Journal of Language Evolution, 2017, 1–16
doi: 10.1093/jole/lzx001
Research article

Research article

The emergence of linguistic structure in an online iterated learning task

Clay Beckner,[1,*] Janet B. Pierrehumbert[1,2,3] and Jennifer Hay[1,4]

[1]New Zealand Institute of Language, Brain & Behavior, University of Canterbury, Christchurch, New Zealand, [2]Oxford e-Research Centre, University of Oxford, UK, [3]Department of Linguistics, Northwestern University, Evanston, IL, USA and [4]Department of Linguistics, University of Canterbury, Christchurch, New Zealand

*Corresponding author: New Zealand Institute of Language, Brain & Behaviour, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand clayton.beckner@canterbury.ac.nz

Abstract

Previous research by Kirby et al. has found that strikingly compositional language systems can be developed in the laboratory via iterated learning of an artificial language. However, our reanalysis of the data indicates that while iterated learning prompts an increase in language compositionality, the increase is followed by an apparent decrease. This decrease in compositionality is inexplicable, and seems to arise from chance events in a small dataset (four transmission chains). The current study thus investigates the iterated emergence of language structure on a larger scale using Amazon Mechanical Turk, encompassing twenty-four independent chains of learners over ten generations. This richer dataset provides further evidence that iterated learning causes languages to become more compositional, although the trend levels off before the 10th generation. Moreover, analysis of the data (and reanalysis of Kirby et al.) reveals that systematic units arise along some meaning dimensions before others, giving insight into the biases of learners.

Key words: iterated learning; semantics; compositionality; morphosyntax; artificial languages.

1. Introduction

This article reports on a large-scale implementation of an iterated artificial language learning task. It makes two substantial contributions to the literature. First, it replicates previously reported results using a statistically much more robust dataset. Learners of our artificial languages show a significant tendency to increase compositionality. Second, it examines the different dimensions of meaning for which compositionality could emerge, showing that some aspects of meaning develop compositionality in advance of others.

The iterated learning paradigm (Kirby 2001; Kirby and Hurford 2002; Kirby et al. 2007) has provided a fruitful methodology for investigating the dynamics of language structure. The core feature of the paradigm is that an agent attempts to learn a language, after which that agent's output is passed on to a new learner, complete with errors or innovations. This transmission process is thereafter iterated. Originally developed as a computational, agent-based model of cultural transmission, iterated learning has been extended into a series of artificial language studies in the laboratory, in which a human participant learns an artificial language, then completes a test round; the results of the test then determine the language taught to the next generation (Cornish 2010; 2011; Cornish et al. 2009;

Kirby et al. 2008, 2015; Reali and Griffiths 2009; Smith and Wonnacott 2010).

Of particular interest to the current study is Kirby, Cornish & Smith (2008) (henceforth KCS), which demonstrated the development of compositional structure in a laboratory experiment, starting from random initial conditions and iterating across multiple generations. In this article, we follow KCS in using the term 'compositionality', to refer to a property of languages in which similar meanings tend to be expressed by similar forms, and complex meanings are expressed by combining these forms in a regular manner (Hockett 1960). Strict 'compositionality' requires both systematicity (e.g. there is a systematic, one-to-one correspondence between forms and basic meanings) and structure (complex meanings can be expressed by combining forms in accordance with grammatical principles of some sort). Kirby et al. (2008) found that artificial languages become increasingly compositional through iteration. As languages are learned by successive generations, the languages increasingly reuse basic forms to express consistent meanings, in different combinations. In KCS, structure arose without conscious intention by the participants; indeed, there was no motivation to willfully change the language, since participants were unaware of the iterated nature of the study. The KCS study provided a proof-of-concept that laboratory study of iterated learning is possible, and indeed some of the languages created by their chains of participants' exhibit striking compositional structure.

The goals of the present study are threefold. First, we use online data collection methods to 'scale up' the KCS investigation, and to solidify the statistical findings in that study regarding compositionality via iterated learning. Toward this end, the current study collects a large dataset using a game-based artificial language experiment on Amazon Mechanical Turk (Von Ahn 2006; Munro et al. 2010). The findings of the KCS study are of theoretical interest, since it is held that compositionality is a defining feature of human languages (e.g. Hockett 1960), although it is also known that idiomaticity and irregularity arise in natural languages over time (Paul 1890; Wray 2002; Vincent 2014; Cuskley et al. 2014). Given that competing tendencies exist in natural languages, how reliably will artificial languages develop compositionality, and how stable is it once it has developed? As a second goal, we investigate the mechanisms whereby compositionality enters a language, by examining the timecourse for its emergence across different semantic dimensions. When compositionality arises, is it driven by an across-the-board tendency toward increased structure? Or are such effects driven by individual components of meaning, and if so, which components? This type of systematic by-dimension analysis has not been performed in previous iterated learning studies. Cornish (2011: 129, n13) does comment on by-dimension trends in the final generation of iterated languages, including an observation that motion tends to be encoded systematically, whereas differences in color or shape are more often ignored, and underspecified in the language's forms. Silvey et al. (2015) investigate asymmetries in underspecification arising from experimental manipulations of semantic dimensions. However, we delve more broadly into comparing the systematicity of different meaning dimensions. We predict that when facing inconsistent form-meaning mappings, language learners will look for 'footholds' where systematic meaning is most expected, and systematicity will arise in these areas first.

Third, we investigate the effect of the size of the 'training bottleneck' in an iterated learning experiment. A common feature of iterated learning models is a data 'bottleneck': agents in the model are trained on a subset of the full range of items, but then tested on the full set. This feature of the model is, at the very least, required in order to observe agents drawing generalizations, although in some cases it is provided with a theoretical interpretation as a selective pressure that shapes the structure of language. Kirby et al. (2007) argue, based on computer simulations, that tighter bottlenecks are associated with increased regularity (i.e. less idiomaticity) as a result of pressures on learnability. However, an excessively tight data bottleneck would also render a language unlearnable, if it fails to represent a sufficient range of training tokens. Cornish (2010) suggests that linguistic structure arises if the data bottleneck is neither too narrow nor too wide. However, few studies with human participants have measured the effect of changing the size of the training bottleneck. Some variations in bottleneck size arise in KCS Experiment 2, but it is difficult to draw conclusions, since the variations are not controlled. Moreover, Cornish (2010; 2011) performs an iterated learning experiment in which *no* data bottleneck is present, which nevertheless leads to the development of systematicity in artificial languages. Cornish argues that participants' memory limitations themselves impose a processing 'bottleneck'; the difficulty of recalling twenty-seven unrelated training forms prompts the introduction of errors (i.e. innovations). The current study examines whether moderate differences in training set sizes (the 'training bottleneck') result in any observable differences across trials.

A number of commentaries have pointed out that the standard iterated learning approach (in which a speaker population consists of individual speaker/learners in succession) has limitations as a general model of language change (Croft 2004; Niyogi and Berwick 2009; Smith 2009; Beckner and Wedel 2009; Theisen-White et al.

2011). However, in the current study, we make no strong claims about the population dynamics of iterated learning, and do not argue that the model provides a one-to-one representation of language change in a group of speakers. Rather, iterated learning approaches provide a helpful empirical tool for exploring cognitive biases that underlie the learning and use of language. Iterating the process of language inputs and outputs provides a glimpse into how the effects of these biases can accrue over time.

1.1 Synopsis and reanalysis of findings by Kirby et al. (2008)

In this section, we review the design of the KCS experiments, and reconsider the quantitative findings to help to motivate the current work. Our quantitative analysis is made possible by publicly available tables of the raw KCS data, online at http://dx.doi.org/10.7488/ds/1586.[1]

For learners in the KCS experiments, participants attempt to learn linguistic expressions for a subset of items selected from a twenty-seven-item meaning space. The full meaning space consists of recombinations of three shapes, three colors, and three motion types. Participants are trained on approximately 50% of the test set; thus, on the first generation, participants are exposed to fourteen out of the twenty-seven possible meaning combinations. However, in the final test round, participants are tested on all twenty-seven items. Learners in the KCS study are not aware that others will be learning from their language output, and there are no pressures toward expressivity imposed on participants. As a result, in KCS Experiment 1, languages exhibit a tendency toward homonymy. That is, language learners tend to create identical forms for distinct stimuli, thus leaving certain dimensions of meaning underspecified. When left unchecked, these processes allow languages to evolve into degenerate states where underspecification runs rampant, such as one language in which twenty-six out of twenty-seven meanings are expressed with a single form (*nepa*).

To counteract such tendencies, in KCS Experiment 2 the training sets were 'filtered' by experimenters on each intergenerational transmission. Thus, in KCS Experiment 2, a pressure toward expressivity is enforced by (1) randomly selecting a training set (fourteen items from a participant's output), then (2) randomly removing all but one duplicate, before passing the training set to a new generation. Thus, in Experiment 2, learners encounter only training sets in which each form has a single meaning. However, the filtering of training data causes the training set size to fluctuate, ranging between eight and fourteen items (mean = 11.55, SD = 1.71).

1.1.1 Compositionality analysis of the KCS data

In both experiments, Kirby et al. find that across generations, languages *increase* in form-meaning compositionality, and *decrease* in intergenerational innovation/error. Quantitative details about the KCS compositionality and error metrics, respectively, are included in Texts S4 and S6 in this article's Supplementary Texts. However, we examine here how robust the evidence is that languages tend toward compositionality, focusing on the results in Experiment 2. That experiment gathered data from forty participants—four different transmission chains, with ten generations each. Since successive generations of the artificial language are not independent (the output of one generation forms the training input for the next), the KCS dataset constitutes just four trials of the paradigm. It is true that on the whole, compositionality increases across the four transmission chains, as shown by a *t*-test (with three degrees of freedom) comparing the initial conditions with the 10th generation. However, this *t*-test is not altogether satisfactory, for several reasons. Comparing only the beginning and end points of a history need not indicate that an overall tendency is at work in a dynamic system. Note that the experiment is initialized such that structure cannot *decrease*—that is, with zero structure—and even without directed change, a random walk model can exhibit an apparent increase from boundary conditions (cf. Gould (1988, 1996), on lower bounds of complexity in biological evolution, and the illusion of goal-directed progress). Moreover, compositionality does not increase monotonically across the ten generations in the study: out of forty trials in Experiment 2, compositionality increases twenty-three times, and decreases seventeen times. This noisiness suggests that more detailed reanalyses are appropriate.

We thus reanalyze the data from KCS Experiment 2 using linear mixed-effects models (Pinheiro and Bates 2000, Baayen 2008), implemented using lmer from the *lme4* package in R (Bates et al. 2015b). Mixed-effects models provide a meaningful way to group data, which is essential since iterated learning studies require repeated measures over the same transmission chain (see Winter and Wieling 2016). Mixed-effects models also take into account variation that exists across the population; the model allows for individual slopes to represent different effects of the model's predictors in different transmission chains. Alternate modeling approaches are possible; we discuss additional modeling details in Supplementary Text S5.

1 The raw data also appeared as a published supplement to KCS. However, the original distribution of the data contains a typographical error, in the 10th generation of the third chain (Experiment 2) (Kirby, pers. comm.). In the online materials linked above, this error has been corrected.

To allow for the possibility that compositionality does not increase linearly across the ten generation series, we investigate both linear terms (*generation*) and higher order terms (*generation* squared, *generation* cubed, etc.) (cf. Winter and Wieling (2016), while noting methodological differences below, and in Supplementary Text S5). Model selection using the Akaike information criterion (AIC) and likelihood ratio tests (Baayen 2008) leads to a quadratic model for the KCS Experiment 2 data.

To reduce the likelihood of Type I errors, we proceed with a maximal random-effects structure, that is, with random intercepts for each transmission chain, and chain-specific random slopes for each predictor (Barr et al. 2013). However, in some cases (as in this first model), the maximal model will not converge; as a minimal step toward parsimony, we thus do not require a correlation parameter between random slopes and intercepts (Bates et al. 2015a).

Table 1 shows the resulting mixed-effects model for the KCS Experiment 2 data. Figure 1a summarizes the raw compositionality scores that form the basis for this model, and Fig. 1b presents the predictions of the quadratic model from Table 1. The Fig. 1b model plot displays both overall estimates (in black) and chain-specific random effects (in gray); this plot makes use of R scripts made available in Winter and Wieling (2016). Note that the effects plotted in this figure (fixed effects, as well as chain-specific random effects) show model predictions, rather than raw compositionality values. While linear mixed-effects models incorporate variability across chains, random effects are '"pulled toward" the fixed-effects estimates', that is, there is 'shrinkage' of individual variation toward the population mean (Pinheiro and Bates 2000: 152; Winter and Wieling 2016).

Overall, the model clearly exhibits an increase in compositionality via iteration, corresponding to the positive, linear term for *generation*. Moreover, by the second generation of iteration, compositionality scores consistently surpass the threshold designating random structure (the horizontal line in Fig. 1). The model, however, also includes a negative quadratic term, corresponding to a downward curve across generations. In the fixed effects (as well as the random effects) of Fig. 1b, this downward curve actually suggests an overall decline in compositionality during the final generations of iteration.[2] Nevertheless, there is no particular theoretical explanation for such a downward trend, and it is likely an artifact of a small dataset. That is, two of the four transmission chains decline markedly in compositionality toward the end of the transmission chain, and the model is unduly influenced by these (presumably chance) events.

Table 1. Linear regression model of compositionality, for our reanalysis of KCS Experiment 2.

	Coef	Std.Error	*t*-value	P-value
(Intercept)	0.226	1.110	0.204	0.842
Generation	2.294	0.424	5.409	<0.001
Generation^2	−0.172	0.043	−3.949	<0.001

Model: compositionality ~ generation + generation^2 + (1|chain) + (0 + generation| chain) + (0 + generation^2 |chain)

Although the basic findings in KCS are highly noteworthy, we would argue that more data is altogether appropriate. The current study thus aims to replicate KCS, with a substantially larger amount of data.

1.1.2 Evaluating individual semantic dimensions in KCS

Next we consider an additional open issue in the KCS data. Earlier we noted that in KCS Experiment 1, participants tend to neglect some meaning dimensions, while others are encoded more systematically. This raises the question of whether there are any patterns of semantic biases across the three dimensions (shape, color, motion) in KCS Experiment 1, and whether related biases may be observed in Experiment 2. Thus we conduct a new analysis of the KCS data to consider this question, as follows.

When the meaning dimension is, for instance, *color*, first consider the level of 'red'. At generation *t* of a language, we calculate the normalized Levenshtein edit distance between all pairs of strings that both contain a meaning of 'red'. This edit distance is subtracted from 1 to represent string similarity, which provides an approximation of how systematically the language represents the meaning 'red'. This process is repeated for the two other levels of the dimension ('blue,' 'black'), and the three

2 Given that a quadratic term necessarily assumes a U-shaped curve, a possible concern is that the downward curve in Fig. 1 is merely an artifact of the polynomial model. To avoid inaccuracies in the tails of the data, *restricted cubic splines* offer an alternative nonlinear approach (Harrell 2001; Baayen 2008; Steyerberg 2009). In such an analysis, the data is divided into intervals (based on a specified number of knots); cubic polynomials are fit to each interval, and smooth transitions are imposed at each boundary. Using the *rcs* function (with three knots), from R's *rms* package (Harrell 2013), we performed a second analysis of the KCS data, and again found that a downward compositionality trend appears in later generations. Since the rcs models yield comparable results, we have elected to focus here on the more widely known approach using polynomials (Mirman et al. 2008; Winter and Wieling 2016). In general, both approaches suggest that the four chains of KCS Experiment 2 have insufficient statistical power.

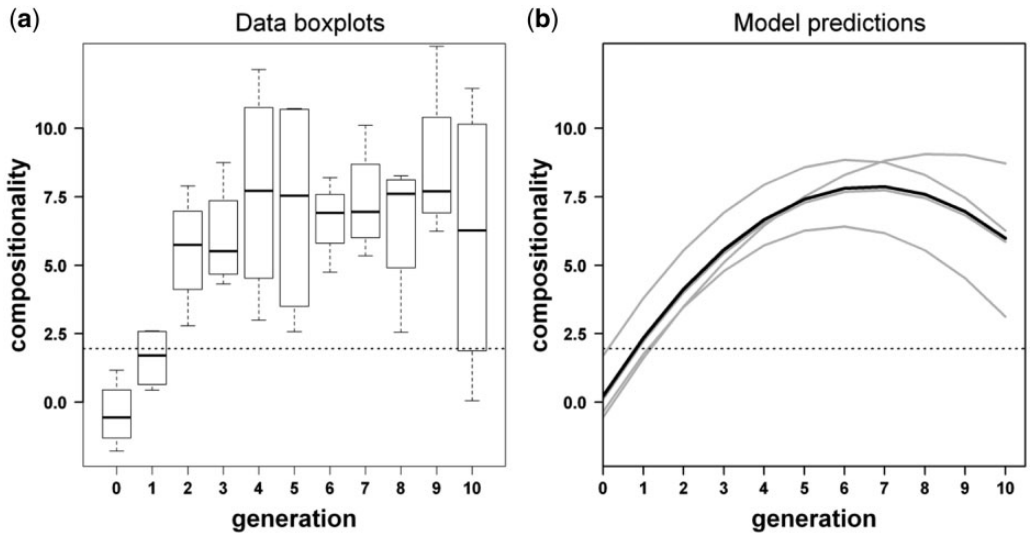

Figure 1. (a) Boxplots of language compositionality over ten generations (plus random initial state), summarizing scores from the four chains of KCS Experiment 2. (b) Plot of the predictions for the regression model shown in Table 1, representing compositionality scores in our first reanalysis of the KCS Experiment 2 data. Fixed effects are shown in black, with random-effect estimates (for the four individual transmission chains) shown in gray. In both plots, the horizontal dotted line represents a 95% confidence interval based on Monte Carlo simulations, corresponding to a threshold between random and compositional languages.

scores are averaged to represent how consistently the language encodes colors at generation *t*.

The same measure is then calculated for the other two dimensions of meaning (shape, motion). The resulting *within-category similarity score* is meaningful insofar as there may be differences between dimensions—that is, at a given generation *t*, some dimensions show more category-internal similarity than others.

The averaged by-dimension results of our reanalysis are shown in Fig. 2a; the average scores suggest that the *motion* dimension leads the way in the development of structure.

The evident pattern is indeed statistically significant, as borne out by a mixed-effects regression analysis, using within-category similarity as the dependent variable. The model's random effects include random intercepts for each transmission chain, with *dimension* nested by chain (since within each chain, similarity scores for individual dimensions represent a repeated measure of interest). The model also includes random slopes for *generation* and *generation^2*. The regression model is presented in Table 2, and a plot of the model's fixed effects is shown in Fig. 2b.

In the KCS data, within-category similarity generally increases, that is, generation is a significant predictor, with a small negative quadratic effect. Moreover, the *motion* of the object (*spiral, bounce*, or *horizontal*) significantly leads the other dimensions, as shown by a significant *generation* × *dimension* interaction. The generation interaction effects for *color* (*red, black*, or *blue*) and *shape* (*circle, square*, or *triangle*) are not significantly different from one another.

These analyses indicate that one dimension in particular (*motion*) tends to lead the way in the emergence of structure in the KCS artificial languages. We take this finding as evidence for a cognitive bias—involving particular attention to motion, and/or a predisposition to impose systematicity in linguistic representations of motion.[3] The reanalysis thus raises the question of whether a more general phenomenon is at work, and whether dimensional biases may be observed in an altogether different semantic space.

In the remainder of this article, we revisit these open questions with a new iterated learning experiment, drawing data from a much larger group of participants.

2. Methods

Our experiment is in many respects fashioned after Experiment 2 in KCS, and has been adapted to online data collection methods using AMT. AMT has proven to be a reliable platform for behavioral research, and provides

3 The bias, however, is not likely to involve any *perceptual* salience of motion; note that in the KCS experiments, movement is not conveyed directly via animations, but is rather represented abstractly with arrows. This element of the KCS design does mean that the semantics of motion in the experiment are not entirely clear; the arrows could also be regarded as static shapes. In part as a result of this uncertainty, we avoided motion as a semantic dimension (see Section 2.1), although animated motions remain a topic for future study.

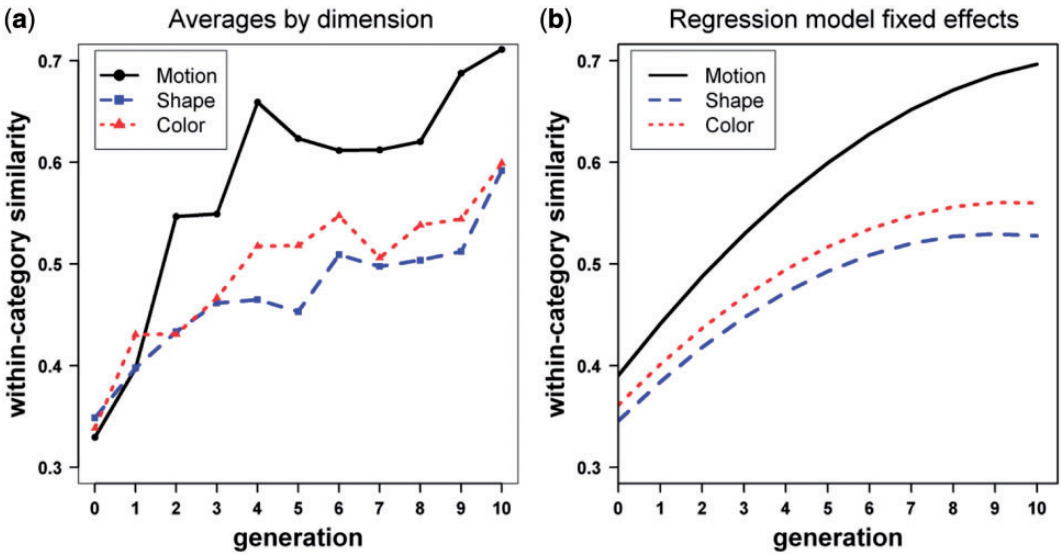

Figure 2. Reanalysis of data from KCS Experiment 2, showing the timecourse of development for within-category similarity for different dimensions of meaning (color vs. shape vs. motion). (a) Amount of string similarity within each semantic dimension, averaged across the four chains of the KCS experiment. (b) Regression estimates (showing fixed effects only) for the three dimensions, using the model presented in Table 2.

Table 2. Mixed-effects regression model for amount of within-category string similarity, in data for our reanalysis of KCS Experiment 2, by dimension across generations. Dimension is a categorical variable, with shape as the reference level.

	Coef	Std.Error	*t*-value	P-value
(Intercept)	0.346	0.040	8.679	<0.001
Generation	0.041	0.008	5.178	<0.001
Generation^2	−0.002	0.001	−2.824	0.009
Dimension = *color*	0.015	0.036	0.431	0.676
Dimension = *motion*	0.044	0.035	1.248	0.242
Generation: dim = *color*	0.001	0.005	0.347	0.736
Generation: dim = *motion*	0.012	0.005	2.529	0.031

Model: in.category.similarity ∼ generation × dimension + generation^2 + (1|chain/dimension) + (0 + generation|chain/dimension) + (0 + generation^2|chain/dimension).

resources for the recruitment of a large and diverse population of participants (Munro et al. 2010; Crump et al. 2013). At the same time, online data collection using AMT presents challenges that are not present when participants are highly educated and motivated university students tested in the lab (as in KCS). The elaborations of the original paradigm described here reflect extensive pilot testing designed to overcome these challenges. We wanted to ensure completion of the task within a maximum time frame and minimize the number of dropouts, while also automatically eliminating the few participants who respond at random without following instructions.

Several innovations of our experiment's setup are similar to (but developed independently from) approaches used in Carr (2012). That study failed to find evidence of increased compositionality via iteration, but introduced several variations on the original KCS design. Our current experiment shares with Carr (2012) (1) a redefined meaning space, which is focused on unambiguously noun-like units; (2) a considerably broader syllable repertoire in the initial state, and (3) the use of a fixed training set size.

2.1 Meanings

As in KCS, participants use artificial languages within a 3 × 3 × 3 meaning space, but the dimensions of meaning are as follows. Referent stimuli encompass three unfamiliar, 'alien' objects (shapes), in three colors (red, green, and blue), and three different numerical configurations (one, two, or three instances of the object). The alien objects used in the current experiment are illustrated in Fig. 3. In discussions later in this text, these alien objects are given the shorthand English names (from left to right) of 'berry', 'phone', and 'key'.

It is arguable whether the linguistic units of interest in this study are morphological or syntactic in nature. The referent stimuli are objects and their properties; the relevant linguistic units are nominal, and could alternately be viewed as a noun phrase, or as a noun stem with affixes. By comparison, recall that the meaning space of the KCS experiments involves types of *motion* for items differing in *shape* and *color*. The productions in the KCS experiments are thus presumably (morpho-)syntactic, involving a subject and a predicate.

The linguistic forms in the current study, as in the KCS experiments, appeared with no internal spaces. The

Figure 3. Illustration of the 'alien' item shapes used in different configurations (varying in number and color) in the experiment. Graphics: Visual Voice (vvlab.net).

instructions to participants generally took measures to avoid referring to a response as a 'word' or a 'name', so as to not specify what type of unit was being learned.[4] In this article, we will generally refer to training forms and test responses as 'linguistic forms', on the assumption that each response could be an assemblage of morphemes, an assemblage of words, or both.

2.2 Forms at initial generation

Linguistic forms were initialized randomly so as to contain no structure. Initial-generation forms were created by concatenating three syllables generated with the following grammar:

$$\text{Syllable} \rightarrow \{\text{t, k, s, v}\}\{\text{i, a, o, u}\}\{\text{n, l, } \varnothing\}$$

This artificial grammar generates linguistic forms such as *vilkantin, tinkalsol*, and *kalvonsi*. The language initialization differs from that in KCS by having forms of approximately constant length, but comprised of a broader range of different syllables. These changes in the protocol were made to avoid possible artifacts from the KCS initialization method; see Supplementary Text S3 for further discussion.

For each run of the experiment, twenty-seven linguistic forms were selected at random from the set of all possible forms, and assigned randomly to the twenty-seven meaning configurations. For different versions of the experiment, the training set size was set to be either $N = 12$ or $N = 15$ items. For the first generation of players, N items were randomly chosen out of the full set of twenty-seven; these constituted the training items for first-generation participants.

All initial-generation lexicons are available as part of this article's Online Supplementary Data.

4 As an exception, the instructions do refer in some instances to 'words', in contexts where the plurality allows for ambiguity about whether an individual response contains more than one word.

2.3 Presentation of stimuli

Participants completed the experiment as a game called 'Teleporters', offered as a paid assignment for workers on AMT. An alien avatar on the screen asked players to attempt to learn an alien language; the full game instructions are presented in Supplementary Text S1.

Players were instructed that learning the language would allow them to 'teleport' to new landscapes throughout the game. Items to be learned were initially presented in a hide-and-seek phase, so as to draw players' attention to the form of the object(s); items had to be located in the midst of the current game landscape. Once the training item was clicked, it was enlarged and presented alongside the appropriate alien linguistic form, for 6 seconds. After every three items, participants were presented with a review 'quiz', to promote attentiveness to the form-meaning pairings (cf. similar forced-choice review strategies in Tily et al. (2011)). After each training block, participants also completed an interim task requiring open-text responses. Full details regarding the sequence of training rounds are provided in Supplementary Text S2.

At the end of the experiment, a 'final exam' asked players to provide open-text responses in the alien language, to all twenty-seven items in the set (including both previously seen and previously unseen items). As in KCS, the instructions never alert participants to the fact that they are being tested on items which they have not encountered during training. However, the instructions do encourage participants to do their best to 'guess' answers ('Just pick the first answer that comes into your mind'). Participants were allowed 24 seconds in which to enter a response.

2.4 Filtering and iteration

The results of each participant's final test were used as the basis for the language to be learned by the next participant in the chain. As in KCS, participants in successive generations were not informed that the language they were learning had been produced by other players. (However, as part of the consent process, participants were informed that their answers during the game could be used to create new versions of the game.)

Before transmission to the next generation, the output languages were adjusted as follows. As in KCS, the training set for each new learner comprised a subset of the language output from the previous generation. Thus, the output of one player's final test is randomly sampled to select the 'seen' items for the next participant in the chain.

However, some selection processes were imposed by experimenters at the intergenerational stage. As in KCS Experiment 2, no two identical forms were allowed in any

training set, to avoid a tendency toward underspecification. If identical forms occurred in the sample, one duplicate was randomly selected to occur in the 'seen' set, and all others were randomly replaced with other items currently in the 'unseen' set.

On this point, the procedures in our study for randomly selecting training items are different from those of KCS. In KCS Experiment 2, when identical forms occurred (and happened to be randomly selected for the next generation), all homonyms but one was removed from the training set. However, these filtered items were not replaced, thus leading to variable training set sizes (dipping as low as eight items out of twenty-seven) depending on how often the language repeated the exact same form. In the current study, we imposed a requirement that training languages must include a fixed number of training items (either $N = 12$ or $N = 15$, out of a total space of twenty-seven possible meaning combinations). Thus, if homonyms were randomly selected as training items for the next generation, all instances but one were assigned to the 'unseen' set. For each removed item, we then selected replacement candidates at random from the same language (while continuing to disallow identical forms as candidates), and used these as training items for the next generation.

We introduced a fixed size in our experiment for several reasons. First, we set out to investigate the effect of controlled adjustments to the training set size, given previous arguments that the 'bottleneck size' may be important to the emergence of structure (Kirby et al. 2007). Additionally, imposing consistency across different chains and generations is helpful for interpretive purposes. One of the KCS metrics—the amount of intergenerational change, by generation—is difficult to evaluate meaningfully if the amount of training input fluctuates. The amount of intergenerational change will inevitably increase if participants encounter a smaller portion of the previous generation's language; holding the training set size provides a consistent measure by condition.

Holding the training set size constant imposes an additional layer of data filtering beyond the filtering implemented by KCS. In our experiment, in cases where the output language contained fewer than N items, we discarded the output and reran the exact same experiment setup with a new participant. Thus, in effect, the filtering processes in our current approach imposes more than one selection pressure on the output. As in KCS, there is a pressure to avoid homonymy, since all participants' output is filtered. Second, the procedure removes altogether any output from participants who have more severe tendencies toward homonymy. However, the removal of at least some participants is unavoidable if we impose the requirement of a fixed training set size, while also filtering homonyms.

2.5 Participants

Participants were paid $3 for completing the experiment as an assignment on AMT. The experiment was limited to native speakers of English, aged 18 and older. Physical presence within the USA was verified by limiting IP addresses via Mechanical Turk, and native language background was verified via participants' self-report in a pre-game questionnaire. The Mechanical Turk assignment was set up such that each participant (identified by a Mechanical Turk ID) could complete the experiment only once.

The target dataset for this experiment includes 12 transmission chains $\times$ 10 generations $\times$ 2 training set sizes, for a total of 240 participants. Thus, there were a total of twenty-four distinct transmission chains initialized with random form-meaning mappings (the zeroth generation); these structureless languages were provided as training for twenty-four participants, and the process iterated for a total of ten generations of participants.

3. Results

Each training set condition ($N = 12$, $N = 15$) had a target of 120 participants. However, the requirement of a fixed training set size led to the rejection of some candidate participants, and replacement with new participants prior to iteration, since these participants failed to provide a sufficient number of distinct answers to be used on the next generation. For the runs with a training set size of twelve, eleven participants were discarded for having fewer than twelve unique linguistic forms in the final testing round (rejection rate: 8.40%). Among participants with a training set of fifteen, sixteen were discarded for providing fewer than fifteen unique forms (rejection rate: 11.76%).

Two participants were replaced for failing to follow experiment instructions. In the training-set-twelve group, we discarded one subject who commented on the task (in English) in seven out of twenty-seven responses, with answers such as 'idontknow'. In the training-set-fifteen group, one participant was discarded for entirely responding with English answers such as 'threebluewrenches'. In addition to the foregoing, a number of participants elected not to complete the experiment. In the training-set-twelve condition, thirty-nine participants began the study but dropped out before finishing (22.8% dropout rate). In the training-set-fifteen condition, sixty-seven participants began the study but dropped out before finishing (32.8% dropout rate).

For the trials with a training set of twelve items, the final set of 120 participants consisted of 72 women and 48 men. The average age of participants in this group is 35 (SD = 9.86). This version of the experiment lasted an average of 22.23 minutes (SD = 3.18).

For the trials with a training set of fifteen items, the final set of 120 participants consists of 56 women and 64 men. The mean age is 35.91 (SD = 11.03); two participants declined to provide their age. The experiment lasted an average of 25.33 minutes (SD = 3.25).

Participant responses in the final test round comprise a total of 6,480 open-text responses (3,240 responses per training-set condition). The full set of responses (twenty-four iterated language histories) may be obtained in this article's Online Supplementary Data.

Note that in a few instances in the dataset, participants failed to provide a response within the 24-second timeout (see Section 2.3). Such nonresponses are flagged as 'NA' in the Supplementary Data. With respect to iteration, these 'NA' items were automatically assigned to the 'unseen' set for the next generation. In the training-set-twelve condition, a total of five nonresponses occurred (out of 3,240 entries); in one instance, two different NAs occurred in a single participant's response set. In the training-set-fifteen condition, a total of nine nonresponses occurred (again out of 3,240 entries). In this condition, one participant provided two different NA responses, and a different participant provided three different NA responses. The four remaining NAs were the only nonresponses for the participant.

3.1 Compositionality

We quantify language compositionality using the same methods as KCS and Cornish (2011), which adapts a test developed by Mantel (1967). The Mantel score is devised so as to quantify the relationship between forms and meanings in a language; across different linguistic items in a compositional language, similarities in meaning should correspond to similarities in form. Monte Carlo methods are used to determine a threshold for identifying likely compositional languages. These quantitative methods are discussed in greater detail in Supplementary Text S4.

Summaries of the compositionality scores across ten generations (plus the random initial state) are presented in the boxplots of Fig. 4. Monte Carlo investigations (see Supplementary Text S4) confirm that as intended, none of the twenty-four languages in the current study have significant compositional structure at the randomly initialized zeroth generation.

As in the original KCS data, variability exists in participants' tendencies to increase or decrease language compositionality from the previous generation. For each of the training size conditions, there are 120 cross-generational transitions (12 chains × 10 generations, including a transition from random initial conditions). In the twelve-item training size condition, compositionality increases 61% of the time (73 participants out of 120). In the fifteen-item training size condition, compositionality increases 56% of the time (67 participants out of 120).

However, the plots in Fig. 4 show that as a general trend, the form-meaning correlations rise above the 95% confidence interval using Monte Carlo methods, and visually suggest an overall increase in compositionality over time (i.e. over successive generations). The structure does seem to emerge somewhat slower than in earlier work. In

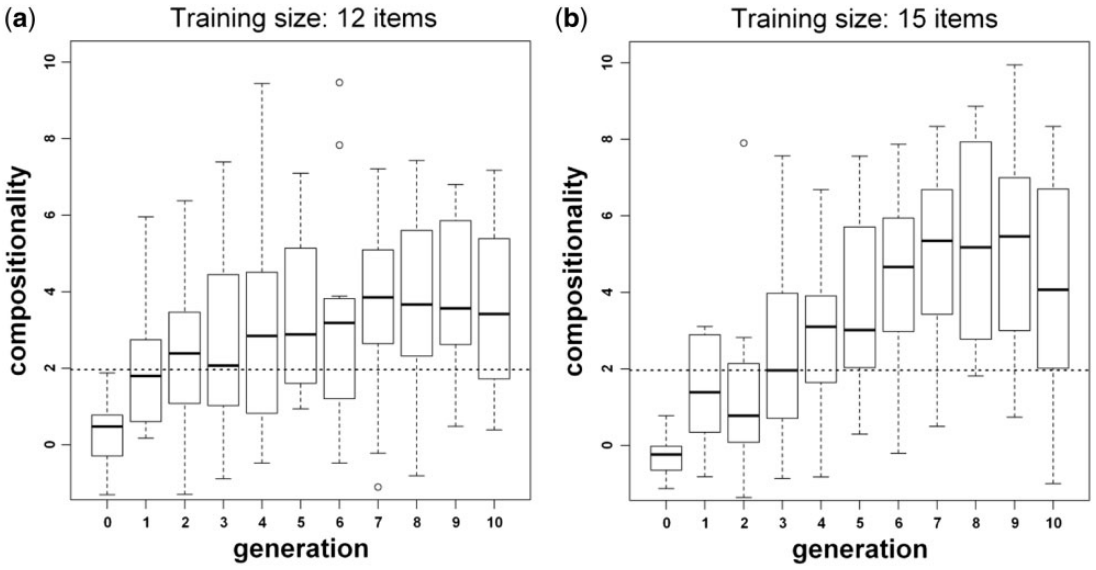

Figure 4. Boxplots of language compositionality over ten generations (plus random initial state). The plots summarize twelve chains using a training set size of twelve (Fig. 4a), and twelve chains using a training set size of fifteen (Fig. 4b). The compositionality metrics displayed are *z*-scored with respect to 1,000 randomized rearrangements of the linguistic forms and meanings. The horizontal dotted line at 1.96 represents the upper bound of the 95% confidence interval for a Monte Carlo simulation. That is, 95% of all random reshufflings of each language result in compositionality scores below this line.

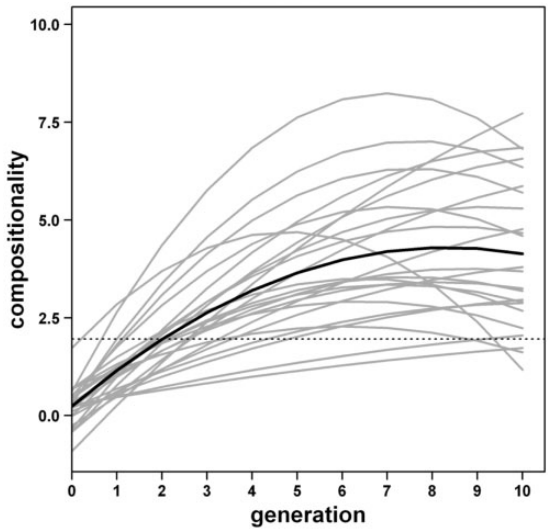

Figure 5. Plot of the compositionality predictions for the model shown in Table 3, based on our dataset of twenty-four transmission chains. Fixed effects are shown in black, with random effects (for the twenty-four individual chains) shown in gray. The horizontal dotted line represents a 95% confidence interval based on Monte Carlo simulations, corresponding to a threshold between random and compositional languages.

KCS Experiment 2, there was a rapid increase in compositionality on early iterations; nonrandom structure was evident on two of the four chains by the first generation, and on all four chains by the second generation. In contrast, in the current dataset, structure seems to be slower to 'get off the ground,' as is evident by examining generations zero, one, and two in Fig. 4. The slow rise above the nonrandom threshold is due to several chains that persisted in structurelessness in early generations; indeed, one chain (with a training set size of twelve) continued for five generations without surpassing the 95% confidence interval for structured mappings. However, across the twenty-four chains in the current experiment, 75% had attained nonrandom structure by the third generation, and 83% by the fourth generation.

The general trend for increasing compositionality in the current dataset can be verified, again using mixed-effects linear regression. Our model investigates the effect of *generation*, as well as higher order *generation* terms to allow for nonlinearity. We performed stepwise regression, examining interactions between *generation* terms and *training size* (*N*). Training set size is not significant, either as a main factor or interaction (with linear or higher order terms), and thus it is dropped from the model. We include random intercepts for each transmission chain, and random slopes for generation and generation^2. The resulting regression model is presented in Table 3, and fixed and random effects are plotted in Fig. 5.

Table 3. Mixed-effects regression model for compositionality

	Coefficient	Std.Error	*t*-value	P-value
(Intercept)	0.240	0.285	0.845	0.403
Generation	0.973	0.157	6.216	<0.001
Generation^2	−0.058	0.014	−4.050	<0.001

Model: compositionality ∼ generation + generation^2 + (1 + generation + generation^2|chain).

Several things are of note in this model. First, the model intercept is not significantly different from zero; this means that in the initial state (when *generation* and *generation^2* are zero), the compositionality is effectively zero. Given our initialization of languages with random structure, this aspect of the model is as expected. However, some variation in intercepts is evident in the random effects, which is consistent with slight variations in the initial conditions for our twenty-four chains. Moreover, the model gives evidence of an overall increase in compositionality via iteration, represented by the positive coefficient for *generation*. However, there is also a small downward curve, represented by the negative value for *generation^2*. In contrast with the smaller KCS dataset (and the rather puzzling result in Fig. 1b), the model of the current dataset does not exhibit a downward trend toward the end of ten generations.[5] However, as shown in Fig. 5, the current analysis also does not suggest an ongoing upward trend in compositionality via iteration; rather, compositionality increases, then reaches a plateau by the seventh generation.

The foregoing analyses provide evidence for an overall increase in language structure across iterated generations. This finding should be accompanied by an acknowledgement of just how noisy these data are. As a general trend, learners indeed build on the advances of previous generations, but it is not uncommon for structure to be lost altogether. Out of 240 trials, there are twenty-five instances in which the structure in a language drops below the 95% confidence threshold for identifying nonrandom languages, even though the language's immediate predecessor

5 We note, however, that the fixed-effect curve in Fig. 5 suggests an almost imperceptible downward trend. That is, if we take the curve literally, and extrapolate beyond the 10th generation, it implies that eventually compositionality will decline markedly. However, as discussed in n2, an unfortunate feature of quadratic models is that they require trends to be parabolic. In the case of this dataset, we would argue that the exhibited trend is for all practical purposes, horizontally asymptotic, and the suggested downward trend is an artifact of polynomial regression. We have verified this impression using a restricted cubic spline analysis (with three knots), which demonstrates that in our dataset, compositionality indeed increases, then levels off.

contained structure, following the same metric. That is, in around 10.4% of trials, participants are trained on a language containing some degree of structure, but during testing all significant compositionality has been lost. By way of comparison, we also note that the KCS datasets also contain trials in which all existing structure is lost, although the rates are lower (7.5% and 2.5% for Experiments 1 and 2, respectively). These striking failures offer a reminder that in this iterated system, each timepoint in a transmission chain is entirely subject to the innovations and errors of an individual participant. These errors are, in turn, subject to the vicissitudes of sampling, and potential lapses in participant attention. Thus, while general trends are as predicted, the compositionality arcs for individual languages over time are far from monotonic.

In sum, the current study's dataset replicates the basic finding of KCS that artificial languages increase compositionality via iterated learning. On a separate point, our dataset replicates an additional KCS result: the amount of transmission error decreases across generations in an iterated chain. This result is not central to the present article's research questions, but for purposes of comparison to KCS we provide details of the analysis in Supplementary Text S6.

3.2 Dimensions of meaning

With respect to increases in compositionality, we now consider in closer detail the ways in which structure arises and persists over time. Consider the language output presented in Table 4. This data is drawn from the ninth generation. (A full history of this language is presented in the Supplementary Data, data training size fifteen, Chain 4).

The hyphens in this table are inserted to mark what is a likely morpheme (or word) boundary. Such divisions are supported the by the offline analyses of five linguists who offered their expertise on this language's likely structure; all five proposed that *shen-*, *div-*, and *lolni-* were meaningful units.[6] This language clearly exhibits structure; indeed, this particular language snapshot represents one of the peaks in compositionality using the metric described in 3.1 (with a computed score of 9.94). However, note that structure does not seem to be evenly distributed across the three dimensions. The shape of the alien object is encoded with almost perfect consistency: *shen-* means 'berry', *div-* (or *dev-*) means 'key', and *lolni-/lolne-* means 'phone'. However, the other dimensions are not encoded in the same way. With respect to number, 'three' is often encoded with *-stra*, although this is not uniform. In one case, *-stra* means 'two' (and there is a very similar *-stro* 'two'), and in three cases, 'three' is expressed with *-trio*.[7] Color does not seem to be encoded at all.

Table 4. Sample output from a run of the Teleporters game, taken from the ninth generation

	'red'	'green'	'blue'	
'berry'	shen-to	shen-ta	shen-to	'1'
	shen-tra	shen-tro	shen-tra	'2'
	shen-trio	shen-trio	shen-trio	'3'
'key'	div-tro	div-tro	div-tro	'1'
	dev-tro	dev-tro	dev-etrio	'2'
	dev-stra	div-stra	dev-stra	'3'
'phone'	lolni-tro	lolni-tro	lolni-to	'1'
	lolne-stra	lolni-tro	lolne-stro	'2'
	lolni-tra	lolni-stra	lolni-stra	'3'

Hyphens are inserted into linguistic forms in the table to aid readability, and to suggest the *de facto* presence of a morpheme/word boundary. However, these hyphens did not appear in the input to (or output from) the actual experiment participant.

More systematically, we can investigate these dimensional asymmetries over time by considering each meaning dimension (*shape*, *color*, *number*) separately, and determining how much within-category similarity exists at different generations. We use the metrics introduced in Section 1.1.2—string similarity is determined across all linguistic forms that share a level for a dimension (e.g. similarity between strings is calculated for all forms containing a meaning of 'red', and so on for 'blue' and 'green'), and the scores are averaged by dimension. Since the current dataset has some missing values, as noted, missing forms are omitted from edit distance calculations and averaging.

The resulting by-dimension mean scores are summarized in Fig. 6a, combining the two training size conditions.

These figures suggest that within-category similarity for *shape* outstrips the within-category similarity for *color* and *number*. There is indeed a significant effect for the shape dimension, compared with other dimensions on the same generation, as verified by linear regression. We perform stepwise linear regression, with within-category similarity as the dependent variable and semantic dimension,

6 More precisely, our query to these linguists asked them to propose slightly idealized versions of the language output, which we explained might contain erroneous transcriptions or lexical exceptions.

7 There are hints of English in the uses of *-trio* in this chain—but its use is inconsistent, and it should be noted that *-trio* is also used in an item with the meaning 'two'. We note that use of quasi-English or mnemonic innovations, whether or not intentional by participants, is a risk in this methodology. For instance, there are possible parallels in the artificial language used as an example by KCS for Experiment 2 (all 'red' items have a prefix of *r-*, 'black' items all have a *n-*prefix (cf. *noir, night, nocturnal*)).

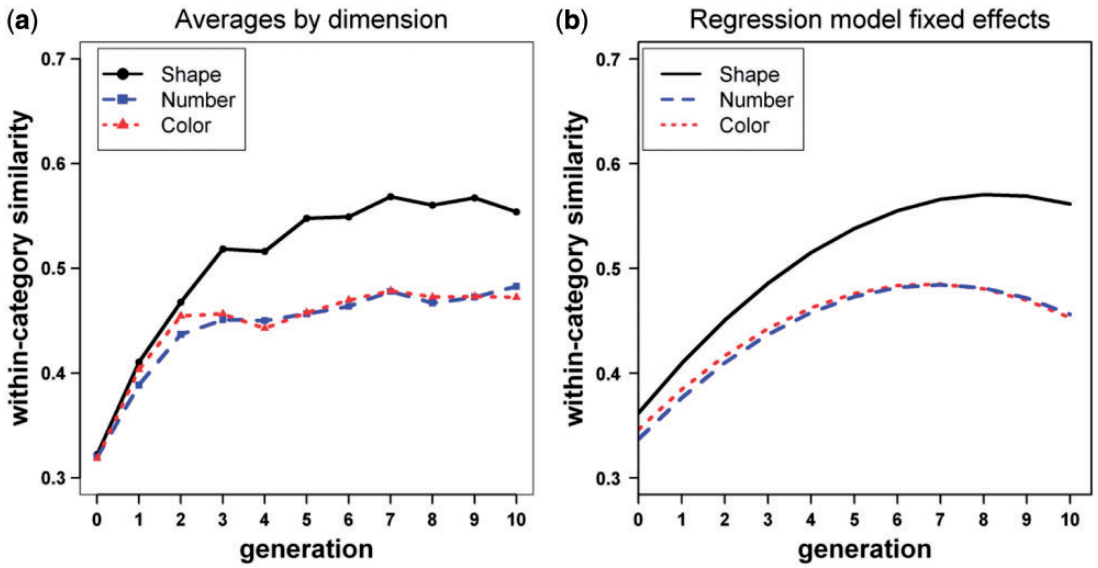

Figure 6. Timecourse of development for within-category similarity for different dimensions of meaning (color vs. shape vs. number), for the twenty-four transmission chains in our experiment. (a) Amount of string similarity within each semantic dimension, averaged across the twenty-four chains of our experiment. (b) Regression estimates (showing fixed effects only) for the three semantic dimensions, drawn from the model presented below in Table 5.

generation, and training set size as independent variables. Training set size is not significant, and thus omitted from the model. The regression model is presented in Table 5.

This model indicates that within-category similarity generally increases over successive generations, as evident from the significant main effect for *generation*. Follow-up regression models of subsets of the data verify that this finding is true for all three meaning dimensions: generation is significant for *shape*, *number*, and *color*, in separate analyses. However, *shape* is different from the other two dimensions, and the model indicates a positive interaction with *generation*. *Number* and *color* are not significantly different from one another. All dimensions exhibit a small downward curve in the fixed-effects estimate, which may be characterized as a leveling-off of within-category similarity by the seventh generation.

Table 5. Mixed-effects linear regression model for amount of within-category string similarity in experiment data, by dimension across generations. In this model, *color* is the reference (default) level; generation:dim = *shape* is significantly different from generation:dim = *color*, but generation:dim = *number* is not significantly different from the default. A separate model with re-ordered categories confirms that generation:dim = *shape* is significantly different from generation:dim = *number*.

	Coef	Std.Error	*t*-value	P-value
(Intercept)	0.346	0.009	37.242	<0.001
Generation	0.041	0.004	11.126	<0.001
Generation^2	−0.003	0.000	−9.658	<0.001
Dimension = *number*	−0.009	0.011	−0.863	0.391
Dimension = *shape*	0.015	0.011	1.438	0.155
Generation: dim = *number*	0.001	0.002	0.508	0.614
Generation: dim = *shape*	0.009	0.002	3.817	<0.001

Model: within.category.similarity ∼ generation × dimension + generation^2 + (1|chain/dimension) + (0 + generation|chain/dimension) + (0 + generation^2 |chain/dimension).

4. Discussion

The current study replicates findings by Kirby et al. (2008), and verifies the utility of iterated learning as a tool for investigating participants' cognitive biases. Our study collects data from a substantially larger participant pool, increasing sixfold the size of the KCS Experiment 2 dataset. The present study's participant pool is, moreover, more demographically diverse than the pool of university students included in the KCS study. Replicating a laboratory study in this way has merit, given the widespread tendency for behavioral experiments to be limited to demographically homogeneous participant pools (Wintre et al. 2001; Henrich et al. 2010).

The current study demonstrates that online data collection is quite feasible in iterated learning experiments. The demands of implementing iterated learning experiments in the laboratory are formidable; the availability of a large online subject pool on AMT was indispensable to scaling up the KCS study.

As in KCS, participants in our study effected an increase in language compositionality, by learning a language that required them to generalize to previously unseen items. The

dataset provides evidence for a cumulative increase across iterated generations of participants. That is, participants tend to build on the progress of earlier generations in developing language systematicity. The data gives evidence that participants have cognitive biases that lead to compositionality, and that artificial languages increase in structure across multiple generations. This is hardly to say that language users always tend toward increasing compositionality. Clearly, there are various biases at work in language change, including, for instance, economy of forms and distinctiveness (Lindblom 1990; Bybee 2007; DuBois 2014). These biases may amplify one another, or compete with one another; additional iterated learning research is required to study the interactions between compositionality and other biases (cf. Kirby 2001; Kirby et al. 2015).

Moreover, in the current study, the *failures* in increasing compositionality are instructive. Although the study's artificial languages do, in general, become increasingly structured via experimental iteration, we note that in a surprising number of cases (10.4%), participants produce a language completely lacking in compositional structure, despite having been trained on a language with some degree of structure. By comparison, the complete loss of compositionality is also observed in the KCS data, but the rate of loss is lower (2.5% for the more easily comparable Experiment 2). We also observed that in the current experiment, structure is slower to emerge in early generations of the transmission chains, that is, there are outliers in which structurelessness persists across several generations.

These patterns hint at methodological considerations regarding adapting behavioral experiments to AMT. Although laboratory experiments are generally found to be replicable on AMT, research also suggests that difficult experimental tasks are disproportionately more challenging for these participants, compared with participants in a controlled laboratory setting (Crump et al. 2013). The artificial language learning task in the current study is indeed quite challenging: the participant attempts to learn a language that typically exhibits limited structure (and may hint at altogether inconsistent analyses), and the final test requires generalizing to previously unseen items. It may be that a higher failure rate is to be expected on AMT, where there are fewer controls on participant distractions, and the participant population is more variable. On the other hand, it is also possible that the noisy trends in our data more accurately represent the population at large, given that our sample is six times as large as that of KCS Experiment 2. That is—noting that the KCS results were easily swayed by chance failures in later generations—it might be that their four chains also benefitted from fortuitous successes early in iteration.

Despite the general increase in compositionality, it hardly seems that compositionality is being maximized in these iterated languages. Indeed, the nonlinear curve of our models' fixed effects does not give evidence of boundlessly increasing compositionality; rather, structure increases for several generations, but then reaches a plateau. In an offline morphological analysis of the data, we can readily identify ways that compositionality could, in principle, have continued to increase. For example, in the language presented in Table 4, there are many inconsistencies present, despite hints of systematicity. Morphologists who review this data propose more idealized, consistent expressions of number (such as *-to* for 'one', *-tro* for 'two', *-stra* for 'three'). Moreover, none of the chains in our experiment come close to any 'ceiling' in compositionality. For a $3 \times 3 \times 3$ meaning space, a perfectly compositional language would receive a z-score of approximately 20 using the KCS metric (see Supplementary Text S4 for details on this point). By comparison, the highest compositionality scores in our data never rise above 10, and in KCS Experiment 2, they never rise above 13. Of course, natural languages are hardly perfectly compositional, since idiomaticity is widespread, and many complex meanings are expressed lexically rather than aggregating component morphemes. However, the artificial languages produced by participants are often unsystematic in seemingly random ways, or give evidence of error in recall.

It is important to note that in the current study (as in KCS Experiment 2), languages' expressivity is imposed by the experimenters' filtering of output, rather than via communicative demands. Adapting one's output for a presumed interlocutor should provide stronger incentives for participants to reuse meaningful units systematically. That is, speakers in interaction must rely on language components they believe to be *shared* when conveying meaning, which provides additional motivation for compositionality (Vincent 2014). Iterated learning studies by Berdicevskis (2012) and Kirby et al. (2015) have demonstrated that the functional pressures of having a communicative partner can, when combined with learnability requirements, lead to the emergence of compositional languages. Future research will investigate whether the intrinsic motivations of communicating with an interlocutor (in contrast with merely having one's language output winnowed) prompts a more persistent upward trend in compositionality.

Of course, in order for learners to advance compositionality further, it may also be necessary to provide not just incentives, but also to allow relevant resources, that is, more time and training. We note that, on the one hand, participants' innovative errors are an important mechanism in iterated learning dynamics; on the other hand, in the face of a challenging language learning task, many of these

errors prove to be unsystematic. On this point, the role of the training set size is in need of further investigation. Our analyses in the current study do not indicate any significant effect from changing the size of the 'data bottleneck', that is, varying the training set size between twelve or fifteen items out of twenty-seven. This null result does not provide any evidence for a strong theoretic interpretation of the data bottleneck, that is, one in which the amount of (non-) exposure is linked to the development of generalizations (cf. Kirby et al. 2007). However, the amount of difference between our experiment conditions is small (11% of the total meaning space), and this difference may be too small to prompt observable differences. Moreover, it may be that, as Cornish (2010; 2011) suggests, a larger training set also prompts the innovation of systematicity due to the difficulty of recalling more items. For human participants (in contrast with earlier computational simulations), there may be a trade-off between having less training (which prompts errors due to insufficient exposure) and having an overabundance of input (which prompts recall errors). Further experimental work is necessary, but it may turn out that the training set size is not of crucial importance in iterated learning methodologies with human participants.

The most interesting findings in the current study involve the particular mechanisms whereby compositionality enters artificial languages. Detailed analyses of the experimental data reveal that compositionality arises asymmetrically with respect to different semantic dimensions. In initial generations of this iterated learning experiment, participants encounter training input that is highly unstructured and inconsistent: there are no mappings between linguistic forms and meanings. Yet, analysis across the set of transmission chains shows that participants do not restructure this language at random to fit a compositional template; rather, they increase structure differentially by semantic dimension. In the present experiment, forms representing the *shape* are generally at the forefront of compositionality.

The literature on the 'shape bias', first described as a feature of word learning among children, suggests a more general foundation for this particular pattern. When young learners (starting around age 2.5) learn a new word for an item, they then extend that new word to other objects which have the same shape, as opposed to other properties such as color or size (see Baldwin (1989); Diesendruck and Bloom (2003); Jones et al. (1991); Landau et al. (1988); Smith et al. (1996); Smith and Samuelson (2006); though see Cimpian and Markman (2005) for an opposing view). A similar shape bias has also been demonstrated to be active among adults (Gentner 1978; Tek et al. 2012). The predisposition to generalize on the basis of shape is particularly high when the items in question are highly dissimilar (Tek et al. 2012), and in the present experiment, the alien 'berry', 'key', and 'phone' objects bear little resemblance to one another. The alien objects may also suggest different functions (such as a food, a tool, and a communication device). The design of the study is not sufficiently fine-grained to evaluate various suggestions about the conceptual and/or functional explanation for the shape bias (Gentner 1978; Booth et al. 2005; Smith and Samuelson 2006).

The foregoing considerations address the semantic asymmetries observed in the nominal units of this study's artificial languages. However, it is likely there is also a broader generalization at work. Recall that in our reanalysis of the KCS data (Section 1.1.2), we also found that systematicity developed with one semantic dimension leading the way. In a parallel regression analysis, we determined that *motion* is systematized in advance of the other two dimensions. There is no significant difference between the other two dimensions. For KCS, these are *color* and *shape*.

Given the shape bias evident in the current article's experiment, it is rather surprising that in the KCS data, systematicity for shape does not develop in advance of systematicity for color. Nevertheless, it does appear that across artificial language experiments, emerging systematicity is organized around a central dimension where learners most expect meaningful units—the type of *motion* in the KCS data, and item *shape* in the current experiment.

We propose that the results may be unified if grammatical relations are taken into account. In the sentence-level KCS languages, systematic representation emerges first with respect to motion 'verbs'. In the nominal units of the current study, systematicity first develops in representing which object shape is being referred to; this shape most closely corresponds to a noun (or noun root), with other semantic dimensions (number, color) functioning as modifiers. In phrasal terms, the 'verb' would be expected to be the head of the unit in the KCS case, and the 'noun' the head of the unit in the current dataset (De Marneffe et al. 2006, 2014; De Marneffe and Manning 2008). It appears that such phrasal heads provide a focus for learners attempting to organize structure in inconsistent language systems, and this focus is evident as dimensional biases in the emergence of structure.

This article has provided a large-scale replication of the widely cited iterated artificial language learning paradigm. With a large statistically robust dataset, we see that, as previously reported, successive generations have a tendency to increase compositionality in artificial languages. This tendency is not manifest equally across all aspects of the language, however. Systematicity develops much earlier for some semantic dimensions than others. This finding points toward future research in linguistic typology which would

be amenable to iterated learning studies, to determine the degree to which these dimensional biases extend to other semantic domains.

Supplementary Data

Supplementary data and Supplementary text materials are available at *Journal of Language Evolution* online.

Acknowledgements

The authors are grateful for help from Chun-Liang Chan, Patrick LaShell, Daniel Gerhard, Péter Rácz, Márton Sóskuthy, Jeremy Needle, R. Alexander Schumacher, Kayo Takasugi, and Simon Kirby. This manuscript also benefitted from feedback from Kenny Smith and two reviewers for *Journal of Language Evolution*.

Conflict of interest statement. None declared.

Funding

This project was made possible through the support of a subaward under a grant to Northwestern University from the John Templeton Foundation (Award ID 36617). J.H. and C.B. were also supported by a Royal Society of New Zealand Rutherford Discovery Fellowship (Grant Number E5909) awarded to J.H. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation or the Rutherford Foundation.

References

Baayen, R. H. (2008) *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Baldwin, D. A. (1989) 'Priorities in Children's Expectations about Object Label Reference: Form Over Color', *Child Development*, 60: 1289–306.

Barr, D. J. *et al*. (2013) 'Random Effects Structure for Confirmatory Hypothesis Testing: Keep It Maximal', *Journal of Memory and Language*, 68/3: 255–78

Bates, D. *et al*. (2015a) Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Bates, D. *et al*. (2015b) 'Fitting Linear Mixed-Effects Models Using lme4', *Journal of Statistical Software*, 67/1: 1–48. doi: 10.18637/jss.v067.i01.

Beckner, C., and Wedel, A. (2009) 'The Roles of Acquisition and Usage in Morphological Change', *Proceedings of the Berkeley Linguistics Society*, 35: 1–12

Berdicevskis, A. (2012) 'Introducing Pressure for Expressivity into Language Evolution Experiments' in T. C. Scott-Phillips, M. Tamariz, E. Cartmill & J. R. Hurford (eds) *The Evolution of Language: Proceedings of the 9th International Conference* (EVOLANG9), Kyoto, Singapore: World Scientific. pp. 64–71.

Booth, A. E. *et al*. (2005) 'Conceptual Information Permeates Word Learning in Infancy', *Developmental Psychology*, 41/3: 491.

Bybee, J. (2007) *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Carr, J. W. (2012) 'The Effects of Modified Variables on an Iterated Learning Model of Linguistic Evolution by Cultural Transmission', in T. C. Scott-Phillips *et al*. (eds) *The Evolution of Language: Proceedings of the 9th International Conference* (EVOLANG9), Kyoto, pp. 416–17. Singapore: World Scientific. doi:10.1142/9789814401500_0058.

Cimpian, A., and Markman, E. M. (2005) 'The Absence of a Shape Bias in Children's Word Learning', *Developmental Psychology*, 41/6: 1003–19.

Cornish, H. (2010) 'Investigating How Cultural Transmission Leads to the Appearance of Design Without a Designer in Human Communication Systems', *Interaction Studies*, 11/1: 112–37.

Cornish, H. (2011) 'Language Adapts: Exploring the Cultural Dynamics of Iterated Learning', Unpublished doctoral dissertation, University of Edinburgh, Edinburgh, UK.

Cornish, H. *et al*. (2009) 'Complex Adaptive Systems and the Origins of Adaptive Structure: What Experiments Can Tell Us', *Language Learning*, 59(S1): 187–205.

Croft, W. (2004) 'Form, Meaning and Speakers: Commentary on Kirby, Smith and Brighton', *Studies in Language*, 28/3: 608–11.

Crump, M. J. *et al*. (2013) 'Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research', *PLOS ONE*, 8/3: e57410.

Cuskley, C. F. *et al*. (2014) 'Internal and External Dynamics in Language: Evidence from Verb Regularity in a Historical Corpus of English', *PLOS ONE*, 9/8: e102882.

De Marneffe, M. C. *et al*. (2006) 'Generating Typed Dependency Parses from Phrase Structure Parses' Nicoletta Calzolari et al. in *Proceedings of the Language Resources and Evaluation Conference*, LREC 2006, European Language Resources Association, pp. 449–54.

De Marneffe, M. C., and Manning, C. D. (2008) 'The Stanford Typed Dependencies Representation' in *COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pp. 1–8. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1608858

De Marneffe, M. C. *et al*. (2014) 'Universal Stanford Dependencies: A Cross-Linguistic Typology' Nicoletta Calzolari et al. in *Proceedings of the Language Resources and Evaluation Conference*, LREC 2014, European Language Resources Association, pp. 4585–92.

Diesendruck, G., and Bloom, P. (2003) 'How Specific is the Shape Bias?', *Child Development*, 74/1: 168–78

DuBois, J. W. (2014) 'Motivating Competitions' in B. MacWhinney *et al*. (eds.) *Competing Motivations in Grammar and Usage*, pp. 263–81. Oxford, UK: Oxford University Press.

Gentner, D. (1978) 'What Looks Like a Jiggy but Acts Like a Zimbo? A Study of Early Word Meaning Using Artificial Objects', *Papers and Reports on Child Language Development*, 15: 1–6.

Gould, S. J. (1988) 'Trends as Changes in Variance: A New Slant on Progress and Directionality in Evolution', *Journal of Paleontology*, 62/3: 319–29.

Gould, S. (1996) *Full House: The Spread of Excellence from Plato to Darwin*. New York: Harmony House.

Harrell, F. E. (2001) *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Harrell, F. E. (2013) rms: Regression modeling strategies. *R package version 4.3-0*.

Henrich, J. *et al*. (2010) 'The Weirdest People in the World?', *Behavioral and Brain Sciences*, 33/2–3: 61–83.

Hockett, C. F. (1960) 'The Origin of Speech', *Scientific American*, 203: 5–12

Jones, S. S. *et al*. (1991) 'Object Properties and Knowledge in Early Lexical Learning', *Cognitive Development*, 8: 113–39.

Kirby, S. (2001) 'Spontaneous Evolution of Linguistic Structure: An Iterated Learning Model of the Emergence of Regularity and Irregularity', *Evolutionary Computation, IEEE Transactions on Evolutionary Computation*, 5/2: 102–10.

Kirby, S. *et al*. (2007) 'Innateness and Culture in the Evolution of Language', *Proceedings of the National Academy of Sciences*, 104/12: 5241–5.

Kirby, S. *et al*. (2008) 'Cumulative Cultural Evolution in the Laboratory: An Experimental Approach to the Origins of Structure in Human Language', *PNAS*, 105/31: 10681–6.

Kirby, S. *et al*. (2015) 'Compression and Communication in the Cultural Evolution of Linguistic Structure', *Cognition*, 141: 87–102.

Kirby, S., and Hurford, J. R. (2002) 'The Emergence of Linguistic Structure: An Overview of the Iterated Learning Model' in A. Cangelosi and D. Parisi (eds) *Simulating the Evolution of Language*, pp. 121–47. London: Springer-Verlag.

Landau, B. *et al*. (1988) 'The Importance of Shape in Early Lexical Learning', *Cognitive Development*, 3/3: 299–321.

Lindblom, B. (1990) 'Explaining Phonetic Variation: A Sketch of the H & H Theory' in W. Hardcastle and A. Marchal (eds) *Speech Production and Speech Modeling*, pp. 403–39. Dordrecht: Kluwer Academic.

Mantel, N. (1967) 'The Detection of Disease Clustering and a Generalized Regression Approach', *Cancer Research*, 27: 209–20.

Mirman, D., *et al*. (2008) 'Statistical and Computational Models of the Visual World Paradigm: Growth Curves and Individual Differences', *Journal of Memory and Language*, 59/4: 475–94.

Munro, R. *et al*. (2010) 'Crowdsourcing and Language Studies: The New Generation of Linguistic Data' in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 122–30. Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=1866696 accessed 22 March 2017.

Niyogi, P., and Berwick, R. C. (2009) 'The Proper Treatment of Language Acquisition and Change in a Population Setting', *PNAS*, 106/25: 10124–9.

Paul, H. (1890) *Principles of the History of Language*, trans. H.A. Strong. College Park: McGrath.

Pinheiro, J., and Bates, D. (2000) *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Reali, F., and Griffiths, T. L. (2009) 'The Evolution of Frequency Distributions: Relating Regularization to Inductive Biases Through Iterated Learning', *Cognition*, 111/3: 317–28.

Silvey, C. *et al*. (2015) 'Word Meanings Evolve to Selectively Preserve Distinctions on Salient Dimensions', *Cognitive Science*, 39 /1: 212–26.

Smith, K. (2009) 'Iterated Learning in Populations of Bayesian Agents' in *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pp. 697–702. Austin, TX: Cognitive Science Society.

Smith, K., and Wonnacott, E. (2010) 'Eliminating Unpredictable Variation Through Iterated Learning', *Cognition*, 116/3: 444–9.

Smith, L. B. *et al*. (1996) 'Naming in Young Children: A Dumb Attentional Mechanism? ', *Cognition*, 60/2: 143–71.

Smith, L. B., and Samuelson, L. (2006) 'An Attentional Learning Account of the Shape Bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005)', *Developmental Psychology*, 42/6: 1339–43.

Steyerberg, W. W. (2009) *Clinical Prediction Models*. Berlin: Springer-Verlag.

Tek, S. *et al*. (2012) 'The Shape Bias is Affected by Differing Similarity Among Objects', *Cognitive Development*, 27/1: 28–38.

Theisen-White, C. *et al*. (2011) 'Integrating the Horizontal and Vertical Cultural Transmission of Novel Communication Systems' in L. Carlson (ed.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, Cognitive Science Society, pp. 956–61.

Tily, H. *et al*. (2011) 'The Learnability of Constructed Languages Reflects Typological Patterns' in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, pp. 1364–9.

Vincent, N. (2014) 'Compositionality and Change' in C. Bowern and B. Evans (eds) *The Routledge Handbook of Historical Linguistics*, pp. 103–23. New York: Routledge.

Von Ahn, L. (2006) 'Games with a Purpose', *Computer*, 39/6: 92–4.

Winter, B., and Wieling, M. (2016) 'How to Analyze Linguistic Change Using Mixed Models, Growth Curve Analysis, and Generalized Additive Modeling', *Journal of Language Evolution*, 1/1: 7–18.

Wintre, M. G. *et al*. (2001) 'Psychologists' Response to Criticisms about Research Based on Undergraduate Participants: A Developmental Perspective', *Canadian Psychology*, 42/3: 216–25.

Wray, A. (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.