



Compression and communication in the cultural evolution of linguistic structure



Simon Kirby^{a,*}, Monica Tamariz^a, Hannah Cornish^b, Kenny Smith^a

^a School of Philosophy, Psychology and Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

^b School of Psychology, University of Stirling, Stirling, United Kingdom

ARTICLE INFO

Article history:

Received 7 February 2014

Revised 9 February 2015

Accepted 29 March 2015

Available online 14 May 2015

Keywords:

Cultural transmission

Language evolution

Iterated learning

ABSTRACT

Language exhibits striking systematic structure. Words are composed of combinations of reusable sounds, and those words in turn are combined to form complex sentences. These properties make language unique among natural communication systems and enable our species to convey an open-ended set of messages. We provide a cultural evolutionary account of the origins of this structure. We show, using simulations of rational learners and laboratory experiments, that structure arises from a trade-off between pressures for compressibility (imposed during learning) and expressivity (imposed during communication). We further demonstrate that the relative strength of these two pressures can be varied in different social contexts, leading to novel predictions about the emergence of structured behaviour in the wild.

© 2015 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Language is unique among the communication systems of the natural world in exhibiting rich combinatorial and compositional structure. Our species can productively construct novel signals on the fly by recombining reusable meaningless elements (speech sounds) to form meaning-bearing units (morphemes and words) which are further recursively combined. Furthermore, the meanings of these complex utterances are derivable in a predictable way from the composition of their subparts. The precise way in which this *combinatorial* and *compositional* structure is realised differs from language to language and is part of the knowledge that each language learner must acquire. Nevertheless, the existence of this kind of systematicity is both universal to all languages – it is one of the fundamental design features of human language

(Hockett, 1960) – and largely absent in the communication of other species.¹

Understanding the origins of this structure is a central goal of cognitive science. A recent productive approach treats it as a consequence of cultural evolution (Christiansen & Chater, 2008). Languages, in common with many other human behaviours, persist through a repeated cycle of learning and production: individuals learn a language by observing the linguistic behaviour of their speech community, and the linguistic behaviour they subsequently produce shapes learning in others. Languages potentially change and evolve as a result of their

¹ Some aspects of this universality remain controversial. For example, the evidence for recursion in some languages is limited (Everett, 2005), although such languages remain open-ended in their productivity. Equally, some aspects of the suite of structural design features can be found in a limited way in other species (Collier, Bickel, van Schaik, Manser, & Townsend, 2014). Bird song is combinatorial (Berwick, Okanoya, Beckers, & Bolhuis, 2011), and bee dance is compositional in a limited sense (von Frisch, 1975). However, bee dance is not culturally transmitted like human language, and neither birds nor bees can exploit these properties to communicate an open-ended set of meanings.

* Corresponding author.

E-mail address: simon@ling.ed.ac.uk (S. Kirby).

transmission, adapting to the biases inherent in the processes of language learning and language use.

In this paper we present computational and experimental models of the processes of language transmission which show that structure (specifically, compositionality) arises from cultural evolution when language is under pressure to be both learnable and expressive: language learning by naïve individuals introduces a pressure for simplicity arising from a domain-independent bias for compressibility in learning, and a pressure for expressivity arises from language use in communication. Crucially, both must be in play: neither pressure alone leads reliably to structure. The structural design features of language are a solution to the problem of being compressible and expressive, a solution delivered by the process of cultural evolution.

1.1. Compressibility and expressivity in language design

The idea that key features of language arise from the trade-off between competing pressures has a long history. Competing motivations of speaker and hearer, for instance, have been a rich explanatory tool for cognitive scientists (e.g. Zipf, 1949; Ferrer i Cancho & Solé, 2003; Piantadosi, Tily, & Gibson, 2012) and linguists seeking explanations for typological universals of language (e.g. Givón, 1979; DuBois, 1987; Kirby, 1997; Jäger, 2007): for example, utterances in a language will tend to minimise effort for the speaker as long as distinctiveness for the hearer is not compromised (Zipf, 1949). This kind of observation can be couched in terms of *compression*, i.e., optimisation of a repertoire of signals such that the energetic cost of unambiguously conveying any meaning is minimised. This leads naturally to the inverse relationship between frequency and length of words identified by Zipf (1936); more generally, it has been suggested that such optimally-compressible signal inventories are a universal feature of natural communication systems across all species (Ferrer i Cancho et al., 2013).

The fact that language is compositional and combinatorial – that it has system-wide structure – also means that languages as whole systems are compressible, i.e., allow the formation of compressed representations. We commonly refer to these representations as *grammars*, which are concise descriptions of the generative system underlying a language. These are compressed to the degree that they are more concise than a simple listing of all the possible utterances in a language. Note that this notion of compressibility is orthogonal to the compressibility of signals themselves.² For example, regular morphological paradigms are highly systematic and therefore highly compressible, but this potentially comes at the cost of less efficient signals, since exploiting unsystematic irregulars might allow shorter forms (e.g., “ran” is shorter than “runned” but leads to a more complex, less compressible morphological paradigm).

For our purposes, it will be useful to consider the compressibility of three classes of languages: *holistic* languages, lacking any of the system-level structure (e.g. compositionality) that characterises natural languages; *structured* languages, which exhibit system-level structure (e.g. where aspects of meaning reliably co-occur with sub-parts of signals); and *degenerate* languages, in which every meaning is associated with a single, shared, maximally ambiguous signal.³ Holistic languages are incompressible: the most concise encoding of a holistic language would be a dictionary that simply listed every signal paired with its meaning, i.e., the ‘grammar’ of this language would simply recapitulate the language in its entirety. Structured languages, in contrast, permit some compression: a grammar which captured the systematic regularities of such a language would be considerably shorter than a dictionary of all the signals in the language. Finally, degenerate languages are maximally compressible, since the entire language can be captured by a single rule stating the identity of the ambiguous signal. Following, e.g., Chater and Vitanyi (2003) and Kemp and Regier (2012), we assume that learners are naturally biased towards simpler, compressible languages, in line with the notion that a preference for simplicity is a fundamental cognitive principle: languages which permit the formation of compressed mental representations are easier to learn than those which do not.

As highlighted by Kemp and Regier (2012), the most compressible languages are not necessarily useful for communication: in particular, a degenerate language is highly compressible but not *expressive*, since it does not allow a speaker to discriminate an intended referent from possible alternative referents in a context. In contrast, less compressible languages (e.g. holistic or structured languages) are expressive to the extent that they provide a unique and unambiguous signal for every meaning. As demonstrated by Regier and colleagues for a range of cases (kinship categories, colour terms, numeral systems: Kemp & Regier, 2012; Xu & Regier, 2014; Regier, Kemp, & Kay, 2015), natural language lexicons exhibit a near-optimal trade-off between these two pressures, being among the most expressive and yet compressible of all possible systems. However, showing that language is near-optimal with respect to these two pressures does not provide an explanatory mechanism for this striking fit between the design and the function of language – the problem of linkage (Kirby, 1999) remains. In this paper we show that cultural evolution, the process by which languages persist through a cycle of learning and use, solves the problem of linkage, and (under some conditions) leads to the emergence of languages which are both highly compressible and highly expressive. Furthermore, we show that this same trade-off between compressibility and expressivity, which has been used to explain the structure of lexicalised concepts in various domains, also explains the existence of structural design features like compositionality.

² A consideration of the possible competition between system-wide compressibility and the compressibility of signals is an obvious extension to the model we present in this paper, as are a range of other possible additional pressures on a language being transmitted (e.g., the structure of the world: Perfors & Navarro, 2014).

³ In reality, languages may not lie cleanly in one or other of these classes. They may exhibit some partial compositionality, or they may be partially degenerate, for example.

1.2. Compression and expression in iterated learning

We use iterated learning as a tool to explore how languages adapt to pressures for compressibility and expressivity. Iterated learning is the process whereby one individual learns by observing the output of learning in another individual, who learned in the same way (Kirby, Cornish, & Smith, 2008). Iterated learning has been studied in models (either mathematical or agent-based) and in experiments with human participants, and provides a framework for studying the cultural evolution of language and other behaviours. Models of iterated learning embodying the assumption that learners prefer compressible (i.e., simple) grammars shown how this pressure can be amplified by the “bottleneck” on the cultural transmission of language. Specifically, the fact that learners must reconstruct their language from a finite subset of data leads to the development, over many episodes of transmission, of increasingly compressible grammars (Kirby, 2002; Brighton, Smith, & Kirby, 2005b). Under the additional assumption that there is an expressivity bias in the learners favouring one-to-one mappings between meanings and signals (Brighton et al., 2005b), these simulations demonstrate the emergence of compositional structure. Other simulation models have focussed on the emergence of combinatorial structure in phonology through processes of repeated reproduction (Oudeyer, 2006; Zuidema & De Boer, 2009; De Boer & Zuidema, 2010; Wedel, 2012). These models too have at their core a bias to favour simple systems (e.g. by merging categories, or reinforcing frequently recurring representations) alongside pressures to maintain distinctiveness of forms.

More recently, this simulation-based modelling work has been complemented by laboratory studies using adult human participants (for review see Scott-Phillips & Kirby, 2010; Kirby, Griffiths, & Smith, 2014). In these studies, building on well-established artificial language learning paradigms (Gomez & Gerken, 2000), researchers observe how an artificial language or communication system is changed by transmission between experimental participants. In one such experiment, Kirby et al., 2008 recreate in the lab a cultural process that closely parallels earlier simulations of iterated learning. They use a *transmission chain method* (Mesoudi & Whiten, 2008) where each participant learns the language produced by the previous participant in the test phase of the experiment. As in the models, an initially holistic language (i.e., a highly incompressible language, lacking the structure characteristic of natural language) becomes ever simpler over generations; the eventual outcome of this process is a highly compressible, largely degenerate language, in which many distinct meanings are conveyed with a small number of highly ambiguous signals (a similar effect can also be seen in the results of Perfors & Navarro, 2014; Silvey, Kirby, & Smith, 2015). Adding an artificial experimental intervention to discourage degenerate languages, for example by the experimenter removing ambiguous strings from the training data (Kirby et al., 2008), leads to the emergence of compositionally structured languages. These results are mirrored by laboratory experiments looking at culturally transmitted sound systems (Verhoef, 2012), in which a similar

manipulation for removing duplicate signals leads to the development of combinatorial structure.

In an alternative experimental approach, Garrod, Fay, Lee, Oberlander, and MacLeod (2007), Fay and Ellison (2013) and Fay, Garrod, Roberts, and Swoboda (2010) use a *closed group* design (Mesoudi & Whiten, 2008) in which participants are required to communicate a set of pre-specified concepts using drawings. Pairs of participants who repeatedly play the game together develop an expressive system of symbol-like graphical representations to communicate these concepts. This system of communication is holistic, since each symbol is an idiosyncratic, stand-alone entity. Theisen-White, Kirby, and Oberlander (2011) present a modified version of this paradigm, introducing aspects of the transmission chain method: an initial pair play a variant of the communication game from Garrod et al. (2007); the drawings produced by that pair during communication are then observed by a fresh pair of participants, who go on to communicate together, and so on. The system of communication is therefore under pressure to be both expressive (communicatively functional) and learnable (easy to reproduce faithfully by the naïve individuals). Theisen-White et al. (2011) find that the sets of drawings become more structured over these chains of transmission: the drawings develop component parts which refer to distinct aspects of meaning.

1.3. Summary of our hypotheses

Although it is hard to directly compare these experiments, they nevertheless suggest a three-way contrast: pressure for compressibility arising from transmission to new learners results in degenerate languages (Kirby et al., 2008; Perfors & Navarro, 2014; Silvey et al., 2015); pressure for expressivity arising from communication leads to holistic systems (Garrod et al., 2007; Fay & Ellison, 2013); pressure from both communication and transmission leads to structure (Theisen-White et al., 2011), an effect which can also be achieved by transmission and an artificial pressure against degeneracy (Kirby et al., 2008). However, no one model or experimental paradigm completely decouples learnability and expressivity: below, we present a model and an experiment which do precisely this, conclusively demonstrating the link between expressivity, learnability and structure, and furthermore showing in general how cultural evolution provides a linking mechanism by which languages can adapt to become highly compressible and highly expressive, as observed by Kemp and Regier (2012), Xu and Regier (2014), Regier et al. (2015). The novelty of our approach lies in the explicit demonstration that cultural evolution can lead to the emergence of language structure under competing pressures from learning and communication.

More specifically, we can set out the following series of predictions that can be tested using computational models and laboratory experiments:

- A pressure from learning alone will lead, over repeated episodes of language transmission, to degenerate languages that are highly compressible, but dysfunctional from the point of view of communication.

- A pressure from communication alone will lead, over repeated episodes of communication, to holistic languages that are expressive, but incompressible.
- Where both learning and communication impact on the cultural transmission of language, we will see the emergence of structured languages which are both expressive and compressible.

We will set out our computational model in the next section, and our laboratory experiment in Section 3. These models and experiments allow us to set up counterfactual situations where one or other pressures on the transmission of language is removed (or at least reduced substantially). Obviously, real language is systematically structured – it exhibits design features like combinatoriality and compositionality. It is also learned anew every generation, and used communicatively. It might appear that testing our hypotheses outside of the lab or computer simulation is therefore impossible. However, in the last section of the paper, we will suggest possible places to look for tentative support in situations of real language transmission under differing social settings.

2. Model

In order to test these hypotheses we first construct a model of the processes of language learning and language use, use this model to test the internal coherence of our theory, and then test the predictions of the model in experiments with human participants (described in Section 3). Our model includes the minimal ingredients required to test our specific hypotheses about how the link between compressibility, expressivity and linguistic structure plays out over cultural transmission: we have a language model, a model of language learning, a model of language use during communication, and a model of transmission in populations. A brief summary is provided here, and full technical details are given in the following sections.

We lay out a very simple model of languages that allows us to differentiate between the three language classes mentioned in the discussion: holistic, structured and degenerate. Following, e.g., Griffiths and Kalish (2007), Real and Griffiths (2009) and Culbertson and Smolensky (2012), we model language learning as a process of Bayesian inference: learners infer a language or languages from observed linguistic behaviour, and we assume that learners have a prior preference for simple, compressible languages, as motivated by our assumption that such languages are in principle easier to learn. We model communication as a process of selecting an utterance to convey a meaning to a communicative partner, and assume that language users have a tendency to avoid utterances which are ambiguous; the strength of this preference is determined by the parameter γ , which we manipulate to remove the pressure for expressivity (by setting γ to 0) or to include it (by setting $\gamma > 0$).

Finally, we model cultural transmission via iterated Bayesian learning (Griffiths & Kalish, 2007; Real & Griffiths, 2009; Burkett & Griffiths, 2010): the data that forms the basis of language learning is itself the product

of language use. We compare iterated learning in two types of population: in *chains*, simulated agents are organised into pairs, are trained on data produced by the previous pair (see below), and then interact to produce data which the next generation in the chain (a new, naïve pair of simulated individuals) are trained on. In *closed groups*, exactly the same regime of training and interaction is observed. However, naïve individuals are not introduced at each generation: rather, the same individuals are trained on their own productions from the previous phase of interaction. This minimal difference between chains and closed groups allows us to manipulate the pressure for learnability. In chains, where naïve individuals are introduced at every generation, the pressure for learnability (i.e., the influence of the prior preference for compressible languages) is likely to be relatively strong. In closed groups with no turnover of the population, there is only one episode of transmission to naïve individuals (at the first point at which the simulated agents encounter the language), and consequently the pressure for simplicity arising from learning is substantially diminished.⁴

Manipulating population type (chain vs. closed group) and expressivity pressure (via the parameter γ) therefore allows us to test the hypotheses outlined above: in particular, we should expect degenerate languages in *chains* with $\gamma = 0$ (pressure for compressibility arising from learning by naïve individuals; no pressure for expressivity); holistic languages in *closed groups* with $\gamma > 0$ (reduced pressure for compressibility due to lack of transmission to naïve individuals; pressure for expressivity during communication); and structured languages in *chains* with $\gamma > 0$ (pressure for compressibility arising from learning by naïve individuals; pressure for expressivity during communication).

The following section goes into the model in some detail, setting out the various components outlined above in turn: the languages, which consist of meaning-form pairs; the hypotheses learners infer, which consist of simple grammars; the prior bias of the learners, favouring simple grammars; the process of inference; and the way in which languages are transmitted within and between pairs of agents (also summarised in Fig. 1). Readers wishing to skip the technical details can safely move on to Section 2.2 to see the results of the simulation model.

2.1. Details of the model

2.1.1. Languages

A language consists of a system for expressing meanings using forms. We consider the simplest possible meanings and forms which are nonetheless capable of evidencing systematic structure: meanings are sets containing feature-values for f features, each taking v possible values. Similarly, forms are of strings of characters of length l , where each character is drawn from some alphabet Σ . We take $f = v = l = |\Sigma| = 2$, which yields a set of

⁴ Another possible approach to manipulating the pressure for compressibility would have been to make the learners completely unbiased. However, we keep the bias constant and change the population type in order to enable direct comparison with our experimental approach covered in Section 3.

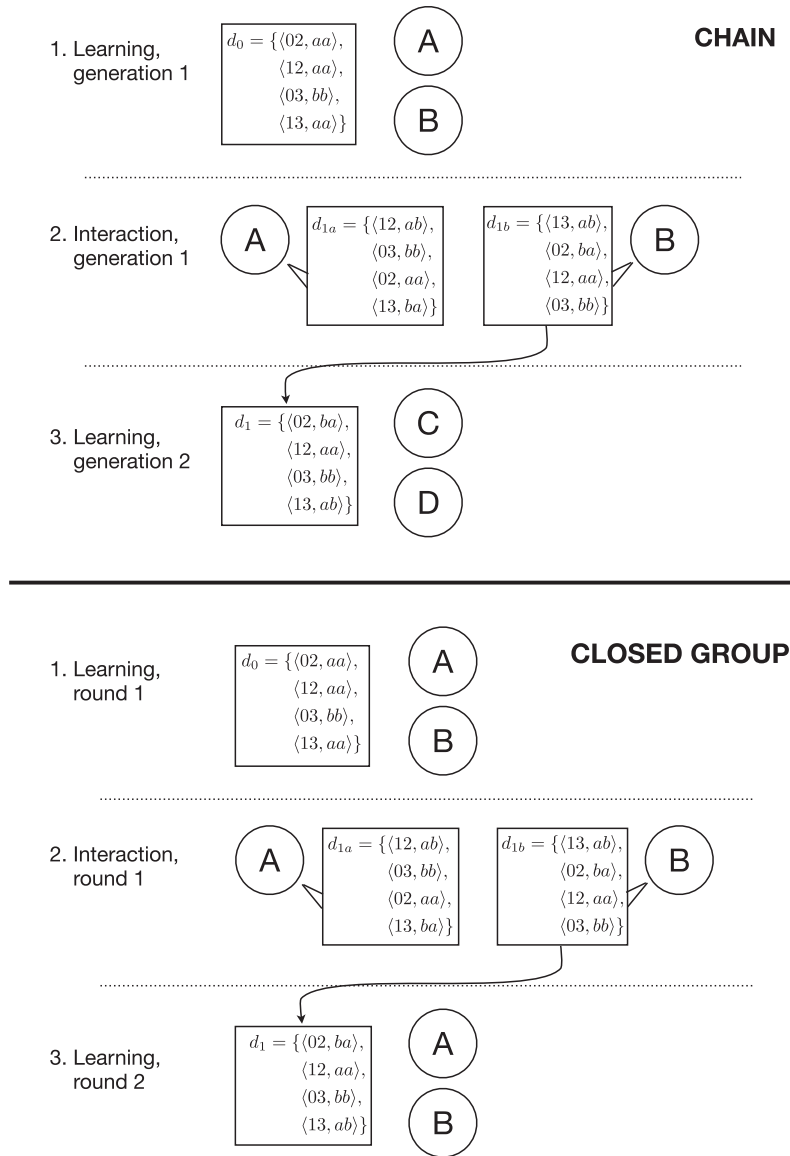


Fig. 1. Simulation procedure in the chain (top) and closed group (bottom) conditions. During learning (1), agents A and B are exposed to some data and sample a hypothesis according to $P(h|d)$. During interaction (2), the agents take turns to produce (meaning, form) pairs according to $P(d|h)$, updating $P(h|d)$ according to their partner's productions. The data produced by one randomly-selected agent during interaction is used (3) to train the next generation of agents. In chains (top) these are fresh, naïve learners; in closed groups (bottom) they are the same two agents.

forms, $\mathcal{F} = \{aa, ab, ba, bb\}$ and a set of meanings, $\mathcal{M} = \{02, 12, 03, 13\}$, where the values for the first feature are drawn from $\{0, 1\}$ and the second feature from $\{2, 3\}$. This gives a space of 256 possible languages, including degenerate, compositionally-structured and holistic mappings. An example degenerate language would be one that associates each meaning with the form *aa*. A structured language would be one where each aspect of the meaning was individually and consistently mapped to a single character, for example $\{\langle 02, aa \rangle; \langle 03, ab \rangle; \langle 12, ba \rangle; \langle 13, bb \rangle\}$. Finally, a holistic language is one in which every meaning has a distinct form but the mapping is not compositional, such as $\{\langle 02, aa \rangle; \langle 03, ab \rangle; \langle 12, bb \rangle; \langle 13, ba \rangle\}$.

2.1.2. Hypotheses

Learners infer a distribution over languages: the space of hypotheses is therefore the space of possible distributions over all 256 languages.⁵ We use a Dirichlet process prior (Burkett & Griffiths, 2010; Ferguson, 1973), characterised

⁵ Various other Bayesian models of language learning (e.g. Griffiths & Kalish, 2007; Real & Griffiths, 2009; Culbertson & Smolensky, 2012) assume that learners infer a single grammar. The generalisation to allowing learners to infer a distribution over languages, rather than a single language, follows techniques provided by Burkett and Griffiths (2010) and is particularly appropriate in our case as it allows learners to track changes in their partners' linguistic behaviour over time.

by concentration parameter α and base distribution G_0 . The parameter α determines how many languages feature in this distribution: low α (we use $\alpha = 0.1$) corresponds to an *a priori* belief that the majority of the probability mass will be on a single language. The base distribution is a distribution over languages, and would be the prior if learners only considered single-language hypotheses.

Our base distribution encodes a preference for simplicity, operationalised as a preference for languages whose description is compressible. Intuitively, degenerate languages permit more compressed descriptions than compositional languages; holistic languages are, by definition, incompressible. We calculate the compressibility of a language by specifying the grammar of that language in a *minimally redundant form* (Brighton, 2002; Dowman, 1998). The coding length of a grammar can then be specified as the number of bits taken to encode it in this minimally redundant form, which we convert into a prior probability, as described below.

2.1.3. A compression-based prior

We treat languages as a rewrite grammar for mapping forms and meanings, then encode that grammar in a minimally redundant form and calculate the minimum number of bits required for that coding. The probability of language l in the base distribution is given by $G_0(l) \propto 2^{-L(l)}$ where $L(l)$ is the number of bits required to code l , and we normalise over all languages. Rewrite rules take one of two forms: $X \rightarrow YZ$ indicates that an item of category X rewrites to items from categories Y and Z (in that order), and inherits the union of the meanings of Y and Z , whereas $X : M \rightarrow F$, where M is a set of possible meanings and F is a form, indicates that category X rewrites to the form F and can take any of the meanings in M . The degenerate language given in Section 2.1.1 is described by the grammar:

$$S : \{02, 03, 12, 13\} \rightarrow aa$$

where S is the start category for the grammar. The compositional grammar is:

$$\begin{aligned} S &\rightarrow AB \\ A : 0 &\rightarrow a \\ A : 1 &\rightarrow b \\ B : 2 &\rightarrow a \\ B : 3 &\rightarrow b \end{aligned}$$

and the holistic grammar is:

$$\begin{aligned} S : 02 &\rightarrow aa \\ S : 03 &\rightarrow ab \\ S : 12 &\rightarrow bb \\ S : 13 &\rightarrow ba \end{aligned}$$

These grammars are encoded as a string of characters that remove any unnecessary redundancies but still allow for reconstruction of the rewrite grammars. In this case, the three encodings are, respectively:

- S02,03,12,13aa
- SAB.A0a.A1b.B2a.B3b
- S02aa.S03ab.S12bb.S13ba

The coding length in bits of a language l is calculated from these strings using $L(l) = -\sum_{i=1}^{|l|} \log_2 p(l_i)$ where $p(l_i)$ is the probability of the i th character in the code for l . This gives code lengths for the three examples of 38.55, 55.20 and 67.29 respectively.

This prior in favour of simpler, more compressible grammars is closely related to the work of Kemp and Regier (2012) and Perfors, Tenenbaum, and Regier (2011), who also assume that learners prefer simpler grammars, operationalised as grammars with fewer rewrite rules (Kemp & Regier, 2012) or grammars with fewer and simpler rewrite rules (Perfors et al., 2011), and yields an intuitively reasonable ranking of languages: degenerate languages are the most compressible, compositional languages far less so, and holistic languages are the least compressible of all. The table below gives coding length (L) and probability in the base distribution (G_0) for example languages, where the forms are given in order for meanings 02, 12, 03 and 13 respectively. (We label all languages that are only partially holistic, compositional or degenerate as “other” in this table.).

type	l	$L(l)$	$G_0(l)$
Degenerate	aa aa aa aa	38.55	0.399
Degenerate	ab ab ab ab	40.55	0.100
Other	aa aa aa ab	52.73	0.0000215
Other	aa aa aa bb	53.49	0.0000127
Compositional	aa ab ba bb	55.20	0.00000388
Other	aa aa ab ba	61.68	0.0000000433
Other	aa aa ab bb	62.17	0.0000000309
Holistic	aa ab bb ba	67.29	0.00000000891

2.1.4. Likelihood

To model language use, we sample a form f from the distribution $P(f|h, t)$, which specifies the probability of f given hypothesis h and a topic $t \in \mathcal{M}$ which the speaker attempts to discriminate from the other meanings in \mathcal{M} . $P(f|h, t) = P(l|h) \cdot P(f|l, t)$: we simply sample a language l from the speaker's hypothesis, then given that language and the topic, sample an utterance. We include a parameterisable preference to avoid ambiguity during this latter step, following the model of pragmatics provided by Frank and Goodman (2012). Assuming some small probability of error on production ϵ :

$$P(f|l, t) \propto \begin{cases} \left(\frac{1}{a}\right)^\gamma (1 - \epsilon) & \text{if } t \text{ is mapped to } f \text{ in } l \\ \frac{\epsilon}{|\mathcal{F}| - 1} & \text{if } t \text{ is not mapped to } f \text{ in } l \end{cases}$$

where we normalise over all possible forms from \mathcal{F} . a is ambiguity, the number of meanings in \mathcal{M} that map to form f in l , and γ specifies the extent to which ambiguous utterances are penalised. If $a = 1$ (f is unambiguous) and/or

$\gamma = 0$ then this yields a model of production where the ‘correct’ form is produced with probability $1 - \epsilon$. However, when $\gamma > 0$ and f is ambiguous (i.e., $a > 1$), then the ‘correct’ mapping from t to f is less likely to be produced (the probability $P(f|l, t)$ is reduced by the factor $(\frac{1}{a})^\gamma$) and the remaining probability mass is spread equally over the other possible forms, leading to increased probability of producing $f' \neq f$. Therefore, $\gamma > 0$ introduces a penalty for languages whose utterances are ambiguous. We use $\epsilon = 0.05$ and vary γ in order to vary expressivity pressure, as discussed above.

Given this model of production, and under the assumption that topics are selected with uniform probability from \mathcal{M} (which is of size $|\mathcal{M}|$), the probability of an individual with hypothesis h producing a given series of (meaning, form) pairs d is

$$P(d|h) = \prod_{(t,f) \in d} \frac{1}{|\mathcal{M}|} P(f|h, t).$$

2.1.5. Inference

The posterior probability of h (a distribution over languages) given data d (a set of (meaning, form) pairs) is

$$P(h|d) \propto P(d|h) \cdot P(h|G_0, \alpha)$$

where $P(d|h)$ is the likelihood function provided in the previous section and $P(h|G_0, \alpha)$ is the Dirichlet process prior over h , characterised by the base distribution G_0 and concentration parameter α . Exact inference over this hypothesis space is intractable: instead, following [Burkett and Griffiths \(2010\)](#), we use a Gibbs sampler based on the Chinese Restaurant Process to sample a hypothesis direct from the posterior. As described below, learners acquire an expanding set of observed utterances during their lifetime: we run the inference over the most recent $r = 80$ observations, in order to improve simulation runtimes.

2.1.6. Transmission in populations

As described above, we compare two types of population: in the *chain* condition, simulated agents are organised into pairs, are trained on data produced by the previous pair (see below), and then interact to produce data which the next generation in the chain (a new, naïve pair of simulated individuals) are trained on. In the *closed group* condition, exactly the same regime of training and interaction is observed. However, naïve individuals are not introduced at each generation: rather, the same individuals are trained on their own productions from the previous phase of interaction.⁶

⁶ For convenience in comparing conditions, we will continue to use the term ‘generation’ rather than the more appropriate ‘round’ for the closed group condition. Training pairs of agents on their own productions ensures that the configuration of the model is identical for both chain and closed group conditions. We ran an additional set of closed group simulations with a modified transmission regime, such that pairs are trained on the initial target language and go on to interact repeatedly but are not retrained on their productions from the last round of interaction (i.e., there is no training phase after generation 1): this produces results which are highly similar to the closed group condition with retraining, showing that the retraining step does not introduce an additional conservative tendency.

During training, the pair are presented with a shared set of $b = 20$ meaning-form pairs, produced by the preceding pair during interaction or (for the first generation only) a shared set of b meaning-form pairs generated from a randomly-selected fully-expressive holistic language (this initialisation with holistic languages is analogous to the set up of the human experiments to follow, but see below, and [Appendix A](#) for discussion of an alternative initialisation). This data is added to each agent’s memory (which will be empty for individuals in chains), and then a hypothesis is sampled from the posterior.

After training, the pair interact for $2b$ rounds. At each round of interaction, one individual acts as speaker and the other as hearer. The speaker samples a single meaning-form pair from their hypothesis according to the likelihood function $P(d|h)$ described above (i.e., a topic t is selected at random from \mathcal{M} , the learner samples a language l from their hypothesis h then, given that language and the topic, samples a form according to $P(f|l, t)$). The hearer adds the observed meaning-form pair to its memory, and samples an updated hypothesis from the posterior. The roles of speaker and hearer then switch, and a new round is played.

The b meaning-form pairs produced by one randomly-selected member of the pair at generation n are used as the training data for the pair at generation $n + 1$. See [Fig. 1](#) for an overview of this setup.

2.2. Results

As described above, we run simulations under three configurations of the model, manipulating population type (chain vs. closed group) and expressivity pressure (by manipulating γ , the penalty for ambiguous utterances during production): in the Learnability Only condition we use the chain population type and $\gamma = 0$; in the Expressivity Only condition we use closed groups with $\gamma = 2$; in the Learnability And Expressivity condition we use chains with $\gamma = 2$.

The results ([Fig. 2](#)) match the predictions of our hypothesis. In the Learnability Only condition, the final distribution is dominated by degenerate languages. As a result of their repeated transmission to fresh, naïve learners, the languages are under pressure to adapt to the learners’ prior preference in favour of compressible languages, and given the absence of countervailing pressures for expressivity, the eventual distribution of languages is dominated by the prior, which favour the most compressible languages: the degenerate ones. In contrast, in the Expressivity Only condition, the initial holistic languages (which are maximally expressive, providing a distinct, unambiguous form for every meaning, but not compressible) persist. The languages in this condition are under little pressure to conform to the prior preference for compressibility of naïve learners, since (after the first ‘generation’ of learning) members of the group approach each fresh bout of learning with overwhelming evidence that the language they are being exposed to is holistic, and constantly replenish their own evidence that the language is holistic during

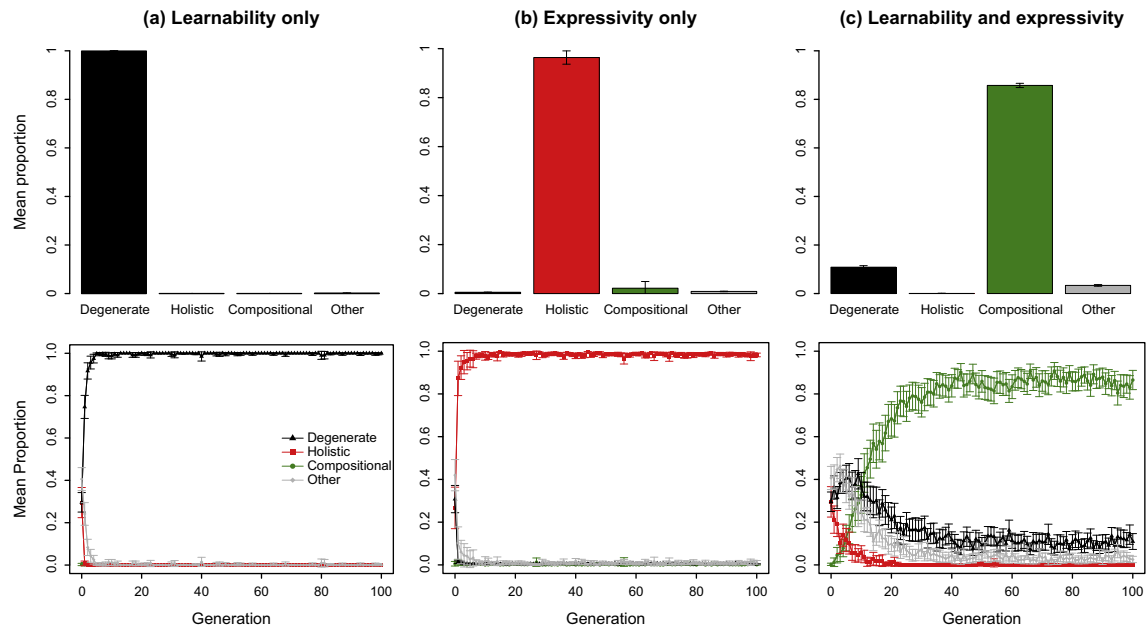


Fig. 2. Final distributions of languages (top; each averaged over generations 50–100 of 100 simulation runs, error bars give 95% confidence intervals) and time courses (bottom) of evolution for three different conditions. When learnability is the only pressure (a), degenerate languages dominate the final distribution, reflecting the prior bias of the learners. When expressivity is the only pressure (b), the original expressive but holistic languages are preserved. When both pressures are at play (c), expressive but compositionally-structured languages emerge.

interaction. Consequently, the initial holistic language is locked in.⁷

Finally, in the Learnability And Expressivity condition, we see structured languages emerge: the final distribution of languages is dominated by languages which are both expressive (in that they provide an unambiguous form for every meaning) and yet relatively compressible (because they are compositionally structured). These languages emerge over cultural evolution as a result of the trade-off between the compressibility preference operating during learning and the expressivity preference operating during communication: the initial holistic languages are highly expressive but difficult for learners to acquire, and the resulting errors made during learning lead to the emergence of more compressible, degenerate languages; however, the expressivity pressure imposed during language use prevents the degenerate languages seen in the Learnability Only condition from taking over, and consequently the structured languages, which are both compressible (and therefore learnable) and expressive increase in frequency. Again, this matches our hypothesis that structure only develops when pressures for both compressibility and expressivity are in play.

The results we describe here obtain when we initialise the simulations with a holistic language. In other words, in line with a number of hypotheses about the nature of protolanguage (e.g., Wray, 1998) we assume a starting point

entirely lacking in structure: a completely incompressible language. However, it is worth considering what would happen if we ran our simulations from the opposite starting point, from the most compressible rather than least compressible languages. Details of this, along with other explorations of the parameter space for the models, are given in Appendix A. In contrast to the results shown in Fig. 2, we find that compositional languages *can* evolve in the Expressivity Only condition if we initialise the population with a degenerate language (see Fig. 5 in Appendix A). The initial degenerate languages are highly unstable due to the expressivity pressure acting on production; however, the noisy mix of communication systems that results from the initial process of eliminating the degenerate starting language provides an opportunity for the prior bias to assert itself as the agents learn from their partner's (variable) productions and their own (variable) behaviour. Nevertheless, it is clear that the crucial contrast between the Expressivity & Learnability condition and the Expressivity Only condition still holds: unstructured, holistic languages never take hold in the former, where the introduction of naïve learners provides a strong and continued pressure against incompressible languages, whereas they constitute a large proportion of the stable languages that emerge in the latter condition irrespective of whether we start the simulations with the most or least compressible languages.

3. Experiments

As discussed in Section 1.2, previous iterated learning experiments with human participants have shown that pressure for learnability alone leads to the emergence of largely degenerate languages (Kirby et al., 2008; Perfors

⁷ Note also that this result holds despite the fact that we set a fairly low memory limit for individuals ($r = 80$) – even if we reduce the memory limit further, so that individual's memory stretches back only as far as it would for a single generation in the chain condition ($r = 40$) we still fail to observe any dominance of structured language. See Appendix A for more details of manipulations of model parameters.

& Navarro, 2014; Silvey et al., 2015), matching the predictions of our model. In order to test the remaining predictions, we developed an experimental method that introduces a natural expressivity pressure arising from communication. Following our modelling approach, we test two conditions: in the *chain* condition, each generation consists of a pair of participants, who are trained on the same target language and subsequently engage in a communicative task, as described below. The utterances that these participants produce during their interaction then form the training data for the next generation in the chain, consisting of a fresh pair of naïve participants. In the *closed group* condition, the same pair of participants remain in the lab throughout, and are re-trained on their own communicative output. The training, communication, and transmission steps are therefore identical across conditions, the two conditions differing only in whether naïve participants are introduced at each generation. In line with the model, and previous experiments, in every case we start with the least structured, least compressible random holistic languages. The logic behind these two experimental conditions is identical to that outlined for the model in the previous section: we should expect holistic languages to persist in *closed groups* (reduced pressure for compressibility due to lack of transmission to naïve individuals; pressure for expressivity during communication); and structured languages in *chains* (pressure for compressibility arising from learning by naïve individuals; pressure for expressivity during communication).

3.1. Method

3.1.1. Participants













A total of 60 student participants (41 female) were recruited through a University of Edinburgh Student And

Graduate Employment service, which advertises temporary employment opportunities for Edinburgh students. Participants were paid at an hourly rate of £6. This experiment was run concurrently with two others, and the pair of participants (from all three experiments) who achieved the highest score in the shortest time received a £25 cash prize. The participants were informed of this in the instructions.

3.1.2. Stimuli and initial languages

Participants were asked to learn a language in which strings of letters (signals) were paired with abstract pictures (meanings). Meanings were drawn from a set of twelve and varied along two dimensions: shape and fill texture. There were three distinct shapes, and four distinct textures. In addition, each of the 12 filled shapes had a unique appendage: the meanings could be described either by referring to these appendages as 12 completely distinct objects, or by referring to a combination of their shape and texture. See Fig. 3 for the full set of meanings.

The initial signals were generated by concatenating 2, 3 or 4 CV syllables, selected randomly with replacement from a set of nine syllables (following the procedure used by Kirby et al., 2008). For each run of the experiment, this set of nine syllables was selected randomly without replacement from a larger set composed of all possible combinations of 8 consonants g, h, k, l, m, n, p, w and 5 vowels a, e, i, o, u. Initial signals were examined and excluded if they contained forms resembling English. Signals were then randomly paired with meanings, and candidate languages thus generated were analysed for structure using the structure measure described below, and were rejected if they returned a significant level of structure. This ensured that the languages given to the first generation of learners were genuinely unstructured with

	pihino		kapa		newhomo
	nemone		gakho		kamone
	piga		wuwele		gaku
	kawake		nepi		hokako













	ege-wawu		mega		gamene-wawu
	ege-wawa		mega-wawa		gamene-wawa
	ege-wuwu		mega-wuwu		gamene-wuwu
	ege		wulagi		gamane

Fig. 3. The language produced at the last generation of a closed group (top) and a chain (bottom). Hyphens added for clarity.

regards to the mapping between signals and meanings. Later generations learnt from the signals produced by the participants in the previous generation, with no restrictions placed upon their composition.

3.1.3. Procedure

The two participants within a pair were trained separately, at two networked computers, with exactly the same learning stimuli. Each generation experienced a *training phase*, during which participants simply observed the meanings and their associated signals (as text) presented on the computer screen, and a *playing phase*. Participants underwent six blocks of training during the training phase: within a single block, each meaning-signal pair appeared once, in random order, with the meaning appearing on screen first (for 1 s) followed by the meaning plus its associated signal (for a further 5 s).

In the playing phase, the participants took turns as *speaker* and *hearer* in a series of communication trials. In each trial, the speaker sees one meaning (the *topic*), and has to type a signal to identify it for the hearer. The hearer sees a context array of six different meanings (including the topic), plus the speaker's signal, and attempts to select the topic, using the mouse. At this point, both speaker and hearer are given feedback: the hearer sees what the correct topic was, and the speaker sees what the hearer picked. If the communication has been successful and the hearer has selected the intended topic, a point is added to the pair's collective score. In total, each participant acts as speaker and hearer twice for each meaning.

In order to transmit the language to the next pair of participants in a chain (or to retrain the same pair in the closed-group condition), the second utterances produced by one participant (randomly selected from the pair) are collected and used as the training language in the next generation. Utterances from a single participant, rather than a mixture of utterances from both, were transmitted to the next generation following existing iterated learning experiments, where input to each participant was produced by a single person. Thus, any potential effects of mixing the input were avoided. We ran four transmission chains and six closed groups, each starting with a different initial language and consisting of six generations.

3.2. Results

Success was quantified as the number of successful interactions during play (i.e., interactions where the hearer successfully identified the topic). The maximum success score is 48 (two blocks of 24 interactions). Transmission error and structure at generation n were evaluated based on the labels that would be used when training generation $n + 1$ (i.e., we applied these measures to the language produced by one participant in each pair). Following the techniques used in Kirby et al. (2008), transmission error was quantified as the normalised Levenshtein distance between the trained signal associated with a given meaning and the signal produced during play for that meaning, averaging across all meanings. The normalised Levenshtein distance is the number of characters that need to be changed, inserted or deleted to transform one character string into

another divided by the length of the longest string. Structure was quantified as the z-score of the Mantel test (Mantel, 1967) between signal-similarities (measured using normalised Levenshtein distance) and meaning-similarities (measured using Hamming distance – the number of features that are different between the two meanings), following the technique used in Kirby et al. (2008). High structure scores indicate languages in which distance between pairs of meanings correlates with the distance between their associated signals to a degree unlikely to arise by chance (specifically, $p < .05$ when the structure scores is greater than 1.96). In other words, similar meanings are associated with similar signals, as observed in compositional languages.

These variables were submitted to mixed-design ANOVAs with Generation as a within-subjects factor (capturing the fact that the languages produced within a single chain are not independent) and Transmission Condition (chain or closed-group) as a between-subjects factor. We used Page's test of trend (Page, 1963) to test for cumulative changes in success, error and structure over generations.

Fig. 3 gives example final languages in both conditions, and Fig. 4 shows the results of our three measures.

For success, there was a main effect of Transmission Condition (i.e., chain vs. closed group: $F(1,8) = 11.5$, $p = .009$): success was higher in the closed group ($M = 41.94$, $SD = 16.22$) than chain ($M = 30.36$, $SD = 14.65$) conditions. There was a main effect of Generation ($F(5,40) = 10.51$, $p < .001$), but no interaction ($F(5,40) = 1.83$, $p = .13$): success at Generation 6 was higher than at Generation 1 ($t(9) = 4.62$, $p = .001$), and increases cumulatively over generations (Page's $L = 910$, $m = 10$, $n = 6$, $p < .01$).

For error, there was a main effect of Transmission Condition ($F(1,8) = 10.76$, $p = .011$), with higher error in the chain ($M = 0.22$, $SD = 0.17$) than in the closed group condition ($M = 0.12$, $SD = 0.12$). There was also a main effect of Generation ($F(5,40) = 13.64$, $p < .001$), and a significant interaction ($F(5,40) = 3.61$, $p = .009$). While t -tests comparing error at first and last generations suggest a significant decrease in Error in the closed group condition ($t(5) = 10.53$, $p < .001$), but not in the chain condition, ($t(3) = 1.01$, $p = .38$), error decreases cumulatively over generations in both the chain ($L = 333$, $m = 4$, $n = 6$, $p < .01$), and closed group ($L = 516$, $m = 6$, $n = 6$, $p < .001$) conditions. Independent-samples t -tests at each generation show that error is lower in the closed group conditions at generation 2 ($t(8) = 3.591$, $p = .007$) but not elsewhere ($t(8) < 2.3$, $p \geq .051$, n.s. with corrected alpha of 0.008): this explains the Generation \times Transmission interaction in the ANOVA, but is not theoretically interesting.

For structure, there was a main effect of Generation ($F(6,48) = 12.47$, $p < .001$) and Transmission Condition ($F(1,8) = 18.03$, $p = .003$) and a significant interaction ($F(6,48) = 8.10$, $p < .001$). Structure cumulatively increases over generations in the chain condition ($L = 534$, $m = 4$, $n = 7$, $p < .001$), yielding higher structure at generation 6 than generation 0 ($t(3) = 4.493$, $p = .021$) but does not increase in the closed-group condition ($L = 708$, $m = 6$, $n = 6$, $p = .38$; t -tests indicate no difference between structure at generation 0 and 6, $t(5) = 1.359$, $p = .232$). Independent-samples t -tests show that there is no

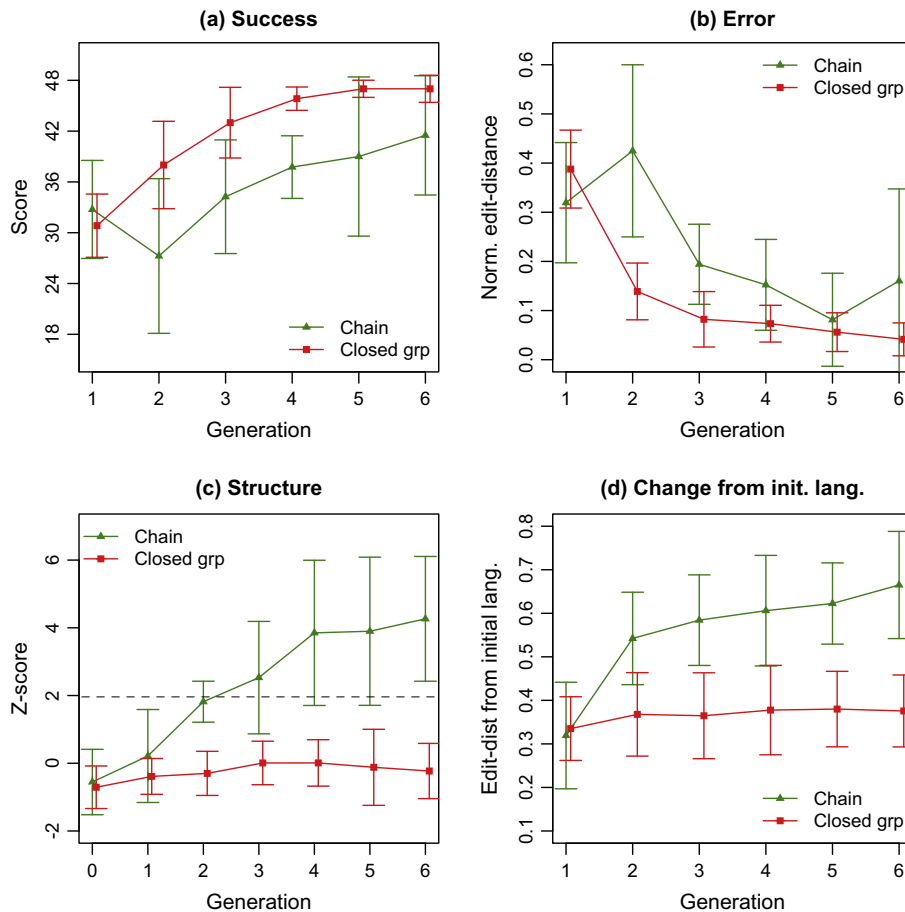


Fig. 4. The effects of Generation and Transmission Condition (chain or closed-group) on (a) communicative success, (b) transmission error, and (c) structure. 95% confidence intervals shown. The 1.96 level, corresponding to $p = .05$, for Structure is shown as a dashed line in (c). In addition, normalised distance from the initial language is shown in (d). Structure increases over generations only in the chain condition.

difference in Structure between conditions at generation 0 ($t(8) = 0.28$, $p = .79$), but a highly significant difference at generation 6 ($t(8) = 4.96$, $p = .001$). Moreover, from generation 3 onwards, the mean Structure score in the chain condition is well above 1.96, indicating that the languages which develop in chains are significantly more structured than expected by chance: in contrast, no language attains this level of structure in the closed group condition.

3.3. Experiment discussion

As discussed earlier, in previous iterated learning experiments where there is only a learnability pressure, languages become less and less expressive over time (Kirby et al., 2008). However, here we see languages that maintain their expressivity (see Fig. 3 for an example of the final language from each condition). In other words, transmission of a language through iterated learning in the presence of a communicative task appears to be sufficient for the emergence of expressive languages, which provide a distinct label for every object.

The languages also become more stable as a result of their transmission, as indicated by the cumulative decrease

in error in both conditions. However, the mechanisms driving this reduction in error differ between conditions. In the closed group condition, error decreases as participants become more and more experienced in using the initial holistic language they are provided with: repeated training and use allows them to master this language, and consequently the language changes relatively little. In the chain condition, however, participants do not have the benefit of repeated rounds of training and interaction: every participant is trained and plays just once. The decrease in error is therefore driven not by increasing familiarity of the participants with a relatively fixed language, but by the language changing to become more learnable. This is the same process of cultural evolution for learnability that we have previously seen in models (Brighton, Kirby, & Smith, 2005a) and experiments (Kirby et al., 2008). This is illustrated in Fig. 4d, where we track distance from the initial language: a 2×6 mixed ANOVA reveals a main effect of Generation ($F(1,618,12,947) = 20.29$, $p < .001$), Transmission Condition ($F(1,8) = 7.78$, $p = .024$), and a significant interaction ($F(1,618,12,947) = 12.07$, $p = .002$). The main effect of Generation is due to languages becoming increasingly different from the initial language over

generations. This increase is cumulative ($L = 885$, $m = 10$, $n = 6$, $p < .001$). The main effect of Transmission Condition is due to languages being more different from the initial language in the chain ($M = 0.56$, $SD = 0.15$) than the closed group ($M = 0.36$, $SD = 0.11$) condition. Distance from the initial language increases cumulatively in both conditions (chain: $m = 4$, $n = 6$, $L = 364$, $p < .001$; closed group: $m = 6$, $n = 6$, $L = 521$, $p < .001$). However, independent-samples t -tests comparing across condition at each generation reveal a significant difference at Generations 5 ($t(3) = 6.72$; $p = .006$) and 6 ($t(3) = 8.07$; $p = .004$).

As the results for Structure show, mode of transmission (chain versus closed group) has a substantial impact on the structure of the evolving languages. In the chain condition, languages are under pressure to be both expressive and learnable: the solution to this problem is for them to exhibit a simple compositional structure, which our structure score picks up on. This compositional structure can be seen in Fig. 3 (bottom): this language marks shape in the first part of the word, fill-pattern in the second, with white fill receiving null marking: a simple and elegant solution to the challenge of being both expressive and learnable. In contrast, we see no increase in structure in the closed group condition: the initial holistic structure is largely preserved, due to the reduced pressure for learnability associated with having the same two participants repeatedly interact with each other.

Our experiments can only explore a small region of the entire space of possible parameters, a space that we are able to explore more fully in simulation (see Appendix A). We have focussed on one particular set of meanings, a particular training and interaction regime, and have used holistic languages as initial input. It would be interesting to see if, in line with the model predictions, more compositional languages might emerge in closed groups if degenerate languages were used as the initial input (although it is possible that the demand characteristics of such training would be somewhat peculiar, since participants would need to be trained extensively on a single word). Scaling up to larger meaning spaces (i.e., containing more than 12 meanings), necessitating larger languages, would also be potentially revealing: we suspect that larger languages will increase the pressure for compressibility, especially in chains (cf. the additional results presented in the Appendix showing that decreasing the amount of training data while holding language size constant increases the pressure for compressibility, which is in line with the prediction). However, expanding the language in this way would pose practical problems for the closed groups purely in terms of the length of experiment required.

4. Conclusions

Our experimental results confirm the predictions of our model exactly. A pressure for expressivity or compressibility alone does not lead to the emergence of structure: only when both pressures are at play does structure reliably emerge. Crucially, we have shown that cultural evolution is a mechanism that can deliver a structured linguistic system where these two pressures from communication and learning interact.

In this paper, we have equated the expressivity pressure with communication and compressibility with learning. The bias we have used in our model is grounded in the general, domain-independent, principle that cognition favours simplicity (Chater & Vitanyi, 2003). An alternative approach might be to include the expressivity pressure as part of the bias of learners, e.g. as preferences for clarity (Slobin, 1977), transparency (Langacker, 1977), or isomorphism (Haiman, 1980). However, the experimental data we review in Section 1.2 shows that iterated learning in human participants leads to the emergence of degenerate languages (i.e., languages which are highly compressible but not highly expressive). This suggests that learning biases favouring expressive languages are weak relative to the biases in favour of compressibility, at least in the types of task which are amenable to iterated learning designs with human participants. Compressibility pressures could also apply during language use. For example, Piantadosi et al. (2012) note that frequent forms are easier to process; highly compressible languages have fewer, more frequent items than less compressible languages, and are therefore more useable in this sense. Again though, the experimental data we present in this paper (specifically, that from closed groups) suggests that the compression pressures which apply during use are relatively weak compared to the expressivity pressures enforcing distinctiveness. Finally, it is worth noting that compressibility and expressivity biases can push in the same direction in some circumstances, e.g. in the task employed by Fedzechkina, Jaeger, and Newport (2012), where learners change languages in ways that improve compressibility and (potential) communicative function. While further consideration of exactly when compression and expression pressures apply would certainly be worthwhile, we nevertheless think that our approach, to assume that these pressures apply primarily during learning (for compressibility) and use (for expressivity), is a reasonable first step.

Our findings also make sense of the distribution of structure in the communication systems of non-human animals. Many small but expressive communication systems exist in nature, a classic example being alarm calling systems, which allow the discrimination of several referents (predators), but do so using vocalisations which are holistic and unlearned (Fitch, 2000). Learned vocal communication systems are witnessed in many species of bird, as well as being patchily distributed among mammals (Fitch, 2000): strikingly, song, the classic example of (combinatorial, not compositional) structure in animal communication, occurs in precisely these species, whose communication system is under cultural selection to be learnable and expressive. This is entirely consistent with the predictions of our model, although we would suggest that the expressivity pressures inherent in communication in these species must be rather different from the expressivity pressure in language. In human communication, the pressure to be expressive derives from the need to discriminate between potential referents in a context of communication, whereas in birdsong a pressure for expressivity may derive from the need to signal individual quality through a large song/syllable repertoire (Collins, 2004).

Our results suggest that the appearance of structure in human language is not inevitable. If there are factors that influence the relative strength of the pressure to be expressive and the pressure to be compressible, then the existence of the structural design features of language may not be universal. Specifically, our models and experiments demonstrate that, in addition to the need to communicate, structure emerges as language is repeatedly transmitted to naïve learners. There is some suggestive evidence that structure in language can be modulated by the composition of populations. For example, there are apparent differences in the structure of emerging sign languages depending on the type of population they emerge in. Meir, Sandler, Padden, and Aronoff (2010) contrast Al-Sayyid Bedouin Sign Language, which emerged in a village with a high incidence of congenital deafness, and Israeli Sign Language, which arose through contact between many deaf learners brought together in schools and clubs. The latter, like most sign languages, exhibits combinatorial structure in its sign inventory, whereas the former surprisingly appears to lack this structural design feature. Similarly, Lupyan and Dale (2010) show that the complexity of morphological structure in language is inversely correlated with size of population. We suggest that examining the impact of naïve learners on the transmission of language might make sense of these findings. For example, in contexts where there is frequent transmission of language between individuals that do not share a long interaction history (cf. Wray & Grace, 2007), where adult second-language learners are involved in transmission (Lupyan & Dale, 2010), or where language emerges through transmission between child learners (Meir et al., 2010), we expect languages to exhibit greater structure.

This work provides a new approach to understanding the structural design features of human language. Cultural evolution responds to a pressure for language to be expressive, driven by the fact that it is used for communication, and a pressure for language to be compressible, driven by the fact that it needs to be learned in order to be transmitted over multiple generations. Linguistic structure is the result.

Acknowledgments

This work was supported by ESRC Grant No. RES-062-23-1537 and by AHRC Grant No. AH/F017677/1. These funding bodies were not involved in the study design; collection, analysis or interpretation of data; writing of the report; or the decision to submit the article for publication. The software for the experiment was created by Sean Roberts (see Roberts, 2012). Many thanks to Amy Perfors, Padraic Monaghan and 2 anonymous reviewers for their helpful comments.

Appendix A. Manipulating model parameters

The results presented in the main text were for simulation runs which were initialised with holistic languages. This assumption is consistent with most previous iterated learning models and experiments (e.g. Kirby, 2002;

Brighton et al., 2005b; Kirby et al., 2008; Perfors & Navarro, 2014; Silvey et al., 2015) and with the argument that, in common with the communication systems of most non-human animals, proto-linguistic systems are likely to have been holistic (Wray, 1998). Initialising our model with degenerate languages⁸ instead yields a pattern of results (Fig. 5) which are slightly different from the results presented in the main text, but still consistent with our theory and hypotheses. As can be seen in Fig. 5, initialising chains with degenerate languages has very little impact on the final distribution of languages: the influence of the initial language quickly disappears as the language is filtered repeatedly through fresh learners, and the final distribution of languages is determined by the prior (when $\gamma = 0$) or the trade-off between the prior and the expressivity pressure (when $\gamma = 2$). The picture is more complex in closed groups with an expressivity pressure ($\gamma = 2$, equivalent to our Expressivity Only condition described above): here, rather than the initial degenerate languages being preserved, a mix of holistic and compositional languages emerge. Due to the expressivity pressure acting on production, the initial degenerate languages are highly unstable: users avoid producing the ambiguous utterances provided by such languages and instead produce forms at random. When they subsequently attempt to update their inferred language from this noisy data (either while learning from their partner's random productions during interaction, or at the episodes of inter-'generation' transmission), due to the uninformative nature of their data, their prior preference for compressibility influences their choice of language; consequently, holistic languages are penalised and compositional languages are favoured as pairs begin to settle on a final, stable language. However, consistent with our theory, the influence of the prior in closed groups is still significantly lower than in chains: holistic languages, although being *a priori* many orders of magnitude less likely than compositional languages (see table in Section 2.1.3), are still common in the final languages emerging in these closed groups.

In addition to the initial languages, there are various other parameters which determine the behaviour of the simulation model, specifically: γ (which determines the extent to which users avoid ambiguous utterances during language use, set to 0 or 2 for the results reported in the body of the paper); ϵ (the noise parameter on language use, set to 0.05 for the results reported in the body of the paper); b (the number of times each agent produces during interaction and therefore the number of data items passed on to the next generation during transmission, set to 20 for the results reported in the body of the paper); α (the concentration parameter in the Dirichlet Process prior over language distributions, influencing whether learners *a priori* expect to infer a single language or multiple languages, set to 0.1 for the results reported in the body of the paper); r (the sliding memory window, determining the number of data items which influence inference, set to 80 for the results reported in the body of the paper). We briefly report the results of manipulating these parameters below: while the details of the simulation results presented in the

⁸ We thank Amy Perfors and an anonymous reviewer for this suggestion.

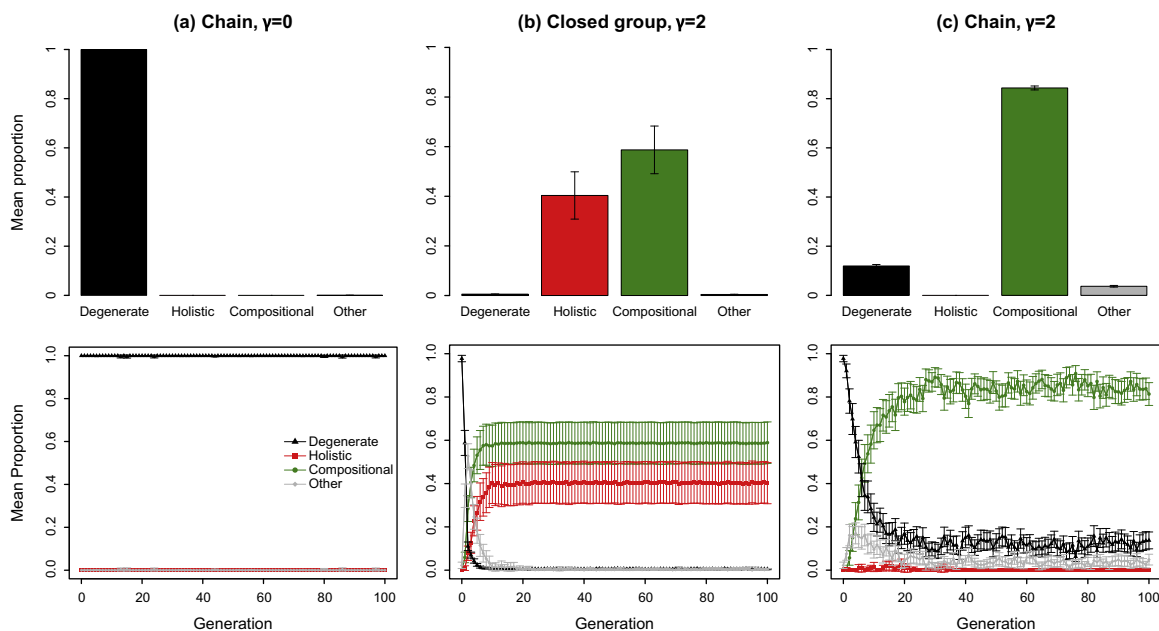


Fig. 5. Simulation runs starting from degenerate languages. Final distributions of languages (top; each averaged over generations 50–100 of 100 simulation runs, error bars give 95% confidence intervals) and time courses (bottom) of evolution for three different conditions. Initialising the simulation runs with degenerate languages makes little difference in chains, but leads to a mix of compositional and holistic languages in closed groups, due to the influence of the prior as the degenerate languages are abandoned.

body of the paper depend on the precise parameter settings used, manipulating these parameters simply serves to alter the balance between pressures for expressivity and compressibility; consequently, the overall pattern of results arising from manipulating these parameters is consistent with our theory that structured languages only reliably emerge when pressures for expressivity and compressibility are at play.

A.1. Manipulating γ , penalty for ambiguity

Reducing γ weakens the pressure for expressive languages and therefore increases the relative influence of the prior leading, ultimately, to the dominance of degenerate languages seen in chains where $\gamma = 0$. The effects of reducing γ are more muted in closed groups, due to the reduced influence of the prior in closed groups: for instance, $\gamma = 1$ in chains results in degenerate languages dominating; however, in closed groups holistic languages still dominate at $\gamma = 1$, although some compositional languages do emerge. Increasing γ above 2 serves to increase the penalty on languages containing ambiguous forms, but since such languages are already rare this has relatively little impact in chains or closed groups.

A.2. Manipulating ϵ , noise on production

Increasing ϵ increases the influence of the prior, since data becomes less informative as to the language which generated it, although these effects are modest. Increasing ϵ has little effect in chains when $\gamma = 0$, since the behaviour here is entirely governed by the prior

anyway. Increasing ϵ (e.g. to 0.2) in chains when $\gamma = 2$ leads to a modest increase in the number of degenerate languages; increasing ϵ (e.g. to 0.2) in closed groups when $\gamma = 2$ leads to a modest increase in the number of compositional languages, as the influence of the prior is slightly increased relative to the expressivity pressure.

A.3. Manipulating b , size of transmission bottleneck

Reducing b increases the influence of the prior, since learners have less data; increasing b reduces the influence of the prior, since learners have more data. These manipulations have little effect in chains when $\gamma = 0$. In closed groups with $\gamma = 2$, reducing b (e.g. to 10) leads to the emergence of some compositional languages, since the influence of the prior increases relative to the influence of expressivity pressures; increasing b (e.g. to 40) has little effect. In chains with $\gamma = 2$, reducing b (e.g. to 10) increases the influence of the prior, leading to the emergence of some degenerate languages, again due to the increased influence of the prior; increasing b (e.g. to 40) leads to the retention and emergence of some holistic languages, as the expressivity pressure begins to outweigh the influence of the prior.

A.4. Manipulating α , concentration parameter of the Dirichlet process prior

Manipulating α , which determines the peakiness of an individual's hypothesised distribution over languages (low values of α favour inferring hypotheses which place the vast majority of the probability mass on one or two

languages; higher values of α lead to hypotheses featuring many languages), indirectly manipulates the influence of the preference for compressibility built into the prior: as shown by Burkett and Griffiths (2010), iterated Bayesian learning converges to distributions of languages which are determined by the base distribution when α is set high enough. Consequently, increasing α leads to the emergence of more degenerate languages in chains: such languages are already common in chains when $\gamma = 0$; increasing α in chains where $\gamma = 2$ leads to a reduction in the number of compositional languages and more degenerate languages emerging, as the balance between compressibility and expressivity is shifted in favour of compressibility; for e.g. $\alpha = 1$ the expressivity pressure is completely overcome and degenerate languages dominate. In closed groups, manipulating α also has the tendency to increase the influence of the prior, but since the prior has less influence in closed groups (as discussed above, due to the lack of transmission to naïve individuals) the effects are more muted: increasing α means that the final languages in closed groups can be either holistic or compositional, with holistic languages being more frequent for low α (as in the results reported in Fig. 2) and compositional languages becoming more frequent as α increases (e.g. holistic and compositional languages are equally frequent in closed groups for $\alpha = 1$).

A.5. Manipulating r , memory size

Reducing r has very little effect on chains. In closed groups, reducing r increases the influence of the prior, since learners retain and are influenced by relatively little data; consequently, for low enough r (e.g. $r = 20$), the difference between closed groups and chains disappears and structured languages emerge in closed groups with $\gamma > 0$, due to the trade-off between compressibility and expressivity pressures discussed in the body of the paper.

References

- Berwick, R., Okanoya, K., Beckers, G., & Bolhuis, J. (2011). Songs to syntax: The linguistics of birdsong. *Trends in Cognitive Science*, 15(3), 13–121.
- Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8, 25–54.
- Brighton, H., Kirby, S., & Smith, K. (2005a). Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In M. Tallerman (Ed.), *Language origins: Perspectives on evolution* (pp. 291–309). Oxford: Oxford University Press.
- Brighton, H., Smith, K., & Kirby, S. (2005b). Language as an evolutionary system. *Physics of Life Reviews*, 2, 177–226.
- Burkett, D., & Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In A. D. M. Smith, M. Schouwstra, B. de Boer, & K. Smith (Eds.), *The evolution of language: Proceedings of the 8th international conference* (pp. 58–65). Singapore: World Scientific.
- Chater, N., & Vitanyi, P. (2003). Simplicity: A unifying principle in cognitive science. *Trends in Cognitive Science*, 7, 19–22.
- Christiansen, M., & Chater, N. (2008). Language as shaped by the brain. *Behavioral and Brain Sciences*, 31, 489–509.
- Collier, K., Bickel, B., van Schaik, C., Manser, M., & Townsend, S. (2014). Language evolution: Syntax before phonology? *Proceedings of the Royal Society B*, 281(20140263).
- Collins, S. (2004). Vocal fighting and flirting: The functions of birdsong. In P. Marler & H. Slabbekoorn (Eds.), *Nature's music: The science of birdsong*. Academic Press.
- Culbertson, J., & Smolensky, P. (2012). A Bayesian model of biases in artificial language learning: The case of a word order universal. *Cognitive Science*, 36, 1468–1498.
- De Boer, B., & Zuidema, W. (2010). Multi-agent simulations of the evolution of combinatorial phonology. *Adaptive Behavior*, 18(2), 141–154.
- Dowman, M. (1998). A cross-linguistic computational investigation of the learnability of syntactic, morpho-syntactic, and phonological structure. Research report, Centre for Cognitive Science, University of Edinburgh, <<http://www.lel.ed.ac.uk/~mdowman/EUCCS-RP-1998-6.pdf>>.
- DuBois, J. (1987). The discourse basis of ergativity. *Language*, 64, 805–855.
- Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology*, 46, 621–646.
- Fay, N., & Ellison, M. E. (2013). The cultural evolution of human communication systems in different sized populations: Usability trumps learnability. *PLoS one*, 8(8), e71781.
- Fay, N., Garrod, S., Roberts, L., & Swoboda, N. (2010). The interactive evolution of human communication systems. *Cognitive Science*, 34, 351–386.
- Fedzechkina, M., Jaeger, T. F., & Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *PNAS*, 109(44), 17897–17902.
- Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1, 209–230.
- Ferrer i Cancho, R., Hernandez-Fernandez, A., Lusseau, D., Agoromoorthy, G., Hsu, M. J., & Semple, S. (2013). Compression as a universal principle of animal behavior. *Cognitive Science*, 37, 1565–1578.
- Ferrer i Cancho, R., & Solé, R. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3), 788.
- Fitch, W. T. (2000). The evolution of speech: A comparative review. *Trends in Cognitive Science*, 4(3), 258–267.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998.
- 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.
- Givón, T. (1979). *On understanding grammar*. Academic Press.
- Gomez, R. L., & Gerken, L. (2000). Infant artificial language learning and language acquisition. *Trends in Cognitive Sciences*, 4(5), 178–186.
- Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31, 441–480.
- Haiman, J. (1980). The iconicity of grammar: Isomorphism and motivation. *Language*, 56(3), 515–540.
- Hockett, C. F. (1960). The origin of speech. *Scientific American*, 203, 88–96.
- Jäger, G. (2007). Evolutionary game theory and typology: A case study. *Language*, 83(1), 74–109.
- Kemp, C., & Regier, T. (2012). Kinship categories across languages reflect general communicative principles. *Science*, 336, 1049–1054.
- Kirby, S. (1997). Competing motivations and emergence: Explaining implicational hierarchies. *Language Typology*, 1(1), 5–32.
- Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford: Oxford University Press.
- Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models* (pp. 173–203). Cambridge: Cambridge University Press.
- Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS*, 105(31), 10681–10686.
- Kirby, S., Griffiths, T. L., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, 28, 108–114.
- Langacker, R. W. (1977). Syntactic reanalysis. In C. N. Li (Ed.), *Mechanisms of syntactic change* (pp. 57–139). Austin, TX: University of Texas Press.
- Lupyan, G., & Dale, R. (2010). Language structure is partly determined by social structure. *PLoS one*, 5(1), e8559.
- Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27, 209–220.
- Meir, I., Sandler, W., Padden, C., & Aronoff, M. (2010). Emerging sign languages. In M. Marschark & P. Spencer (Eds.), *Oxford handbook of deaf studies, language, and education* (Vol. 2, pp. 267–280). Oxford: Oxford University Press.
- Mesoudi, A., & Whiten, A. (2008). The multiple roles of cultural transmission experiments in understanding human cultural evolution. *Philosophical Transactions of the Royal Society B*, 363, 3489–3501.
- Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech*. Oxford University Press.

- Page, E. B. (1963). Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58(301), 216–230.
- Perfors, A., & Navarro, D. J. (2014). Language evolution can be shaped by the structure of the world. *Cognitive science*, 38(4), 775–793.
- Perfors, A., Tenenbaum, J., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, 118, 306–338.
- Piantadosi, S., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, 122, 280–291.
- Real, F., & Griffiths, T. L. (2009). The evolution of frequency distributions: Relating regularization to inductive biases through iterated learning. *Cognition*, 111, 317–328.
- Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–264). Hoboken, NJ: Wiley.
- Roberts, S. (2012). The iterated learning model server: Technical report. Tech. rep., Language Evolution and Computation Research Unit: Edinburgh, UK, <<http://www.lel.ed.ac.uk/lec/ilms/>>.
- Scott-Phillips, T., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Science*, 14(9), 411–417.
- Silvey, C., Kirby, S., & Smith, K. (2015). Word meanings evolve to selectively preserve distinctions on salient dimensions. *Cognitive Science*, 39, 212–226.
- Slobin, D. I. (1977). Language change in childhood and history. In J. Macnamara (Ed.), *Language learning and thought* (pp. 185–221). London: Academic Press.
- Theisen-White, C., Kirby, S., & Oberlander, J. (2011). Integrating the horizontal and vertical cultural transmission of novel communication systems. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd annual conference of the cognitive science society* (pp. 956–961). Austin, TX: Cognitive Science Society.
- Verhoef, T. (2012). The origins of duality of patterning in artificial whistled languages. *Language and Cognition*, 4(4), 357–380.
- von Frisch, K. (1975). *The dance language and orientation of bees*. Harvard University, MA: Harvard University Press.
- Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, 4(4), 319–355.
- Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, 18, 47–67.
- 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, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In *Proceedings of the 36th annual meeting of the cognitive science society*. Austin, TX: Cognitive Science Society.
- Zipf, G. (1936). *The psychobiology of language*. London: Routledge.
- Zipf, G. (1949). *Human behavior and the principle of least effort*. New York: Addison-Wesley.
- Zuidema, W., & De Boer, B. (2009). The evolution of combinatorial phonology. *Journal of Phonetics*, 37(2), 125–144.