Cognition 182 (2019) 151–164

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Compositional structure can emerge without generational transmission

Limor Raviv[a,*], Antje Meyer[a,b], Shiri Lev-Ari[a,c]

[a] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[b] Radboud University Nijmegen, the Netherlands
[c] Royal Holloway University of London, Egham, UK

ARTICLE INFO

Keywords:
Language evolution
Iterated learning
Communication
Input variability
Artificial language experiments
Compositionality

ABSTRACT

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.

1. 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, & MacLeod, 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.,

* Corresponding author at: Max Plank Institute for Psycholinguistics, P.O. Box 310, 6500 AH Nijmegen, the Netherlands.
E-mail address: limor.raviv@mpi.nl (L. Raviv).

https://doi.org/10.1016/j.cognition.2018.09.010
Received 25 September 2017; Received in revised form 13 September 2018; Accepted 13 September 2018

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 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 are highly compressed and easy to learn, they are 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, 2017; 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, & Ozyurek, 2004), and work on pidgins has suggested that new child learners are required in order to develop recursion (Bickerton, 1984). 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, Staib, Fusaroli, and Tylén (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 communicating 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 (Gómez, 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, 2016; 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.

2. 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.

2.1. Method

2.1.1. 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 and 2:45 h). 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).

2.1.2. 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 2.55 cm^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 (0°, 30°, 45°, 60°, 90°, 120°, 135°, 150°, 180°, 210°, 225°, 240°, 270°, 300°, 315°, 330°)[1], 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.

[1] 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.

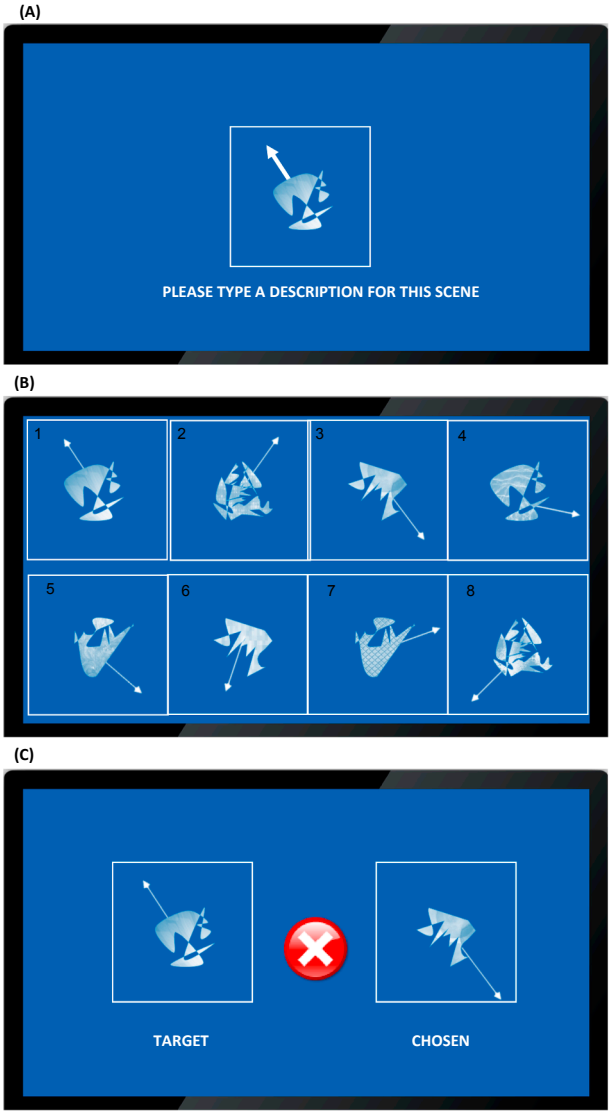

Fig. 1. Example of the computer interfaces in a single game in the communication phase. Arrows illustrate the shapes' direction of movement on the screen. The producer saw the target scene on their screen (A) and typed a description for it using their keyboard. Once the guesser saw the description (presented 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).

2.1.3. 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 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 they thought would be understood by their partner, they could choose to use that label instead.

During the communication phase (rounds 1–7), 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 Fig. 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 Fig. 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 Fig. 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.

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.

2.2. 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 can be found at https://osf.io/wht86/.

2.2.1. 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 ($\beta = 0.2$, SE = 0.06, z = 3.1, $p < 0.01$; see Table 1 and Fig. 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****

** < 0.01.

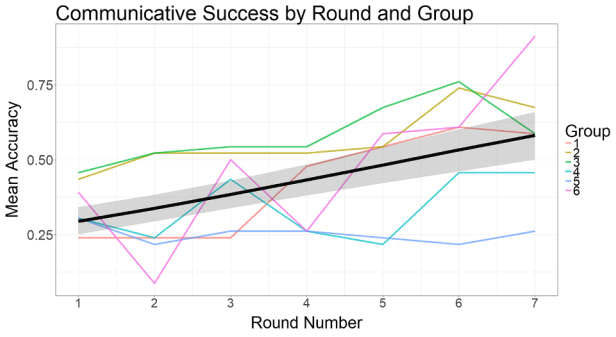

Fig. 2. Summary statistics of mean accuracy by Round Number. The colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represents the model's standard error. Round Number ranged from 1 (the first communication round) to 7 (the last communication round). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2.2. 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 Levenstein distances between all labels produced by different participants for that scene. The normalized Levenstein 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 ($\beta = 0.02$, SE = 0.01, t = 2, $p = 0.067$; see Table 2 and Fig. 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 Fig. 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**·

· < 0.1.
*** < 0.001.

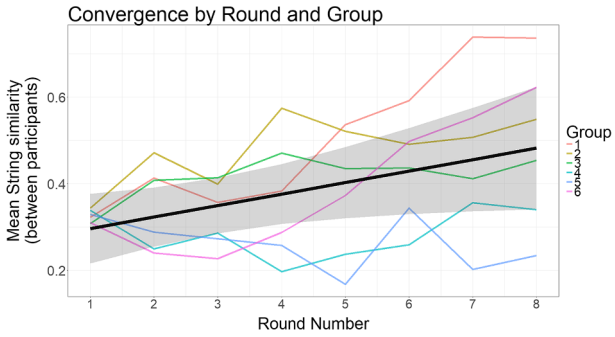

Fig. 3. Summary statistics of mean convergence by Round Number. Higher string similarities between participants indicate greater convergence. The different colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represents the model's standard error. Round Number ranged from 1 (first communication round) to 8 (the final test round). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2.3. 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 Levenstein 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 ($\beta = 0.028$, SE = 0.01, t = 2.19, $p = 0.06$; see Table 3 and Fig. 4). Interestingly, examining the rate of stabilization for scenes as they entered the game revealed that newer scenes stabilized faster (Fig. 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.

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**

$^{\cdot}$ < 0.1.
*** < 0.001.

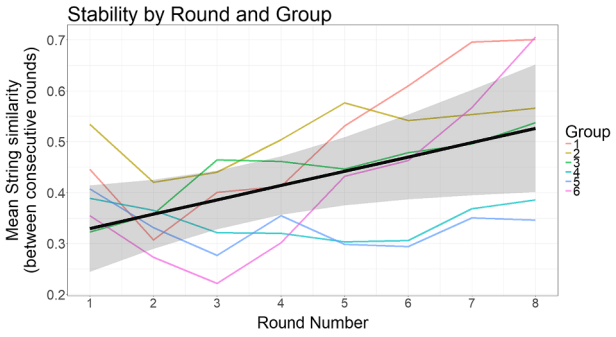

Fig. 4. Summary statistics of mean stability by Round Number. Higher string similarity between consecutive rounds indicate greater stability. The different colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represents 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

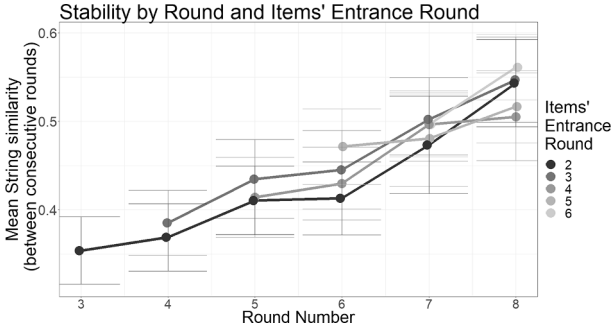

Fig. 5. Summary statistics of mean stability by Round Number and Items' Entrance Round for all labels that were introduced after the initial round. Higher string similarities between consecutive rounds indicate greater stability. Items' Entrance Round reflects the point in time at which the item was 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).

2.2.4. 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. The model showed that structure increased significantly over rounds ($\beta = 1.19$, SE = 0.1, t = 4.4, $p < 0.01$; see Table 4). 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 Fig. 6 shows, there was a high degree of compositional structure in this experiment, with some groups reaching correlations as high as 0.6.

Fig. 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 language experiments. Importantly, Stadler's segmentation algorithm identified all the sub-strings indicated in Fig. 7[2].

As can been seen, the language 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.

As can be seen in Fig. 7, 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.

Table 4
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

** < 0.01.
*** < 0.001.

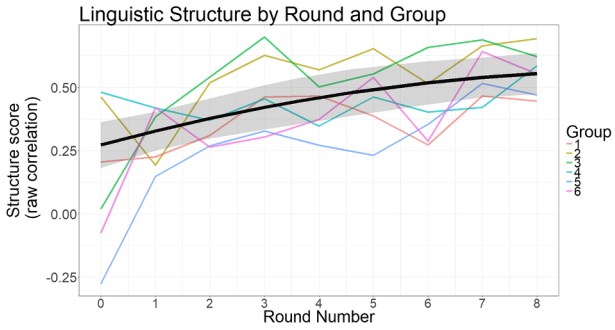

Fig. 6. Summary statistics of the label-meaning correlations by Round Number. The different colored lines represent the six groups. The black line represents the model's estimate for the effect of Round Number, and its shading represent the model's standard error. Round number ranged from 0 (the group naming phase) to 8 (the final test phase). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2.5. 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.[3]

3. 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.

3.1. Method

3.1.1. 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 six small groups of four participants, and

[2] Since the labels used to refer to Shape 2 had more variation in its final letter (i.e., "nena" and "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

[3] These 18 additional groups were run using the same paradigm to test other hypotheses (see Discussion) and will be reported elsewhere (Raviv, Meyer & Lev-Ari, under review). Importantly, this specific analysis is not reported anywhere else. We are happy to share the data from these additional groups upon request.

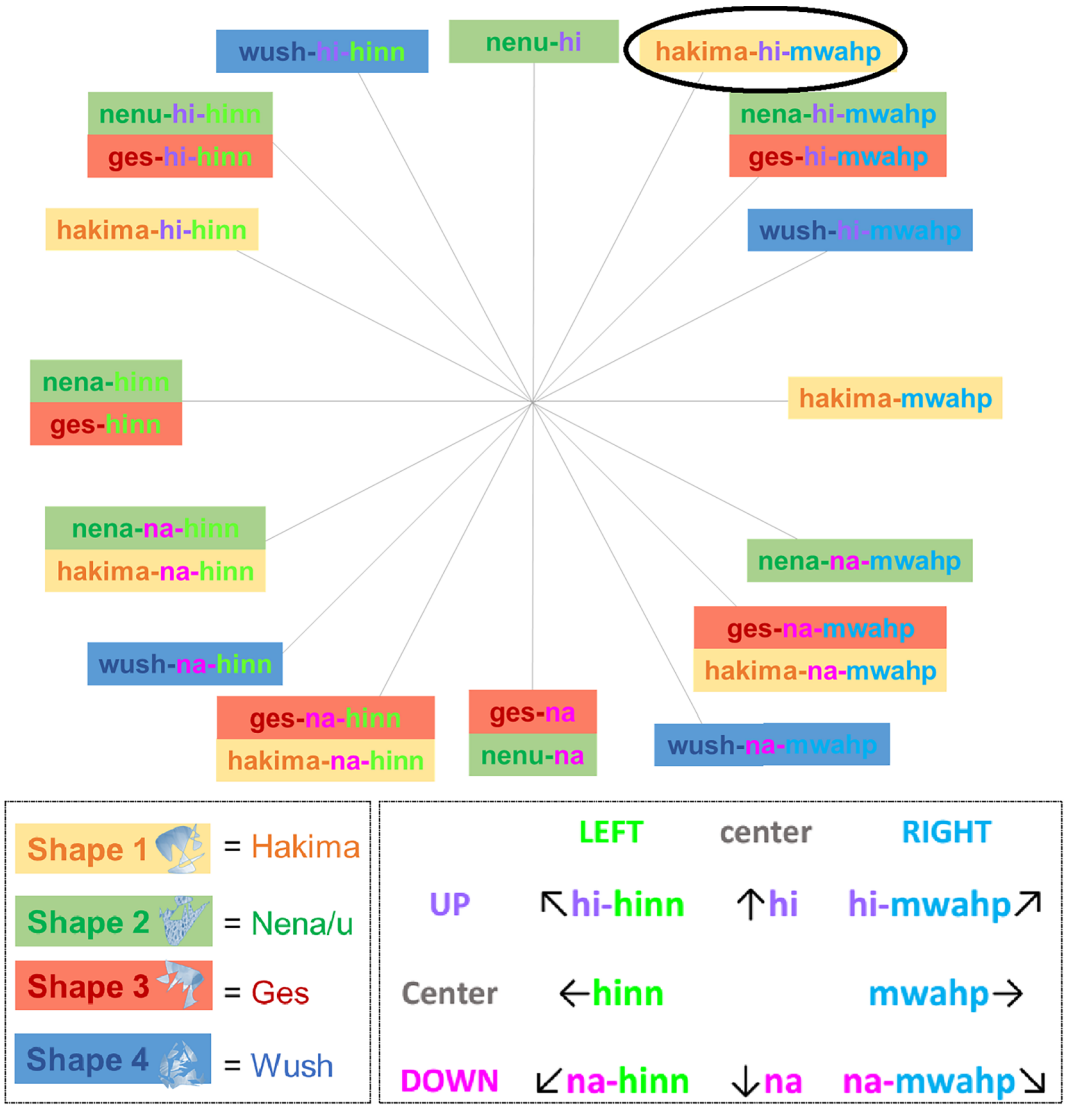

Fig. 7. An example of a compositional language, produced in the final test phase by a participant in Group 6, along with a "dictionary". Different box colors represent the four different shapes which appeared in the scenes, and the grey axes indicate the direction in which the shape was moving on the 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 60° 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".

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 and 2:45 h). 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 and 5:15 h, including a lunch break). All participants were native Dutch speakers and were recruited using the participant database of the Max Planck Institute for Psycholinguistics. Ethical approval was granted by the Faculty of Social Sciences of the Radboud University Nijmegen.

3.1.2. Stimuli

Identical to the stimuli used in Study 1.

3.1.3. 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).

3.2. 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 ($\beta = 4.65$, $SE = 0.3$, $t = 15.4$, $p < 0.001$), only the additional eight rounds ($\beta = 0.77$, $SE = 0.2$, $t = 3.7$, $p < 0.005$), or all 16 rounds together ($\beta = 5.6$, $SE = 0.4$, $t = 13.6$, $p < 0.001$). Notably, this increase in structure leveled off in later rounds: the quadratic term was significant during the first eight rounds ($\beta = -0.74$, $SE = 0.2$, $t = -3.6$, $p < 0.005$), and also when all rounds were taken into account ($\beta = -2.4$, $SE = 0.2$, $t = -11.1$, $p < 0.001$). 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 meanings, or both. First, we ran separate models that added to the model with Round 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 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 and 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).

Table 5
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 (quad)	**−0.8866**	**0.447**	**−1.9824**	**0.0481***

* < 0.05.
*** < 0.001.

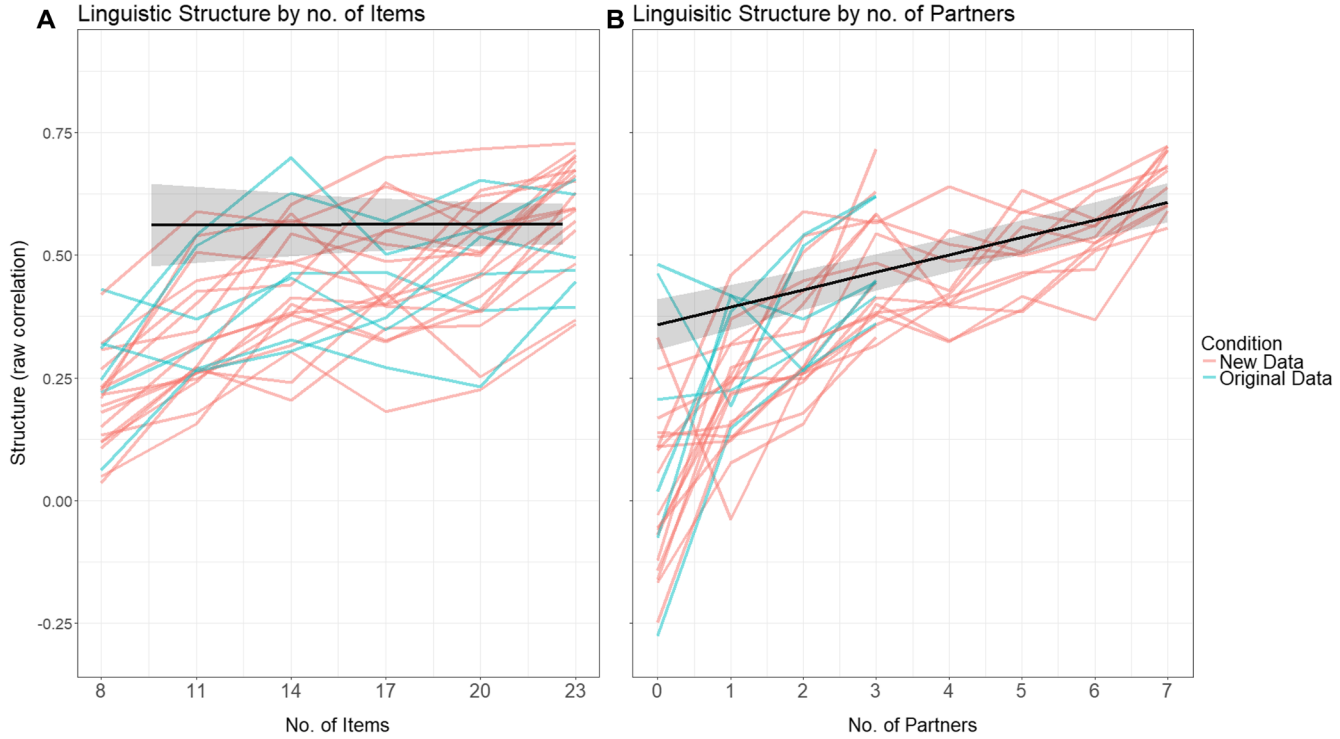

Fig. 8. Summary statistics of structure score by the number of items (A) and the number of partners (B) to which participants were exposed. The colored lines represent the different groups in the meta-analysis. The black line represents the models' estimate, and its shading represent the models' standard 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). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

All models showed both linear and quadratic effects of Round, 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 ($\beta = 0.04$, $SE = 0.005$, $t = 7.05$, $p < 0.001$), and the Number of Scenes did too albeit with a smaller effect size ($\beta = 0.007$, $SE = 0.002$, $t = 2.9$, $p < 0.01$). Importantly, the full model was favored compared to the model that included only the Number of Scenes ($\Delta AIC = 26$, $p < 0.0001$), but was similar to a model that included only the Number of Partners ($\Delta AIC = 2$, $p = 0.96$). 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 5; Fig. 8): When all factors were included in the model, structure scores significantly increased with the Number of Partners ($\beta = 0.04$, $SE = 0.006$, $t = 6.37$, $p < 0.001$) but not with the Number of Scenes ($\beta = 0.0001$, $SE = 0.002$, $t = 0.04$, $p = 0.96$). 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.

4. 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 (see Fig. 7). 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 et al., 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). At the moment, there is no published experimental work on differences in linguistic structure between groups of different sizes, but computational models (e.g., Dale & Lupyan, 2012; Reali & Griffiths, 2009) predict that it can have dramatic effects on linguistic structure. 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. We are currently examining how differences in population size and network configuration may affect the emergence of compositionality (Raviv, Meyer & Lev-Ari, under review).

5. 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

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

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2018.09.010.

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.
Arends, J., & Bruyn, A. (1994). Gradualist and developmental hypotheses. *Pidgins and creoles: An introduction: Vol. 15*, (pp. 111–120). John Benjamins Publishing.
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.
Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2016). lme4: Mixed-effects modeling with R; 2010. URL: < http://lme4 > .
Beckner, C., Blythe, R., Bybee, J., Christiansen, M. H., Croft, W., Ellis, N. C., ... 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.
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.
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.
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.
Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences,* 39.
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. *Proceedings of the 6th international conference on the evolution of language* (pp. 35–42). .
Deacon, T. (1997). *The symbolic species.* London: Penguin.
Dunbar, R. I. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*(4), 681–694.
Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science, 34*(3), 351–386.
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.
Galantucci, B., & Garrod, S. (2011). Experimental semiotics: A review. *Frontiers in Human Neuroscience, 5*, 11.
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.
Goldberg, A. E. (1999). The emergence of the semantics of argument structure constructions. *The Emergence of Language,* 197–212.
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., 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. 158–164). MIT Press.
Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). New York: 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.
Hockett, C. F. (1960). The origin of speech. *Scientific American, 203*(3), 88–97.
Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences, 3*(7), 272–279.
Kennedy, P. (1992). *A guide to econometrics.* Oxford: Blackwell.
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., 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. *Simulating the evolution of language* (pp. 121–147). London: Springer.
Kirby, S., Tamariz, M., Cornish, H., & Smith, K. (2015). Compression and communication in the cultural evolution of linguistic structure. *Cognition, 141*, 87–102.
Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5*(11), 831–843.
Labov, W. (2007). Transmission and diffusion. *Language,* 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.
Lev-Ari, S., & Shao, Z. (2017). How social network heterogeneity facilitates lexical access and lexical prediction. *Memory & Cognition, 45*(3), 528–538.
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.
Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PloS one, 5*(1), e8559.
Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.
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.
Motamedi, Y., Schouwstra, M., Smith, K., & Kirby, S. (2016). Linguistic structure emerges in the cultural evolution of artificial sign languages. *The evolution of language: Proceedings of the 11th international conference (EVOLANG11)* (pp. 493–495). .
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 B, 367*(1597), 1829–1836.
Nilsenová, M., & Nolting, P. (2010). Linguistic adaptation in semi-natural dialogues: age comparison. *Proceedings of the 13th international conference on text, speech and dialogue* (pp. 531–538). Berlin: Springer (September).
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.
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.
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, March 30–April 1, 2017*.
Raviv, L., Meyer, A. & Lev-Ari, S. (under review). Larger communities create more structured languages.
Reali, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition, 111*(3), 317–328.
Roberts, S., & Winters, J. (2012). Social structure and language structure: The new nomothetic approach. *Psychology of Language and Communication, 16*(2), 89–112.
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).
Samuelson, L. K., & Smith, L. B. (1999). Early noun vocabularies: Do ontology, category structure and syntax correspond? *Cognition, 73*(1), 1–33.
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.
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.
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.
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.
Verhoef, T., Walker, E., & Marghetis, T. (2016). Cognitive biases and social coordination in the emergence of temporal language. *The 38th annual meeting of the cognitive science society (CogSci 2016)* (pp. 2615–2620). .
Winters, J., Kirby, S., & Smith, K. (2015). Languages adapt to their contextual niche. *Language and Cognition, 7*(3), 415–449.
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.
Xu, J., Dowman, M., & Griffiths, T. L. (2013). Cultural transmission results in convergence towards colour term universals. *Philosophical Transactions of the Royal Society B, 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.
Zlatev, J. (2008). From proto-mimesis to language: Evidence from primatology and social neuroscience. *Journal of Physiology-Paris, 102*(1), 137–151.