. Although larger communities indeed entail a larger pool of phonetic variability due to the mere existence of more individuals (and the fact that variation in pronunciation is always present, even within a single individual), they do not necessarily entail more morpho-lexical variability. A big community could be comprised of hundreds of thousands of people speaking the exact same dialect (i.e., lots of phonetic variability, but no morpho-lexical variability), whereas a much smaller community could be comprised of a few dozens of individuals each speaking a completely different dialect.

Crucially, the main definition of input variability relies on *heterogeneity* (Nettle, 2012). That is, the degree of variability in the community cannot simply be equated with population size: while large communities indeed tend to be more heterogeneous, this is not always true. In other words, one can argue that the group size effects found in Chapter 3 were not driven directly by the number of individuals in the group, but rather by the fact that there is typically more heterogeneity in larger groups. Indeed, computational simulations have shown that the effects of input variability on speech perception and lexical production are affected by the degree of heterogeneity in the population rather than simply by population size (Lev-Ari, 2018; Lev-Ari & Shao, 2017). This implies that increasing the degree of heterogeneity in a given group while keeping population size constant would still yield similar results to Chapter 3, namely, that more heterogeneous groups would also create more systematic languages. As such, the number of individuals in the

community is only one possible source for differences in morpho-lexical input variability, and it is important to distinguish between this frequent implementation of variability and other possible causes for variability.

Heterogeneity can stem from many social and psychical factors beyond population size, such as the individuals' age, gender, native language, cultural background, etc. Each of these factors can potentially influence speakers' lexical choices, and therefore affect their variants and how different they are from those of other speakers. For example, age heterogeneity in the community can affect its members' linguistic behaviors: Lev-Ari & Shao (2017) found that interacting with people from a wider age range improves individuals' lexical prediction, lexical access, and lexical use. Similarly, it is possible that variability in cultural backgrounds can affect morphological simplifications, such that more cultural diversity would create a stronger pressure for developing systematic and regular grammars. This idea is in line with claims that the languages with the simplest morphology in the world are creole languages, which are formed as the result of extensive language contact between individuals speaking different native languages (McWhorter, 2001; Parkvall, 2008). I discuss these possibilities (and others) in detail in the next section.

3 Methodological challenges and future directions

In the process of piloting the group communication paradigm and collecting data for the experimental chapters presented in this dissertation, I have tested over 380 participants: over 100 individuals and over 45 different groups of various sizes. This was a rewarding effort given the data I gathered and the conclusions that were eventually drawn from it, yet it was also highly demanding and challenging to implement. In the upcoming section, I lay out some of the general methodological challenges I faced throughout my PhD. I also present in detail an experiment aimed at testing the role of non-native speakers in the community, which was originally planned as a part of this dissertation, but could not have been completed within the time limits of my PhD due to theoretical issues. Finally, given that the paradigm introduced in this dissertation opens the door to many more exciting studies, I make several suggestions for future work.

3.1 General methodological challenges

When embarking on this project in early 2016, my goal was to directly test the evolution of languages in a community-like setting. Specifically, I wanted to bring groups of multiple participants to the lab and mimic communication in social (yet controlled) environments. The idea was to go beyond previous language evolution experiments, which typically included only chains of single individuals (e.g., Carr, Smith, Cornish, & Kirby, 2016; Kirby et al., 2008; Verhoef, 2012), or pairs of interacting participants (e.g., Eryilmaz & Little, 2016; Galantucci, 2005; Kirby et al., 2015; Roberts & Galantucci, 2012; Winters, Kirby, & Smith, 2015). Even in experiments that looked at dyadic communication, participants were often seated in separated rooms or in completely different sites, and communicated via a computer interface – never experiencing the presence of another person. At that point in time, very few language evolution studies included face-to-face interaction (e.g., Christensen, Fusaroli, & Tylén, 2016; Perlman, Dale, & Lupyan, 2015; Tan & Fay, 2011), and very few studies included interaction between more than two people (Atkinson, 2016; Caldwell & Smith, 2012; Fay et al., 2010; Galantucci et al., 2012). To extend on previous work, the paradigm needed to include both face-to-face interaction and multiple participants. Testing multiple participants was desired since it would allow us to directly probe communicative pressures involved in social interaction and communal settings, and to manipulate specific features such as group size and network connectivity. Face-to-face interaction was desired since it was a naturalistic feature that allowed paired participants to make use of non-verbal signs such as facial expressions to determine how certain/confident their partner was when typing a label or making a guess, and allowed all participants to truly feel as if they are a part of a group.

The first challenge I encountered was in the choice of stimuli. Specifically, I wanted to create a semi-structured meaning space which lent itself to compositionality, but did not impose it. That is, the meaning space itself needed to be structured to some extent to begin with in order to promote and motivate the creation of structured symbols. Having a semi-structured meaning space was therefore an important feature of the design of this study, as of other iterated learning studies: it is meant to simulate a fundamental property of our environment, in which elements in the real world repeat in many different combinations and in various contexts (e.g., *I eat, I eat cake, the cake is big, the tower is big, I eat the big cake on the tower*, etc). This fact about the world is what gives rise to languages' productivity to begin with, and allows for the reuse of

linguistic elements over multiple different interactions. As individuals are exposed to more and more input, they are able to detect such repeating elements, which in turn could promote the development of more productive and predictable labeling over time. In other words, the emergence of compositional structure in natural and artificial languages is motivated by the fact that the environment itself is partly compositional. Without this feature, each event would *require* unique labeling. Even in this extreme scenario, humans' tendency to form categories in the absence of a-priori structure suggests that linguistic structure could emerge nonetheless, perhaps as a means to relieve the memory load required in memorizing a holistic language. Indeed, languages frequently categorize continuous and unstructured meaning spaces into categories (e.g., the color space). In order to promote compositionality while not imposing a specific categorization, I chose to create a meaning space with three dimensions: one categorical (shape), one continuous (motion), and one abstract and unstructured (fill pattern). This meaning space was selected after a pilot version that included items that varied in size instead of motion. The choice of size as a semantic feature what not ideal given that people are much more likely to categorize novel items based on their shape, and much less likely to do so based on size. I therefore chose to replace size with motion, which is a more salient feature and much more likely to be encoded linguistically.

Importantly, the selected meaning space was only partly structured: the dimension of motion was continuous rather than categorical, and all scenes had a unique idiosyncratic feature that allowed for the individualization of scenes based on their texture. Because motion was a continuous and unstructured feature, participants were not obligated to categorize it in any particular way, and indeed, different groups categorized the motion space differently. Moreover, participants could differ in the way they categorized what was "new" and what was "recombination" with respect to motion. For example, if a participant already had a label for the directions ↖ and ↗, and now saw a known shape moving ↑ for the first time, they could either decide that it is a combination of ↖ and ↗ (because it was in the middle), or think of it as a completely new direction and assign it a different label. As such, compositionality could arise in different ways across different groups. In addition, each scene had a unique fill pattern. Participants could have chosen to focused on fill pattern and use unique, holistic labels to describe each scene (which in the beginning of the experiment, many of them did), and could have theoretically also categorized these patterns into specific groups (e.g.,

darker vs. lighter), despite the fact that the pattern had no objective categorization to begin with. As such, while the meaning space lent itself to predictable, compositional structure, this structure was not deterministic. In order to ensure that my findings hold for different meaning spaces and are not restricted to one set of categories, I created three versions of the stimuli that were balanced across experimental conditions (e.g., big group vs. small groups). These three versions differed in their distribution of shapes and angles, and required making different differentiations. As a result, groups differed in how fine-grained their categories were, and often developed categories only for semantic features that were directly contrasted. This point strengthens the idea that the communicative context in which languages emerge shape their final structure. However, while the specific structure of the meaning space could have influenced participants' categorizations, the effect of group size was significant across all versions. That is, big groups created more structured languages above and beyond the specific version of the stimuli, suggesting that the results of Chapters 2-4 are generalizable despite exposure to slightly different meaning spaces.

The second challenge I encountered was deciding on the best starting conditions for the paradigm (i.e., Round 0). I wanted participants to have some common ground to start with, such that they would have a few shared labels for the first eight scenes. However, I was not sure what would be the right way to create this common ground. Specifically, should participants first be trained on a random seed language, or should they start by creating their own labels? When I began my PhD project, all iterated language learning studies showing the emergence of compositional structure with human participants used random and holistic labels as the seed language (i.e., participants were first trained on a given set of computer-generated labels, with no structure). In contrast, similar studies looking at gestures and drawings have allowed participants to create their own initial labels at the start of the experiment. I was not sure which of these starting conditions is preferable, and decided to run a pilot study to examine these two starting conditions. On the one hand, starting the communication game by first training all participants on a carefully controlled input language would provide a comparable starting point for all groups, and would ensure that all groups started out without linguistic structure. In addition, it would prevent a scenario in which participants would be unable to inhibit their use of Dutch language, or unwilling to actively produce labels. On the other hand, I had reasons to believe that allowing participants to create their own initial labels would be

advantageous for the current communication paradigm. First, the goal of the paradigm was for participants to innovate and change the language as the experiment progressed, and I suspected that learning a seed language would bias them to memorize and use specific pre-given labels, assuming those were "correct". Second, I believed that computer-generated seed languages were not a natural starting point, given that real proto-languages were created by people. As such, I predicted that allowing participants to come up with their own initial labels would be a much more realistic scenario, and would potentially result in labels that sound more natural (e.g., in terms of their shared cognitive biases, Dutch phonotactics, etc.). Third, I worried that providing an input language would impose an additional learning challenge at the start of the experiment, and predicted that participants would remember self-generated labels better than computer-generated labels due to a better representation of their own made-up labels in memory. Indeed, the pilot study confirmed that participants in the computer-generated language condition had more difficulties in remembering the initial labels compared to those in the self-generated language condition. Moreover, communicative success was already higher in the self-generating condition in the first round, and participants reported enjoying the game rather than struggling. Importantly, none of the self-generated initial languages was structured, providing a similar starting point for all groups. As such, I preferred to implement Round 0 in this format, and allow participants to create their own languages at the beginning. As such, Round 0 was considered the first round for all analyses but communicative success, as there was no communication in that round, only generating labels followed by passive exposure to these labels.

Finally, brining multiple participants to the lab simultaneously proved to be challenging. It required much coordination (and luck), and often resulted in last-minute cancellations and rescheduling of groups. It was quite problematic to find four or eight different participants who were all available at exactly the same time and for a relatively long duration (i.e., at least 4.5 hours for the experiments in Chapters 3 and 4), and it was often the case that allocated time slots were only partly full. Moreover, since the experiment depended on all individuals being present on time, even when enough people signed up it in advance, it was never clear whether they would actually show up and if the experiment would take place. Although I sent multiple confirmation and reminder emails, and although I tried to have at least one backup participant at every time slot, I still encountered multiple situations where groups had to be cancelled in the

last minute, often when several of the participants were already present and waiting for the experiment to start. Besides causing inconvenience (and at times, frustration) to participants and myself, data collection also took much longer than expected. It was also quite costly in terms of participant payment, room setup, and equipment. Moreover, the experimental setup required the presence of an experimenter (myself) in the room at all times, who needed to closely monitor participants' behavior in order to ensure that they are not talking, pointing, or gesturing, and that their languages do not include Dutch or English words. This was also physically and mentally demanding, especially when testing multiple groups on consecutive days. Nevertheless, I believe this was all worth the trouble given the quality of the collected data and the increased ecological validity of the paradigm compared to previous studies.

One possible way of overcoming some of the general challenges raised above is to turn to online experimental settings. The main advantages of shifting to an online paradigm lay in the ability to recruit many more participants from various locations, and test them simultaneously without the need to bring them physically to the lab. The availability of individuals may be higher when tested from home, the chances of no-shows may be reduced, and the testing costs may be relieved (e.g., participants would be using their own computers instead of laptops provided by the institute). Moreover, it would allow data collection from much larger groups (e.g., 20 or 50 participants). In fact, the technical group at the Max Planck Institute had developed an online pilot version of the experiment used in Chapters 3 and 4, which could be activated on remote computers by entering a link. However, I ended up not using this online setup due to several crucial disadvantages. First, I couldn't be sure that participants would indeed be focused on the experiment, and actually follow the instructions. For example, participants tested remotely may be engaged in other activities during the experiment, and may write down the words they encountered while communicating with different people, eliminating the crucial memory constraint relevant for creating a systematicity pressure. Second, participants may be more likely to stop the experiment in the middle as they get distracted or engaged with other things (e.g., go make a cup of coffee and never come back). Finally, and most importantly, the experiment would lose its naturalistic feature of face-to-face interaction, which was considered to be one of the strengths of the current paradigm. Such disadvantages may be inevitable if researchers are interested in testing larger groups. Nevertheless, some issues could be potentially addressed by using designated crowdsourcing websites such as mTurk or

CrowdFlower, which allow filtering of participants based on criteria like drop-out rates.

3.2 Planned experiment: The role of non-native speakers

The classic contrast between esoteric and exoteric communities relies on three main features: community size, network structure, and the degree of language contact. The latter feature, also thought of as the proportion of non-native speakers or of adult second-language (L2) learners in the community, is considered to be a crucial parameter shaping cross-linguistic structural differences, and is argued to be the main driver of morphological simplification (Bentz et al., 2015; Bentz & Winter, 2013; Dale & Lupyan, 2012; Lupyan & Dale, 2010; Nettle, 2012; Trudgill, 1992, 2002, 2008, 2009). In fact, the inverse correlation found in typological comparisons between linguistic complexity and group size is often attributed to a larger proportion of non-native speakers in larger communities, with group size treated as a mere proxy for the proportion of adult L2 learners. The argument goes as follows: given adults' greater difficulty in learning a second language (and especially a morphologically-complex language), the learnability pressure on exoteric languages increases since there is a high proportion of them in the community. This can lead exoteric languages to lose complex and/or irregular morphological systems over time, and result in more simplified and systematic languages. This process it attributed to the fact that native speakers' tend to accommodate to non-native speech even when it includes mistakes (Atkinson et al., 2018; Loy & Smith, 2019), and to the idea that children of non-native speakers (and children in their close surrounding) will encounter the non-native variants and therefore acquire a simplified variation of the language (Lupyan & Dale, 2015; Nettle, 2012). In contrast, the main language learners in esoteric communities are children, who are presumably not biased against complexity, and may even benefit from having more redundant linguistic cues in the process of language acquisition (Dale & Lupyan, 2012). Consequently, esoteric languages have higher chances of sustaining elaborate inflectional systems over time, and may be more likely to develop rich and non-transparent structures.

My original PhD research proposal included an experiment that directly tests this assumption, and aimed at isolating the role of non-native speakers and the mechanisms that lead to linguistic systematization over time. The experiment was carefully planned and piloted with several

participants, yet was not feasible to complete. In the following paragraphs, I explain the rationale and design of this experiment in detail, present my predictions for it, and discuss the methodological issues that prevented the execution of this experiment, which are closely related to the findings of Chapter 5.

The goal of the experiment was to examine the influence of L1 vs. L2 learners on the systematization of linguistic structure using a replacement paradigm (Caldwell & Smith, 2012). Specifically, I wanted to test whether introducing new learners to the community causes linguistic simplification in an established (yet partly-irregular) language, and test whether this process is affected by the identity and prior knowledge of these learners (L2 vs. naïve participants). In brief, the goal of the experiment was to compare the languages of different four-person groups after replacing two existing members with two newcomers in two different conditions (see Figure 1). Condition 1 (esoteric community) planned to introduce two naïve learners with no prior knowledge of the language and/or the meaning space used in the experiment. Condition 2 (exoteric community) planned to introduce two experienced learners who were already familiar with the meaning space used in the experiment, yet trained on another language with a different grammar and a different lexicon.

At the beginning of the experiment, a group of four participants would be trained on an input language, which they should learn perfectly (Figure 1A). The plan was to have two possible input languages that would be inspired by languages created by real participants in Chapters 3 and 4, and that will be matched on their initial structure score and their expected final structure score after all possible simplifications. Both languages would be semi-structured and contain some irregularities (making them less predictable and allowing for systemization to occur), but would differ in their labels, irregularities and the relevant semantic dimensions for categorization. For example, the grammar of one language would be based on shape and all possible ranges of motion (up-down, left-right, and their combination), while the other grammar will be based on shape, texture (i.e., stripes, bubbles or empty), and a simple up-down motion.

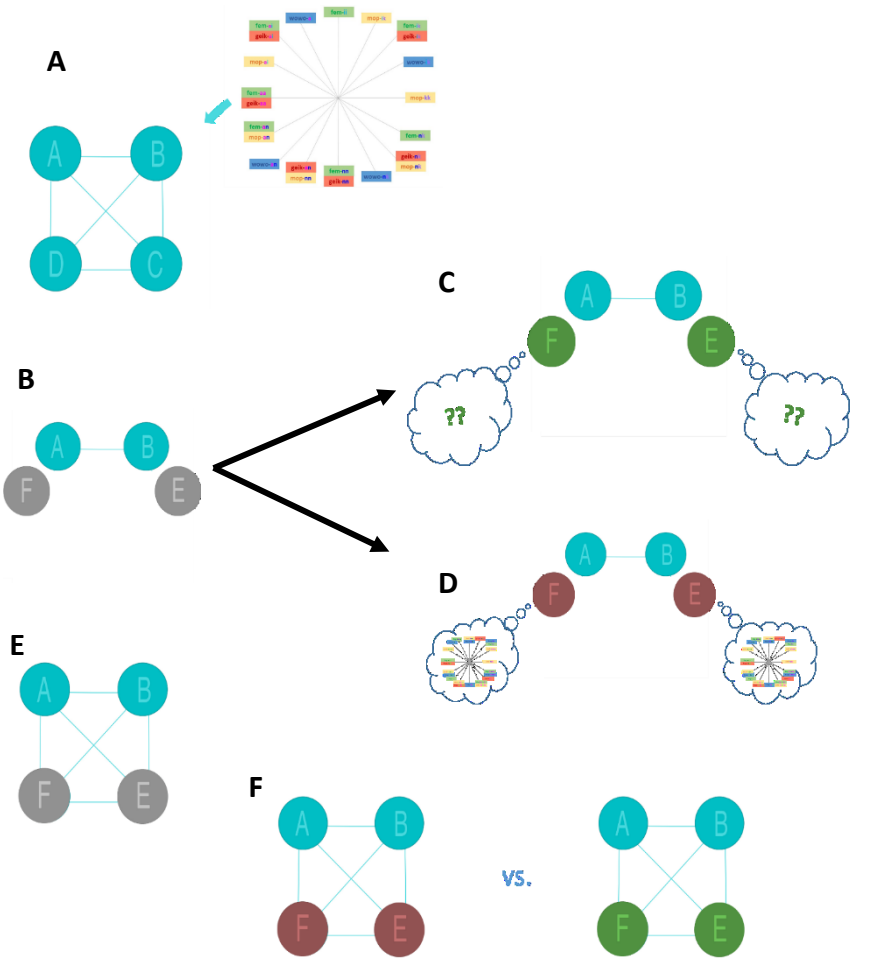


Figure 1. The design of the planned experiment. In the beginning, groups of four participants are trained on an input language and use it to interact with each other for several rounds (A). Then, two members are removed from the group and replaced with two newcomers (B), who need to learn the input language by observing the remaining two participants interact with each other for several rounds. We manipulate whether the newcomers are naïve learners with no prior linguist model (C), or experienced learners who have already acquired a different input language (D). After learning from observation, the newcomers are integrated in the group (E) and need to communicate with all other members for several rounds. Finally, we compare the final languages used across both conditions (F).

Participants would be trained on the languages in several blocks, using an identical procedure to the one used in the learning experiment presented in Chapter 5. All learning blocks would include passive exposure to the label-meaning mappings, followed by a guessing phase and a production phase with feedback. After learning the language to a near-ceiling level (~90% accuracy), participants would use the input language to communicate with the other members of their community for three communication rounds, which would be identical to the communication rounds used in Chapters 3 and 4. Once all possible pairs in the group have interacted, participants would have completed an individual test phase where they would reproduce all labels in their input language – confirming that no significant changes have happened to the language during communication, and that participants were indeed using the languages they learned.

After the learning and communication phase, two participants would have been removed from the group and replaced with two new members (Fig 1B). The new members will either be naïve participants playing the game for the first time (Fig 1C), or experienced members who have just completed the learning and communication phase with another group using the second input language (Fig 1D). The newcomers will need to learn the target language of the new group by observing the remaining two members interact with each other for three rounds. No talking, signing or explicit teaching would be allowed in this observation phase. After learning from observation, the newcomers will enter the group as full-fledged members and will communicate with its other two members for six additional rounds (Fig 1E), in which all group members (new and old) will interact with each other exactly twice. Finally, all participants would complete an individual test phase to see what variation of the language they are using, and those languages would be compared across conditions (Fig 1F).

I planned to manipulate learners' prior linguistic knowledge as a proxy for L1 vs. L2 learning: “native” child learners were simulated by naïve participants, who are exposed to the community, language, meaning space and experimental design for the first time, while “non-native” adult learners were simulated by experienced participants, who had already acquired a different language and used it to successfully communicate in another group. This manipulation was motivated by several studies that have attributed age-related differences in language learning to differences in prior knowledge of one's first language (Baptista et al., 2016; Brooks & Kempe, 2019; Hernandez et al., 2005). Generally speaking, adults

learning a new language suffer from high levels of competition between the newly learned L2 and their prior knowledge of their L1, which they have already consolidated (unlike children, who still show flexible mappings). For example, adult learners seem to rely heavily on positive transfer or overlap between the L2 and their L1: they tend to show better learning of features that are similar between their native language and the second language, and have more trouble in learning lexical and phonological distinctions in the second language when they mismatch with their native language (Baptista et al., 2016; Hernandez et al., 2005; Potter et al., 2016; Schepens et al., 2020). I reasoned that if some differences between children and adults' language learning can be explained by their different levels of prior linguistic knowledge, this manipulation should invoke the desired effects.

Accordingly, experienced participants who have already played the game with a different group and have fully established word-to-meaning mappings would have a harder time learning a new language to describe the same meanings (in comparison to naïve participants). In other words, experienced participants who have already acquired a language in a different group would potentially struggle more with learning yet another language, especially if the new language would include a partially different logic (e.g., categorization based on different axis or on texture). Experienced learners may also be biased in favor of their existing linguistic system and its relevant categories. For example, a participant who learned a language with morphemes for shape and motion might have learned to ignore scenes' texture, and would struggle to learn a language which treats texture as a relevant grammatical feature. Naïve learners, on the other hand, have no prior model of the artificial language, no established labels for the scenes, and no preferences as to the relevant semantic categories of the languages. As such, they are presumably less biased when learning the new language. In other words, naïve learners, like children, would need to acquire the language (as well as their knowledge about the "world" of the experiment and the relevant categories of scene) from scratch, and may therefore be more accepting of complex and elaborate systems.

Of course, prior knowledge is only one possible explanation of age-related differences in language learning, and there are many other developmental differences that influence L1 vs. L2 learning outcomes. Specifically, children and adults may be exposed to different types of input, with adult second-language learners generally argued to learn from inferior input in quantity and quality (Atkinson et al., 2018; Bentz &

Berdicevskis, 2016). Additionally, children's inferior working memory (Gathercole et al., 2004) can lead them to overgeneralize more compared to adults (Hudson Kam & Newport, 2009; Wonnacott et al., 2013). Such developmental differences would obviously not be captured in the planned design, yet I believed that this manipulation will be enough to simulate the presence of non-native speakers in the community.

If these assumptions hold, then introducing experienced learners into the community would result in more simplifications in the form of more regularizations of inconsistencies and increased linguistic structure, while introducing naïve learners would result in preservation of the existing linguistic structure. Importantly, by the end of the experiment, there would be significant differences in linguistic structure of the input language across conditions. If prior linguistic knowledge is indeed making it harder for experienced participants to learn a new language, they should favor more simplified and transparent structures compared to naïve participants, given that such structures are easier to learn (see Chapter 5). As such, the final languages in this condition should be more systematic and compositional (e.g., irregulars should be regularized). This change could be the result of the non-natives introducing more simplifications, and/or the result of the original members of the group (i.e., the native speakers) accommodating to them and adopting their simplifications (Atkinson et al., 2018; Bentz & Berdicevskis, 2016; Loy & Smith, 2019).

However, it was also possible that having no prior knowledge or having newcomers of *any* type would shape languages into being more learnable in general. In this case, both conditions would have shown the same pattern of results, namely, simplifications due to learning. Interestingly, there was also a chance that languages with naïve learners would actually develop *more* linguistic structure and regularities compared to languages with experienced learners, given that experienced learners may be used to the idea that their target languages are imperfect and contain irregularities. This possibility was supported by the literature of emerging sign languages, home-signing and Creoles, which has suggested that learners with no prior linguistic model can lead to more regularization (Bickerton, 1984; Goldin-Meadow & Mylander, 1990; Senghas et al., 2004).

Since no specific hypotheses have been formulated in the literature about the effect of L2 learning on the degree of convergence, stability and communicative success in the community, I had no clear predictions regarding possible differences in these three parameters across conditions. Nevertheless, it was possible that there would be differences in the

degrees of communicative success and convergence due to existing members' tendency to accommodate more or less to certain newcomers. Although all groups would have been given the exact same instructions (to "communicate as successfully as possible with the newcomers"), it was possible that existing members would be more tolerant of mistakes made by naïve participants rather than "immigrants", or would show more tolerance towards non-natives in general (Lev-Ari, 2015a). Moreover, it was possible that the presence of two L2 speakers who speak the same language would preserve divergence in the group, since those speakers could revert to using their original input language whenever paired with each other. It was also possible that learning difficulties in early rounds would affect accuracy. For example, experienced participants may find it harder to learn the language and to abandon their old labels, possibly leading to less communicative success. Alternatively, experienced participants may find it easier to learn some aspects of the language given their prior knowledge and familiarity with the meaning space, possibly leading to more communicative success (especially if the L1 speakers will accommodate to them and simplify the language even more). This planned experiment could have resulted in many interesting findings, and could have shed light on the mechanisms involved in language simplifications in the presence of non-native speakers (e.g., are long-lasting simplifications the result of accommodation by native speakers, and/or by second-language learners difficulties in learning?).

However, I was unable to carry out this experiment as planned due to two main issues: it was not possible to get the first "native" learners to learn the language to a near-ceiling level in the allocated time, and consequently, it was not possible to get the first "native" members of the community to communicate with each other without changing that language, well before the introduction of the newcomers. In most pilots, at least one member of the group failed to learn the language to a sufficient level, and started introducing changes and regularities already in the first communication rounds. This was problematic, since a prerequisite of the manipulation was that the input language was well learned by everyone in the group. If some of the first participants did not learn the L1 language well, then any adaptations and simplifications could be the result of these participants' difficulties and learning biases. That is, when one of the L1 learners struggled with the language to begin with, it was highly likely that they would introduce L2-like mistakes and simplifications already in the first communication rounds, leading to changes (and perhaps ceiling

effects) before the manipulation even took place, making it extremely hard to compare the languages of different groups across different conditions.

Since the input language had to be partly structured and to contain irregularities and inconsistent patterns in order to allow for regularizations to take place, participants' difficulty with learning them was, in fact, unsurprising. The findings of Chapter 5 predict and explain this pattern, given that learners were shown to struggle with learning semi-structured languages in a similar learning paradigm. Specifically, learners in Chapter 5 who were trained on medium and mid-high structured languages (which were relatively similar in their linguistic structure scores to the input languages selected for this experiment, i.e., structure scores of 0.5-0.6) did not reach ceiling levels of accuracy in a similar time frame. It is of course possible that the first participants would have all learned the language well enough given more training time, yet the replacement procedure was already too long, and it was problematic to introduce more learning rounds (especially since some participants already learned the language well).

Although I was not able to solve these problems in the time constraints of my PhD project, a possible and highly implementable solution is to have the first participants trained on the language online prior to coming to the lab. Participants could get a link to an online learning experiment days in advance, and could only be allowed to participate in the actual experiment if they had passed a certain threshold of learning. I hope that I will be able to carry out this version of the experiment in the future.

3.3 Future work

Drawing on the paradigm introduced above, an interesting follow-up experiment would be to examine the scenario of **creole formation** in the lab. Specifically, an additional condition could be included, in which all four group members are first trained on four different input languages, and then come together to form one group so that they need to develop a new system to understand each other – a Lingua Franca. One possible prediction is that languages developed under such high-contact and high-heterogeneity conditions would be highly systematic and regular, and encompass the "best" of each language. Such an experiment could also manipulate the number of individuals with the same input language, examining whether the structure of the emerging creole is based on the dominant language in the community, often referred to as the Lexifier in creole literature (Blasi et al., 2017).

Other possible venues for future research are related to social diversity and to different sources of input heterogeneity. For example, does **gender diversity** in the community affect language evolution? Do communities that are more balanced in the number of men and women show more stable development? Are there specific tendencies that can be identified with male vs. female speakers? In most cases, languages evolve in communities that are balanced in terms of gender. However, it has been claimed that women and men have different preferences when it comes to linguistic variants (Al-Ali & Arafa, 2010; Barbu et al., 2014; Haas, 1979; Rosenhouse, 1998), even to the degree of having different dialects (i.e., "genderlects") within a community (Yokoyama, 1999). Based on language change theories, women have a more prominent role in the creation and spread of new linguistic innovations (Labov, 2007; Milroy & Milroy, 1993), but balanced communities may be more favorable for maintaining equilibrium. This potential effect of gender on language diversity and change has not been experimentally investigated, and could have important implications for promoting gender equality and respect to individual differences. This idea can be tested experimentally, by examining whether languages evolve differently in gender-homogenous communities (i.e., most members of the community are either male or female) compared to gender-balanced communities (i.e., half the population is male and the other half is female).

Similarly, it is possible to examine how **age diversity** impacts language evolution and change. Does the proportion of older vs. younger people affects the formation of new languages, and the rate of language change? Do age-homogenous communities comprised of mostly younger speakers (similar to the community in which Nicaraguan sign language has emerged) show more rapid adaptations? There is a basic intuition that older people are more conformist and more resistance to change, while younger people tend to quickly adapt new variants and cause rapid language change. This phenomenon has been documented in a few case-studies across languages, showing that children and adolescences are responsible for the creation of new linguistic innovations (Kerswill & Williams, 2000). Interestingly, theories of language change suggest that while younger people may innovate more, adults are responsible for the spread and fixation of these innovations: adults are seen as the main relevant adapters and trend-setters of language change, thanks to their social influence in society compared to children and teens (Labov, 2007; Roberts & Winters, 2012). However, these ideas were never tested systematically. The group communication paradigm introduced in this

dissertation can be adopted to examine how languages change as a function of the proportion of younger vs. older people in the community. For example, it is possible to manipulate age diversity as a function of real-age (i.e., testing groups with older and younger participants) and/or as function of prior experience with the language (i.e., testing groups in which some participants are highly experienced with the language while others are not). This will allow us to tease apart the role of age vs. prior experience in the creation and spread of new linguistic variants, and see if flexibility and stubbornness are underlined by prior knowledge or by cognitive maturation. Together, this future work could greatly promote our understanding of the role of age differences in the community, and can help explain trends of language change, adaptation and innovation.

Finally, it would be interesting to test the role of **social prestige** and how it can affect local and global alignment in a community. The literature of accommodation has long argued that people are more or less likely to align with others according to different social parameters, and that people preferentially copy from individuals of higher social, political, or economic status (Chartrand & van Baaren, 2009; Giles, Coupland, & Coupland, 1991; Lev-Ari, 2015b). Such ideas could be tested by introducing social prestige to the existing group communication paradigm, for example by announcing one participant as the leader, the "king", the "best", etc. The prestige participant could in fact have no additional advantages besides this biased impression, or alternatively, could have some concrete benefits. For example, the leader could be trained to objectively be the best user of the language, or to be the most worthy partner (e.g., successful interaction with them could be rewarded with twice as many points).

From a theoretical perspective, natural selection favors agents that are able to identify and copy models that show better-than-average information (Henrich & Gil-White, 2001). Therefore, higher alignment and more local convergence (within pairs) may be expected with high-status people, who are considered to be preferred models (either because they show high-performance or because they have greater respect in the community). In other words, participants may align more with such favored individuals. This can either increase the overall convergence in the group (as the majority of participants will align with the same favored model), or may actually hinder global convergence if, at the same time, participants adopt more egocentric and uncooperative behaviors and align less with the other "disfavored" agents (Galantucci et al., 2012).

With respect to Chapter 4, introducing a social prestige bias to the design could have additional effects of stabilization and conventionalization trends. Computational models that looked at agent prestige (by manipulating the weights given to different agents) showed that the presences of highly connected agents, i.e., "hubs", significantly increases the spread of innovations and degree of convergence in the language (Baxter, 2016; Fagyal et al., 2010; Zubek et al., 2017). Specifically, Baxter (2016) argued that differences in the degree of social influence can dramatically affect the mean time to reach consensus, and claimed that when social interactions are symmetric (so that all agents are weighed equally), the details of the network structure have no effect on the mean time to reach fixation or on the probability that a particular variant fixes. In contrast, the mean time to fixation can be dramatically affected by the presence of large disparities in the influence of different speakers. Fagyal et al. (2010) further claimed that having a prestige bias in the network is crucial for convergence: in a symmetrical network (where everybody's equal) or in a random network (where agents' influence is distributed randomly) there is much less consistency in the selection of the preferred model, resulting in more changes and less chances to establish a norm. The authors also argue that a leader's prestige is important for stability over time, as hubs are the propagators and enforcers of norms (though not the innovators). Given these models, it is possible that convergence would be faster (and perhaps greater) when some participants are more valued than others. As for stability, groups with a social prestige bias should fixate faster on a language, resulting in less changes in later rounds.

Interestingly, this manipulation may or may not affect linguistic structure as well. There is little reason to assume that social biases mediate the emergence of compositionality, as no study has directly examined the effects of prestige on grammatical structure. Nevertheless, it is possible to make several predictions based on intuitive arguments. First, it is possible that the effects of social biases are restricted to convergence and propagation trends. However, even if this is the case, having highly influential agents in the group could still indirectly affect structure by increase the chances that a given compositional innovation will spread to the entire community (i.e., increasing the overall chances for compositionality to be picked up). Alternatively, it could be that prestige will have a negative effect on structure, as participants may be more likely to simply adopt whatever language the popular agent would be using, rather than developing a predictable language with systematic structure to

facilitate convergence in the entire community. Whatever the case, examining such effects would have important implications for our theories on community structure, language structure, and linguistic diversity.

4 Conclusion

Reflecting back on this thesis, the take-home message can be simply summarized as "languages adapt to fit their social environments". The original goal of my PhD project was to tease apart and evaluate different social factors that may affect language evolution and diversity, which I have done using a novel group communication paradigm. Across five chapters, I have shown that the structure of languages can be shaped by communicative pressures, and especially by the size of the community in which they evolved. I have also demonstrated that languages with different degrees of systematic structure vary in their ease of learnability and suggested that this variation is relevant for language use and communication between strangers. I believe that the work presented in this thesis is merely a first step in understanding the social origins of cross-linguistic differences, and opens the door to a promising line of exciting research.

References

Acerbi, A., Kendal, J., & Tehrani, J. J. (2017). Cultural complexity and demography: the case of folktales. *Evolution and Human Behavior*, *38*(4), 474-480.

Ackerman, F., & Malouf, R. (2013). Morphological organization: The low conditional entropy conjecture. *Language*, *89*(3), 429–464.

Al-Ali, M. N., & Arafa, H. M. (2010). An experimental sociolinguistic study of language variation in Jordanian Arabic. He Buckingham Journal of Language and Linguistics, 3, 220–243.

Allcott, H., Karlan, D., Möbius, M. M., Rosenblat, T. S., & Szeidl, A. (2007). Community Size and Network Closure. American Economic Review, 97(2), 80–85.

Apfelbaum, K. S., & McMurray, B. (2011). Using Variability to Guide Dimensional Weighting: Associative Mechanisms in Early Word Learning. *Cognitive Science*, *35*(6), 1105–1138.

Arends, J., & Bruyn, A. (1994). Gradualist and developmental hypotheses. In *Pidgins and creoles: An introduction* (Vol. 15), (pp. 111-120). John Benjamins Publishing.

Armon-Lotem, S., Haman, E., López, K. J. de, Smoczynska, M., Yatsushiro, K., Szczerbinski, M., van Hout, A., Dabašinskienė, I., Gavarró, A., Hobbs, E., & Kamandulytė-Merfeldienė, L. (2016). A large-scale cross-linguistic investigation of the acquisition of passive. *Language Acquisition*, *23*(1), 27–56.

Aronoff, M., Meir, I., Padden, C. A., & Sandler, W. (2008). The roots of linguistic organization in a new language. *Interaction studies*, *9*(1), 133-153.

Atkinson, M. (2016). *Sociocultural determination of linguistic complexity*. PhD thesis. School of Philosophy, Psychology, and Language Sciences University of Edinburgh.

Atkinson, M., Kirby, S., & Smith, K. (2015). Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages. *PloS One*, *10*(6), e0129463.

Atkinson, M., Mills, G. J., & Smith, K. (2018). Social Group Effects on the Emergence of Communicative Conventions and Language Complexity. *Journal of Language Evolution*, *4*(1), 1-18.

Atkinson, M., Smith, K., & Kirby, S. (2018). Adult Learning and Language Simplification. *Cognitive Science*, *42*(8), 2818–2854.

Bahlmann, M. D. (2014). Geographic Network Diversity: How Does it Affect Exploratory Innovation? *Industry and Innovation*, *21*(7–8), 633–654.

Baptista, M., Gelman, S. A., & Beck, E. (2016). Testing the role of convergence in language acquisition, with implications for creole genesis. *International Journal of Bilingualism*, *20*(3), 269–296.

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512.

Barbu, S., Martin, N., & Chevrot, J.-P. (2014). The maintenance of regional dialects: A matter of gender? Boys, but not girls, use local varieties in relation to their friends' nativeness and local identity. *Frontiers in Psychology*, *5*, 1251.

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387–414.

Barking, M. (2016). The Impact of Word Order and Case Marking on Word and Structure Learning-An Artificial Language Learning Experiment. *Student Undergraduate Research E-Journal!*, *2*.

Baronchelli, A., Felici, M., Loreto, V., Caglioti, E., & Steels, L. (2006). Sharp transition towards shared vocabularies in multi-agent systems. *Journal of Statistical Mechanics: Theory and Experiment*, *2006*(06), P06014.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). lme4: Mixed-effects modeling with R; 2010. *URL: http://lme4.*

Baxter, G. J. (2016). Social Networks and Beyond in Language Change. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, & B. Job (Eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, (pp. 257–277). Springer.

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.

Beckner, C., Pierrehumbert, J. B., & Hay, J. (2017). The emergence of linguistic structure in an online iterated learning task. *Journal of Language Evolution*, lzx001.

Bentz, C., Alikaniotis, D., Cysouw, M., & Ferrer-i-Cancho, R. (2017). The Entropy of Words—Learnability and Expressivity across More than 1000 Languages. *Entropy*, *19*(6), 275.

Bentz, C., & Berdicevskis, A. (2016). Learning pressures reduce morphological complexity: Linking corpus, computational and experimental evidence. In *26th International Conference on Computational Linguistics (COLING 2016),* (pp. 222-232)

Bentz, C., Dediu, D., Verkerk, A., & Jäger, G. (2018). The evolution of language families is shaped by the environment beyond neutral drift. *Nature Human Behaviour*, *2*(11), 816–821.

Bentz, C., Verkerk, A., Kiela, D., Hill, F., & Buttery, P. (2015). Adaptive Communication: Languages with More Non-Native Speakers Tend to Have Fewer Word Forms. *PLOS ONE*, *10*(6), e0128254.

Bentz, C., & Winter, B. (2013). Languages with More Second Language Learners Tend to Lose Nominal Case. *Language Dynamics and Change*, *3*(1), 1–27.

Berdicevskis, A. (2012). Introducing pressure for expressivity into language evolution experiments. In *The Evolution of Language*, (pp. 64–71).

Bergmann, C., & Cristia, A. (2018). Environmental Influences on Infants' Native Vowel Discrimination: The Case of Talker Number in Daily Life. *Infancy*, *23*(4), 484–501.

Bergmann, C., Cristia, A., & Dupoux, E. (2016). Discriminability of sound contrasts in the face of speaker variation quantified. *38th Annual Conference of the Cognitive Science Society*, 1331–1336.

Bickerton, D. (1984). The language bioprogram hypothesis. *Behavioral and brain sciences*, *7*(2), 173-188.

Bickerton, D. (2007). Language evolution: A brief guide for linguists. *Lingua*, *117*(3), 510-526.

Birdsong, D. (2006). Age and Second Language Acquisition and Processing: A Selective Overview. *Language Learning*, *56*(s1), 9–49.

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In *Metacognition: Knowing about Knowing* (pp. 185–205). MIT Press.

Blasi, D. E., Michaelis, S. M., & Haspelmath, M. (2017). Grammars are robustly transmitted even during the emergence of creole languages. *Nature Human Behaviour*, *1*(10), 723–729.

Bleses, D., Basbøll, H., & Vach, W. (2011). Is Danish difficult to acquire? Evidence from Nordic past-tense studies. *Language and Cognitive Processes*, *26*(8), 1193–1231.

Bleses, D., Vach, W., Slott, M., Wehberg, S., Thomsen, P., Madsen, T. O., & Basbøll, H. (2008). Early vocabulary development in Danish and other languages: A CDI-based comparison. *Journal of Child Language*, *35*(3), 619–650.

Blom, E., Paradis, J., & Duncan, T. S. (2012). Effects of input properties, vocabulary size, and L1 on the development of third person singular–s in child L2 English. *Language Learning*, *62*(3), 965-994.

Blume, J. D., & Royall, R. M. (2003). Illustrating the law of large numbers (and confidence intervals). *The American Statistician*, *57*(1), 51-57.

Boas, F., Teit, J. A., Farrand, L., Gould, M. K., & Spinden, H. J. (1917). *Folk-tales of Salishan and Sahaptin tribes* (Vol. 11). American Folk-Lore Society.

Bohn, M., Kachel, G., & Tomasello, M. (2019). Young children spontaneously recreate core properties of language in a new modality. *Proceedings of the National Academy of Sciences, 116(51), 26072-26077*.

Bradlow, A.R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729.

Braginsky, M., Yurovsky, D., Marchman, V., & Frank, M. C. (2019). Consistency and variability in word learning across languages. *Open Mind: Discoveries in Cognitive Science*, *3*, 52–67.

Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of Gender-like Noun Subclasses in an Artificial

Language: The Contribution of Phonological Markers to Learning. *Journal of Memory and Language*, *32*(1), 76–95.

Brooks, P. J., & Kempe, V. (2019). More Is More in Language Learning: Reconsidering the Less-Is-More Hypothesis. *Language Learning*, *69*(S1), 13–41.

Caldwell, C. A., & Millen, A. E. (2010). Conservatism in laboratory microsocieties: Unpredictable payoffs accentuate group-specific traditions. *Evolution and Human Behavior*, *31*(2), 123–130.

Caldwell, C. A., & Smith, K. (2012). Cultural Evolution and Perpetuation of Arbitrary Communicative Conventions in Experimental Microsocieties. *PLoS ONE*, *7*(8), e43807.

Carr, J. W., Smith, K., Cornish, H., & Kirby, S. (2017). The cultural evolution of structured languages in an open-ended, continuous world. *Cognitive science*, *41*(4), 892-923.

Centola, D. (2010). The Spread of Behavior in an Online Social Network Experiment. *Science*, *329*(5996), 1194–1197.

Chartrand, T. L., & van Baaren, R. (2009). Human Mimicry. *Advances in experimental social psychology, 41*, 219-274

Christensen, P., Fusaroli, R., & Tylén, K. (2016). Environmental constraints shaping constituent order in emerging communication systems: Structural iconicity, interactive alignment and conventionalization. *Cognition*, *146*, 67–80.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62.

Clopper, C. G., & Pisoni, D. B. (2004). Effects of Talker Variability on Perceptual Learning of Dialects. *Language and Speech*, *47*(3), 207–238.

Cornish, H. (2010). Investigating how cultural transmission leads to the appearance of design without a designer in human communication systems. *Interaction Studies*, *11*(1), 112–137.

Cornish, H., Tamariz, M., & Kirby, S. (2009). Complex Adaptive Systems and the Origins of Adaptive Structure: What Experiments Can Tell Us. *Language Learning*, *59*, 187–205.

Coward, F. (2010). Small worlds, material culture and ancient Near Eastern social networks. *Social Brain, Distributed Mind*, 449–484.

Crystal, D. (1987). *The Cambridge encyclopedia of language*. Cambridge University Press.

Culbertson, J., & Kirby, S. (2016). Simplicity and specificity in language: Domain-general biases have domain-specific effects. *Frontiers in psychology*, *6*, 1964.

Dale, R., & Lupyan, G. (2012). Understanding the origins of morphological diversity: The linguistic niche hypothesis. *Advances in Complex Systems*, *15*(03n04), 1150017.

De Beule, J., & Bergen, B. K. (2006). On the emergence of compositionality. In *Proceedings of the 6th International Conference on the Evolution of Language,* 35-42.

Deacon, T. (1997). *The symbolic species: the coevolution of language and the brain*. W.W. Norton & Co.

DeKeyser, R. M. (2013). Age Effects in Second Language Learning: Stepping Stones Toward Better Understanding. *Language Learning*, *63*, 52–67.

DeKeyser, R. M. (2005). What Makes Learning Second-Language Grammar Difficult? A Review of Issues. *Language Learning*, *55*(S1), 1–25.

Derex, M., & Boyd, R. (2016). Partial connectivity increases cultural accumulation within groups. *Proceedings of the National Academy of Sciences*, *113*(11), 2982–2987.

Dressler, W. U. (2003). Morphological typology and first language acquisition: Some mutual challenges. *Mediterranean Morphology Meetings*, *4*(0), 7-20–20.

Dressler, W. U. (2010). A typological approach to first language acquisition. *Language Acquisition across Linguistic and Cognitive Systems*, *52*, 109–124.

Dryer, M.S., & Haspelmath, M. (2017). WALS Online (Leipzig: Max Planck Institute for Evolutionary Anthropology) Available at: http://wals.info/.

Dunbar, R. I. M. (2017). Group size, vocal grooming and the origins of language. *Psychonomic Bulletin & Review*, *24*(1), 209–212.

Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and brain sciences*, *16*(4), 681-694.

Eryilmaz, K., & Little, H. (2016). Using leap motion to investigate the emergence of structure in speech and language. *Behavior Research Methods, 49(5), 1748-1768.*

Estes, K. G., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental Psychology*, *51*(11), 1517–1528.

Estes, W. K., & Burke, C. J. (1953). A theory of stimulus variability in learning. *Psychological Review*, *60*(4), 276–286.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, *32*(5), 429-448.

Everett, C. (2013). Evidence for Direct Geographic Influences on Linguistic Sounds: The Case of Ejectives. *PLOS ONE*, *8*(6), e65275.

Everett, C., Blasi, D. E., & Roberts, S. G. (2016). Language evolution and climate: the case of desiccation and tone. *Journal of Language Evolution*, *1*(1), 33-46.

Everett, C., Blasi, D. E., & Roberts, S. G. (2015). Climate, vocal folds, and tonal languages: Connecting the physiological and geographic dots. *Proceedings of the National Academy of Sciences*, *112*(5), 1322–1327.

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.

Fay, N., Arbib, M., & Garrod, S. (2013). How to Bootstrap a Human Communication System. *Cognitive Science*, *37*(7), 1356–1367.

Fay, N., & Ellison, T. M. (2013). The Cultural Evolution of Human Communication Systems in Different Sized Populations: Usability Trumps Learnability. *PLOS ONE*, *8*(8), e71781.

Fay, N., Garrod, S., & Roberts, L. (2008). The fitness and functionality of culturally evolved communication systems. *Philosophical*

Transactions of the Royal Society B: Biological Sciences, *363*(1509), 3553–3561.

Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The Interactive Evolution of Human Communication Systems. *Cognitive Science*, *34*(3), 351–386.

Fay, N., Kleine, N. D., Walker, B., & Caldwell, C. A. (2019). Increasing population size can inhibit cumulative cultural evolution. *Proceedings of the National Academy of Sciences*, 201811413.

Fehér, O., Kirby, S., & Smith, K. (2014). Social influences on the regularization of unpredictable linguistic variation. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, 2187–2191.

Fehér, O., Wonnacott, E., & Smith, K. (2016). Structural priming in artificial languages and the regularisation of unpredictable variation. *Journal of Memory and Language*, *91*, 158-180.

Ferdinand, V., Kirby, S., & Smith, K. (2019). The cognitive roots of regularization in language. *Cognition*, *184*, 53–68.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, *44*(3), 677-694.

Fusaroli, R., & Tylén, K. (2012). Carving language for social coordination: A dynamical approach. *Interaction studies*, *13*(1), 103-124.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science*, *29*(5), 737–767.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in human neuroscience*, *5,* 11.

Galantucci, B., Theisen, C., Gutierrez, E. D., Kroos, C., & Rhodes, T. (2012). The diffusion of novel signs beyond the dyad. *Language Sciences*, *34*(5), 583–590.

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The Role of Single Talker Acoustic Variation in Early Word Learning. *Language Learning and Development*, *11*(1), 66–79.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from?. *Cognitive Science*, *31*(6), 961-987.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. Developmental Psychology, 40(2), 177-190.

Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ*, *5*, e3209.

Gibson, E., Futrell, R., Piandadosi, S. T., Dautriche, I., Mahowald, K., Bergen, L., & Levy, R. (2019). How Efficiency Shapes Human Language. *Trends in Cognitive Sciences, 23(4), 389-407*.

Gil, D. (2001). Creoles, Complexity and Riau Indonesian. *Linguistic Typology*, *5*, 325–371.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In *Contexts of Accommodation: Developments in Applied Sociolinguistics* (pp. 1–69). Cambridge University Press.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826.

Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. *The emergence of language*, 197-212.

Goldfield, B. A., & Reznick, J. S. (1990). Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of child language*, *17*(1), 171-183.

Goldin-Meadow, S., & Mylander, C. (1990). Beyond the input given: The child's role in the acquisition of language. *Language*, 323-355.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, *13*(5), 431-436.

Gong, T., Baronchelli, A., Puglisi, A., & Loreto, V. (2012). Exploring the roles of complex networks in linguistic categorization. *Artificial Life*, *18*(1), 107–121.

Gong, T., Ke, J., Minett, J. W., & Wang, W. S. (2004). A computational framework to simulate the coevolution of language and social structure. In J. Pollack, M. Bedau, P. Husbands, T. Ikegami & R. A. Watson (Eds.), *Artificial Life IX: Proceedings of the 9th International Conference on the Simulation and Synthesis of Living Systems*, (pp. 158–163). MIT Press.

Gong, T., Shuai, L., & Zhang, M. (2014). Modelling language evolution: Examples and predictions. *Physics of life reviews*, *11*(2), 280-302.

Granito, C., Tehrani, J., Kendal, J., & Scott-Phillips, T. (2019). Style of pictorial representation is shaped by intergroup contact. *Evolutionary Human Sciences*, *1,* e8.

Granovetter, M. (1976). Network Sampling: Some First Steps. *American Journal of Sociology*, *81*(6), 1287–1303.

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, *1*(1983), 201–233.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics*, *59*(5), 675–692.

Green, P., MacLeod, C.J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, **7**(4), 493–498. https://CRAN.R-project.org/package=simr.

Haas, A. (1979). Male and female spoken language differences: Stereotypes and evidence. *Psychological Bulletin*, *86*(3), 616–626.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). *Multivariate Data Analysis, 3rd edition*. Macmillan.

Halekoh, U., & Højsgaard, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models–the R package pbkrtest. *Journal of Statistical Software*, *59*(9), 1-32.

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, *9*(7), 342–348.

Haznedar, B. (2006). Persistent problems with case morphology in L2 acquisition. *Interfaces in Multilingualism: Acquisition and Representation*, (pp. 179–206).

Hengeveld, K., & Leufkens, S. (2018). Transparent and non-transparent languages. *Folia Linguistica*, *52*(1), 139–175.

Henrich, J. (2004). Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case. *American Antiquity*, *69*(2), 197–214.

Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, *22*(3), 165–196.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, *9*(5), 220–225.

Hewitt, B. G. (1995). *Georgian: A structural reference grammar* (Vol. 2). John Benjamins Publishing.

Hudson Kam, C. L., & Chang, A. (2009). Investigating the cause of language regularization in adults: Memory constraints or learning effects? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(3), 815–821.

Hudson Kam, C. L., & Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, *59*(1), 30–66. https://doi.org/10.1016/j.cogpsych.2009.01.001

Hudson Kam, C. L., & Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, *1*(2), 151–195.

Hockett, C. F. (1960). The Origin of Speech. *Scientific American*, *203*(3), 88–97.

Imedadze, N., & Tuite, K. (1992). The acquisition of Georgian. *The Crosslinguistic Study of Language Acquisition*, *3*, 39–109.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in cognitive sciences*, *3*(7), 272-279.

Johnson, A. W., & Earle, T. K. (2000). *The Evolution of Human Societies: From Foraging Group to Agrarian State*. Stanford University Press.

Jones, M., Vinson, D., Clostre, N., Zhu, A. L., Santiago, J., & Vigliocco, G. (2014). The Bouba Effect: Sound-Shape Iconicity in Iterated and Implicit Learning. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society (CogSci 2014)*.

Joseph, J. E., & Newmeyer, F. J. (2012). 'All Languages Are Equally Complex': The rise and fall of a consensus. *Historiographia Linguistica*, *39*(2–3), 341–368.

Ke, J., Gong, T., & Wang, W. S. (2008). Language change and social networks. *Communications in Computational Physics*, *3*(4), 935–949.

Kempe, M., & Mesoudi, A. (2014). An experimental demonstration of the effect of group size on cultural accumulation. *Evolution and Human Behavior*, *35*(4), 285–290.

Kempe, V., & Brooks, P. J. (2008). Second language learning of complex inflectional systems. *Language Learning*, *58*(4), 703–746.

Kempe, V., & Brooks, P. J. (2018). Linking Adult Second Language Learning and Diachronic Change: A Cautionary Note. *Frontiers in Psychology*, *9, 480*.

Kempe, V., & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition*, *20*(4), 543–587.

Kennedy, P. (1992). *A Guide to Econometrics*. Blackwell.

Kerswill, P., & Williams, A. (2000). *Creating a New Town koine: Children and language change in Milton Keyens*. *Language in society*, *29*(1), 65-115

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, E.J. (Eds.), *Linguistic Evolution Through Language Acquisition: Formal and Computational Models*, (pp. 173–204). Cambridge 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.

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.

Kirby, S., Griffiths, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current opinion in neurobiology*, *28*, 108-114.

Kirby, S., & Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the evolution of language,* (pp. 121-147). Springer.

Kirby, S., Smith, K., & Brighton, H. (2004). From UG to universals: Linguistic adaptation through iterated learning. *Studies in Language*, *28*(3), 587–607.

Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition*, *141*, 87–102.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kramer, S. N. (1968). The "Babel of Tongues": A Sumerian Version. *Journal of the American Oriental Society*, *88*(1), 108–111.

Krupnik, I., & Müller-Wille, L. (2010). Franz Boas and Inuktitut terminology for ice and snow: From the emergence of the field to the "Great Eskimo Vocabulary Hoax." In *SIKU: Knowing our ice,* (pp. 377–400). Springer.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature reviews neuroscience*, *5*(11), 831-843.

Kurvers, R. H. J. M., Krause, J., Croft, D. P., Wilson, A. D. M., & Wolf, M. (2014). The evolutionary and ecological consequences of animal social networks: Emerging issues. *Trends in Ecology & Evolution*, *29*(6), 326–335.

Labov, W. (2007). Transmission and diffusion. *Language*, *83*(2), 344–387.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299-321.

Laufer, B. (2009). Why are some words more difficult than others? - Some intralexical factors that affect the learning of words. *IRAL - International Review of Applied Linguistics in Language Teaching*, *28*(4), 293–308.

Lev-Ari, S. (2018). The influence of social network size on speech perception. *Quarterly Journal of Experimental Psychology*, *71*(10), 2249-2260.

Lev-Ari, S. (2016). How the size of our social network influences our semantic skills. *Cognitive science*, *40*(8), 2050-2064.

Lev-Ari, S. (2015a). Comprehending non-native speakers: Theory and evidence for adjustment in manner of processing. *Frontiers in Psychology*, *5, 1546*.

Lev-Ari, S. (2015b). Selective Grammatical Convergence: Learning From Desirable Speakers. *Discourse Processes*, *53*(8), 657–674.

Lev-Ari, S., & Shao, Z. (2017). How social network heterogeneity facilitates lexical access and lexical prediction. *Memory & cognition*, *45*(3), 528-538.

Lewis, M., & Frank, M. C. (2016). Linguistic niches emerge from pressures at multiple timescales. In *The 38th annual meeting of the cognitive science society (CogSci 2016).*

Lewis, M., Simons, G.F., & Fennig, C.D. (2017). Ethnologue: Languages of the world (SIL international Dallas, TX).

Liu, B. S.-C., Madhavan, R., & Sudharshan, D. (2005). DiffuNET: The impact of network structure on diffusion of innovation. *European Journal of Innovation Management*, *8*(2), 240-262.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242-1255.

Lou-Magnuson, M., & Onnis, L. (2018). Social Network Limits Language Complexity. *Cognitive Science*, *42*(8), 2790–2817.

Loy, J. E., & Smith, K. (2019). *Syntactic adaptation may depend on perceived linguistic knowledge: Native English speakers differentially adapt to native and nonnative confederates in dialogue* [Preprint]. https://doi.org/10.31234/osf.io/pu2qa

Lupyan, G., & Dale, R. (2016). Why are there different languages? The role of adaptation in linguistic diversity. *Trends in cognitive sciences*, *20*(9), 649-660.

Lupyan, G., & Dale, R. A. (2015). The role of adaptation in understanding linguistic diversity. *Language Structure and Environment: Social, Cultural, and Natural Factors*, (pp. 287–316).

Lupyan, G., & Dale, R. (2010). Language Structure Is Partly Determined by Social Structure. *PLoS ONE*, *5*(1), e8559.

MacWhinney, B. (1978). The Acquisition of Morphophonology. *Monographs of the Society for Research in Child Development*, *43*(1/2), 1–123.

Maddieson, I., & Coupé, C. (2015). Human spoken language diversity and the acoustic adaptation hypothesis. *The Journal of the Acoustical Society of America*, *138*(3), 1838–1838.

Maffi, L. (2005). Linguistic, Cultural, and Biological Diversity. *Annual Review of Anthropology, 34,* 599-617.

Maher, J. C. (2017). *Multilingualism: A Very Short Introduction*. Oxford University Press.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Sommers, M. S. (1989). Effects of Talker Variability on Recall of Spoken Word Lists. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *15*(4), 676–684.

Martín, J. S., Walker, B., Fay, N., & Tamariz, M. (2019). *Network connectivity dynamics affect the evolution of culturally transmitted variants* [Preprint]. ArXiv:1902.06598.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*(1), 1-51.

McFadden, T. (2003). On Morphological Case and Word-Order Freedom. *Proceedings of the Twenty-Ninth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Phonetic Sources of Phonological Patterns: Synchronic and Diachronic Explanations*, 295–306.

McWhorter, J. (2007). *Language interrupted: Signs of non-native acquisition in standard language grammars*. Oxford University Press.

McWhorter, J. H. (2001). The worlds simplest grammars are creole grammars. *Linguistic Typology*, *5*(2–3), 125–166.

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.

Milroy, J., & Milroy, L. (1993). Mechanisms of change in urban dialects: The role of class, social network and gender. *International Journal of Applied Linguistics*, *3*(1), 57–77.

Milroy, J., & Milroy, L. (1985). Linguistic change, social network and speaker innovation. *Journal of Linguistics*, *21*(02), 339–384.

Mirolli, M., & Parisi, D. (2008). How producer biases can favor the evolution of communication: An analysis of evolutionary dynamics. *Adaptive Behavior*, *16*(1), 27-52.

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347.

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.

Motamedi, Y., Schouwstra, M., Smith, K., & Kirby, S. (2016). Linguistic structure emerges in the cultural evolution of artificial sign languages. In *The Evolution of Language: Proceedings of the 11th International Conference* (EVOLANG11), (pp. 493-495).

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378.

Munsinger, H., & Kessen, W. (1966). Stimulus variability and cognitive change. *Psychological Review*, *73*(2), 164–178.

Nettle, D. (1999). Using social impact theory to simulate language change. *Lingua*, 108(2), 95–117.

Nettle, D. (2012). Social scale and structural complexity in human languages. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *367*(1597), 1829–1836.

Nilsenová, M., & Nolting, P. (2010, September). Linguistic adaptation in semi-natural dialogues: age comparison. In *Proceedings of the 13th International Conference on Text, Speech and Dialogue,* (pp. 531–538). Springer.

Nölle, J., Staib, M., Fusaroli, R., & Tylén, K. (2018). The emergence of systematicity: How environmental and communicative factors shape a novel communication system. *Cognition, 181,* 93-104.

Papagno, C., & Vallar, G. (1992). Phonological Short-term Memory and the Learning of Novel Words: The Effect of Phonological Similarity

and Item Length. *The Quarterly Journal of Experimental Psychology Section*, *44*(1), 47–67.

Parkvall, M. (2008). The simplicity of creoles in a cross-linguistic perspective. *Language Complexity: Typology, Contact, Change*, 265–285.

Parodi, T., Schwartz, B. D., & Clahsen, H. (2004). On the L2 acquisition of the morphosyntax of German nominals. *Linguistics*, 669–706.

Perfors, A. (2016). Adult Regularization of Inconsistent Input Depends on Pragmatic Factors. *Language Learning and Development*, *12*(2), 138–155.

Perlman, M., Dale, R., & Lupyan, G. (2015). Iconicity can ground the creation of vocal symbols. *Royal Society Open Science*, *2*(8), 150152.

Perlman, M., Little, H., Thompson, B., & Thompson, R. L. (2018). Iconicity in Signed and Spoken Vocabulary: A Comparison Between American Sign Language, British Sign Language, English, and Spanish. *Frontiers in Psychology*, *9, 1433*.

Perry, L. K., Axelsson, E. L., & Horst, J. S. (2016). Learning what to remember: Vocabulary knowledge and children's memory for object names and features. *Infant and Child Development*, *25*(4), 247-258.

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological science*, *21*(12), 1894-1902.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3, Pt.1), 353–363.

Post, B., Marslen-Wilson, W. D., Randall, B., & Tyler, L. K. (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition*, *109*(1), 1–17.

Potter, C. E., Wang, T., & Saffran, J. R. (2016). Second Language Experience Facilitates Statistical Learning of Novel Linguistic Materials. *Cognitive Science*, *41*, 913-927.

R Core Team (2016). A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; http://www.R-project.org/

Raviv, L. & Arnon, I. (2018). Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning. *Cognition, 181,* 160-173.

Raviv, L., Meyer, A. & Lev-Ari, S. (2017). Compositional structure can emerge without generational transmission. *Paper presented at the 30th CUNY Conference on Human Sentence Processing,* Cambridge MA, USA.

Raviv, L., Meyer, A., & Lev-Ari, S. (2019a). Compositional structure can emerge without generational transmission. *Cognition*, *182*, 151-164.

Raviv, L., Meyer, A. S., & Lev-Ari, S. (2019b). Larger communities create more systematic languages. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1907): 20191262

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.

Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, *111*(3), 317-328.

Regier, T., Carstensen, A., & Kemp, C. (2016). Languages Support Efficient Communication about the Environment: Words for Snow Revisited. *PLOS ONE*, *11*(4), e0151138.

Richtsmeier, P. T., Gerken, L., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition*, *111*(3), 372–377.

Roberts, G. (2010). An experimental study of social selection and frequency of interaction in linguistic diversity. *Interaction studies*, *11*(1), 138-159.

Roberts, G., & Galantucci, B. (2012). The emergence of duality of patterning: Insights from the laboratory. *Language and Cognition*, *4*(04), 297–318.

Roberts, S. G., & Winters, J. (2012). Social structure and language structure: The new nomothetic approach. *Psychology of Language and Communication*, *16*(2), 89–112.

Rosenhouse, J. (1998). Women's speech and language variation in Arabic dialects. *Al-'Arabiyya*, *31*, 123–152.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental science*, *12*(2), 339-349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy : The Official Journal of the International Society on Infant Studies*, *15*(6).

Saldana, C., Fagot, J., Kirby, S., Smith, K., & Claidière, N. (2019). High-fidelity copying is not necessarily the key to cumulative cultural evolution: A study in monkeys and children. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1904), 20190729.

Samara, A., Smith, K., Brown, H., & Wonnacott, E. (2017). Acquiring variation in an artificial language: Children and adults are sensitive to socially conditioned linguistic variation. *Cognitive Psychology*, *94*, 85–114.

Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond?. *Cognition*, *73*(1), 1-33.

Schepens, J., van Hout, R., & Jaeger, T. F. (2020). Big data suggest strong constraints of linguistic similarity on adult language learning. *Cognition*, *194*, 104056.

Scott-Phillips, T. C. (2017). A (simple) experimental demonstration that cultural evolution is not replicative, but reconstructive—and an explanation of why this difference matters. *Journal of Cognition and Culture*, *17*(1-2), 1-11.

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in cognitive sciences*, *14*(9), 411-417.

Seidl, A., Onishi, K. H., & Cristia, A. (2014). Talker Variation Aids Young Infants' Phonotactic Learning. *Language Learning and Development*, *10*(4), 297–307.

Selten, R., & Warglien, M. (2007). The emergence of simple languages in an experimental coordination game. *Proceedings of the National Academy of Sciences*, *104*(18), 7361–7366.

Senghas, A., & Coppola, M. (2001). Children creating language: How Nicaraguan Sign Language acquired a spatial grammar. *Psychological science*, *12*(4), 323-328.

Senghas, A., Kita, S., & Ozyurek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in Nicaragua. *Science*, *305*(5691), 1779-1782.

Sinkeviciute, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, *41*(04), 795–820.

Slobin, D. (1985). *The crosslinguistic study of language acquisition* (Vol. 2). Hillsdale/Mahwah, NJ: Erlbaum.

Smith, K. (2011). Learning Bias, Cultural Evolution of Language, and the Biological Evolution of the Language Faculty. *Human Biology*, *83*(2), 261–278.

Smith, K., Brighton, H., & Kirby, S. (2003). Complex systems in language evolution: the cultural emergence of compositional structure. *Advances in Complex Systems*, *6*(04), 537-558.

Smith, K., & Wonnacott, E. (2010). Eliminating unpredictable variation through iterated learning. *Cognition*, *116*(3), 444-449.

Sommers, M. S., & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406–2416.

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(2), 231–249.

Spike, M. (2017). Population size, learning, and innovation determine linguistic complexity, In *The 39th annual meeting of the cognitive science society (CogSci 2017)*.

Spike, M. (2016). *Minimal requirements for the cultural evolution of language*. PhD thesis, The University of Edinburgh.

Stadler, K. (under preparation). Find a segmentation that maximises the overall string coverage across all signals. http://kevinstadler.github.io/cultevo/reference/ssm.compositionality.html

Sumner, M. (2011). The role of variation in the perception of accented speech. *Cognition*, *119*(1), 131–136.

Sweet, H. (1899). *Thepractical Study of Languages: A Guide for teachers and Learners*. Oxford (reimp. en 1964).

Tal, S., & Arnon, I. (2019). Redundant morphological marking facilitates children's learning of a novel construction. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci 2019)*.

Tan, R., & Fay, N. (2011). Cultural transmission in the laboratory: Agent interaction improves the intergenerational transfer of information. *Evolution and Human Behavior*, *32*(6), 399–406.

Tamariz, M., Brown, J. E., & Murray, K. M. (2010). The role of practice and literacy in the evolution of linguistic structure. *Proceedings of the 8th International Conference on the Evolution of Language*, 313–320.

Tamariz, M., & Smith, A. (2008). Regularity in mappings between signals and meanings. *Proceedings of the 7th International Conference on the Evolution of Language*, 315–322.

Teit, J. A. (1917). Kaska tales. *The Journal of American Folklore*, *30*(118), 427–473.

Theisen, C. A., Oberlander, J., & Kirby, S. (2010). Systematicity and arbitrariness in novel communication systems. *Interaction Studies*, *11*(1), 14-32.

Trehub, S. E. (2015). Cross-cultural convergence of musical features. *Proceedings of the National Academy of Sciences*, *112*(29), 8809-8810.

Trudgill, P. (2009). *Sociolinguistic typology and complxification*. In G. Sampson, D. Gil & P. Trudgill (Eds.), *Language Complexity as an Evolving Variable*, (pp. 98–109). Oxford University Press.

Trudgill, P. (2008). On the role of children, and the mechanical view: A rejoinder. *Language in Society*, *37*(02), 277–280.

Trudgill, P. (2005). Linguistic and social typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology*, *8*(3), 305–320.

Trudgill, P. (2002). Linguistic and Social Typology. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change* (pp. 707–728). Blackwell.

Trudgill, P. (1992). *Dialect Typology and Social Structure*. Walter de Gruyter.

Van Heugten, M., Bergmann, C., & Cristia, A. (2015). The Effects of Talker Voice and Accent on Young Children's Speech Perception. In *Individual differences in speech production and perception* (pp. 57–88). Peter Lang.

Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, *4*(4), 357–380.

Verhoef, T., Walker, E., & Marghetis, T. (2016). Cognitive biases and social coordination in the emergence of temporal language. In *The 38th annual meeting of the cognitive science society (CogSci 2016)*.

Vogt, P. (2007). Group size effects on the emergence of compositional structures in language. In *European Conference on Artificial Life* (pp. 405-414). Springer.

Vogt, P. (2009). Modeling interactions between language evolution and demography. *Human Biology*, *81*(2), 237–258.

Vukatana, E., Graham, S. A., Curtin, S., & Zepeda, M. S. (2015). One is Not Enough: Multiple Exemplars Facilitate Infants' Generalizations of Novel Properties. *Infancy*, *20*(5), 548–575.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*, 440-442.

Wichmann, S., & Holman, E. W. (2009). Population size and rates of language change. *Human biology*, *81*(3), 259-274.

Wichmann, S., Stauffer, D., Schulze, C., & Holman, E. W. (2008). Do language change rates depend on population size? *Advances in Complex Systems*, *11*(03), 357–369.

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *The Language Learning Journal*, *40*(1), 125–137.

Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition*, *7*(3), 415-449.

Wonnacott, E., Brown, H., & Nation, K. (2013). *Comparing generalisation in children and adults learning an artificial language*. Poster presenter at *Child Language Seminar, Manchester, UK.*

Wonnacott, E., & Newport, E. L. (2005). Novelty and Regularization: The Effect of Novel Instances on Rule Formation. *Proceedings of the 29th Annual Boston University Conference on Language Development*.

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.

Xanthos, A., Laaha, S., Gillis, S., Stephany, U., Aksu-Koç, A., Christofidou, A., Gagarina, N., Hrzica, G., Ketrez, F.N., Kilani-Schoch, M., & Korecky-Kröll, K (2011). On the role of morphological richness in the early development of noun and verb inflection. *First Language*, *31*(4), 461–479.

Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society B: Biological Sciences, 280*(1758), 20123073.

Xu, Y., Malt, B. C., & Srinivasan, M. (2017). Evolution of word meanings through metaphorical mapping: Systematicity over the past millennium. *Cognitive psychology*, *96*, 41-53.

Yokoyama, O. T. (1999). Russian Genderlects and Referential Expressions. *Language in Society*, *28*(3), 401–429.

Zlatev, J. (2008). From proto-mimesis to language: Evidence from primatology and social neuroscience. *Journal of Physiology-Paris*, *102*(1), 137-151.

Zubek, J., Denkiewicz, M., Barański, J., Wróblewski, P., Rączaszek-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.

Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In *Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference*, (pp. 51-58). MIT Press.

English Summary

Why are there so many different languages in the world? How much do languages differ from each other in terms of their linguistic structure and their learnability? And how do such differences come about?

One possibility is that differences between languages (i.e., linguistic diversity) stem from differences in the social environments in which languages evolve. In this doctoral thesis, I tried to shed light on the social origin of language diversity. I did this by experimentally examining the formation and acquisition of new languages created in real-time under different social conditions. I developed a group communication experiment, where groups of participants needed to create a new artificial language to communicate with each other about various scenes with moving objects. I tested how the process of language evolution was shaped by the fact that the languages developed in different communities, with different population sizes and different types of social network structure. I also tested whether languages that evolved under these different conditions differed from each other in how easily they were learned and used by new people, who had not been involved in their development.

In the first chapter, I showed that groups playing this game developed languages with systematic grammars. In the second and third chapters, I varied the size of the groups (big vs. small) and how well-connected people in the groups were to each other (dense vs. sparse). I asked how these changes affected the emerging languages. The results showed that big groups developed more systematic and structured languages, and did so faster and more consistently than small groups. In contrast, there was no evidence that the network structure in the groups played a similar role: Densely connected and sparsely connected groups all reached similar levels of systematic structure. In the last chapter, I tested whether the languages previously created in the group communication experiment differed from each other in how easily they were learned and used by new individuals. The results showed that more systematic languages were learned better and faster compared to languages with many irregularities. Moreover, participants who learned more systematic languages were better able to label scenes they had not seen before, and could better communicate about the scenes with strangers they had not met before. In sum, the studies in this thesis show how community structure affects the development and nature of language structure in the lab.

סיכום בעברית

מדוע יש כל כך הרבה שפות שונות בעולם? עד כמה שפות באמת שונות אחת מהשניה מבחינת המבנה הדקדוקי שלהן? ואיך הבדלים אלו נוצרו מלכתחילה?

אחת התשובות האפשריות והמעניינות ביותר לשאלה זו היא שהבדלים בין שפות (או גיוון לשוני) נובעים מהבדלים בסביבה החברתית שבה כל שפה התפתחה. בעבודת הדוקטורט שלי ניסיתי לשפוך אור על המקור החברתי של גיוון לשוני, ולבחון באמצעות ניסויים בתנאי מעבדה את תהליך ההיווצרות וההשתנות של שפות חדשות בזמן אמת תחת תנאים חברתיים מגוונים. פיתחתי ניסוי תקשורת קבוצתי שבו על קהילות קטנות של משתתפים ליצור ביחד שפה מלאכותית חדשה על מנת לתקשר זה עם זה על מגוון של תרחישים, שבהן צורה לא מוכרת נעה במרחב לכיוונים שונים. בחנתי כיצד התהליך של היווצרות שפה (או אבולוציה של שפה) מעוצב ומשתנה על ידי העובדה ששפות נוצרות בקהילות שונות, עם מספר אחר של דוברים ומבנה אחר לחלוטין של רשתות חברתיות. בנוסף, בחנתי כיצד שפות שנוצרו תחת תנאים שונים נבדלות זו מזו גם בקלות שבהן הן נקלטות ונלמדות ע"י דוברים חדשים שלא היו חלק מתהליך היצירה של השפה. שאלתי האם ישנן שפות שקל יותר ללמוד מאחרות מבחינת קצב ורמת הדיוק של רכישת השפה, והאם ישנן שפות שהינן יותר נוחות ואפקטיביות לשימוש.

בפרק הראשון של הדוקטורט, הראיתי שקהילות קטנות ששיחקו את משחק התקשורת הקבוצתי פיתחו שפה עם מבנה דקדוקי שיטתי. בפרק השני והשלישי בחנתי כיצד תהליך היווצרותה של שפה מושפע ע"י שינוי של גודל הקבוצה (גדולה מול קטנה) ורמת הקישוריות בקבוצה (האם אנשים יותר או פחות מקושרים זה לזה). תוצאות המחקר הראו ששפות אשר התפתחו בקבוצות הגדולות היו בעלות מבנה דקדוקי שיטתי יותר ביחס לאלו שהתפתחו בקבוצות הקטנות, ושהמבנה הדקדוקי בקבוצות גדולות נוצר מהר יותר ובאופן עקבי יותר. לעומת זאת, לא ניכרו עדויות לכך שמבנה הרשת החברתית של הקהילה משפיע על היווצרות הדקדוק באותה הדרך: שפות שנוצרו ע"י קהילות צפופות ומקושרות היטב היו בעלות אותה דרגת שיטתיות כמו שפות שנוצרו ע"י קהילות דלילות יותר שבהן לא כולם תיקשרו זה עם זה. בפרק האחרון בחנתי האם השפות שהתפתחו בניסוי ע"י קהילות בגדלים שונים נבדלות זו מזו בקלות שבהן הן נלמדות ע"י דוברים חדשים. תוצאות המחקר הראו ששפות בעלות מבנה שיטתי יותר נלמדו במהירות ובקלות רבה יותר ביחס לשפות שהכילו הרבה יוצאי דופן. בנוסף, משתתפים שלמדו שפות שיטתיות יותר הצליחו להשתמש בשפה באופן יותר יעיל על מנת לסווג ולכנות תרחישים חדשים שמעולם לא ראו, ויכלו להבין משתתפים אחרים שמעולם לא פגשו בצורה ברורה יותר. לסיכום, המחקרים בדוקטורט זה מראים כי המבנה החברתי של הקהילה יכול להשפיע על אופן ההיווצרות ועל טבע המבנה הדקדוקי של שפות חדשות.

Nederlandse Samenvatting

Waarom zijn er zoveel verschillende talen op de wereld? Hoeveel verschillen talen van elkaar in hun structuur en leergemak? En hoe ontstaan deze verschillen?

Een mogelijke verklaring is dat verschillen in talen (linguïstische diversiteit) ontspringen uit de verschillen in sociale omgeving waarin talen ontwikkelen. Met dit proefschrift probeerde ik licht te werpen op de sociale oorsprong van verschillen tussen talen. Ik onderzocht hoe mensen, onder verschillende sociale omstandigheden, nieuwe talen ontwikkelden en leerden. Ik ontwierp een real-time groepscommunicatie-spel als experiment, waarin groepen deelnemers samen een nieuwe, kunstmatige taal ontwikkelden. Deze taal moesten ze gebruiken om te communiceren over beelden van bewegende objecten. Ik onderzocht hoe de evolutie van de taal gevormd werd door de grootte van de groepen en de structuren van sociale netwerken binnen de groepen. Ook bekeek ik of talen die op deze manier ontstonden verschillen in hoe makkelijk ze door nieuwe mensen, die niet betrokken waren bij het ontstaan van deze talen, geleerd en gebruikt worden.

In het eerste hoofdstuk toonde ik aan dat groepen die dit spel speelden talen ontwikkelden met systematische grammatica. In het tweede en derde hoofdstuk varieerde ik de groepsgrootte (groot vs. klein) en de mate waarin mensen verbonden waren met anderen (grote groepsdichtheid vs. kleine groepsdichtheid). Ik onderzocht hoe verschillen in deze factoren de opkomende talen beïnvloedden. De uitkomsten toonden aan dat grote groepen meer systematiek en structuur in hun talen aanbrachten. Ook gebeurde dit sneller en consistenter dan in kleinere groepen. Ik vond echter geen bewijs dat de netwerkstructuur binnen groepen een vergelijkbare rol speelde. Groepen die nauw verbonden waren bereikten vergelijkbare niveaus van systematiek en structuur in talen als groepen die weinig verbonden waren. In het laatste hoofdstuk testte ik of de talen die ontwikkeld werden in het groepscommunicatie-spel verschilden in leergemak en bruikbaarheid door nieuwe individuen. De resultaten toonden aan dat systematischere talen beter en sneller aangeleerd werden dan talen met veel onregelmatigheden. Daarnaast pasten de deelnemers die systematischere talen leerden deze beter toe bij het omschrijven van beelden die ze nog niet eerder gezien hadden. Ook communiceerden ze makkelijker met onbekenden over deze beelden. Kortom, de onderzoeken in dit proefschrift tonen hoe groepsstructuur de ontwikkeling en eigenschappen van taalstructuur beïnvloedt binnen een gecontroleerde setting.

Acknowledgements

So many amazing people have supported me in the last four years. I have been surrounded with a group of incredible and kind colleagues, who helped me and inspired me in so many ways. I was (and still am) surrounded with a group of fun and loving friends, who made me smile and feel loved even in hard times. And on top of this, I have been blessed with a beautiful family, who was there for me in every step of the way, feeling so close and connected even from afar. I feel extremely lucky and privileged to have had all these amazing people in my life. My heart is full of love and gratitude to all of them. And I guess it's not trivial to have so many people to value, look up to, and adore, which is why I appreciate this opportunity to thank them all. This means it's going to be a very long list of acknowledgments (bear with me here!), but also a long celebration of gratitude, connection, and inspiration. Thank you all, dear ones!

First and foremost, I would like to thank my supervisor, **Antje**, who means far more to me than just a supervisor. Thank you for believing in me and offering me the unique opportunity to do a PhD in your department. Thank you for supporting my ambitious ideas and plans even when they were not aligned with your own interests. Thank you for investing your time, resources, and effort in making sure this project succeeds. And for making sure I succeed and stay sane and happy along the way. Thank you for giving me the freedom and independence to work under the conditions that make me most productive, for accommodating to my needs, and for supporting me throughout this journey. Thank you for all you helpful (and quick!) feedback, for your kind words, for your helpful comments and clever suggestions, for your professional advice, and for your encouragement. Most importantly, thank you for genuinely caring about me. Thank you for being there for me in my periods of grieve and loss, for making me feel welcome to talk to you at any time. I really appreciated and valued our conversations about life, science, and everything in between. Thank you for doing your absolute best to comfort me and help me out whenever you could, and for giving me a hug when I needed one. Our relationship has grown stronger and closer as years passed, and I am so grateful for having you in my life. I wish for everyone to have such an amazing and supportive supervisor.

I also want to thank my co-supervisor **Shiri**. Thank you for supporting this project, and for your dedicated involvement. I am grateful for your

contribution to chapters 2-4, as well as for your insightful feedback on the rest of this thesis.

I would also like to thank the members of my reading committee. Thank you **Asli** for all your support and care from the moment I arrived to Nijmegen. It was always a pleasure to talk to you and to exchange ideas and thoughts. Thank you for your positive energy, your encouragement, and your valuable comments. Thank you **Caro**, for your brilliant advice on various topics, for your active support as a director in promoting a better life for PhDs, and for helping out in times of conflict. Thank you **Mark** for your inspiring talks and ideas, and for having your door open for questions and guidance. I would also like to thank you for introducing me to Zotero at the beginning of my PhD, and teaching me how to use it to make my workflow so much better - that tool really revolutionized the way I work, and made my life easier!

A special thank you goes to **Caitlin**, who did an absolutely amazing job at programming the group communication experiment used throughout this entire thesis. I really appreciate all your hard work and the time you spend on this, and I greatly value your consistency, effort, and creativity in making this experiment come to life. I also want to thank you for being kind, open, and helpful during this process, making me feel supported and welcome, even when things were hard and less convenient. Thank you so much! This whole project would not have been possible without you.

I would also like thank the Technical group for their constant support throughout this challenging project. To **Peter**, for your incredible hard work and investment in creating a beautiful online version for my experiment. To **Reiner**, for your time, effort, and care in allocating resources to this project. To **Johan**, for making sure everything works smoothly and efficiently, and always with a smile. To **Alex**, for your ongoing technical support, kindness, and availability. And to **Jeroen**, for all your help, patience, and fun moments of laughter. I would also like to thank **Jan** and **Rober** for making my working environment pleasant and comfortable.

I also want to thank the Operations team for all their wonderful work and for taking care of me and my fellow PhDs. It was great to feel that the administration is truly a team, and that you all truly care for us. Thank you **Angela**, for being the first friendly face I saw at the MPI, for arranging so many administrative things on my behalf, and for always being there for me. Thank you for listening and for supporting me, for sharing your

thoughts with me, and for making sure I was doing ok, work-wise and life-wise. Thank you **Laura**, for truly caring about my personal well-being as well as the well-being of my fellow PhDs. Thank you so much for fighting for us to get better working conditions, and for always being kind, honest, and welcoming. Thank you **Anique**, for your joyful energy and useful advice, and for thinking outside the box. It was a pleasure to interact with you on any topic. Thank you **Luise**, for always being so helpful and informative, and for your support in arranging business trips and reimbursements (a truly important thing for any student). Thank you **Marjolein**, for giving me wonderful advice on public relation and communication, and for trying your best to make sure I am a happy camper.

I also want to thank **Karin** and **Meggie**, the library team, for ensuring that we have everything we need to do quality research, including rare materials that you always managed to obtain quickly. Thank you both for supporting the goal of open science, and for making sure our work is visible and accessible to all. And most importantly, thank you for doing all that with a smile and enthusiasm. I would also like to thank the IMPRS coordinator **Kevin**. Thank you for your care, dedication, and support from the very first day of my PhD. Thank you for providing me with ample opportunities to learn new things and to teach others. Thank you for organizing social events and for arranging awesome, enriching, and innovative activities. Thank you **Thea**, for taking wonderful care of my vegan belly and for making sure there are suitable options at the canteen. And thank you **Maud**, for being a friend and a sister, for sharing pain and joy with me, and for giving me a warm loving hug when I come to work.

Thank you to all my friends and colleagues at the Psychology of Language department. Thank you **Annelies**, for supporting me in the tedious process of recruiting and contacting participants, and for making sure the visualizations and translations go smoothly and professionally. Thank you **Birgit**, for helping me in the first steps of testing and sharing the load with me. Thank you **Phillip**, for giving me useful coding and statistical advice, and for helping me improve my workflow (and my whisky collection). Thank you **Laurel**, for your enthusiastic and charismatic personality that lights up the room, for your extremely important stash of candy, and for always offering to listen, advise, and help out. Thank you **Falk**, for sharing inspiring ideas and giving me valuable feedback, as well as stopping for fun chats in the hallway. Thank you **Eirini**, for your fun and lively conversations, for sharing happy moments, and for your spot-on cultural tips (Belinda left an impact on

me). Thank you **Saoradh**, for being an inspiration of wisdom, class, and elegance, and for always being so sweet and open. Thank you **Federica**, for fun and silly chats, and for always smiling at me, no matter how busy you are. Thank you **Sophie**, a new yet lovely addition to the list of awesome people I'm happy to know, for your plant-caring skills, for your areal talent, and for your fresh point of view. Thank you **Jeroen** and **Ashley**, for sharing your wisdom and insights with me, for listening and supporting me whenever in need, and for many beautiful afternoons in your sunny garden. Thank you **Greta**, for being a ray of sunshine and optimism in my life, for always being supportive and kind, and for many lovely moments of happy giggles alongside deep conversation. Thank you **Amie**, for being my super-chatty-high-pitch soul buddy, and for all your support and care throughout the years. I miss you! Thank you for all the fun outings together, for buying me furniture, for your linguistic editing and fancy stats advice, and for your generally lovely and happy spirit. Thank you **Marianne** and **Li**, my extremely talented students, who made taught me more about my topic and my personality than I expected. Thank you for helping me grow as a person, for making me a better supervisor, and for honoring me with the chance to support you in your own journey towards success.

A super special thank you goes to my amazing friends **Sara**, **Joe** and **Mischa**, without whom I wouldn't have made it through this period, at least not as happy, fulfilled, and sane. I am grouping you guys together because to me, you three represent a tight hub of friendship, support, and love, and I just realized that I wanted to thank you all for the same things, which are beyond important to me. Thank you so much for caring for me, and for really loving me as I am. Thank you for being a shoulder to cry on, partners to gossip with, and honest soulmates that always told me the truth and made me feel safe. Thank you for sharing your feelings with me, for opening up your homes and your hearts, and for letting me comfort you too whenever I could. You three made me a better person, no doubt about it. I am so grateful for the countless fun evenings we had together, and for the beautiful memorable conversations we had, which I will cherish forever. Thank you for your clever academic advice, for your technical support, for feeding me lots of yummy food, and for just being inspiring humans. I love you three very very much, and your support and friendship is priceless to me. I feel blessed to have you as friends, and I hope we stay connected for many more years to come.

I also want to give a giant thank you to my awesome paranymphs and super close friends, Aitor and Markus. Thank you **Aitor**, for making my

life happier, funnier, and more interesting since the moment you arrived to Nijmegen. Thank you for the fun times, sassy talks, deep conversations, and lekker vegan brunches. Thank you for sharing your beautiful mind with me, and for letting me in. I really value our connection, and I am happy to have you by my side. Thank you **Markus**, for four years of inspiration and close friendship. Thank you for so many amazing and innovative conversations about life, brains, consciousness, technology (or should I say, Sci-Fi), and evolution. These really sparked my joy and interest in science and humanity. Thank you for sharing your feelings and thoughts with me (as well as your delicious bread), and for letting me be a part of your life. Thank you for all the fun evenings we had together, and for truly being there for me in times of struggle. Thank you for introducing me to your amazing partner Nadine, who has become my close and dear friend. Speaking of whom, thank you **Nadine**, for countless fun times in the Netherlands and in Israel, for inspiring talks, for being my happy vegan yoga buddy, and for the beautiful and intimate connection we share.

There are so many more MPI and IMPRS people I'd like to thank, who have made my time in Nijmegen more wonderful. Thank you to my fellow PhD representatives **Miguel**, **Merel M.**, **Merel W.**, **Julia E.**, and **Julia M.**, for standing shoulder to shoulder with me and trying your best to make PhDs' lives better. Thank you for investing your time, effort, and minds in this noble goal, and for you active and social engagement. Thank you **Christina**, for your inspiring attitude, rigorous approach to science, valuable opinions, and lots of laughter. Thank you **Lara**, for being my housemate and close friend in times of confusion and growth. Thank you **Arnold**, for your true friendship and lovely personality, for our Czech chitty-chat, and for happy Amsterdam bunny time. Thank you **Marieke**, for our beautiful and meaningful connection, for sharing our fears and dreams, and for your easy and caring persona. Thank you **Francie**, for many fun outings, for honest sharing, and for womanhood empathy beyond expected. Thank you **Ksenia**, for being such a lovely and honest person, for your inspiring stories, and for supporting me in all work and life matters. Thank you **Middy**, for always having your door open for me, for letting me feel safe to share my thoughts and feelings with you, and for giving me great advice. Thank you **Arushi** and **Devansh**, for a long and close friendship full of fun and good tastes. I am grateful to have you both in my life, and for being able to share our experiences, doubts, and successes. Thank you to the Evolution group, who gave me a home in the first year of my PhD and made me feel supported and encouraged. Thank

you **Alan**, for so many hilarious and funny moments, while at the same being deeply sincere, genuine, and honest. Thank you **Bill**, for your enormous support in teaching me new skills, and for many happy high-teas with you and lovely **Laura**. Thank you **Kevin**, for your incredible engagement in my project and for developing tools I still use today. Thank you **Tessa**, for inspiring me with your beautiful mind and professional success, and for guiding me along the way. Thank you **Andrea**, for our wonderful connection, for your active attempts to promote my career, and for generally awesome times around the Netherlands. Thank you **Rabia**, for the mutual support, for the safe space to offload and recharge, and for our promising and exciting collaboration. Thank you **Timo**, for all your valuable and insightful feedback, for making me a better presenter and a better human being. Thank you **Katja**, **Hannah**, **Patricia**, **Marlou, Dilay**, **Arianna**, **Paulo**, **Beyza**, **Brigitte**, **Ezgi**, **Rowan**, **Ine**, **Weng**, **Midas**, **Rebecca**, **David**, **Hans-Rutger**, **Valeria**, **Edith**, **Nanjo**, **Florian**, **Linda**, **Else**, **Naomi**, **Zeynep**, **Sonja**, **Ellen**, **Laura**, **Monique**, **Katherine**, **James**, and so many other people I adore and cherish their company but did not mention.

I also want to thank my Israeli colleagues and friends from the Hebrew University, who supported me academically and mentally for the past six years, from my Master's until today. Thank you **Inbal**, for being my mentor in every sense of the word. Thank you for teaching me how to write, how to publish, how to ask interesting research questions, how to formalize clear and convincing arguments, and how to do all this with enthusiasm, rigor, and kindness. Thank you **Naomi**, for your ongoing support and inspiration, for being my dear friend, colleague, and teacher all at once. Thank you for all the lovely conversations, for your valuable feedback, and for letting me know your beautiful family. Thank you **Shira**, for your joyful spirit and curiosity for life, for being able to share my feelings with you and always get support and empathy, and for all the times we shared ideas, experimental plans, and life goals.

I would also like to give a giant thank you to my closest friends in the Netherlands. First, to my Dutch-Israeli lady hub, **Sagi** and **Shira**, for being unbelievably supportive and so much fun to be around. I am extremely blessed to have you two in my life, and I can't imagine my time here without you. You have become my true soulmates, חברות אמת, my Hebrew sanctuary, my cultural home, my gezellig בוביק, and my source for comfort, happiness, and self-analysis. Thank you so much for being my friends. Thank you **Ankie**, for your lovely and beautiful presence, and for sharing your yoga, your touch, and your amazing heart with me. Thank

you **Hadar** and **Anne**, for being my true friends in both Israel and NL. I feel lucky to have been able to share experiences with each other in two different continents, and for so many years. !!מתה עליכם. Thank you for all the support, love, shakshuka, and fun times from childhood till now. Importantly, I want to thank the amazing Dagani family, **Tzuf**, **Raz**, **Tete**, and **Reizi**, for giving me a true sense of belonging across space and time. You are my second home and second family in this country, and I thank you so much for supporting me and caring for me in ways beyond my imagination. תודה וחיבוק ענקי על כל התמיכה והאהבה שלכם. Thank you for being a home not only to me, but also to so many animals in need. This shows how truly beautiful your hearts are. I love you all so much, and I am so happy to be a part of your lives. I also want to take this opportunity to thank the ones I lost along the way, but have supported me nonetheless – Amit, Buffy, my grandmother, my grandfather, Karel. I am grateful for the lessons you taught me while still being a part of my life. I am a better person because of you, and I will always remember the good times.

Thank you to my amazing friends back home, who even from a distance of thousands kilometers felt so close to me. Thank you **Ben** and **Dorit**, העברה בנדורית המקורית!, for years of intimate, meaningful, and profound connection. Thank you for challenging me as a thinker and as a feeler with lots of deep conversations, and for always being there for me in my journey for self-love and growth. I learned so much from you two, and you are my inspiration for true love and respectful relationship. Thank you **Rona**, קוקילה שלי, for your beautiful company, laughter, help, and advice throughout the years. Your support was so valuable to me, in every single topic: you taught me how to dress, cook, be a more empathic human, and take better care of myself. I also really value your help in designing some my beautiful PhD-related visualizations. You made everything look much nicer! And I really love you. Thank you **Tali**, for being a critical voice of sanity, wisdom, and objectivity in my life. Thank you for your smart and realistic perspective, for your care to me, for answering silly medical questions in the middle of the night, and for always coming to see me even when you're pregnant, sick, and hours' drive from me. You are an amazing friend.

Last but not least, I want to thank my incredible family. Words simply cannot describe how I love them, how much they mean to me, and how close our connection is. To my amazing **mother Amira**, אימוש יקרה שלי, thank you for inspiring me already from childhood to be a strong and independent woman, while also being social, energetic, gentle, and kind to others. Thank you for showing me strength, empathy, wisdom, and

intelligence. Thank you for loving me unconditionally, and for supporting me in my journey, even if it meant being far away from you for so many years. Thank you for checking in on me every day, and for spoiling me beyond repair whenever I go back home to visit. To my amazing **father Ofer**, אבאל'ה אהוב שלי, thank you for providing me with so much love and support, for challenging my decisions, and for making me feel like I can always count on your for help. Thank you for all the inspiring conversations about technology, science, human nature, and feelings. Thank you both, אמאבא מדהימים שלי, for raising me to be the person I am today. I hope you are proud of me at least as I am proud of you. I am extremely lucky to have you as my parents. And thank you **Doron** and **Dana**, my second patents, for always making me feel at home, and for making my parents happy. Thank you both for your smart advice, fresh perspective, and all our memorable and lovely moments together. Thank you to my beautiful **sister Keren**, חמשוצ קטן וחמוד, who is actually one of my best and closest friends in the whole world. Our connection is so important to me, and I am really grateful for being able to share with you everything that's going on inside me and in my life. I have learned so much from you, and your advice and support has kept me going in times of crisis as well as times of joy. I love you so much. Thank you also to my brilliant cousins **Omer** and **Matan**, for your exciting, inspirational, and stimulating existence, for being my true friends and my mentors. I realize how lucky I am to have you as sources for both knowledge and spirituality. אתם מדהימים. To my sweet aunt and confidant **Niti**, ניתוש יקרה, thank you for being there for me whenever I need, for advising me, for supporting me, and for caring about my heart, skin, and vegan belly. Thank you **Ronit** and **Yali**, for so many valuable advices and heartwarming conversations. A giant thank you to the rest of my amazing family: **Michal**, **Yaron**, **Ofir**, **Omri**, **Maya**, my grandmother **Raya**, and her savior **Slava**. Thank you all so much for your support, care, and beautiful impact on my life. I love you all.

Curriculum vitae

Limor Raviv was born in 1985 in Petah-Tikva, Israel. She obtained her bachelor's degree in Generative Linguistics and Cognitive Science from The Hebrew University of Jerusalem in 2013. She then completed her Master's degree in Cognitive Science at the Hebrew University in 2015 as part of the Language Learning and Processing Lab. In February 2016, Limor began her PhD project in the Psychology of Language Department at the Max Planck Institute for Psycholinguistics, which she completed in January 2020.

Publications

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

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

Havron, N., **Raviv, L.**, & Arnon, I. (2018). Literate and preliterate children show different learning patterns in an artificial language learning task. *Journal of Cultural Cognitive Science,* 2, 21-33. doi:10.1007/s41809-018-0015-9

Raviv, L., & Arnon, I. (2018). Systematicity, but not compositionality: Examining the emergence of linguistic structure in children and adults using iterated learning. *Cognition*, 181, 160-173. doi:10.1016/j.cognition.2018.08.011

Raviv, L., & Arnon, I. (2018). The developmental trajectory of children's auditory and visual statistical learning abilities: Modality-based differences in the effect of age. *Developmental Science*. 21(4): e12593. doi:10.1111/desc.12593

Raviv, L., Meyer, A., & Lev-Ari, S. (2018). The role of community size in the emergence of linguistic structure. In *Proceedings of the 12th International Conference (EVOLANGXII)*. Toruń, Poland: NCU Press. doi:10.12775/3991-1.096.

Raviv, L., & Arnon, I. (2016). Language evolution in the lab: The case of child learners. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci 2016). Austin, TX: Cognitive Science Society*.

Raviv, L., & Arnon, I. (2016). The developmental trajectory of children's statistical learning abilities. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci 2016). Austin, TX: Cognitive Science Society*.

Raviv, L., Meyer, A., & Lev-Ari, S. (Submitted). *The role of social network structure in the emergence of linguistic structure.*

Raviv, L., de Heer Kloots, M. & Meyer, A. (Submitted). *What makes a language easy to learn? A preregistered study on how systematic structure and community size affect lang*uage *learnability*.

van der Ham, S., **Raviv, L.**,. & de Boer, B. (Submitted). *Modality-based differences in Distributional Learning: Categorization and production of signals in the auditory, visual and tactile*

MPI Series in Psycholinguistics

1. The electrophysiology of speaking: Investigations on the time course of semantic, syntactic, and phonological processing. *Miranda van Turennout*
2. The role of the syllable in speech production: Evidence from lexical statistics, metalinguistics, masked priming, and electromagnetic midsagittal articulography. *Niels O. Schiller*
3. Lexical access in the production of ellipsis and pronouns. *Bernadette M. Schmitt*
4. The open-/closed-class distinction in spoken-word recognition. *Alette Haveman*
5. The acquisition of phonetic categories in young infants: A self-organising artificial neural network approach. *Kay Behnke*
6. Gesture and speech production. *Jan-Peter de Ruiter*
7. Comparative intonational phonology: English and German. *Esther Grabe*
8. Finiteness in adult and child German. *Ingeborg Lasser*
9. Language input for word discovery. *Joost van de Weijer*
10. Inherent complement verbs revisited: Towards an understanding of argument structure in Ewe. *James Essegbey*
11. Producing past and plural inflections. *Dirk Janssen*
12. Valence and transitivity in Saliba: An Oceanic language of Papua New Guinea. *Anna Margetts*
13. From speech to words. *Arie van der Lugt*
14. Simple and complex verbs in Jaminjung: A study of event categorisation in an Australian language. *Eva Schultze-Berndt*
15. Interpreting indefinites: An experimental study of children's language comprehension. *Irene Krämer*
16. Language-specific listening: The case of phonetic sequences. *Andrea Weber*
17. Moving eyes and naming objects. *Femke van der Meulen*

18. Analogy in morphology: The selection of linking elements in Dutch compounds. *Andrea Krott*
19. Morphology in speech comprehension. *Kerstin Mauth*
20. Morphological families in the mental lexicon. *Nivja H. de Jong*
21. Fixed expressions and the production of idioms. *Simone A. Sprenger*
22. The grammatical coding of postural semantics in Goemai (a West Chadic language of Nigeria). *Birgit Hellwig*
23. Paradigmatic structures in morphological processing: Computational and cross-linguistic experimental studies. *Fermín Moscoso del Prado Martín*
24. Contextual influences on spoken-word processing: An electrophysiological approach. *Daniëlle van den Brink*
25. Perceptual relevance of prevoicing in Dutch. *Petra M. van Alphen*
26. Syllables in speech production: Effects of syllable preparation and syllable frequency. *Joana Cholin*
27. Producing complex spoken numerals for time and space. *Marjolein Meeuwissen*
28. Morphology in auditory lexical processing: Sensitivity to fine phonetic detail and insensitivity to suffix reduction. *Rachèl J. J. K. Kemps*
29. At the same time...: The expression of simultaneity in learner varieties. *Barbara Schmiedtová*
30. A grammar of Jalonke argument structure. *Friederike Lüpke*
31. Agrammatic comprehension: An electrophysiological approach. *Marlies Wassenaar*
32. The structure and use of shape-based noun classes in Miraña (North West Amazon). *Frank Seifart*
33. Prosodically-conditioned detail in the recognition of spoken words. *Anne Pier Salverda*
34. Phonetic and lexical processing in a second language. *Mirjam Broersma*
35. Retrieving semantic and syntactic word properties. *Oliver Müller*

36. Lexically-guided perceptual learning in speech processing. *Frank Eisner*
37. Sensitivity to detailed acoustic information in word recognition. *Keren B. Shatzman*
38. The relationship between spoken word production and comprehension. *Rebecca Özdemir*
39. Disfluency: Interrupting speech and gesture. *Mandana Seyfeddinipur*
40. The acquisition of phonological structure: Distinguishing contrastive from non-contrastive variation. *Christiane Dietrich*
41. Cognitive cladistics and the relativity of spatial cognition. *Daniel B.M. Haun*
42. The acquisition of auditory categories. *Martijn Goudbeek*
43. Affix reduction in spoken Dutch. *Mark Pluymaekers*
44. Continuous-speech segmentation at the beginning of language acquisition: Electrophysiological evidence. *Valesca Kooijman*
45. Space and iconicity in German Sign Language (DGS). *Pamela Perniss*
46. On the production of morphologically complex words with special attention to effects of frequency. *Heidrun Bien*
47. Crosslinguistic influence in first and second languages: Convergence in speech and gesture. *Amanda Brown*
48. The acquisition of verb compounding in Mandarin Chinese. *Jidong Chen*
49. Phoneme inventories and patterns of speech sound perception. *Anita Wagner*
50. Lexical processing of morphologically complex words: An information-theoretical perspective. *Victor Kuperman*
51. A grammar of Savosavo, a Papuan language of the Solomon Islands. *Claudia Wegener*
52. Prosodic structure in speech production and perception. *Claudia Kuzla*

53. The acquisition of finiteness by Turkish learners of German and Turkish learners of French: Investigating knowledge of forms and functions in production and comprehension. *Sarah Schimke*

54. Studies on intonation and information structure in child and adult German. *Laura de Ruiter*

55. Processing the fine temporal structure of spoken words. *Eva Reinisch*

56. Semantics and (ir)regular inflection in morphological processing. *Wieke Tabak*

57. Processing strongly reduced forms in casual speech. *Susanne Brouwer*

58. Ambiguous pronoun resolution in L1 and L2 German and Dutch. *Miriam Ellert*

59. Lexical interactions in non-native speech comprehension: Evidence from electro-encephalography, eye-tracking, and functional magnetic resonance imaging. *Ian FitzPatrick*

60. Processing casual speech in native and non-native language. *Annelie Tuinman*

61. Split intransitivity in Rotokas, a Papuan language of Bougainville. *Stuart Robinson*

62. Evidentiality and intersubjectivity in Yurakaré: An interactional account. *Sonja Gipper*

63. The influence of information structure on language comprehension: A neurocognitive perspective. *Lin Wang*

64. The meaning and use of ideophones in Siwu. *Mark Dingemanse*

65. The role of acoustic detail and context in the comprehension of reduced pronunciation variants. *Marco van de Ven*

66. Speech reduction in spontaneous French and Spanish. *Francisco Torreira*

67. The relevance of early word recognition: Insights from the infant brain. *Caroline Junge*

68. Adjusting to different speakers: Extrinsic normalization in vowel perception. *Matthias J. Sjerps*

69. Structuring language. Contributions to the neurocognition of syntax. *Katrien R. Segaert*
70. Infants' appreciation of others' mental states in prelinguistic communication: A second person approach to mindreading. *Birgit Knudsen*
71. Gaze behavior in face-to-face interaction. *Federico Rossano*
72. Sign-spatiality in Kata Kolok: how a village sign language of Bali inscribes its signing space**.** *Conny de Vos*
73. Who is talking? Behavioural and neural evidence for norm-based coding in voice identity learning. *Attila Andics*
74. Lexical processing of foreign-accented speech: Rapid and flexible adaptation. *Marijt Witteman*
75. The use of deictic versus representational gestures in infancy. *Daniel Puccini*
76. Territories of knowledge in Japanese conversation. *Kaoru Hayano*
77. Family and neighbourhood relations in the mental lexicon: A cross-language perspective. *Kimberley Mulder*
78. Contributions of executive control to individual differences in word production. *Zeshu Shao*
79. Hearing speech and seeing speech: Perceptual adjustments in auditory-visual processing. *Patrick van der Zande*
80. High pitches and thick voices: The role of language in space-pitch associations. *Sarah Dolscheid*
81. Seeing what's next: Processing and anticipating language referring to objects. *Joost Rommers*
82. Mental representation and processing of reduced words in casual speech. *Iris Hanique*
83. The many ways listeners adapt to reductions in casual speech. *Katja Poellmann*
84. Contrasting opposite polarity in Germanic and Romance languages: Verum Focus and affirmative particles in native speakers and advanced L2 learners. *Giuseppina Turco*

85. Morphological processing in younger and older people: Evidence for flexible dual-route access. *Jana Reifegerste*
86. Semantic and syntactic constraints on the production of subject-verb agreement. *Alma Veenstra*
87. The acquisition of morphophonological alternations across languages. *Helen Buckler*
88. The evolutionary dynamics of motion event encoding. *Annemarie Verkerk*
89. Rediscovering a forgotten language. *Jiyoun Choi*
90. The road to native listening: Language-general perception, language-specific input. *Sho Tsuji*
91. Infants'understanding of communication as participants and observers. *Gudmundur Bjarki Thorgrímsson*
92. Information structure in Avatime. *Saskia van Putten*
93. Switch reference in Whitesands. *Jeremy Hammond*
94. Machine learning for gesture recognition from videos. *Binyam Gebrekidan Gebre*
95. Acquisition of spatial language by signing and speaking children: a comparison of Turkish sign language (TID) and Turkish. *Beyza Sümer*
96. An ear for pitch: on the effects of experience and aptitude in processing pitch in language and music. *Salomi Savvatia Asaridou*
97. lncrementality and Flexibility in Sentence Production. *Maartje van de Velde*
98. Social learning dynamics in chimpanzees: Reflections on (nonhuman) animal culture. *Edwin van Leeuwen*
99. The request system in Italian interaction. *Giovanni Rossi*
100. Timing turns in conversation: A temporal preparation account. *Lilla Magyari*
101. Assessing birth language memory in young adoptees**.** *Wencui Zhou*
102. A social and neurobiological approach to pointing in speech and gesture. *David Peeters*

103. Investigating the genetic basis of reading and language skills. *Alessandro Gialluisi*
104. Conversation Electrified: The Electrophysiology of Spoken Speech Act Recognition. *Rósa Signý Gisladottir*
105. Modelling Multimodal Language Processing. *Alastair Smith*
106. Predicting language in different contexts: The nature and limits of mechanisms in anticipatory language processing. *Florian Hintz*
107. Situational variation in non-native communication. *Huib Kouwenhoven*
108. Sustained attention in language production. *Suzanne Jongman*
109. Acoustic reduction in spoken-word processing: Distributional, syntactic, morphosyntatic, and orthographic effects. *Malte Viebahn*
110. Nativeness, dominance, and the flexibility of listening to spoken language. *Laurence Bruggeman*
111. Semantic specificity of perception verbs in Maniq. *Ewelina Wnuk*
112. On the identification of FOXP2 gene enhancers and their role in brain development. *Martin Becker*
113. Events in language and thought: The case of serial verb constructions in Avatime. *Rebecca Defina*
114. Deciphering common and rare genetic effects on reading ability. *Amaia Carrión Castillo*
115. Music and language comprehension in the brain. *Richard Kunert*
116. Comprehending Comprehension: Insights from neuronal oscillations on the neuronal basis of language. *Nietzsche H.L. Lam*
117. The biology of variation in anatomical brain asymmetries. *Tulio Guadalupe*
118. Language processing in a conversation context. *Lotte Schoot*
119. Achieving mutual understanding in Argentine Sign Language. *Elizabeth Manrique*
120. Talking Sense: the behavioural and neural correlates of sound symbolism. *Gwilym Lockwood*
121. Getting under your skin: The role of perspective and simulation of experience in narrative comprehension. *Franziska Hartung*

122. Sensorimotor experience in speech perception. *Will Schuerman*
123. Explorations of beta-band neural oscillations during language comprehension: Sentence processing and beyond. *Ashley Lewis*
124. Influences on the magnitude of syntactic priming. *Evelien Heyselaar*
125. Lapse organization in interaction. *Elliott Hoey*
126. The processing of reduced word pronunciation variants by natives and foreign language learners: Evidence from French casual speech. *Sophie Brand*
127. The neighbors will tell you what to expect: Effects of aging and predictability on language processing. *Cornelia Moers*
128. The role of voice and word order in incremental sentence processing. *Sebastian Sauppe*
129. Learning from the (un)expected: Age and individual differences in statistical learning and perceptual learning in speech. *Thordis Neger*
130. Mental representations of Dutch regular morphologically complex neologisms. *Laura de Vaan*
131. Speech production, perception, and input of simultaneous bilingual preschoolers: Evidence from voice onset time. *Antje Stoehr*
132. A holistic approach to understanding pre-history. *Vishnupriya Kolipakam*
133. Characterization of transcription factors in monogenic disorders of speech and language. *Sara Busquets Estruch*
134. Indirect request comprehension in different contexts. *Johanne Tromp*
135. Envisioning Language - An Exploration of Perceptual Processes in Language Comprehension. *Markus Ostarek*
136. Listening for the WHAT and the HOW: Older adults' processing of semantic and affective information in speech. *Juliane Kirsch*
137. Let the agents do the talking: on the influence of vocal tract anatomy on speech during ontogeny and glossogeny. *Rick Janssen*
138. Age and hearing loss effects on speech processing. *Xaver Koch*

139. Vocabulary knowledge and learning: Individual differences in adult native speakers. *Nina Mainz*
140. The face in face-to-face communication: Signals of understanding and non-understanding. *Paul Hömke*
141. Person reference and interaction in Umpila/Kuuku Ya'u narrative. *Clair Hill*
142. Beyond the language given: The neurobiological infrastructure for pragmatic inferencing. *Jana Bašnáková*
143. From Kawapanan to Shawi: Topics in language variation and change. *Luis Miguel Rojas-Berscia*
144. On the oscillatory dynamics underlying speech-gesture integration in clear and adverse listening conditions. *Linda Drijvers*
145. Understanding temporal overlap between production and comprehension. *Amie Fairs*
146. The role of exemplars in speech comprehension. *Annika Nijveld*
147. A network of interacting proteins disrupted in language-related disorders. *Elliot Sollis*
148. Fast speech can sound slow: Effects of contextual speech rate on word recognition. *Merel Maslowski*
149. Reason-giving in everyday activities. *Julija Baranova*
150. Speech planning in dialogue - Psycholinguistic studies of the timing of turn taking. *Mathias Barthel*
151. The role of neural feedback in language unification: How awareness affects combinatorial processing. *Valeria Mongelli*
152. Exploring social biases in language processing. *Sara Iacozza*
153. Vocal learning in the pale spear-nosed bat, *Phyllostomus discolor*. *Ella Lattenkamp*
154. The effect of language contact on speech and gesture: The case of Turkish-Dutch bilinguals in the Netherlands. *Elif Zeynep Azar*
155. Language and society: How social pressures shape grammatical structure. *Limor Raviv*