Review

Language as an evolutionary system

Henry Brighton [*,1], Kenny Smith [2], Simon Kirby

*Language Evolution and Computation Research Unit, School of Philosophy, Psychology, and Language Sciences,
The University of Edinburgh, 40 George Square, Edinburgh, EH8 9LL, UK*

Abstract

John Maynard Smith and Eörs Szathmáry argued that human language signified the eighth major transition in evolution: human language marked a new form of information transmission from one generation to another [Maynard Smith J, Szathmáry E. The major transitions in evolution. Oxford: Oxford Univ. Press; 1995]. According to this view language codes cultural information and as such forms the basis for the evolution of complexity in human culture. In this article we develop the theory that language also codes information in another sense: languages code information on their own structure. As a result, languages themselves provide information that influences their own survival. To understand the consequences of this theory we discuss recent computational models of linguistic evolution. Linguistic evolution is the process by which languages themselves evolve. This article draws together this recent work on linguistic evolution and highlights the significance of this process in understanding the evolution of linguistic complexity. Our conclusions are that: (1) the process of linguistic transmission constitutes the basis for an evolutionary system, and (2), that this evolutionary system is only superficially comparable to the process of biological evolution.

* Corresponding author. Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, D-14195 Berlin, Germany. Tel.: +49-(0)30-82406-350; Fax:+49-(0)30-82406-394.
   E-mail addresses: hbrighton@mpib-berlin.mpg.de (H. Brighton), kenny@ling.ed.ac.uk (K. Smith), simon@ling.ed.ac.uk (S. Kirby).

Contents

1. Introduction

[. . .] if we view life on the largest scale, from the first replicating molecules, through simple cells, multicellular organisms, and up to human societies, the means of transmitting information have changed. It is these changes that we have called the 'major transitions': ultimately, they are what made the evolution of complexity possible [64, p. 3].

At some point in the last five million years the arrival of human language signified what Maynard Smith and Szathmáry consider to be the eighth major transition in evolution [63,64]. Language, when compared with every known communication system in the natural world, exhibits unsurpassed complexity. It allows an indefinite number of concepts to be expressed by combining a discrete set of units. This is why, for Maynard Smith and Szathmáry, "[t]he analogy between the genetic code and human language is remarkable" [64, p. 169]. Furthermore, human language, as a means of transmitting information, exhibits defining characteristics of major evolutionary transitions. Firstly, language provided a new medium for information transmission across generations. Secondly, the mechanism used to solve the transmission problem is quite unlike those which preceded it. But in what sense does language carry information? And what mechanisms underly and influence this mode of information transmission? This article brings together a body of work that responds to these questions.

Language is undoubtedly required to support many cultural artifacts and practices such as, for example, religion. It is this kind of complexity that Maynard Smith and Szathmáry appeal to when framing language as a major transition in evolution. The complexity of human society and culture rests on the productivity of language, and how it enables complex informational structures to withstand repeated cultural transmission from one generation to the next. We will show in this article that there is another sense in which language can be considered in evolutionary terms. Firstly, we will argue that the complexity we see in human languages is determined to a significant degree by the manner in which they are transmitted. Secondly, we will show how the transmission of language is achieved using the mechanisms of language learning and language production. These mechanisms impose constraints on transmission, such that languages can be said to undergo adaptation as a result of their transmission. This process is termed *linguistic evolution* [14,15,25,26,34]. In short, language can properly be regarded as an evolutionary system.

The hypothesis that language should be understood in these evolutionary terms rests on the assumption that languages code information that determines the manner in which they are processed by the cognitive system. This assumption is intimately related to a central question in linguistics and cognitive science: to what degree is language an expression of the genes? Section 2 focuses on this question by first considering the position known as *strong* or *Chomskyan nativism* (see [29,72]). This position is based on the hypothesis that the essential properties of languages we see are innately specified to a theoretically significant degree (e.g., [22,23]), and as such represents one extreme position on an unresolved empirical question [32,76]. We will then briefly outline alternatives to this position which propose that at least some of the hallmarks of language are learned through inductive generalisations from data. This alternative standpoint opens up a set of fundamental questions relating to the question of how language evolved in humans. We argue that, given this standpoint, linguistic evolution forms a significant part of the explanation for the evolution of linguistic complexity.

Section 3 describes the interdisciplinary approach to evolutionary linguistics we adopt to develop and test the theory of linguistic evolution. Our discussions will draw on concepts taken from fields such as complex systems, computational learning, artificial life, and linguistics. We will then work toward strengthening the view of language as an evolutionary system, developing our argument using two recent computational models of linguistic evolution. Both models deal with the cultural evolution of compositional structure, a hallmark of language and a test-case which demonstrates the explanatory potential of our approach. In Section 4 we present a simple associative model of language learning (as presented in [87]) which allows us to establish a link between a particular aspect of cultural transmission (a *transmission bottleneck*) and the evolution of compositional structure. In Section 5 this model is extended to consider the role that the biases of language learners play in this evolutionary process, and in par-

ticular investigates the impact of psychologically-plausible learning biases on the evolutionary process. Section 6 takes a complementary perspective, and analyses the process of learning in terms of induction and compression according to a normative theory of induction called the minimum description length principle [9,11,12].

The insights of these models will be used to refine the theory of linguistic evolution developed in Section 7. Here, we consolidate the insights of the models and develop the argument for viewing language as an evolutionary system. In broad terms, then, this article will demonstrate recent progress in understanding language in evolutionary terms, and in particular, understanding language *itself* as an evolutionary system. The overarching theme in this discussion will be the hypothesis that language adapts to aid its own survival by evolving certain types of structural complexity. In order to understand this phenomenon we need a theory detailing the novel medium of linguistic transmission, and the mechanisms that underly this transmission.

2. Background: Explaining the complexity of language

Language is a system relating form and meaning. Individual languages achieve this relationship in different, but tightly constrained ways. That is to say that variation exists across languages, but the object of study for many linguists are the common structural hallmarks we see across the world's languages. Why do all languages share these properties? A widespread hypothesis is that language, like the visual system, is an expression of the genes:

> It is hard to avoid the conclusion that a part of the human biological endowment is a specialized 'language organ', the faculty of language (FL). Its initial state is an expression of the genes, comparable to the initial state of the human visual system, and it appears to be a common human possession to close approximation [23, p. 85].

To support this view, we can note that children master complex features of language on the basis of surprisingly little evidence. The *argument from the poverty of the stimulus* states that the knowledge of language children attain is surprising precisely because it *cannot* be derived solely from information made available by the environment (e.g., [21,29,76,98]). This view can be traced back to Plato (427BC–347BC), who noted that humans come to know more than that suggested by the evidence they encounter, with language being just one example of this general phenomenon.

If knowledge of language is in this sense innate, then why do languages exhibit so much variation? The modern debate on the innateness of language attempts to resolve this problem by suggesting that the framework for linguistic development is innate, while the linguistic environment merely serves to steer an internally directed course of development:

> [the environment] provides primary linguistic data that enable the linguistic system to develop, just as it provides light and food that enable the visual and motor systems to develop [95, p. 523].

In this sense, languages themselves are not encoded entirely in the genes, but the fundamental, abstract properties of language are. How can we gain an understanding of these innately specified hallmarks of language? One possibility is that linguists, by conducting a thorough analysis of the world's languages,

can propose a set of descriptive statements which capture these hallmarks of language. For example, we may identify properties common to all languages we encounter, properties that conform to a certain statistical distribution, or implicational hierarchies of properties that fit with known languages. Collectively, such descriptive statements constitute a theory of *language universals* (e.g., [28,31,70]). Linguistic universals define the dimensions of variation in language. Modern linguistic theory rests on the assertion that it is these dimensions of variation that are genetically determined.

As an explanatory framework this approach to explaining why language exhibits specific structural characteristics is very powerful. One of its strengths is that, by coupling universal properties of language tightly to a theory of innate constraints, our analysis of the structural hallmarks of language must centre on a wholly psychological (i.e., cognitive, mentalistic, or internalist) explanation. As a consequence, we can understand why languages have certain structural characteristics and not others by understanding those parts of the human cognitive system relevant to language. In other words, our object of study has been circumscribed to encompass a physical organ: the brain. As we have seen, this position is largely based on the argument from the poverty of the stimulus. One outcome of this hypothesis is that children do not *learn* language in the usual sense, but rather they acquire it as a result of the internally directed processes of maturation. For example, Chomsky states that "it must be that the basic structure of language is essentially uniform and is coming from inside, not from outside" [23, p. 93].

This claim is controversial, and will impact heavily on the discussion to come. Nevertheless, to characterise the traditional position, we should note that language is often considered part of our biological endowment, just like the visual system. The intuition is that one would not want to claim that we learn to see, and in the same way, we should not claim that we learn speak.

2.1. Language learning under innate constraints

Linguistic nativism, at least in the extreme form presented above, is far from being universally accepted (for a good coverage of the debate, see [29,32,46]). An alternative to this hypothesis is that the structure of language is, to some extent, learned by children: humans can arrive at a sophisticated knowledge of language without the need to have hard-wired (genetically determined) expectations for all the dimensions of linguistic variation. This is the view that we will adopt throughout this article. We assume that, to some degree, language is learned through inductive generalisations from linguistic data, and therefore deviate, as do many others, from Chomsky's position that knowledge of language goes "far beyond the presented primary linguistic data and is in no sense an 'inductive generalisation' from these data" [21, p. 33].

To what degree is it true that language is learned through inductive generalisations? Frustratingly, there is little concrete evidence either way. Linguistics lacks a rigorous account of which (if any) aspects of language are acquired on the basis of innate constraints. General statements such as "linguistic structure is much more complex than the average empiricist supposes" [97, p. 283], and "the attained grammar goes orders of magnitude beyond the information provided by the input data" [98, p. 253] abound, and these claims are to some extent backed up with specific examples designed to show how children's knowledge of language extends beyond what the data suggests (e.g., [3,30,48,56,57]). Nevertheless, many still argue that the required information is in fact present in the linguistic data [75,76], and to claim that it is not is "unfounded hyperbole" [75, p. 508].

It should be noted that, despite the debate being dominated by extremes, the issue is not one of denying that language has an innate biological basis. Only humans can acquire language, so any theory of

language must consider an innateness hypothesis of some form. The real issue is the degree to which language acquisition is a process of induction from data within constraints:

> [...] our experience forms the basis for generalization and abstraction. So induction is the name of the game. But it is also important to recognize, that induction is not unbridled or unconstrained. Indeed, decades of work in machine learning makes abundantly clear that there is no such thing as a general purpose learning algorithm that works equally well across domains. Induction may be the name of the game, but constraints are the rules that we play by [38].

As for the degree to which these constraints are language-specific, and can rightfully be considered genetically determined, the issue is an open empirical question:

> I would also take it to be a matter for empirical investigation the extent to which it is necessary to attribute certain properties of grammars to the emergence in the course of learning of 'innate structures', on the one hand, or to the application of specific learning procedures to a body of linguistic data, on the other [32, p. 11].

In the light of this debate, we make an assumption that will be carried through the remainder of the article: if we deviate from the position that language acquisition in *no* sense involves inductive generalisations, then we must acknowledge that the linguistic environment must supply *information*. This information impacts on how languages are represented and processed within the cognitive system— linguistic data contains information about the structure of the mental grammar required to produce that data. In other words, in addition to its more obvious communicative content, language encodes information about the structure of language.

2.2. Towards an evolutionary explanation

The degree to which language is learned through a process of inductive generalisation has a profound effect on the framework we use to explain why language has the structure that it does [14]. If induction plays a role in determining knowledge of language, then environmental considerations must be taken seriously; any linguistic competence acquired through learning will be determined to a significant degree by the structure, or information, present in the environment in the form of linguistic data. The environment must be supplying information in order for induction to occur. We must therefore explain why the linguistic environment is the way it is: how did this information, or linguistic structure, come to exist? To address this issue we will argue for an evolutionary perspective, and seek to explain *how*, from a non-linguistic environment, linguistic structure can develop through linguistic evolution. In short, this view casts doubt on the view that the hallmarks of language are, as Chomsky claims, "coming from inside, not from outside." Necessarily, if inductive generalisations made from data contained in the environment determine the kind of linguistic structure embodied in language, then a wholly psychological theory of linguistic structure must be inadequate—the environment, in the form of linguistic data, plays a crucial role. How languages themselves can come to carry this information is the issue we turn to next.

3. Linguistic evolution: From theory to models

Linguistic evolution is the process by which languages themselves evolve as a result of their transmission (e.g., [14,24,25,34,49]). Unlike the innate communication systems of, for example, vervet monkeys [20] and bees [96], human language is, as discussed above, learned, and therefore potentially undergoes change as a consequence of its cultural transmission from one generation to another. The most obvious example of this is language change, as witnessed on a historical time scale (e.g., [1,45,58])—for example, the change from Old to modern English.

Linguistic evolution is an instance of the more general phenomenon of cultural evolution which figures in explanations of a number of human cognitive domains (e.g., [7,16,33,36,92]). It is not clear to what extent linguistic evolution mirrors the processes of cultural evolution in the wider sense. Indeed, it is possible that language-specific constraints govern the processing and transmission of language. For this reason, we will begin by assuming that linguistic evolution may differ from other instances of cultural evolution. Our starting position is therefore conservative: we do not set out to explain a general theory of cultural evolution but rather seek to develop a theory which is specifically linguistic.

The process of linguistic evolution has been repeatedly proposed as a source of linguistic complexity (e.g., [24,34,49]). Deacon, for example, states:

> Grammatical universals exist, but I want to suggest that their existence does not imply that they are prefigured in the brain like frozen evolutionary accidents [. . .] they have emerged spontaneously and independently in each evolving language, in response to universal biases in the selection processes affecting language transmission [34, pp. 115–116].

Deacon's position could be taken as an extreme—it may not be the case that all universals can be described in this way. The key point in the present discussion is that linguistic evolution occurs as a result of language being transmitted from one generation to another, and that this linguistic evolution may offer an explanation for linguistic universals. In order to provide a firmer footing for the discussion that follows, we now turn to a more formal characterisation of the process of linguistic evolution.

3.1. Iterated learning: A model of language transmission

Language is transmitted culturally through, firstly, the production of utterances by one generation and, secondly, the induction of a grammar by the next generation, based on these utterances. This cycle, of repeated production and induction, is crucial to understanding the diachronic process of language change (as argued by, e.g., [1,45]). Several models have demonstrated how phenomena of language change can be understood in terms of this characterisation of linguistic transmission [27,42,68]. Such models are designed to inform our understanding of how full-blown human languages undergo structural change over time—for example, these models could inform an enquiry into the morphological change that characterised the history of, say, English (as in [42]).

Of more importance here are studies that focus specifically on the cultural evolution of linguistic complexity from non-linguistic communication systems [4,9,51,52,87]. In principle, the same general

mechanisms which explain the change in languages in recent times should also offer an account of change in the linguistic system at a greater time-depth—in other words, uniform processes acting on languages should be capable of explaining both language *evolution* (a qualitative shift from a non-linguistic to a linguistic system), as well as language *change* (subsequent quantitative shifts).[3]

Much of the work focusing on the emergence of linguistic systems through linguistic evolution has been consolidated under a single computational modelling framework termed the *Iterated Learning Model* [9,51,87,89]. In this article we will use particular examples of the Iterated Learning Model to test aspects of the theory of linguistic evolution.

An iterated learning model consists of a model of a population composed of a number of agents (simulated individuals), typically organised into generations. Language is transmitted from generation to generation within this population. For a language to be transmitted from one agent to another, it must be externalized by one agent (through language production), and then learned by another (through language acquisition). An agent therefore must have the ability to learn from examples of language use. Learning results in the induction of a hypothesis on the basis of data. This hypothesis represents the agent's knowledge of language. Using the hypothesis, an agent also has the ability to produce examples of language use itself. Agents, therefore, have the ability to interrogate an induced hypothesis to yield examples of language use. Within this general setting, we can explore how the process of linguistic evolution is related to the mechanisms of hypothesis induction (language acquisition), and hypothesis interrogation (language production).

In the introduction to this article we described how human language signified a major transition in evolution as it represents a novel medium for information transmission, both in the sense intended by Maynard Smith and Szathmáry [63]—language transmits information about non-linguistic culture—and in the sense that language encodes information about its own structure. The mechanisms that underly and influence the transmission of this latter type of information are the mechanisms of language acquisition and language production that we describe here.

Within the framework of the Iterated Learning Model, various treatments of population size, population turnover, and social network structure are possible.[4] Throughout this article we will consider the simplest case, where each generation contains a single agent. The first agent in the simulation, Agent 1, is initialised with knowledge of language h_1, the precise nature of which will depend on the learning model used. This hypothesis will represent knowledge of some language L_{h_1}. Agent 1 then produces some set of utterances L'_{h_1} by interrogating the hypothesis h_1. This newly constructed set of utterances will be a subset of the language L_{h_1}. These utterances are then passed to the next agent to learn from, the first agent playing no further part in the simulation. This process is illustrated in Fig. 1. The important point is that, under certain circumstances, the language will change from one generation to another; it will evolve and undergo adaptation.

---

[3] See Newmeyer [67] for discussion of the *uniformitarian dogma* in linguistics. The approach to linguistic evolution we describe here rejects uniformity of state, but accepts uniformity of process. In other words, we assume that the form of languages has qualitatively changed although the mechanisms of cultural evolution driving this change have remained constant.

[4] Indeed, data from language genesis, change and death suggest that such population factors have significant impact on the linguistic system—see Smith and Hurford [88] for a pilot study applying iterated learning to an investigation of such issues.
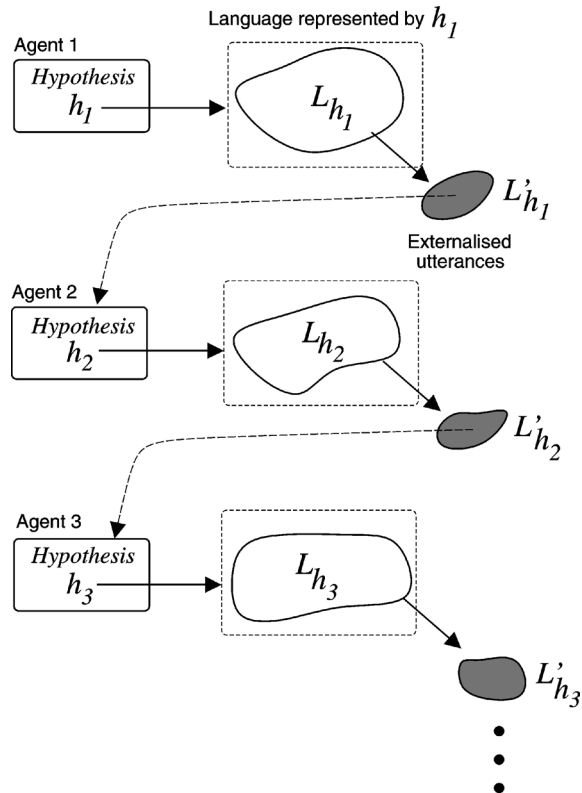
Fig. 1. The iterated learning model. The first agent has knowledge of language represented by a hypothesis h_1. This hypothesis itself represents a language L_{h_1}. Some subset of this mapping, L'_{h_1}, is externalized as linguistic performance for the next agent to learn from. The process of learning results in a hypothesis h_2. The process is then repeated, generation after generation.

3.2. The language model

Before proceeding to a fully-specified Iterated Learning Model we must introduce our language model. The particular model we introduce will figure in both models featured later in the paper. The discussion surrounding the language model will also allow us to define the feature of language we will be investigating throughout this article. This is a property of language—a linguistic universal—termed *compositionality*.

A model of language needs to capture the fact that a language is a particular relationship between sounds and meaning. The level of abstraction we will aim for captures the property that language is mapping from a "characteristic kind of semantic or pragmatic function onto a characteristic kind of symbol sequence" [73, p. 713]. When we refer to a model of language, we will be referring to a set of possible relationships between, on the one hand, entities representing *meanings*, and on the other, entities representing *signals*. Throughout this article we will consider meanings as multi-dimensional feature structures, and signals as sequences of symbols.

Meanings are defined as feature vectors representing points in a *meaning space*. Meaning spaces will be defined by two parameters, F and V. The parameter F defines the dimensionality of the meaning

space—the number of features each meaning has. The parameter V defines how many values each of these features can accommodate.[5] For example, a meaning space $\mathcal{M}$ specified by $F = 2$ and $V = 2$ would represent the set:

$$\mathcal{M} = \big\{(1, 1), (1, 2), (2, 1), (2, 2)\big\}.$$

Notice that meanings represent structured objects of a fixed length, where the values associated with each feature are drawn from a set. We will further assume, in the interest of simplicity, that no graded notion of similarity applies within feature values—feature values are unordered, and the only notion of similarity is one of identity.

Signals are defined as strings of symbols drawn from some alphabet Σ. Signals can be of variable length, from length 1 up to some maximum $l_{\max}$. For example a signal space $\mathcal{S}$, defined by $l_{\max} = 2$ and $\Sigma = \{a, b\}$, would be:

$$\mathcal{S} = \{a, b, aa, ab, ba, bb\}.$$

Again, we assume that no notion of similarity other than identity applies to members of Σ—for example, a is no more similar to b than it is to z.

We now have a precise formulation of the meanings and signals. Of great importance to following discussion will be the kinds of structural relationships which can exist between meanings and signals. It is the nature of the relationship between meanings and signals that makes human language so distinctive. Accordingly, it is crucial to be aware that the model of meanings and signals we have introduced will restrict the set of mappings that are possible. By building an abstract model of language we are necessarily simplifying the range of linguistic phenomenon we seek to explain. For example, recursive structures found in language cannot occur within this model of language (see Kirby [52] for a model which considers this aspect of language. However, as it stands, the model of language presented above can capture a key feature of language we will be focusing on: compositionality.

Compositionality is a property of the mapping between meanings and signals.[6] A compositional mapping is one where the meaning of a signal is some function of the meaning of its parts and the way in which they are combined (e.g., [53,100]). Such a mapping is possible given the model of language developed so far. Consider the language $L_{\text{compositional}}$:

$$L_{\text{compositional}} = \big\{\langle\{1, 1\}, ac\rangle, \langle\{1, 2\}, ad\rangle, \langle\{2, 1\}, bc\rangle, \langle\{2, 2\}, bd\rangle\big\}.$$

This language has compositional structure due to the fact that each meaning is mapped to a signal such that parts of the signal (some sub-string) correspond to parts of the meaning (a feature value). The string-initial symbol a, for example, represents feature value 1 for the first feature.

An instance of a language with no compositional structure whatsoever is also of interest. We will term such relationships *holistic* languages:[7] signals map to meanings in such a way that no systematic

---

[5] Specifying a meaning space using just two parameters is of no intrinsic importance; it serves only to simplify notation. We could just as well define a meaning space using the parameter F, along with an extra $F V$ parameters detailing the number of values each individual feature can take.

[6] Contra a frequently-stated view, it is not a property of meanings alone, nor indeed a property of signals alone.

[7] Strictly speaking, we should use the term *holistic communication system* since one of the defining features of language is compositionality. Nevertheless, we will continue to abuse the term *language* in this way in the interest of convenience.

relationship exists between parts of the signal and parts of the meaning, as in L_{holistic}:

$$L_{\text{holistic}} = \big\{ \langle \{1, 1\}, \mathsf{ppld} \rangle, \langle \{1, 2\}, \mathsf{esox} \rangle \langle \{2, 1\}, \mathsf{q} \rangle, \langle \{2, 2\}, \mathsf{dr} \rangle \big\}.$$

A holistic language is usually constructed by pairing each meaning with a random signal, and consequently holistic languages may also be referred to as random languages in the discussion that follows.

The morphosyntax of language exhibits a high degree of compositionality. For example, the relationship between the string *John walked* and its meaning is not completely arbitrary. It is made up of two components: a noun (*John*) and a verb (*walked*). The verb is also made up of two components: a stem and a past-tense ending. The meaning of *John walked* is thus a function of the meaning of its parts and the way in which they are combined. The compositional structure of language makes the interpretation of previously-unencountered utterances possible—knowing the meaning of the basic elements and the effects associated with combining them enables a user of a compositional system to deduce the meaning of an infinite set of complex utterances.

3.3. How iterated learning models inform theory and explanation

Given the model of language described above we can begin to describe in more depth how iterated learning models can be used to explore theories of linguistic evolution.

A model of language defines a space which contains all the possible languages (relationships between meanings and signals) that the model can accommodate. We will refer to this space as the *language space*. Each simulation run of an Iterated Learning Model represents a trajectory through the language space. As the language is transmitted and evolves the system may enter different regions of the language space. Iterated learning models are informative when, irrespective of the initial language, certain regions of the state space represent *attractors*—regions of the space that the system will always settle in. If an iterated learning model consistently results in trajectories which focus on an attractor, the model will have shown how the process of linguistic evolution mediates between a set of initial conditions and a final region of the state space, and therefore by implication, some structural property of language we are interested in.

For example, Fig. 2 depicts two sets of experiments under different conditions—say, different assumptions regarding the linguistic capacities of the simulated individuals. In both conditions we see a pair of
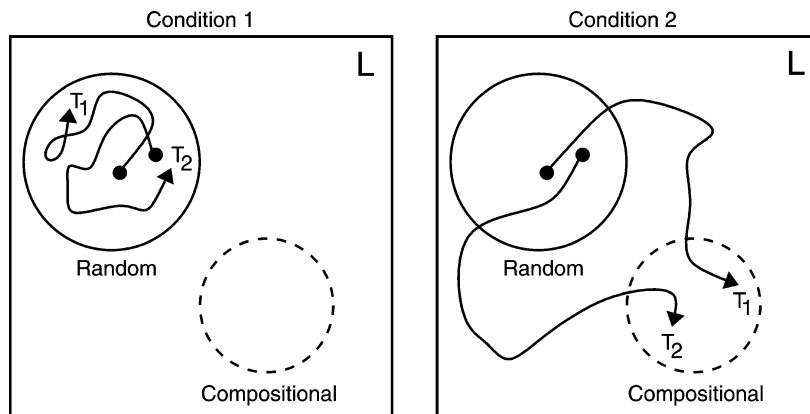

Fig. 2. Trajectories through a language space, given two different sets of experimental conditions.

trajectories with an initial starting position of a random (holistic) language. The language space, represented schematically, has two regions of interest: a large region representing random languages, and a smaller region of languages which have the property of compositionality, a key feature of linguistic structure. Under experimental condition 1 (the left-hand diagram in Fig. 2) the compositional region of the language space is not visited. In contrast, under condition 2 we see consistent evolution from random initial languages to compositional languages.

Given such a series of outcomes of the model, we can begin to identify which properties of the simulated agents (in our example), in combination with the process of linguistic evolution, leads to the evolution of compositionality: linguistic structure develops as a result of how the language is transmitted. By systematically investigating different experimental conditions pertaining to the capacities of simulated agents, the environment of cultural transmission and so on, we can begin to refine our understanding of the characteristics of this evolutionary process.

Iterated learning models will not necessarily result in perfectly stable states. When we refer to the stability of a state of the model, with respect to a linguistic property of interest, we typically refer to Liapounov stability, also known as "start near, stay near stability" (e.g., [40, p. 27]). Our characterisation of stability allows the possibility that the model may enter a specific region of the language space and remain within it, even though no particular language can be said to be stable. For example, the subspace representing compositional languages may be a stable region for a certain model. Similarly, natural languages we observe are stable in the sense that they always conform to linguistic universals, although they undoubtedly undergo change over time.

3.4. Contrasting perspectives on the mechanisms driving linguistic evolution

We have discussed how an iterated learning model can be employed to shed light on an explanation of linguistic evolution, introduced the key components of the iterated learning model, and focused on how the language model can accommodate the linguistic phenomenon of compositionality. The next step is to focus on the agents within the iterated learning model, as properties of the agents, in interaction with the mediating cultural dynamic, determine how languages themselves evolve.

Agents are composed of a learning algorithm and a production algorithm. We will present two models which focus on two sources of insight into the process of linguistic evolution, and which take their inspiration from two complementary conceptual frameworks. First of all, we present an associative model of learning. This model is useful because it allows parallels to be drawn with known psychological processes of language acquisition, therefore providing insights into how such processes drive linguistic evolution. The second model we present views the process of learning from a perspective of general considerations of data compression and Bayesian inference. In addition to providing a theoretically well-grounded approach to induction, the second model facilitates an investigation into the role of invention and innovation in linguistic evolution.

These two contrasting approaches are not mutually exclusive. Indeed, part of the motivation for pursuing these lines of enquiry is to build a solid picture of information transmission through language transmission, based on formal notions of learning as compression while at the same time maintaining clear parallels with what we know about language processing in humans. This pluralistic approach is motivated by the fact that "some of the most important formal properties of a theory are found by contrast, and not by analysis" [39, p. 30]. In this sense, we aim to investigate linguistic evolution through iterated learning in the light of models which approach the problem from complementary perspectives.

4. The cultural evolution of linguistic structure: The role of the transmission bottleneck

In Section 4.1 we develop a simple associative model of language acquisition, which will be used to investigate two factors. Firstly, in Section 4.2, we will use the model to demonstrate a basic result linking the cultural transmission of language with an aspect of linguistic structure, namely compositionality. This constitutes one of the core findings of research on linguistic evolution. Secondly, in Section 5, we will use this model to investigate the role of learner biases in the evolution of compositional structure.

4.1. An associative matrix model of learning and production

As discussed in Section 3.2, languages can be viewed as a system mapping between a space of meanings and a space of signals. One of the simplest ways to model a linguistic agent capable of manipulating a system of meaning-signal mappings is to use an association matrix—a matrix specifying association strengths between meanings and signals, where entry a_{ij} in the matrix gives the strength of association between meaning i and signal j. An agent's production and reception behaviour is determined by association strengths in that agent's matrix, and learning involves adjusting association strengths according to some learning procedure. This approach is frequently used to study the evolution of signalling systems, where meanings and signals are unstructured, atomic entities (see, e.g., [44,71,84,86,90]).

A minimal elaboration to this basic scheme permits such a model to be used to model the learning of associations between structured meanings and structured signals [87]. A linguistic agent is defined by an association matrix A. Entries in A give the strength of association between both partial and complete meanings and signals (as defined below). As in the simpler model, production and reception behaviour are determined by the association matrix, and learning involves adjusting association strengths.

4.1.1. Representation

As summarised in Section 3.2, meanings are vectors in an F-dimensional space where each dimension has V values. *Components* of meanings are vectors such that each feature of a component has either the same value as the meaning in question, or a wildcard. More formally, if c_m is a component of meaning m, then the value of the jth feature of c_m is:

$$c_m[j] = \begin{cases} m[j] & \text{for specified features,} \\ * & \text{for unspecified features} \end{cases} \tag{1}$$

where $*$ represents a wildcard. Similarly, components of signals of length l are (possibly partially specified) strings of length l. We impose the additional constraint that a component must have a minimum of one specified position—$c_m\left[j\right] \neq *$ for all j.

Each row of an A matrix corresponds to a component of a meaning, and there is a single row in A for each component of every possible meaning. Similarly, each column in A corresponds to a component of a signal. The entry a_{ij} in matrix A therefore gives the strength of association between a meaning component i and a signal component j.

4.1.2. Learning

Prior to learning, all entries in *A* have a value of 0. During a learning event, a learner observes a meaning-signal pair $\langle m, s \rangle$.[8] The meaning *m* will specifies a set of meaning components $\mathcal{C}_m$ and the signal *s* specifies a set of signal components $\mathcal{C}_s$. The learner then updates its *A* matrix according to the learning procedure:

$$\Delta a_{ij} = \begin{cases} \alpha & \text{if } i \in \mathcal{C}_m \text{ and } j \in \mathcal{C}_s, \\ \beta & \text{if } i \in \mathcal{C}_m \text{ and } j \notin \mathcal{C}_s, \\ \gamma & \text{if } i \notin \mathcal{C}_m \text{ and } j \in \mathcal{C}_s, \\ \delta & \text{if } i \notin \mathcal{C}_m \text{ and } j \notin \mathcal{C}_s. \end{cases} \tag{2}$$

This is exactly equivalent to the learning procedure from Smith [86], but with respect to components of meanings and signals, rather than unanalysed meanings and signals. The key point is that assignment of values to α, β, γ and δ specifies a particular way of learning, or updating association strengths. Different assignments yielding a range of possible ways of learning, an issue we will turn to in Section 5.

4.1.3. Production and reception

An *analysis* of a meaning or signal is an ordered set of components which fully specifies that meaning or signal. More formally, an analysis of a meaning *m* is a set of *N* components $\{c_m^1, c_m^2, \ldots, c_m^N\}$ that satisfies two conditions:

(1) If $c_m^i[j] = *$, $c_m^k[j] \neq *$ for some choice of $k \neq i$,
(2) If $c_m^i[j] \neq *$, $c_m^k[j] = *$ for any choice of $k \neq i$.

The first condition states that an analysis may not consist of a set of components which all leave a particular feature unspecified—an analysis fully specifies a meaning. The second states that an analysis may not consist of a set of components where more than one component specifies the value of a particular feature—analyses do not contain redundant components. Valid analyses of signals are similarly defined.

During the process of producing utterances, agents are prompted with a meaning and required to produce a meaning-signal pair. In order to retrieve a signal *s* based on an input meaning *m* every possible signal $s_j \in \mathcal{S}$ is evaluated with respect to *m*. For each of these possible meaning-signal pairs $\langle m, s_j \rangle$, every possible analysis of *m* is evaluated with respect to every possible analysis of s_j. The evaluation of a meaning analysis-signal analysis pair yields a score *g*, as defined by Eq. (3). The meaning-signal pair which yields the analysis pair with the highest *g* is returned as the agent's production for the given meaning. The score for a meaning analysis (which consists of a set of meaning components) paired with a signal analysis (a set of signal components) is given by:

$$g\left(\{c_m^1, c_m^2, \ldots, c_m^N\}, \{c_s^1, c_s^2, \ldots, c_s^N\}\right) = \sum_{i=1}^{N} \omega(c_m^i) \cdot a_{c_m^i, c_s^i}, \tag{3}$$

where *N* is the number of components in the analysis of meaning and signal, $a_{c_m^i, c_s^i}$ gives the strength of the association between the *i*th component of the meaning analysis and the *i*th component of the signal analysis and $\omega(x)$ is a weighting function which gives the non-wildcard proportion of *x*.

---

[8] We therefore assume that learners have the capacity to identify the intended meaning of an utterance, as well as the signal produced for that meaning. This is sometimes called the assumption of *explicit meaning transfer* [83].

4.2. Transmission bottlenecks and the pressure to generalise

Using the extended *A* matrix model outlined above, we will consider the impact of a *transmission bottleneck* on a population's communication system.

The transmission bottleneck reflects the fact that in nature languages cannot be transmitted in totality from one individual to another. Languages are capable of expressing an infinite range of concepts, and any member of this infinite array of expressions is interpretable in turn. Acquiring a language therefore entails the acquisition of a system for producing and understanding such an infinite set of meaningful utterances. However, the system for generating this infinite set of utterances must be acquired from a finite set of data—it is necessarily true that language learners do not see all the sentences of a language during the language learning process, because this would take an infinite amount of time. This transmission bottleneck is one aspect of the poverty of the stimulus problem, which is typically advanced as an argument suggesting that linguistic structure must be largely prespecified in language learners.

We will test the consequences of the transmission bottleneck using an implementation of the Iterated Learning Model with the processes of language learning and production modelled using the associative matrix model outlined above. Recall that in an Iterated Learning Model, language is transmitted between generations via production and learning. Learners observe *e* meaning-signal pairs produced by the individual at the previous generation. If these *e* observations are selected so that the learner observes every meaning from the space of possible meanings at least once, paired with its associated signal, then the learner observes the complete language of the previous generation. We will call this the *no bottleneck* case—note that this no bottleneck condition cannot apply in the case of natural language. In contrast, if the meanings expressed in these *e* observations are selected purely at random, then the learner may not observe the complete language of the previous generation—they may only observe a subset of that language, and when called upon to produce they may be required to produce a signal for a meaning which they themselves never observed expressed. We will call this (more realistic case) the *bottleneck* condition.

We will begin by considering a single learning rule, and investigating the effect of a transmission bottleneck, before returning to the issue of learning strategies in Section 5. We will begin by considering the associative learning rule defined by $\alpha = 1, \beta = -1, \gamma = -1, \delta = 0$—connection strengths between meaning and signal components which occur together are strengthened, and connection strengths between meaning and signal components which differ in their occurrence are decreased ($\beta = -1, \gamma = -1$).[9] The measures of interest is the compositionality of the emergent languages. Our measure of compositionality is given in Appendix A—this ranges from ≈ 0 for a holistic system, to ≈ 1 for a compositional language.

Fig. 3 shows compositionality over time, in the two experimental conditions. The graphs plot the mean and standard deviation of compositionality, averaged over 100 runs in each condition, against time in generations, with runs allowed to proceed until a stable system emerges.[10]

As can be seen from Fig. 3, the presence or absence of a transmission bottleneck has a significant impact on the population's language. In the initial generation in both experimental conditions, compositionality is at baseline levels, reflecting the random nature of the initial languages—the initial generation of each population produces a random set of utterances, due to initial association strengths of 0. In the

---

[9] Other simulation parameters: $F = 3$, $V = 4$, $l_{max} = 3$, $|\Sigma| = 8$, $e = 100$ in the no bottleneck condition, $e = 32$ in the bottleneck condition.

[10] Mean and standard deviation of 100 runs are plotted for all results in Sections 4 and 5.
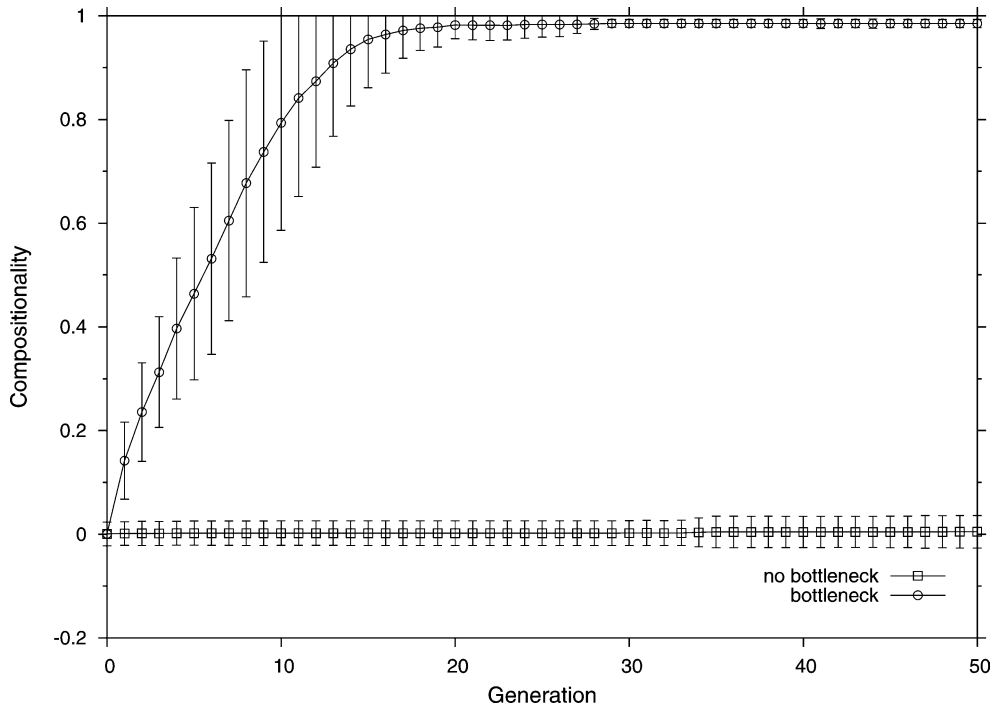
Fig. 3. The impact of a transmission bottleneck on the structure of language.

no bottleneck condition compositionality remains at base-line levels over multiple generations of cultural transmission. In contrast, in the bottleneck condition a highly compositional language evolves—all simulation runs converge on languages with compositionality of approximately 1.

The results for the no bottleneck condition reflect the persistence of the initial, holistic system of meaning-signal mappings—compositional mappings do not emerge when there is no bottleneck on transmission. In the bottleneck condition, this is not the case—a system of meaning-signal mappings evolves in which the structure of a meaning is transparently reflected in the structure of the signal associated with that meaning. Table 1 shows fragments of languages evolving during a particular simulation run—an initial holistic language, and a final compositional language. In short, compositionally-structured languages evolve through cultural processes, but only when there is a bottleneck on transmission. Why is this?

When there is no bottleneck on transmission, learners observe the complete language of the previous generation. This can simply be memorised. The system embodied in the random meaning-signal pairs produced by the initial generation will be a holistic one, and this system will be preserved over time.

However, holistic mappings cannot persist in the presence of a bottleneck. The meaning-signal pairs of a holistic language are arbitrary with respect to structure—the structure of a meaning is in no way reflected in the structure its associated signal. As such, the meaning-signal pairs of a holistic language must be observed if they are to be reproduced. When a learner only observes a subset of a holistic language then certain meaning-signal pairs will not be observed and therefore will not be preserved—the learner, when called upon to produce, may produce some other signal for that meaning, resulting in a change in the language—holistic languages are not stable when there is a bottleneck on transmission.

Table 1
Fragments of initial and final languages from one simulation run

Meaning	Signal in initial language	Signal in final language
(3, 3, 3)	db	def
(3, 3, 2)	cfc	ded
(3, 2, 3)	cfh	daf
(3, 2, 2)	deg	dad
(2, 3, 3)	fbg	fef
(2, 3, 2)	gae	fed
(2, 2, 3)	chg	faf
(2, 2, 2)	cbc	fad
(1, 3, 3)	cgg	gef

Compositionality of initial language: -0.025. Compositionality of final language: 0.991.

In contrast, compositional languages are *generalisable*, due to their structure. In a compositional language there is a regular relationship between feature values and parts of signal—for example, as can be seen Table 1 above, in one compositional language from one simulation run, value 1 for feature 1 maps to string-initial g, value 2 for feature 2 maps to string-medial a, and value 2 for feature 3 maps to string-final d. This structure in the mapping allows learners to generalise from observed meaning-signal pairs in order to produce the appropriate signal for meanings which they were not exposed to during learning. For example, the regularities sketched out above allow us to (correctly) predict that the signal for meaning (1, 2, 2) should be gad, even though meaning (1, 2, 2) is not included in our sample of this language.

The potential of compositional languages to be generalised allows such languages to remain relatively stable over repeated episodes of cultural transmission, even when the learner only observes a subset of the language of the previous generation. Holistic languages cannot be stable under such conditions. The transmission bottleneck therefore introduces a pressure for languages to be generalisable. Over time, languages adapts to this pressure, eventually becoming highly compositional, highly generalisable and consequently highly stable.

This, then, constitutes a basic result for investigations into the cultural evolution of language—a bottleneck on cultural transmission introduces pressure for language to be generalisable, and language adapts to this pressure over time. This result has been demonstrated using a fairly wide range of models of language, language learning and iterated learning—see, e.g., [4,10,50,52,87]. Compositionality represents an adaptation, by language, to the circumstances of its transmission. This explanation linking a feature of linguistic transmission (the transmission bottleneck) with a particular aspect of linguistic structure (compositionality) therefore constitutes an example of how certain linguistic universals can arise as a consequence of linguistic evolution, rather than being prespecified in the genes.

5. The cultural evolution of linguistic structure: The role of the language learner

The results outlined in the previous section were for a particular learning rule—a particular system of altering association strengths in a matrix based on observed meaning-signal pairs, given by a particular assignment of values to the parameters α, β, γ and δ. There are obviously alternative assignments,

leading to alternative learning strategies. To what extent is the result linking the transmission bottleneck with the compositionality of the evolved language dependent on this *particular* learning strategy?

A brief survey of the literature suggests that this result has generality beyond the particular learning model used here. Various other, often significantly different, learning models yield the same basic result—to name but a few, heuristic driven induction of context-free grammars [52], exemplar-based learning [4], and MDL-based induction of finite state transducers ([10] and see Section 6). However, the learning strategies implemented in this apparently disparate selection of models may in fact share some fundamental properties in common, and these shared properties may be crucial to the evolution of compositional structure.

The associative matrix model of learning allows us to make a systematic investigation of alternate learning strategies, and their consequences for the evolution of linguistic systems.

5.1. Learning strategies supporting generalisation

The main result from the previous section was that the transmission bottleneck introduces a pressure to generalise, and that the evolving linguistic system adapts to this pressure. A necessary precondition for this is that the language learners are *capable* of generalising—capable of identifying, extracting and exploiting the regularities in a compositional language. Given that this proves to be the case, the associative matrix learning according to the rule given above must be capable of generalisation, and this capacity for generalisation must correspond to the particular values used for α, β, γ and δ. By pinpointing the locus of this capacity to generalise, we can switch it on or off, and verify that it is in fact a prerequisite for the evolution of compositional structure.

Details of the process of working out learning bias with respect to generalisation are necessarily somewhat involved, and we would refer the interested reader to Smith [85] for details. Briefly: the ability of an associative matrix learner to generalise depends on the relationship between the values assigned to learning rule parameters α and δ. This relationship determines the learner's preference for using *componential* analyses—for producing meaning-signal pairs by associating parts of meaning (feature values or collections of feature values) with parts of signal, rather than atomistically associating unanalysed meanings with an unanalysed signals. The capacity to extract and use componential analyses is the basis of the capacity to generalise.

To illustrate this, let us return to the example compositional language given in Table 1. As discussed above, this language exhibits regularities: value 1 for feature 1 maps to string-initial g, value 2 for feature 2 maps to string-medial a, and value 2 for feature 3 maps to string-final d. Identifying these regularities corresponds to making a componential analysis of the system of the meaning-signal mappings, and allows us to generalise to the unseen meaning $(1, 2, 2)$, and others.

In contrast to making such a componential analysis, we could simply memorise the associations between complete meanings and complete signals embodied in the fragment of compositional language given in Table 1—meaning $(3, 3, 3)$ maps to signal def, meaning $(3, 3, 2)$ maps to ded, and so on. This is the atomistic approach. In this case, due to our failure to make a componential analysis, we cannot generalise to the signal associated with unseen meanings such as $(1, 2, 2)$—although the data we have observed has structure, failure to analyse this data componentially means that generalisation is impossible.

Whether or not a learner takes the atomistic approach, the componential approach, or alternates between the two, depends on the relationship between α and δ in the learning rule used by that learner. Briefly:

$\alpha > \delta$: Preference for the componential analysis. This gives the capacity to generalise.

$\alpha = \delta$: Neutral between the componential and atomistic approaches. This leads to inability to generalise reliably—while the componential approach may be used on one occasion to allow a generalisation to be made, on another occasion the atomistic method may be used, leading to failure to generalise.

$\alpha < \delta$: Preference for the atomistic analysis. This leads to the inability to generalise.

$\alpha > \delta$ is therefore a requirement which must be in place if a learner is capable of generalising, and we should expect that this learner capacity is required if compositional languages are to evolve as shown in the previous section. Note, however, that this capacity for generalisation does not *guarantee* the emergence of compositional structure—as shown in the previous section, even given a learner capable of generalising, compositionality will not emerge if there is no bottleneck on transmission. Further note that the preference for the componential analysis resulting from $\alpha > \delta$ does not mean that this approach will be used if the data does not contain regularities.

What happens over time in an Iterated Learning scenario, where learners do not have the capacity to generalise? Fig. 4 shows the results of simulations for two learning rules:

- The standard rule, which supports generalisation: $\alpha = 1$, $\beta = -1$, $\gamma = -1$, $\delta = 0$. This was the rule used in the previous section.
- A modified variant of this rule, incapable of reliably generalising: $\alpha = 1$, $\beta = -1$, $\gamma = -1$, $\delta = 1$.
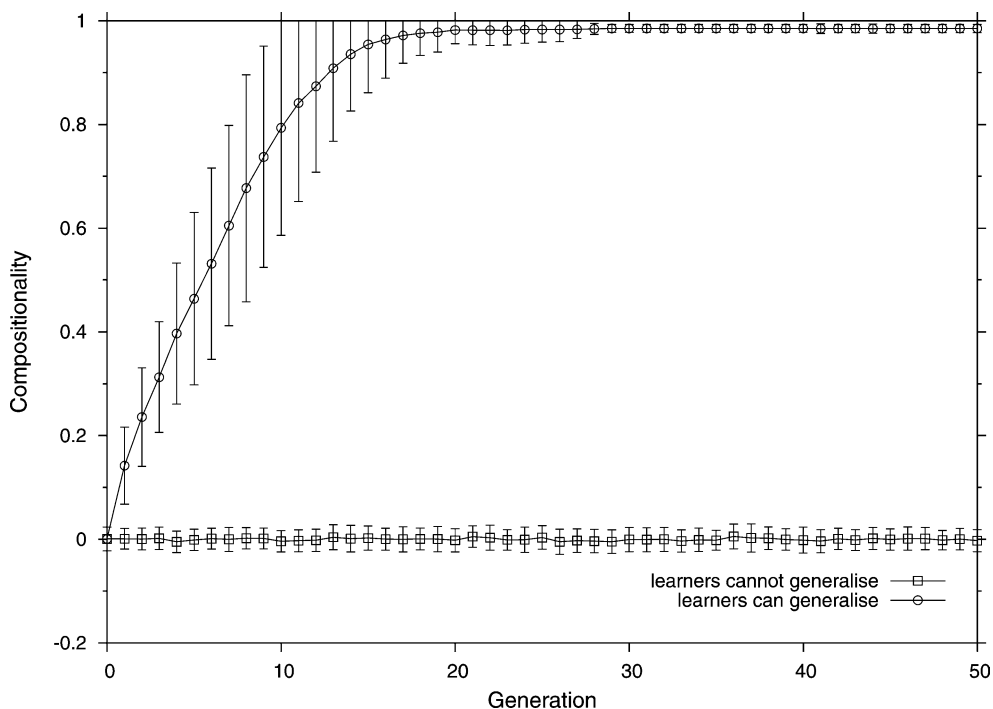

Fig. 4. The importance of the capacity to generalise.

For both sets of results in Fig. 4 there is a bottleneck acting on transmission.

As shown in Fig. 4, when learners do not have the capacity to generalise we no longer see the evolution of compositional languages—compositionality remains at random levels over time when agents learn according to such rules, in spite of the pressure to generalise introduced by the transmission bottleneck. This is in contrast to the behaviour when learners have the capacity to generalise arising from $\alpha > \delta$, also shown in Fig. 4 for comparison. In other words, a learner capacity to generalise is, as expected, a prerequisite for the evolution of compositional structure through cultural processes.

5.2. Learning strategies supporting communicative function

Up to this point we have only considered the *structure* of the evolving languages, and demonstrated a link between the transmission bottleneck and the evolution of compositional structure. In a sense, compositionality is *functional* from the point of view of languages themselves—compositionality allows a language or a subregion of a language to survive repeated passage through the transmission bottleneck. We have said nothing about an alternative notion of functionality—the functionality that a language provides to users of that language.

There are several ways in which language could be useful to language users. It could be that language is useful in as much as it allows language users to communicate with other language users—this is the type of functionality which, for example, Pinker and Bloom [73] suggest is responsible for the biological evolution of the human capacity for language. An alternative function of language could be the extent to which it allows language users to signal their social identity, so as to affiliate themselves with certain social groups and dissociate themselves from others. This proposed function of language, and the associated notions of prestige, covert prestige, and acts of identity, forms the basis of much research in sociolinguistics (e.g., [54,94]).

Could functionality of this sort could also drive the evolution of linguistic systems? In the general case, Boyd and Richerson [7] argue that any culturally transmitted system can evolve under "natural selection of cultural variants", such that variants of cultural traits which maximise the probability of an individual surviving long enough to transmit their trait culturally and/or reproducing disproportionately frequently and transmitting their variant to their offspring, will come to dominate in a population. Boyd and Richerson provide a number of domains in which empirical evidence suggests that this kind of evolution might be observed, as do Mesoudi et al. [66]. Dealing specifically with language, the models of Martin Nowak and colleagues (see [69] for review) reflect the assumption that reproductive fecundity impacts on the likelihood of individual's linguistic system being culturally transmitted. Kirby [49] examines in detail how linguistic function can affect linguistic structure through cultural processes, natural selection, and the interaction between culture and biological evolution.

In the results outlined in Section 4, a consideration of communicative function played no role in the evolution of compositional languages—compositionality represents an adaptation by the language itself to the pressure for generalizability introduced by the transmission bottleneck. It might also be the case that compositionality proves to be functional from the perspective of language users. For example, the relative stability across generations of a compositional language, even in the face of a transmission bottleneck, may be useful for language users. As such, considerations of communicative function could play a role in driving the evolution of compositionality, assuming that a bottleneck is present and that the capacity to generalise is in place and so on. However, there is no need to appeal to this notion of functionality to explain the cultural evolution of compositional structure—compositionality evolves in

response to the pressure introduced by the transmission bottleneck, and it is therefore unnecessary to invoke an explanation which appeals to the combination of a transmission bottleneck plus a pressure for communication.

However, we have not yet addressed the functionality of the compositional systems which evolve under the pressure arising from the transmission bottleneck. It could be that these languages are compositional in structure but useless in terms of communication, in which case there may indeed be a role for communicative function to play in our explanation.

Our measure of communicative accuracy is given in Appendix A. Informally, communicative accuracy between two individuals is the probability, averaged over all meanings, of one of those individuals producing a signal for a given meaning, and the other individual interpreting the received signal as conveying the same meaning that the speaker intended it to. This evaluates to 1 for communicatively optimal systems, and $1/|V^F|$ for a random system. For the results presented here, communicative accuracy is measured across generations—it is an evaluation of the probability with which the individual from generations n and $n + 1$ will successfully communicate.[11] A similar measure to the one used here is often applied to an evaluation of communicative accuracy within generations, where each generation consists of multiple individuals (see, e.g., [44,71,86]). A within-generation measure of communicative accuracy, where an individual's ability to communicate with itself using their signalling system, yields qualitatively similar results to those given here using the across-generation measure.

Fig. 5 shows communicative accuracy over time, given the standard learning rule used in Section 4 ($\alpha = 1$, $\beta = -1$, $\gamma = -1$, $\delta = 0$), in the bottleneck and no bottleneck conditions. In both cases, systems which are optimal for communication emerge and remain stable. This reflects the convergence of a conventionalised meaning-signal mapping which is passed down intact from generation to generation, even in the presence of a transmission bottleneck. Note that the systems in the two experimental conditions are structurally rather different—as demonstrated earlier, they differ in their degree of compositionality—but they both perform optimally in terms of communicative function. This is despite the fact that there is no pressure for agents to communicate—individuals are not rewarded for more successful communication, and individuals do not take communicative function into account during learning.

What, then, drives the evolution of these optimal communication systems? The explanation must reside in the process of language learning—by a similar method to that used in the previous section, we can identify and experimentally vary the learning bias which leads to the evolution of optimal communication.[12]

It is worth considering the space of possible systems of meaning-signal mappings with a view to their communicative function. Meaning-signal mappings can embody many-to-one mappings (a) (Table 2), one-to-one mappings (b), or one-to-many mappings (c), regardless of whether we consider the system at the level of mappings between complete complex meanings and complete signals, or between particular feature values and signal substrings.

Many-to-one mappings, where several distinct meanings (or subparts of meaning) map to a single signal are suboptimal in terms of communication, because the intended meaning of the ambiguous signal cannot be reliably retrieved. The optimal communication systems which evolve in the simulations illustrated in Fig. 5, as we might expect, do not contain many-to-one mappings.

---

[11] Note, therefore, that an evaluation of communicative accuracy at the initial generation of the population is impossible, as there is no preceding generation to communicate with.

[12] We would refer the reader to [84] for a similar analysis for the case of unstructured signalling systems
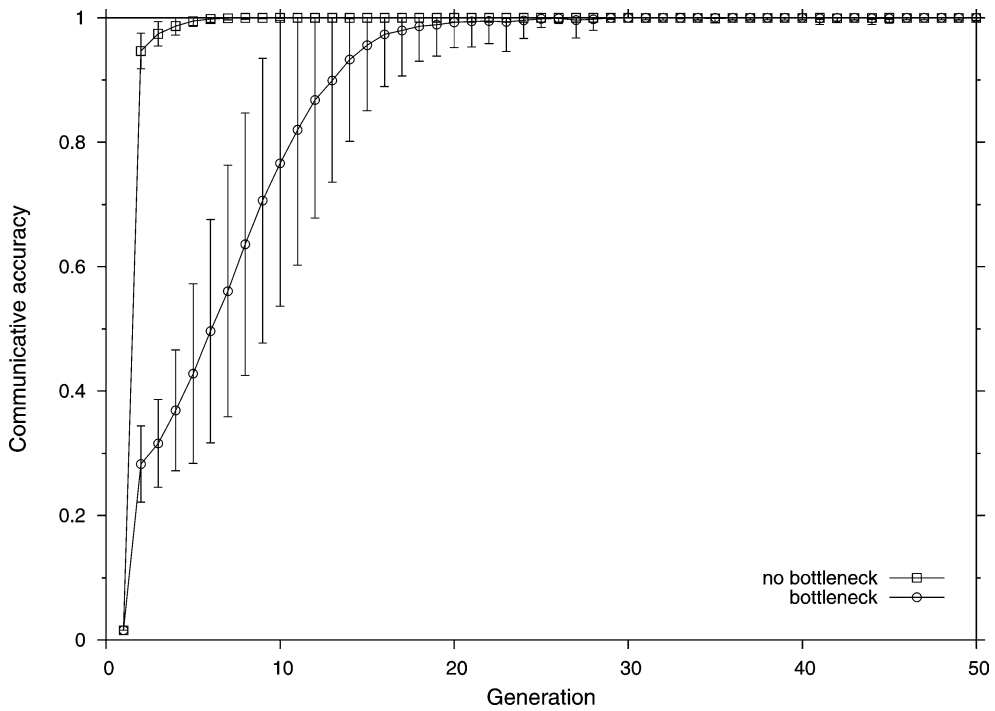
Fig. 5. Evolution of optimal communication systems, regardless of presence or absence of a transmission bottleneck.

Table 2
Various types of meaning-signal mapping

	many-to-one	one-to-one	one-to-many
whole–whole	(3, 2, 1) (3, 3, 3) ⟶ def (1, 1, 1)	(3, 3, 3) ⟶ def	(3, 3, 3) ⟶ esox / def / ef
part–part	(∗, 2, ∗) (3, ∗, ∗) ⟶ d** (∗, 1, 1)	(3, ∗, ∗) ⟶ d**	(3, ∗, ∗) ⟶ **ox / d** / ef

The holistic languages which evolve in the absence of a transmission bottleneck, the systems avoid many-to-one mappings at the level of whole meanings and whole signals—in no case is there a complete meaning which maps to the same complete signal as another complete meaning. For example, if meaning

$(1, 2, 3)$ maps to signal cbb then no other meanings map to the signal cbb. There is no prohibition on distinct feature values appearing to co-occur with particular signal substrings, as such part-part mappings are not exploited by learners learning a holistic system.

In contrast, the compositional systems which evolve in the presence of a transmission bottleneck the systems avoid many-to-one mappings at the level of individual features values and signal substrings—no two values for a given feature map to the same signal substring. For example, in the compositional language in Table 1, feature 1 value 3 maps to string-initial d, and no other value for any other feature maps to string initial d. A consequence of the absence of many-to-one mappings at this level, in combination with the compositionality of the mapping, is the absence of many-to-one mappings at the level of whole meanings and whole signals.

One-to-one and one-to-many mappings are unproblematic from the point of view of communication—in both cases the intended meaning can be retrieved from the observed signal. We might therefore expect the evolved systems to contain examples of both one-to-one and one-to-many mappings. However, they are in fact exclusively one-to-one—one-to-many mappings are not observed in the final systems. We will return to this point in Section 5.3.

What drives the elimination of many-to-one mappings from the evolving linguistic systems? This is a consequence of the learning bias of learners using the standard learning rule $\alpha = 1$, $\beta = \gamma = -1$, $\delta = 0$—learning according to this weight update rule biases learners against acquiring many-to-one mappings, such that many-to-one mappings (either between complete meanings and signals or parts thereof) are less likely to be successfully learned than mappings which are not many-to-one.

As demonstrated in Smith [85], a learner's bias with respect to many-to-one mappings depends on the relationship between γ and δ. Briefly:

$\delta > \gamma$: Bias against many-to-one mappings—many-to-one mappings are less likely to be successfully learned.
$\delta = \gamma$: Neutrality — many-to-one mappings are learnable.
$\delta < \gamma$: Bias in favour of many-to-one mappings—systems involving many-to-one mappings are more likely to be successfully learned.

These learner biases introduce a further pressure acting on the language system during its cultural transmission, and language changes over repeated learning episodes in response to these learner biases, with mappings of the disfavoured types being eliminated. In the simulation results shown in Fig. 5, the learner bias against many-to-one mappings leads to the elimination of such mappings, with convergence on a stable, unambiguous linguistic system. Such a system allows optimal communication across generations.

The bias with respect to many-to-one mappings is independent from the capacity of a learning rule to generalise. Recall from Section 5.1 above that individuals learning using the rule $\alpha = 1$, $\beta = -1$, $\gamma = -1$, $\delta = 1$ are incapable of generalising (as $\alpha = \delta$), and consequently compositional systems do not emerge in such populations (recall Fig. 4). However, this learning strategy results in learners disfavouring many-to-one mappings (as $\delta > \gamma$). Fig. 6 shows the compositionality and communicative accuracy of the evolving systems in populations learning according to this rule, in both the bottleneck and no bottleneck conditions.

As expected, in both cases compositional systems do not evolve—the populations converge on holistic systems. However, those holistic systems offer some degree of communicative functionality. In the case

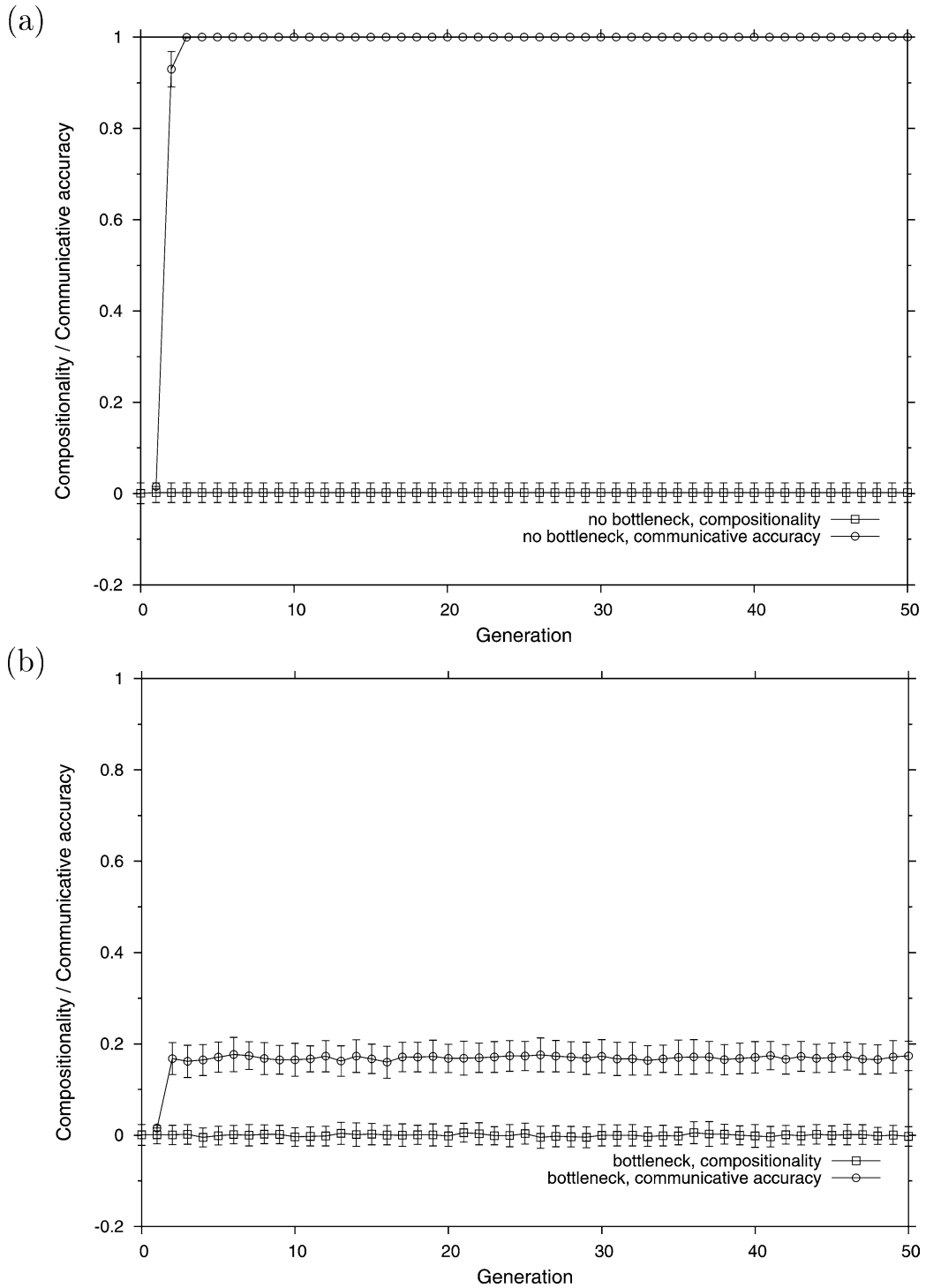*H. Brighton et al. / Physics of Life Reviews 2 (2005) 177–226*

(a)

(b)

Fig. 6. Evolution of languages without the capacity for generalisation, but with a bias against ambiguity, in the no bottleneck (a) and bottleneck (b) conditions.
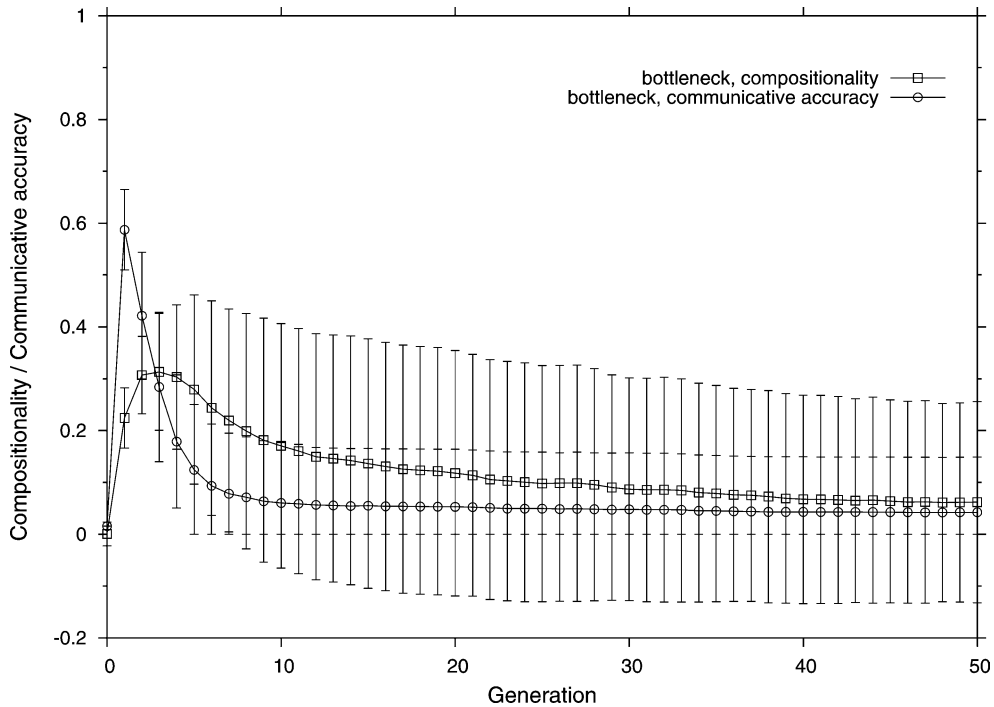
Fig. 7. Evolution of languages with the capacity for generalisation, but without a bias against ambiguity.

where there is no bottleneck on transmission, optimal communication systems rapidly evolve. In the case where there is a bottleneck on transmission, communication systems evolve which give communicative accuracy of slightly less than 0.2—the functionality of these systems across generations is suboptimal due to their instability, but is still greater than chance, reflecting the one-to-one nature of that proportion of mapping which is stable across generations. In other words, in both conditions populations of such learners evolve systems which are non-compositional (as they are incapable of generalising), but which tend to embody a system of one-to-one mappings.

The converse dissociation, where learners are capable of generalising but not biased against many-to-one mappings, is also possible. Learners using the rule $\alpha = 1$, $\beta = -1$, $\gamma = 0$, $\delta = 0$ have such a combination of biases. Fig. 7 shows the compositionality and communicative accuracy of the evolving systems in populations learning according to this rule, in both the bottleneck and no bottleneck conditions.

Given such a combination of biases, we might expect the emergence of a system which is compositional (due to the learner capacity to generalise and the transmission bottleneck), but which only offers intermediate levels of communicative function (due to the lack of any learner bias against many-to-one mappings). As can be seen from Fig. 7, such populations in fact converge on a linguistic system which is useless for communication (in fact, maximally ambiguous, where every meaning maps to a single signal) and, in spite of their capacity for generalisation, non-compositional. This behaviour is a consequence of the interaction between the capacity to generalise and the neutrality with respect to ambiguity.

To see how this is so, consider a scenario where a population of learners learning according to such a strategy is presented with a perfectly compositional, perfectly unambiguous language—this is not the

scenario in the model, but is a useful example. In such a population, a many-to-one mapping between parts of meaning and parts of signal will occur by chance (either due to the randomness in production as a consequence of the transmission bottleneck, or due to noise). For example, suppose the target language is the perfectly compositional, unambiguous L_c.

$$L_c = \{\langle\{1, 1\}, \mathsf{ac}\rangle, \langle\{1, 2\}, \mathsf{ad}\rangle, \langle\{2, 1\}, \mathsf{bc}\rangle, \langle\{2, 2\}, \mathsf{bd}\rangle\}.$$

A learner exposed to the subset L'_c will be capable of reconstructing L_c via generalisation.

$$L'_c = \{\langle\{1, 1\}, \mathsf{ac}\rangle, \langle\{1, 2\}, \mathsf{ad}\rangle, \langle\{2, 1\}, \mathsf{bc}\rangle\}.$$

However, consider a learner exposed to the noisy subset L_{noise}.

$$L_{\mathrm{noise}} = \{\langle\{1, 1\}, \mathsf{ac}\rangle, \langle\{1, 2\}, \mathsf{ad}\rangle, \langle\{2, 1\}, \mathbf{ac}\rangle\}.$$

The learner has to decide what values 1 and 2 for feature 1 should map to. Should they both map to a? Or should they map to distinct characters? Learners with the bias against many-to-one mappings will select the latter option, and will generalise to produce a perfectly compositional (although perhaps changed) language, such as L_{new}.

$$L_{\mathrm{new}} = \{\langle\{1, 1\}, \mathsf{ac}\rangle, \langle\{1, 2\}, \mathsf{ad}\rangle, \langle\{2, 1\}, \mathbf{fc}\rangle, \langle\{2, 2\}, \mathbf{fd}\rangle\}.$$

In contrast, a learners with no bias against many-to-one mappings will take the first option and generalise to produce $L_{\mathrm{ambiguous}}$:

$$L_{\mathrm{noise}} = \{\langle\{1, 1\}, \mathsf{ac}\rangle, \langle\{1, 2\}, \mathsf{ad}\rangle, \langle\{2, 1\}, \mathbf{ac}\rangle, \langle\{2, 2\}, \mathbf{ad}\rangle\}.$$

This language is clearly ambiguous, and in particular the ambiguity has spread form the signal associated with meaning (2, 1) to the signal associated with (2, 2). In this way, randomly-occurring ambiguities rapidly spread in populations of learners who are not biased against ambiguity but are capable of generalising. Note that the ambiguous system above is *not* compositional, according to our measure—the ambiguity destroys the structure-preserving nature of the meaning-signal mapping, as different meanings now map to similar signals.

The biases of language learners with respect to ambiguity (many-to-oneness) therefore interact with their capacity to generalise, or lack thereof, and impact on the structure and functionality of a population's linguistic system. The results presented in this section show that a learner bias against ambiguity leads to a linguistic system which is communicatively functional. However, and perhaps more surprisingly, such a bias is also a prerequisite for the cultural evolution of compositional structure—without a learner bias against many-to-one mappings, languages in which the structure of signals reflect the structure of meanings do not arise. As such, our explanation for linguistic structure also offers an explanation for linguistic function—the model results presented here suggest that the prerequisites for compositionality also deliver communicative function as a side-effect, without the necessity for any explicit pressure for communication.

5.3. Learning strategies supporting language learning

As discussed above, many-to-one mappings are inherently bad for communication, as they introduce ambiguity. In contrast, one-to-many meaning-signal mappings do not endanger communication—a producer's intended meaning can always be retrieved by a receiver who knows the system. We might

therefore expect learner biases with respect to one-to-many mappings to be irrelevant for the structure and function of a population's linguistic system.

This proves not to be the case. As demonstrated in Smith [84], and in Smith [86] for the case of simple signalling systems, a learner's bias with respect to one-to-many mappings depends on the relationship between α and β. Importantly, rules where $\alpha > \beta$ have a bias against one-to-many mappings—this is the case with all learning rules used so far. The biases of rules where $\alpha \leqslant \beta$ biases tend to be rather idiosyncratic and depend on the values of γ and δ, but generally speaking such rules are not biased against one-to-many mappings and may indeed be biased in favour of such mappings.

Having a bias against one-to-many mappings turns out to be crucial if a system of meaning-signal mappings is to be learnable *at all*. To see that this is the case, consider the case where a learner attempts to acquire a mapping between a single meaning and several signals. During learning the learner observes the meaning paired with one of the possible signals. During production the learner must then decide which signal to produce for the meaning, given their observations. The sensible behaviour is to reproduce the observed signal—this is what learners using rules where $\alpha > \beta$ tend to do. However, this behaviour implies a bias against one-to-many mappings—learning according to such a procedure means (defeasibly) discounting the possibility that the meaning maps to some other signal or signals. Alternative learning strategies, where all possible signals are produced with equal probability for the particular meaning, or where some other signal or signals are produced for the meaning, result in failure to reproduce the meaning-signal pair observed by the learner. As such, a bias against one-to-many mappings is required if a linguistic system is to be learned. The consequence of this bias over cultural time is that one-to-many mappings will be eliminated over time—in combination with a learner bias against many-to-one mappings, this leads to emergent systems which map elements of meaning to elements of signal in a perfectly transparent, one-to-one fashion.

5.4. Learning bias in humans

The simulation results presented above highlights three elements of learning as being important in the evolution of linguistic structure:

(1) Learners must have the capacity to generalise.
(2) Learners must be biased against acquiring one-to-many meaning-signal mappings.
(3) Learners must be biased against acquiring many-to-one meaning-signal mappings.

Without (1) compositional structure cannot evolve. Without (2), a system of meaning-signal mappings cannot be acquired. Without (3), neither compositional nor communicatively functional linguistic systems can evolve.[13] However, given all three components, linguistic systems evolve which are functional both from the perspective of the linguistic system (the linguistic system is stable from generation to generation, even in the presence of a transmission bottleneck), and from the point of view of language users (the linguistic system allows perfect communication between individuals).

---

[13] At least without some further pressure for function, such as natural selection acting on cultural transmission. It may be, however, that linguistic evolution resulting from learner biases tends to drown out linguistic evolution driven by natural selection [86]. In other words, having the wrong learning bias makes it difficult to evolve functional communication systems through cultural processes, even with explicit selection for communication.

What biases do human language learners bring to the language acquisition task?

There is a body of evidence from the developmental linguistics literature which suggests that child language learners do indeed possess the biases outlined above in (1)–(3)—this evidence is reviewed briefly below. Children's capacity to generalise is uncontroversial (Section 5.4.1). The claim that human language learners are biased against acquiring one-to-many meaning-signal mappings is supported by a good deal of evidence, and is more or less widely accepted (Section 5.4.2). Research supporting the claim that children are also biased against acquiring many-to-one mappings is at an earlier stage of development, but some recent experimental work suggests that such a bias may indeed be present (Section 5.4.3).

5.4.1. Capacity to generalise

Human language learners have the capacity to generalise. Indeed, it would be extraordinary were this not the case. Firstly, the capacity to exploit similarity structure in the environment in order to generalise to novel situations is a basic property of learning, certainly in connectionist architectures such as the brain [37,79]. In other words, this is not necessarily a language-specific or indeed species-specific capacity. Secondly, the infinite expressivity of human language tells us that human language learners *must* be making generalisations over the data they observe—were language learning merely to proceed by memorisation, with no generalisation, no human with a finite lifespan could come to command an infinitely expressive language.

Thirdly, in addition to these general arguments, specific experimental evidence demonstrates the capacity of children to generalise in a linguistic context. In English, the plural form of nouns (excluding irregulars) is formed by the addition of a suffix *-s* to the noun stem, yielding, for example, *dogs* from *dog*. Furthermore, the realization of the *-s* morpheme depends on the preceding consonant. The unmarked allomorph (variant) of *-s* is realized as /z/, as in *dogs*. If the *-s* morpheme suffixes to a stem ending in a voiceless stop, it is realized as /s/, as in *cats*. Finally, if the plural morpheme appears after a sibilant then it is realized as /Iz/, as in *horses*.

Berko [5] experimentally tested the ability of children to produce the plural forms of nonsense nouns. For example, the child was presented with a toy, told "This is a wug", where "wug" is a nonsense noun, and then shown two such creatures and prompted "Here are two . . .". As *wug* is a nonsense noun, invented by the experimenter, we can be sure that the child will never have come across the plural form of this noun. None the less, Berko found that children aged 4 to 5 can reliably produce the appropriate plural form for such novel words—they are capable of generalising from observed plural forms to novel plurals. Furthermore, children produced the plural using the appropriate allomorph (/z/ in the case of *wugs*), and are therefore capable of making the relatively subtle generalisations involving the three allomorphs of the plural morpheme, although they were most successful with the more common /z/ allomorph and least successful with the /Iz/ allomorph.

5.4.2. Bias against one-to-many mappings

One-to-many mappings exist at several possible levels in natural language:

- At the level of the inflectional affix. For example, the three allomorphs of the English *-s* constitute a one-to-many mapping from meaning (plural number) to form (/z/, /s/, or /Iz/).
- At the level of the free morpheme or word. For example, the English words *dog*, *hound*, and *mutt* are (arguably) three forms which express the same meaning. One-to-many mappings at the level of the word usually termed *synonyms*.

- At the level of the sentence. For example "Charles gave the cake to Bethany", "Charles gave Bethany the cake" and "Bethany was given the cake by Charles" express the same proposition, and are paraphrases of one another.

Based on such instances of one-to-many mappings, we might conclude that human language learners do not possess a bias against one-to-many mappings. However, the end-state of the language acquisition process does not necessarily give us a perfect insight into the biases at play during the process of acquisition. It is possible that child language learners *do* bring some bias against one-to-many mappings to the acquisition of language, but that competing pressures result in an adult competence which contains one-to-many mappings. This in fact seems to be the case—while a full review of the historical and developmental linguistics literature is beyond the scope of this paper, we will present two types of evidence which show the existence of such a bias.

Firstly, there is a historical tendency for languages to loose one-to-many mappings over time, which we would expect if language learners bring a bias against such mappings to the language acquisition task. To give a specific example, whereas only *-s* suffixation for the plural is productive in modern English, Old English had several productive possibilities, including *-en* suffixation (as fossilised in modern *ox-oxen*) and the umlaut marking (as in the modern *goose-geese*). The transition from Old English to modern-day English has involved a reduction in the number of different strategies for expressing the plural—a reduction in one-to-many mappings. More generally, Mańczak [59], based on a survey of historical grammars and etymological dictionaries, presents a number of "laws of analogical evolution" for morphological change, the first of which is that "[t]he number of morphemes having the same meaning more often diminishes than increases" [59, p. 284]—languages tend to lose one-to-many mappings in the morphological system.There is historical evidence suggesting that language learners are biased against one-to-many mappings.

If this is rather circumstantial, more solid examples of a bias against one-to-many mappings are available. Markman and Wachtel [62], following Kagan [47], tested children's behaviour on potentially synonymous nonsense words. In Markman and Wachtel's study, children were shown a single familiar object (for example, a plate) and an unfamiliar object (e.g., a radish rosette maker) and asked by a puppet frog to "Show me the fendle" where "fendle" (or similar) is a nonsense word. Children reliably respond by giving or showing the unfamiliar object. Results from a control group study, where children were asked simply to "Show me one", indicated that this preference was not due to a preference on the part of children to respond with the unfamiliar object—children only exhibit such a preference when prompted with a novel word.

Markman [60–62] proposes that this behaviour is due to a Mutual Exclusivity (ME) bias in children—"children should be biased to assume, especially at first, that terms [words] are mutually exclusive" and "each object will have only one label" [60, p. 188]. Note that this is not an inviolable principle, but a tendency or bias that can be overridden given sufficient evidence. The child in the task above reasons, via Mutual Exclusivity, that the novel word *fendle* cannot refer to the familiar object, as this would result in a one-to-many mapping (the plate object / concept would maps to two words, *plate* and *fendle*). The child therefore infers that the new word must refer to the unfamiliar object, and responds appropriately. Mutual Exclusivity is a bias against one-to-many mappings between meanings and signals.

5.4.3. Bias against many-to-one mappings

As with one-to-many mappings, many-to-one mappings potentially exist at several levels in natural language:

- *Affixes*. For example, the suffix *-s* in English expresses both plurality when suffixed to nouns, and present tense (among other things) when affixed to verbs—this is an instance of a many-to-one mapping between meaning (plural number or present tense) and form.
- *Words*. For example, the English word *bank* expresses several meanings, including: a financial institution (when used as a noun); the ground adjacent to a waterway (noun); and the action of tilting while turning (verb). Such ambiguous words are termed *homonyms*.
- *Sentences*. A single sequence of words may be used to express several distinct propositions, as a consequence of containing homonymous lexical items, or as a consequence of being parsable in several ways. A classic example of the latter, structural, ambiguity is given by the sentence "The boy saw the man with the telescope", which can be used to express two distinct propositions: the state of affairs where the boy uses the telescope to see the man, and the state of affairs where the man has the telescope.

As was the case with one-to-many mappings, the fact that examples of many-to-one mappings are easy to find in natural languages might make us pessimistic about finding a bias against such mappings in language learners. However, once again the point stands that many-to-one mappings could still be prevalent in the end-state of language learning, and indeed language evolution, in spite of a learner bias against such mappings, as a consequence of competing pressures. One obvious competing pressure in the case of many-to-one mappings is the pressure for reuse of affixes and words. Given the necessarily finite capacity of human memory, and the additional pressures imposed by articulatory and acoustic factors, any learner bias against many-to-one mappings is unlikely to have reduced influence.

As was the case for one-to-many mappings, there is evidence that human language learners are biased against acquiring many-to-one mappings from meanings to signals, although this evidence is rather more scarce. At the general level, Slobin claims, under the guise of the maxim "be clear", (e.g., [80–82]) that children "strive to maintain a one-to-one mapping between underlying semantic structures and surface forms" [81, p. 186]. Slobin explicitly links the prevalence of many-to-one mappings with difficulty of acquisition. To repeat Slobin's example: the Serbo-Croat inflectional system is "a classic Indo-European synthetic muddle . . . there are many irregularities, a great deal of homonymy, and scattered zero morphemes" [81, p. 191]. Slobin suggests that such many-to-one mappings explain why the Serbo-Croat system is mastered relatively late by child language learners.

More recently, and perhaps more promisingly, the types of experiments used to demonstrate the Mutual Exclusivity bias have been adapted and applied to the study of homonymous lexical items. These experimental studies provide the strongest evidence, to date, that children are biased against acquiring many-to-one mappings from meaning to signal, at least on the level of lexical items. The original study is detailed in Mazzocco [65], with a subsequent study by Doherty [35].

In Doherty's study, children are presented with a story in which key word is used several times in a context which is intended to give strong clues as to the meaning of that word. For example, one story relates to Hamish accompanying his mother to the zoo, the crucial passage running "At the zoo they saw a strange *blas*/*cake* from Brazil. Hamish thought the *blas*'s/*cake*'s long nose looked funny" [35, p. 213]. The key word is either *blas* or *cake*, with the form of the key word alternating between subjects.

After exposure to the story, the children were shown a selection of photographs and asked "Which one is the [key word] in the story?". In the case of the example above, the set of illustrations includes pictures of a cake and a tapir (a long-nosed South American mammal). Doherty found that children are highly successful at identifying the referent of the novel nonsense word *blas*—the context of the story enables them to correctly identify this as referring to the tapir. However, children have low success rates in identifying the referent of the word *cake* in this story. This word is used in a homonymous way— whereas the context of the story strongly suggests that *cake* refers to the long-nosed South American mammal, children already know that *cake* means a kind of food. As such, identifying the tapir as the referent of *cake* in the story would mean accepting a many-to-one mapping from meanings to signals (the cake and tapir concepts would both map to the ambiguous word *cake*). Children fail to identify the referent of the homonymous *cake* because they are biased against many-to-one mappings in the lexicon.

5.4.4. The origins of learning biases

The review above suggests that human language learners possess all the capacities/biases which the computational model highlights as being key to the cultural evolution of compositional structure. In other words, our explanation linking compositional structure with linguistic evolution holds up when we look in more detail at the issue of the learning strategies which must be involved.

Why do humans have such learning biases? An intriguing possibility is that these biases have evolved because of the type of linguistic structure they underpin—in other words, the particular learning strategy applied by humans to the language acquisition task has evolved because it yields a language which is stable over time (despite of the transmission bottleneck), and communicatively functional. Such a hypothesis has been tested for the case of biases for the acquisition of simple signalling systems [44,86].

When considering the evolution of these capacities and biases in our own species, we might also won- der to what extent these are present in other species—can the uniqueness of human language be explained in terms of the uniqueness of these learning capacities to our species? The capacity to generalise is al- most certainly not unique to humans, being (as discussed above) a general principle of systems that learn. Comparative evidence on the biases of non-human species with respect to one-to-many and many-to-one mappings is rather scarce. However, such biases in human language learners are often described as a consequence of a sophisticated theory of mind (see [6] for review), probably unique to humans.

6. Compression, innovation, and linguistic evolution

The model of learning used in the previous section was geared toward exploring how learning biases, which have fairly obvious parallels in human language acquisition, impact on linguistic evolution. In this section a model of learning based on a normative theory of induction (the minimum description length principle) is explored—rather than focusing on the learning task in terms of psychological principles rooted in child language acquisition, the agents will instead induce the most likely hypothesis in the hypothesis space, given some body of data. This allows us to test our theory of linguistic evolution using a theoretically well-grounded model of induction, and, additionally, refine our understanding of the role of innovation in the evolution of linguistic structure. Innovation—production of novel linguistic structures—plays a role in the associative model outlined in the previous sections, but the procedures of

induction and invention are intimately connected. In the model presented in this section, a more rigorous separation of induction and invention is possible.

The issue we now turn to concerns the following question: When are inductive generalisations justifiable? A feature of the models discussed so far is that so long as structure is present in the data, we consider the learner justified in harnessing this structure, and making generalisations from it. But is this policy always justifiable? If no readily interpretable constraints guide the inductive process, then we really have no way of answering this question. To gain firmer theoretical support for the phenomenon of cumulative linguistic evolution, we need to understand the learning process in terms of a theory of induction. Otherwise, we may be faced with the conclusion that linguistic evolution is only possible when we consider learning algorithms with an inductive bias which is at odds with normative theories of induction.

6.1. Learning based on a simplicity principle

Induction is the task of choosing a hypothesis from a set $\mathcal{H} = \{H_1, H_2, \ldots\}$ in the light of some data D. A central problem in achieving this task stems from the realization that, in the general case, there will be infinitely many candidate hypotheses consistent with the data. To specify which hypothesis is appropriate always requires some criterion on which to judge competing hypotheses. The minimum description length (MDL) principle is one such criterion [55,77]. The MDL principle provides a means of judging, given a hypothesis space $\mathcal{H}$ and some data D, which member of $\mathcal{H}$ represents the most likely hypothesis given that D was observed. This judgement represents a point in a trade-off between complexity and simplicity. An overly complex hypothesis which fits the data perfectly typically suffers from the problem of over-fitting: incidental or noisy characteristics are captured by the hypothesis and are taken to be features of the underlying distribution. An overly simple hypothesis, on the other hand, may suffer from under-fitting: the hypothesis may be too general and fail to capture the characteristics of the data. The MDL principle provides a means of judging which hypothesis represents the best point in the simplicity/complexity trade-off. Importantly, this "best" position picks out the hypothesis which is both the most probable hypothesis and the hypothesis which leads to the shortest redescription of the data.

The crucial observation is that regularity in the data can be used to compress the data. The MDL approach represents a general principle, in that it provides a means by which to judge competing hypothesis in contexts such as learning (e.g., [78]) and model selection in the wider sense (e.g., [41,74]). Of great relevance to this discussion is the fact that MDL also features prominently as a principle in understanding hypothesis selection performed by the cognitive system on many levels [17,18] including that of language acquisition [19,99]. In short, the minimum description length principle offers a theoretically well-founded basis on which to perform hypothesis selection.

Formally, the MDL principle states that the most likely hypothesis is the one which minimises the sum of two quantities. The first quantity is the length, in bits, of encoding the hypothesis. The second quantity is the length, in bits, of the encoding the data, when represented using this hypothesis. To formalise this statement, we require an optimal encoding scheme for the hypotheses, C_1, and an encoding scheme for data represented in terms of the hypothesis, C_2. Furthermore, the only relevant issue for hypothesis selection is the *length* of these encodings: L_{C_1} and L_{C_2}. Given the set of hypotheses $\mathcal{H}$, and the observed data, D, the MDL principle selects a member of $\mathcal{H}$, H_{MDL}, as follows:

$$H_{\mathrm{MDL}} = \min_{H \in \mathcal{H}} \big\{ L_{C_1}(H) + L_{C_2}(D|H) \big\}. \tag{4}$$

This expression states that the best pattern to explain the data is the one which, when chosen, leads to the shortest coding of the data. The coding is achieved using a combination of the chosen hypothesis and a description of the data using this hypothesis. This is to say that, given the hypothesis and the description of the data represented in terms of this hypothesis, the observed data can be described exactly. Note that, in line with the discussion above, picking the smallest hypothesis—the hypothesis with the smallest encoding length—will not necessarily achieve this goal. Small hypotheses may be too general and lead to an inefficient recoding of the data. Similarly, a very specific hypothesis will describe the data verbatim and fail to reveal the structural characteristics of the data, in the same way that atomistic analyses in the associative model of learning failed to exploit structure in the data. The best solution represents a trade-off between these two poles, and the MDL principle tells us how to judge competing hypotheses with respect to this trade-off.

To transfer this discussion into a model and test the impact of learning based on the MDL principle requires us to construct a hypothesis space $\mathcal{H}$, and coding schemes over these hypotheses. Recall that the data we refer to in this discussion are collections of utterances whose form is determined by the language model introduced in Section 3.2. One example is the following set of utterances, L_{comp}:

$$L_{\text{comp}} = \{\langle\{1, 2, 2\}, \text{adf}\rangle, \langle\{1, 1, 1\}, \text{ace}\rangle, \langle\{2, 2, 2\}, \text{bdf}\rangle,$$
$$\langle\{2, 1, 1\}, \text{bce}\rangle, \langle\{1, 2, 1\}, \text{ade}\rangle, \langle\{1, 1, 2\}, \text{acf}\rangle\}.$$

In order to apply the MDL principle to the selection of hypotheses given some arbitrary series of utterances, we consider a hypothesis space composed of finite state unification transducers, or FSUTs[14] [9]. These transducers relate meanings to signals by representing a network of states and transitions. A number of paths exist through the transducer. Each path begins at the *start state*. These paths always end at another privileged state termed the *accepting state*. A path through the transducer is specified by a series of transitions between states; each of these transitions relates part of a signal to part of a meaning. For example, consider the transducer shown in Fig. 8(a). It depicts a transducer which represents the language L_{comp}. This transducer—termed the *prefix tree transducer*—corresponds to the maximally specific hypothesis: it describes the data verbatim, and therefore does not capture any structure present in the language. It is the largest consistent hypothesis in $\mathcal{H}$ that can be used to describe the data L_{comp}, and only L_{comp}. Given a transducer and a signal, the associated meaning can be derived by following a path consistent with that signal, and collecting the meanings associated with each transition taken. Similarly, given a meaning, the signal can be derived by following a path consistent with the meaning, and concatenating each symbol encountered along the path.

Given some observed utterances, the space of candidate hypotheses will consist of all FSUTs consistent with the observed utterances. By consistent, we mean that the candidate hypotheses are always able to generate, at a minimum, all the observed utterances. We are interested in situations within which a transducer is capable of generating utterances for meanings it has never observed; in such a situation, the transducer can be said to have generalised.

If structural regularity exists in the observed language the prefix tree transducer can be used to derive further, more general, transducers that are also consistent with the observed data. Such derivations are achieved by applying compression operations on the transducer. Compression operators, when applica-

---

[14] A FSUT is a variation on the basic notion of a finite state transducer (e.g., [43]). Our use of such transducers was inspired by and extends the work of Teal and Taylor [91].
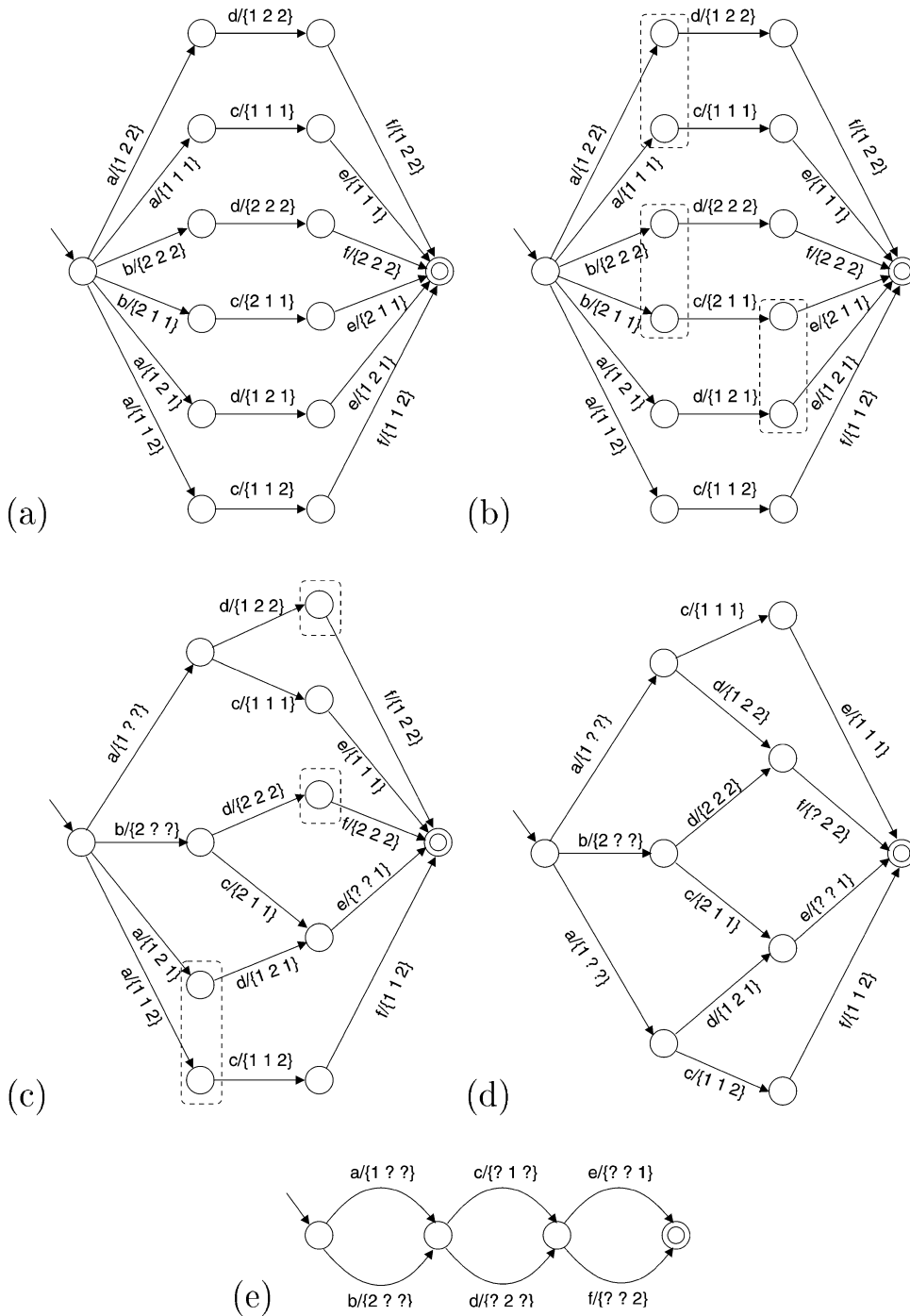
Fig. 8. Given the compositional language L_{comp}, the Prefix Tree Transducer shown in (a) is constructed. By performing edge and state merge operations, outlined in (b) and (c), the transducer can be compressed. The transducer shown in (d) is compressed, but does not lead to any generalisations. The transducer in (e) is fully compressed, and generalises to L_{comp}^{+}. Note that ? indicates a wildcard feature value.

ble, can introduce generalisations by merging states and edges. Given a prefix tree transducer—which is simply a literal representation of the observed data—only two operators, *state merge* and *edge merge*, are required to derive all possible consistent transducers. For the details of how states and edges are merged, as well as the details of the encoding schemes C_1 and C_2, we refer the reader to Brighton [9,10].

The important feature of the FSUT model, in combination with the MDL principle, is that compression can lead to generalisation. For example, Fig. 8(b) and (c) illustrate some possible state and edge merge operations applied to the prefix tree transducer representing L_{comp}. The transducer resulting from these merge operations is show in Fig. 8(d). Fig. 8(e) depicts the fully compressed transducer, which is found by performing additional state and edge merge operations. Note that further compression operations are possible, but would result in the transducer becoming inconsistent with the observed language.

By applying the compression operators, all consistent transducers can be generated. Some of these transducers will be more compressed that others, and as a result, they are more likely to generalise than others. Note that if L_{comp} was an instance of a random (holistic) language, then few, if any, compression operations would be applicable; regularity is required for compression to be possible.

Generalisation can lead to the ability to express meanings which were not included in the observed linguistic data. For example, a close inspection of the compressed transducer shown in Fig. 8(e) reveals that meanings which are not present in L_{comp} can be expressed. The *expressivity* of a transducer is simply the number of meaning that can be expressed. The language L_{comp}^+, shown below, contains all the meaning-signal pairs which can be expressed by the fully compressed transducer in the above example.

$$L_{\text{comp}}^+ = \{\langle\{1, 2, 2\}, \text{adf}\rangle, \langle\{1, 1, 1\}, \text{ace}\rangle, \langle\{2, 2, 2\}, \text{bdf}\rangle, \langle\{2, 1, 1\}, \text{bce}\rangle,$$
$$\langle\{1, 2, 1\}, \text{ade}\rangle, \langle\{1, 1, 2\}, \text{acf}\rangle, \langle\{2, 1, 2\}, \text{bcf}\rangle, \langle\{2, 2, 1\}, \text{bde}\rangle\}.$$

In this case, compression led to generalisation, and the expressivity of the transducer increased from 6 meanings to 8 meanings. By compressing the prefix tree transducer, the structure in the compositional language is made explicit, and as result, generalisation occurs. Compression is not possible when structure is lacking in the observed data, and the result will be that meanings not included in the observed data cannot be expressed.

At this point it is worth highlighting how the FSUT model relates to the discussion of the one-to-one bias discussed in Section 5.2. First consider that, in this model, a one-to-many mapping cannot occur as production is deterministic: even if multiple signals are consistent with a single meaning, only one of these signals will ever be produced. Hence, a one-to-many relationship between meanings and signals is not possible due to a bias imposed by the deterministic production mechanism. Second, we need to consider many-to-one mappings. Because the set of observed meanings is always a sample, the maximally general coding of meanings (i.e., maximal use of wildcards) required to represent a many-to-one mapping will eventually be deviated from. Why is this? At some point the set of meanings supporting the many-to-one relationship will be under-represented such that the production of one or more members of this set will be performed via invention. As a result, the language will deviate from the many-to-one relationship. The bias in this model against many-to-one mappings is therefore a combination of the sampling process imposed by the bottleneck, and the inductive bias. Here, we see how both the model of induction and the model of production can influence the structural characteristics of evolved languages.

We now have a hypothesis space over which we can apply the MDL principle. The hypothesis chosen by a learner in our model in light of data D is the one with the smallest description length, H_{MDL}. This search for this hypothesis is performed using a hill-climbing search described in Brighton [10,11].

6.2. *The evolutionary consequences of the simplicity principle and random invention*

With these model components in place, we are now in a position to assess whether induction based on the MDL principle within the Iterated Learning Model leads to linguistic evolution. We will focus on the case where there is a bottleneck on transmission, with only minimal changes to other components of the Iterated Learning Model.[15]

In the new model, each simulation run must be initialised with a random language. In the associative matrix model detailed above, this was achieved by simply allowing the initial agent to produce at random, according to their matrix of associations of strength 0. In the new model this is not possible, as the initial agent has no FSUT to produce with. Consequently, a random initial language is generated according to the parameter values, and the initial agent learns based on this language. Fig. 9 shows the resulting transducer. Note that negligible compression occurs, and as a result the transducer does not generalise to novel meanings: 32 utterances were given as input, and each of these is encoded by a single path through the transducer.. The language represented by the transducer is holistic and the linguistic structure we seek to explain is therefore lacking. Can a structured mapping which leads to generalisation evolve through cultural adaptation?
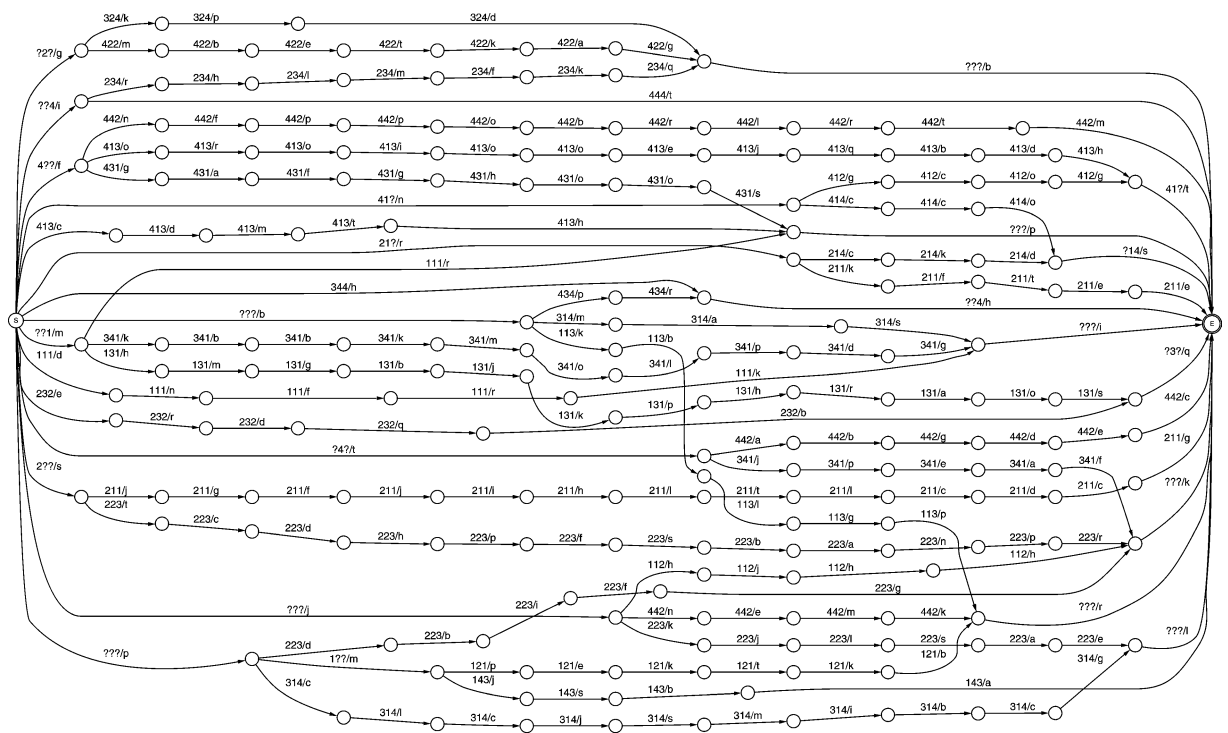

Fig. 9. A transducer H_{MDL} induced from a random initial language. Negligible compression occurs.

We must now consider a crucial aspect of the model which was largely side-stepped in the associative network model presented in Sections 4, 5: the issue of invention. Invention occurs when an agent is prompted to produce a signal for a meaning which it has no signal for—that is, the meaning was not observed in conjunction with a signal during learning, and also cannot be expressed as a result of any generalisation occurring due to compression. According to this definition, true invention never occurs in the associative network model. In the associative model, the learner simultaneously maintains a set of weighted relationships between all possible meanings and all possible signals, including meanings and signals not observed. As such, generalisation based on regularities in the data, and innovation of a new signal for a particular meaning are indistinguishable—both proceed via the same winner-take-all process, as a consequence of the weights in the system of associations. These innovations (generalisations or 'true' inventions) are therefore a consequence of the agent's learning behaviour, with, for example, agents learning with a bias against many-to-one mappings tending to innovate in ways which avoid producing such mappings.

The separation between inductive generalisation and true invention in the current model is much cleaner. For example, the transducer in Fig. 9 can only express the meanings which were present in the observed data. However, within the Iterated Learning Model, individuals will be required to express meanings which were not in the set of utterances which they observed during learning, and we must therefore define an invention procedure. A number of invention strategies are possible—initially we will adopt a policy of random invention, where a random signal is generated for novel meanings.

Fig. 10(a), (b) depicts the process of a 200 generation run of the new model. Fig. 10(a) depicts compression rate, α, as a function of iterations. The compression rate measures the relative size of the prefix tree transducer, H_{prefix}, and the chosen hypothesis H_{mdl}, and is defined as:

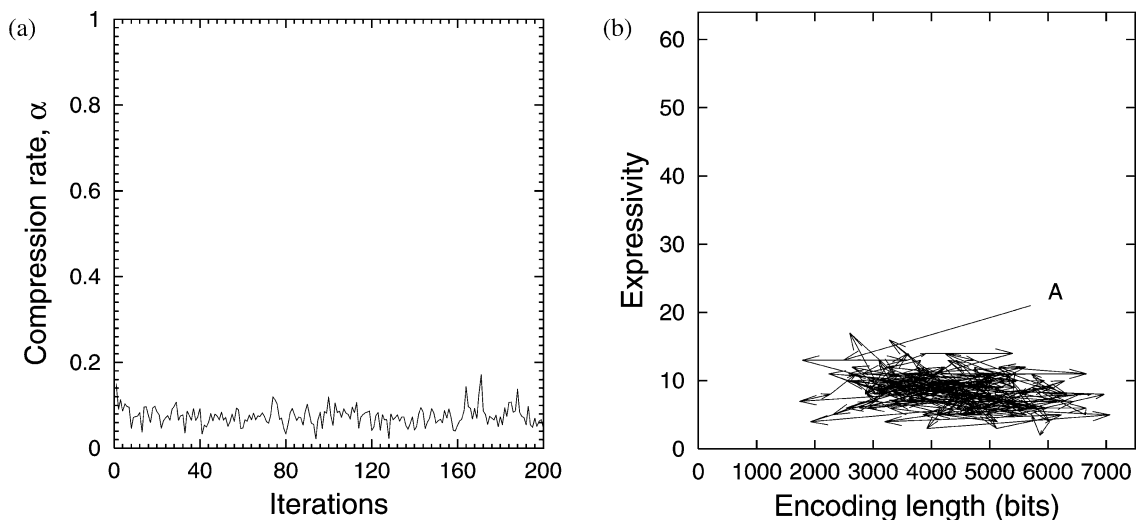$$\alpha = 1 - \frac{|H_{\text{mdl}}|}{|H_{\text{prefix}}|}.$$

Fig. 10. Linguistic evolution resulting from partially random invention.

A high compression rate means that the language is compressible. As can be seen from the figure, the compressibility of the language changes very little over time—the initial random language undergoes no significant adaptation and remains unstructured and therefore uncompressible ($\alpha \approx 0.06$). Fig. 10(b) highlights this fact, by showing the transitions through a state space depicting the expressivity of the language as a function of the encoding length of the language. Here, we see that from the initial state, labelled *A* and corresponding to the transducer depicted in Fig. 9, the systems follows an unordered trajectory through the sub-space of small inexpressive transducers. Because the language remains unstructured, generalisation is not possible and expressivity remains low. Similarly, unstructured languages cannot be compressed, and therefore the encoding length remains relatively high.

The key point here is that a cumulative evolution of structure does not occur as it did in Section 4: the model as it stands fails to match the predictions of our theory, or indeed our findings from the associative learning model. The reason for this failure is that the mechanisms supporting linguistic evolution—language learning and language production—are somehow failing to lead to the cumulative evolution of structure. In fact, the source of the problem is the way in which linguistic innovation via invention is modelled.

6.3. Invention based on simplicity principle

The MDL principle can tell nothing about the process of production—unlike the associative model of learning, the model of learning used here can tell us only which hypotheses should be induced. The process of interrogating the hypothesis with novel meanings to yield signals is not fully defined, and needs to be developed. Our first attempt at an invention mechanism—invention of random strings—proved to be in some way deficient.

To address this problem, a more principled invention mechanism is proposed, where the invented signal is a derived using the induced hypothesis itself, rather than being constructed at random—in the same way that invention is achieved in the associative matrix model, the invented signal will be constrained by structure present in the hypothesis, which is in turn determined by the data observed during learning.

The new invention method exploits the structure already present in the hypothesis by using those parts of the transducer consistent with the novel meaning to construct part of the signal. This approach is detailed in Brighton [10,11], but the essentials of the process can be summarised as follows. An invented signal is selected such that the invented signal, if it were seen in conjunction with the novel meaning during the learning phase, would not lead to an increase in the MDL of the induced hypothesis. This invention procedure therefore proposes a signal which in some sense matches the structure of hypothesis. If such a signal cannot be found, then no signal is produced. In short, the invention procedure, rather than being random, now takes into account the structure present in the hypothesis.

Fig. 11 illustrates the process of a second Iterated Learning simulation, incorporating the new invention procedure. this evolutionary trajectory is typical of such simulation runs. Strikingly, Fig. 11 reveals a very different evolutionary trajectory to that shown in Fig. 10, as a consequence of the alternative invention procedure. Fig. 11(a) shows a transition from low to high rates of compressibility. Fig. 11(b) illustrates an entirely different trajectory through state space, one where a series of transitions lead to small, stable, and expressive hypotheses. Starting at an expected expressivity of approximately 22 meanings (point *A*), the system follows an L-shaped trajectory. There are two distinct jumps to a stable state where we find small hypotheses capable of expressing all 64 meanings. The induction and invention processes consistently direct linguistic evolution toward compositional systems.
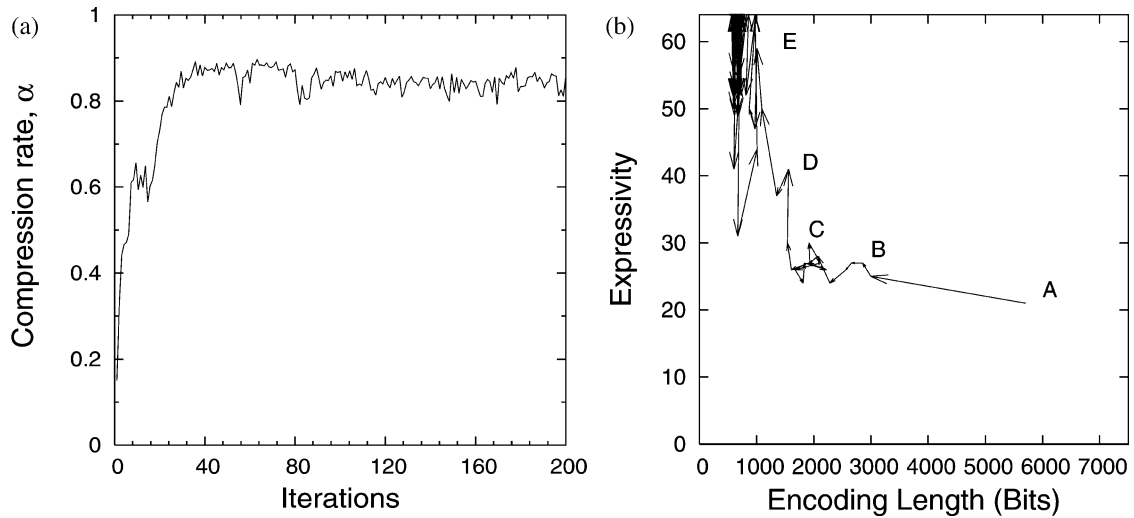
Fig. 11. Linguistic evolution arising from the application of the intelligent invention scheme.

The first significant transition through the state space takes the system from the bottom-right end of the L-shape (point *A*) to the bend in the L-shape (points *B* and *C*), where expressivity increases slightly, but the minimum description length of the language decreases by a factor of 3. From requiring approximately 6000 bits to encode the evolving language, linguistic evolution results in transducers being induced with an MDL of approximately 2000 bits. The lack of increase in expressivity is a reflection of the transducers organising themselves in such a way that significant compression results, but an increase in expressivity is not achieved. The second transition, leading to the top of the L-shape (through point *D* to point *E*), is very different in nature. Here, for a small decrease in the MDL of the developing language, a significant increase in expressivity occurs. This is an important transition, as it results in the system entering a stable region of the state space. Although a few deviations away from this stable region occur early on, the system settles into a steady state characterised by high expressivity.

Fig. 9(a) depicts the transducer corresponding to point *A* in Fig. 11(b), while Fig. 12(a)–(d) depicts the transducers at points *B*, *C*, *D*, and *E*. Fig. 12(a) represents the transducer corresponding to point *B*. In this transducer, we see the beginnings of significant structure emerging. The first symbol in each signal appears to discriminate between feature values in the second feature. This structural relationship acts as a seed for further discrimination, which will ultimately result in generalisation. Between point *B* and point *C*, the evolution of the language becomes increasingly more evident. Point *D*, shown in Fig. 12(c), corresponds to a transducer where further discrimination occurs, and certain meanings can be expressed even though they were not observed—significant generalisation is occurring. Fig. 12(d) illustrate the occurrence of further discrimination and generalisation, as the state of the system climbs up to and moves around a stable region of the state space.

This second model again demonstrates how the mechanisms of induction and production can lead to the evolution of generalisable, compositional structure. This is made possible by linguistic evidence—utterances—coding information that determines the induction of a hypothesis capable of generalisation, where the linguistic evidence comes to have this structure as a consequence of those same mechanisms of
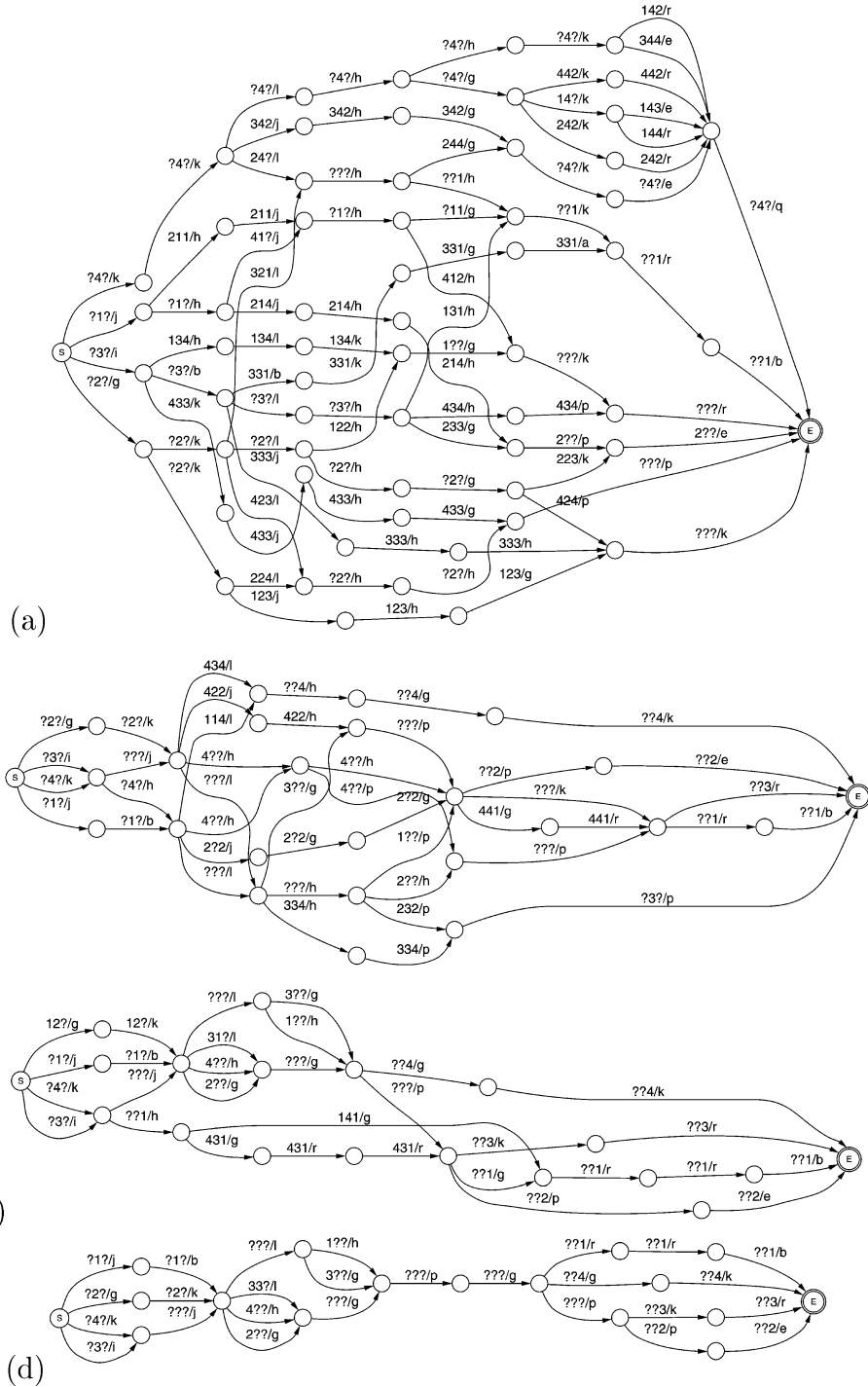
Fig. 12. Languages arising during linguistic evolution driven by MDL induction and intelligent invention. In (a), structure is evident as certain paths merge. In (b), an intermediate stage is shown where significant compression is evident but generalization is not possible. In (c), (d) further compression is possible, and novel meanings can be expressed.

induction and production. The new model highlights the fact that learning is just one of the mechanisms driving linguistic evolution: the issue of production and, in particular, invention, plays a key role.

### 6.4. Historical accidents and the evolution of structural complexity

Before moving on to discuss the nature of language as an evolutionary system in more general terms, it is worth considering the nature of stable states in this model, as they provide examples of the linguistic complexity not seen in the associative model. Fig. 13 shows two stable states. Fig. 13(a) depicts a transducer for a meaning space defined by $F = 3$ and $V = 2$ along with the grammar, G_1, which describes how signals are constructed for each of the 8 meanings. Similarly, Fig. 13(b) depicts the transducer and the corresponding grammar, G_2, for a meaning space defined by $F = 3$ and $V = 3$ which comprises 27 meanings.

Optimal transducers, those with the lowest description length given the parameter values, are those where a single symbol is associated with each feature value of the meaning space. Even though the minimum description length principle would prefer these transducers, they do not occur in the model. A close inspection of the transducers shown in Fig. 13 demonstrates that features are coded inefficiently: variable length strings of symbols are used, rather a single symbol, and some feature values are associated with redundant transitions which carry no meaning. In Fig. 13, for example, *all* meanings are expressed with signals containing a redundant symbol (the second symbol d). These imperfections are frozen accidents: the residue of production decisions made before stability occurred. The imperfections do not have a detrimental impact on the stability of the language, and they therefore survive repeated transmission due to being part of the compositional relationship coded in the language. This phenomenon is an example of how the process of linguistic evolution leads to complexity which is not a direct reflection of the learning bias: transducers with lower description length exist. The evolved transducers serve the function of stability despite this deviation from the "optimal" transducer, and this is why such languages persist.

7. Conclusion: Language as an evolutionary system

An essential distinction underlying the picture developed so far contrasts the *capacity for language* with *languages themselves*. Without doubt the capacity for language is a biologically determined competence, and this competence is specific to humans. Languages themselves, on the other hand, are not biologically determined in the same sense: they result from an interaction between the capacity for language and the linguistic environment. When we talk about languages themselves we refer to a particular relationship between meanings and signals. When we talk about the capacity for language we refer to a computational system that processes languages. Maynard Smith and Szathmáry argue that this genetically determined capacity for language was the foundation on which the eighth major transition in evolution was based. This transition allowed information transmission to occur *through language*: languages provide a substrate for the transmission of information. How can this be? Like DNA, language provides a system for composing an indefinite number of messages from finite means:

Both the genetic and linguistic systems are able to transmit an indefinitely large number of messages by the linear sequence of a small number of distinct units. In genetics, the sequence of four bases enable the specification of a large number of proteins, and these, by their interactions, can specify

(a)
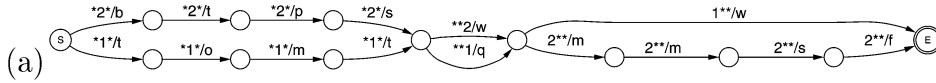
$$G_1: \quad \text{S}/x,y,z \to \text{A}/x \ \text{B}/y \ \text{C}/z$$

$$\text{B}/2 \to \text{btps}$$

$$\text{A}/1 \to \text{w}$$

$$\text{C}/1 \to \text{w}$$

$$\text{A}/2 \to \text{mmsf}$$

$$\text{C}/2 \to \text{q}$$

$$\text{B}/1 \to \text{tomt}$$

(b)
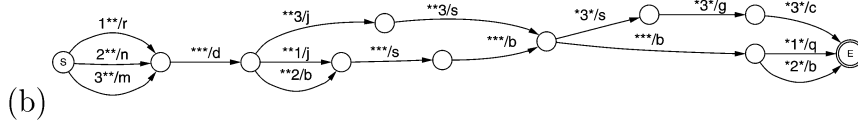
$$G_2: \quad \text{S}/x,y,z \to \text{A}/x \ \text{d} \ \text{B}/y \ \text{C}/z \qquad \text{B}/y \to \text{D}/y \ \text{sb} \qquad \text{C}/3 \to \text{sgc}$$

$$\text{A}/1 \to \text{r} \qquad \qquad \text{B}/3 \to \text{js} \qquad \text{C}/z \to \text{b} \ \text{E}/z$$

$$\text{A}/2 \to \text{n} \qquad \qquad \text{D}/1 \to \text{j} \qquad \text{E}/1 \to \text{q}$$

$$\text{A}/3 \to \text{m} \qquad \qquad \text{D}/2 \to \text{b} \qquad \text{E}/2 \to \text{b}$$

Fig. 13. Two evolved languages: (a) shows a transducer, and the corresponding grammar, containing redundant transitions, variable length signals, and several syntactic categories; (b) shows a language with variable length substrings.

an indefinitely large number of morphologies. In language, the sequence of some 20 or 30 distinct unit sounds, or phonemes, specify many words, and the arrangement of these words in grammatical sentences can convey an indefinitely large number of meanings [64, p. 139].

This system of information transmission was novel in the sense that it introduced a entirely different physical medium over which information could be transmitted (see [63, p. 12]). The information transmitted by language, according to this view, is not information relating to the essential characteristics of language itself, but rather informational structures that are expressed *using* language. Maynard Smith and Szathmáry envisage language as transmitting a message which carries information about cultural artifacts such as traditions, religions, and so on:

It is impossible to imagine our society without language. The society we live in, day and night, depends on it [...] on detailed social contracts, which could not exist without language [64, p. 149].

This is the complexity which they seek to explain, which it turn depends on the complexity of language and its provision of a system in which indefinitely many meanings can be expressed. Put simply, language is the mode of information transmission which allowed the evolution of complexity in human culture.

In contrast, in this paper we have focused on the information that language carries *about its own construction*. Any behaviour that is transmitted through iterated learning must, whatever else it does, provide information to learners sufficient for that behaviour's survival over time. It is this fact that is key to understanding linguistic transmission in an evolutionary sense. Ultimately, iterated learning leads to *adaptation*.

Recall that the transmission of language is subject to constraints; the channel through which linguistic information is transmitted is restricted. Constraints such as the transmission bottleneck determine the kind of information that can be transmitted successfully. As a result, the transmission bottleneck induces an evolutionary dynamic—certain kinds of information (linguistic structure) can survive the transmission bottleneck whereas others cannot.[16]

To illustrate this point, consider the experiments discussed in Sections 4 and 6. Here, data drawn from a compositional mapping between meanings and signals codes linguistic information that can be transmitted successfully. But why is this? In general, data conforming to some regular pattern, such as a compositional relationship, can justify a general statement about the data. The statement can be general in the sense that it identifies a relationship which extends beyond observed instances to include other, unseen data. Contrast this situation with one where the data does not conform to any regular pattern. Here, the induction of a general statement or pattern would not, by and large, be justifiable.

The crucial issue here is that general statements or patterns are consistent with many bodies of evidence, while specific patterns are not. Consequently, these general patterns are re-codeable using different data: the same pattern can justifiably be induced despite support for this pattern coming from a different body of data.

Why is this important? As long as there is a limit on the data available to the learner—as long as there is a transmission bottleneck—the more general a pattern is, the more likely it is to be transmitted successfully. Given variation in the degree to which information can successfully be transmitted, we can rightly talk of certain information coded in a language as being adaptive to the problem of linguistic transmission. This observation forms the basis for the claim that languages adapt to be transmissible. We have seen this in the models presented in this paper. Over the course of many generations, the language changes as it is transmitted, and we see an evolution from unstable languages to stable ones.

One way to understand this process of adaptive evolution is from the perspective of *competition*. We can think of the data available to the learner as a finite resource—a window within which a linguistic generalisation or regularity must be expressed if it is to survive. Crucially, there are likely to be many different regularities that must compete for this resource.

It is clear from the simulation models discussed in Sections 4 and 6 that, even given these restricted models of language, the information coded in linguistic data and interpreted by the language learner

---

[16] It should be stressed, however, that the iterated learning models presented here demonstrate that linguistic structure *is not* an inevitable outcome of iterated learning. The presence of transmission bottleneck, and particular kinds of bias in both learning and production, are examples of constraints that must be in place.

potentially encodes more than one system of regularity. There are several ways in which this can be the case.

Firstly, the data in interaction with the learning device may encode a system of regularity which covers some subpart of the overall system of meaning-signal mappings. As such, a language can consist of several systems of regularity. Such a situation is common-place in natural language. As mentioned earlier, the system for forming the plural in Old English consists of several systems of regularity (the *-s*, *-en* and umlaut systems), each responsible for some subpart of the mapping between meaning and signal. To take another example, Latin has five nouns declensions: every noun belongs to one of these noun declensions, and each declension has an associated system of case endings. These systems of case endings are regular but, importantly, differ across declensions. Thus we see *competing systems of regularity* in natural language, and we see similar competition between generalisations during the evolution of linguistic structure in the models presented in this paper.

The examples above occur when two or more systems of regularity compete for essentially the same role. However, language also embodies multiple systems of regularity which are coexistent, rather than competing, but which are intimately and reciprocally connected. For example, the rich case structure of Latin, through phonological change, became ambiguous during the history of French. As a result of this ambiguity, development from Old to New French saw the abandonment of the case system in favour of word order [93]. Changes in a language's phonological system often have a cascade of consequences for other parts of the linguistic system—a change in one system of regularity results in a change in a system of regularity operating at another level. Such examples, where small changes lead to subsequent restructuring, is a characteristic feature of the process of language change. Such interactions are not observed in the types of models presented here, where linguistic structure is essentially modelled at a single level, but models could be designed specifically to address such questions.

Finally, one system of regularity may be subsumed within another. For example, the structural regularity in English that Verb precedes Object inside the verb phrase is subsumed within the generalisation that the word order of English is Subject-Verb-Object. A clear analogue of this type of multiple regularity is witnessed in the evolutionary trajectories of languages in our models. For example, a system where meaning fragment $(3, 3, *)$ maps to string-initial de, meaning fragment $(3, 2, *)$ maps to string-initial da and so is subsumed in a system where $(3, *, *)$ maps to string-initial d, $(*, 2, *)$ maps to string-second e and so on. In this case, the level of regularity which wins out—which is actually utilised by the linguistic agents—is determined by factors such as the transmission bottleneck. Less general regularities are likely to be replaced by wider-ranging regularities.

In addition to these multiple systems of regularity evolving, competing and interacting, different systems may evolve at different speeds. We see this occurring in, for example, Fig. 11(b), where the second feature of the meaning space (dealt with by the first two transitions of the transducer) is at an advanced stage of regularity in comparison to the other two features. This observation demonstrates that, even for very simple models of language, we must consider the message being transmitted as a composite message relating to more than one system of regularity. Given a model of language capable of capturing a wider range of linguistic phenomenon, and without doubt in the case of natural language, the message will be a significantly more complex: multiple systems of regularity will vie for transmission, and indirectly influence the transmission success of still other systems of regularity.[17]

---

[17] See Brighton and Kirby [13] for a model that makes this kind of competition overt.
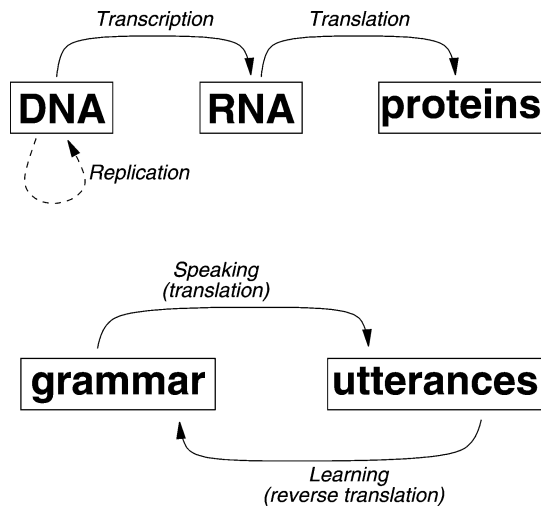
Fig. 14. How the mechanisms driving biological evolution and linguistic evolution differ in the manner in of transmission. Biological evolution proceeds through the direct, but selective, replication of DNA. Linguistic evolution transmits information through the twin processes of speaking (translation) and learning (reverse translation).

The process of linguistic evolution, in the light of this argument, might be understood in terms of the differential replication of multiple systems of regularity. Those adopting the memetic approach take precisely this route, seeking to identify the units of cultural evolution ("memes"), and viewing cultural evolution as an instance of a more general Darwinian process (see, e.g., [2,33]).

Are we then close to finding a strong, and potentially fruitful, analogy between linguistic and genetic evolution, as suggested by memeticists? Or do there remain theoretically-significant differences between these two evolutionary processes, which justify a separate treatment?

One of the problems with a direct analogy is that there are fundamental differences in the mechanism of replication in each case. Fig. 14 illustrates this point. DNA persists by a process of direct copying governed by a selective mechanism that prunes lines of inheritance. Linguistic knowledge, on the other hand, must persist through a repeated cycle of production and induction. We can think of the task of the learner as akin to that of the reverse engineer, trying to figure out what the blueprints are for a device while only being able to look at its behaviour. In the system of biological evolution there is no such reverse engineer—the blueprints are passed on directly every generation.

The consequences of this difference are profound. Whereas the engine driving biological adaptation is the survival and reproductive success of the organism in the environment, linguistic evolution arises from the mechanisms of replication themselves. A linguistic regularity survives because it has properties that make its faithful replication easy. The survival of a set of genes has little to do with its replicability and everything to do with features of the organism those genes code for.

It is interesting to note that these fundamental differences are not immediately obvious when we look at the behaviour of the two kinds of evolutionary system. In particular, much of the time it seems entirely reasonable to treat linguistic evolution as a process driven by *selection* just as genetic evolution is. If two linguistic rules are competing to be expressed in the limited data available to the learner, and only one is successfully induced, perhaps because it was relevant to a larger range of meaning, is this not a perfect example of selective evolution? The problem with this approach is that there may be times where

competition leads to the induction of totally new rules in one step. Linguistic paradigms can change through a process of "reanalysis", the output of a number of different rules can lead to the origins of a new rule that subsumes them all—indeed, this process is fundamental to the evolution of increasingly general regularity in the iterated learning models we have described.

Recognising that linguistic transmission leads to adaptive linguistic evolution *and* that this evolution has important differences from other evolutionary systems is, we believe, the first step to a truly explanatory account of why language is the way it is. Computational and mathematical models of iterated learning offer a principled way of exploring the properties of this adaptive system, helping us understand how our biologically-provided learning biases shape linguistic evolutionary dynamics, and how those dynamics ultimately give rise to the complex structure that is the hallmark of human language.

Appendix A. Measuring compositionality and communicative accuracy

A.1. Measuring compositionality

As discussed in the text, a compositional language is one in which the meaning of a signal is a function of the meaning of its parts and the way in which they are combined. One consequence of this is that compositional languages are *topographic* mappings between meanings and signals [51]. Neighbouring meanings will share structure, and that shared structure in meaning space will map to shared structure in the signal space. Consequently, meaning which are near one another in the meaning space (according to some measure of semantic distance) will tend to map to signals which are near to one another in signal space.

For example, the sentences *John walked* and *Mary walked* have parts of an underlying semantic representation in common (the notion of someone having carried out the act of walking at some point in the past) and will be near one another in semantic representational space. This shared semantic structure leads to shared signal structure (the inflected verb *walked*)—the relationship between the two sentences in semantic and signal space is preserved by the compositional mapping from meanings to signals. A holistic language is one which does not preserve such relationships—as the structure of signals does not reflect the structure of the underlying meaning, shared structure in meaning space will not necessarily result in shared signal structure, and consequently holistic mappings will not preserve topography.

The compositionality measure used in this paper captures this notion, and is based on the measure developed in Brighton [8] for Euclidean meaning and signal spaces. The measure of compositionality is simply the degree of correlation between the distance between pairs of meanings and the distance between the corresponding pairs of signals. In topographic mappings there will be a positive correlation between the distance between pairs of meanings and the distance between the corresponding pairs of signals. If shared structure does not necessarily lead to shared signal structure then there will be no correlation.

In order to evaluate the compositionality of an agent's communication system, the production process is applied to every $m \in \mathcal{E}$ to produce the set $\mathcal{O}$, the observable meaning-signal pairs produced by that agent. In order to measure the degree of compositionality we measure the degree to which the distances between all the possible pairs of meanings correlates with the distances between their associated pairs of signals. More formally, we first take all possible pairs of meanings $\langle m_i, m_{j \neq i} \rangle$, where $m_i \in \mathcal{M}$ and $m_j \in \mathcal{M}$. We then find the signals these meanings map to in the set of observable meaning-signal pairs

$\mathcal{O}$, $\langle s_i, s_j \rangle$. This will give us a set of n meaning-meaning pairs and a set of n signal-signal pairs. Let $\Delta m_n = \mathrm{HD}(m_i, m_j)$ be the Hamming distance[18] between the two meanings in the nth pair of meanings and $\Delta s_n = \mathrm{LD}(s_i, s_j)$ be the Levenstein distance[19] between the nth pair of signals. Furthermore, let $\overline{\Delta m} = \frac{\sum_{i=1}^{n} \Delta m_n}{n}$ be the average inter-meaning Hamming distance and $\overline{\Delta s} = \frac{\sum_{i=1}^{n} \Delta s_n}{n}$ be the average inter-signal Levenstein distance. We can then compute the Pearson correlation coefficient for the distance pairs $\langle m_n, s_n \rangle$, which gives the compositionality of a set of observable behaviour, $C(\mathcal{O})$:

$$C(\mathcal{O}) = \frac{\sum_{i=1}^{n}(\Delta m_i - \overline{\Delta m})(\Delta s_i - \overline{\Delta s})}{\sqrt{(\sum_{i=1}^{n}(\Delta m_i - \overline{\Delta m})^2 \sum_{i=1}^{n}(\Delta s_i - \overline{\Delta s})^2)}},$$

$C(\mathcal{O}) \approx 1$ for a compositional system and $C(\mathcal{O}) \approx 0$ for a holistic system.

A.2. Measuring communicative accuracy

An individual's A matrix therefore defines that individual's production behaviour p and reception behaviour r. If p is interpreted as a probabilistic function $p(s_j|m_i)$, which gives the probability of producing signal s_j given meaning m_i, and r is similarly interpreted as a probabilistic function $r(m_i|s_j)$ then the communicative accuracy between a speaker P using production function $p(s|m)$ and a hearer R using reception function $r(m|s)$ is given by:

$$ca'(P, R) = \frac{\sum_{i=1}^{|\mathcal{M}|} \sum_{j=1}^{|\mathcal{S}|} p(s_j|m_i) \cdot r(m_i|s_j)}{|\mathcal{M}|} \tag{A.1}$$

assuming all meanings are equally frequent and equally important. In other words, the communicative accuracy between speaker P and receiver R is the average probability of the speaker producing a signal for a given meaning m_s, and the hearer interpreting the received signal as meaning $m_h = m_s$. The two-way communicative accuracy between two individuals A and B acting in turn as speaker and hearer is then:

$$ca(A, B) = \frac{ca(A, B) + ca'(B, A)}{2}.$$

This is the measure of communicative accuracy employed throughout the paper.

References

[1] Andersen H. Abductive and deductive change. Language 1973;40:765–93.
[2] Aunger R, editor. Darwinizing culture: The status of memetics as a science. Oxford: Oxford Univ. Press; 2000.
[3] Baker CL. Introduction to generative-transformational syntax. Englewood Cliffs, NJ: Prentice-Hall; 1978.
[4] Batali J. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe E, editor. Linguistic evolution through language acquisition: Formal and computational models. Cambridge: Cambridge Univ. Press; 2002. p. 111–72.

---

[18] The Hamming distance between two meanings is the number of features which have different feature values.

[19] The Levenstein distance between two signals is the minumum number of deletions, insertions, and substitutions required to transform one string into another.

[5] Berko J. The child's learning of English morphology. Word 1958;14:150–77.

[6] Bloom P. How children learn the meanings of words. Cambridge, MA: MIT Press; 2000.

[7] Boyd R, Richerson PJ. Culture and the evolutionary process. Chicago, IL: Univ. of Chicago Press; 1985.

[8] Brighton H. Experiments in iterated instance-based learning. Tech. Rep., Language Evolution and Computation Research Unit; 2000.

[9] Brighton H. Compositional syntax from cultural transmission. Artificial Life 2002;8(1):25–54.

[10] Brighton H. Simplicity as a driving force in linguistic evolution. Ph.D. Thesis, The University of Edinburgh; March 2003.

[11] Brighton H. Linguistic evolution and induction by minimum description length. In preparation.

[12] Brighton H, Kirby S. The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In: Kelemen J, Sosík P, editors. Advances in artificial life: Proceedings of the 6th European conference on artificial life. Berlin: Springer-Verlag; 2001. p. 592–601.

[13] Brighton H, Kirby S. Understanding linguistic evolution by visualizing the emergence of topographic mappings. Artificial Life. In press.

[14] Brighton H, Kirby S, Smith K. Cultural selection for learnability: Three principles underlying the view that language adapts to be learnable. In: Tallerman M, editor. Language origins: Perspectives on evolution. Oxford: Oxford Univ. Press; 2005.

[15] Briscoe E, editor. Linguistic evolution through language acquisition: Formal and computational models. Cambridge: Cambridge Univ. Press; 2002.

[16] Cavalli-Sforza LL, Feldman MW. Cultural transmission and evolution: A quantitative approach. Princeton, NJ: Princeton Univ. Press; 1981.

[17] Chater N. The search for simplicity: A fundamental cognitive principle? Quarterly Journal of Experimental Psychology A 1999;52:273–302.

[18] Chater N, Vitányi PMB. Simplicity: A unifying principle in cognitive science? Trends in Cognitive Sciences 2003;7(1):19–22.

[19] Chater N, Vitányi PMB. A simplicity principle for language acquisition: Re-evaluating what can be learned from positive evidence; 2004 [manuscript under review].

[20] Cheney D, Seyfarth R. How monkeys see the world: Inside the mind of another species. Chicago, IL: Univ. of Chicago Press; 1990.

[21] Chomsky N. Aspects of the theory of syntax. Cambridge, MA: MIT Press; 1965.

[22] Chomsky N. Recent contributions to the theory of innate ideas. Synthese 1967;17:2–11.

[23] Chomsky N. In: Belletti A, Rizzi L, editors. On nature and language. Cambridge: Cambridge Univ. Press; 2002.

[24] Christiansen M. Infinite languages, finite minds: Connectionism, learning and linguistic structure. Ph.D. Thesis, University of Edinburgh; 1994.

[25] Christiansen M, Kirby S. Language evolution. Oxford: Oxford Univ. Press; 2003.

[26] Christiansen MH, Kirby S. Language evolution: Consensus and controversies. Trends in Cognitive Sciences 2003;7(7):300–7.

[27] Clark R, Roberts I. A computational model of language learnability and language change. Linguistic Inquiry 1993;24(2):299–345.

[28] Comrie B. Language universals and linguistic typology. 2nd ed. Oxford: Blackwell; 1989.

[29] Cowie F. What's within? Nativism reconsidered. Oxford: Oxford Univ. Press; 1999.

[30] Crain S. Language acquisition in the absence of experience. Behavioral and Brain Sciences 1991;14.

[31] Croft W. Typology and universals. Cambridge: Cambridge Univ. Press; 1990.

[32] Culicover PW. Syntactic nuts: Hard cases, syntactic theory, and language acquisition. Oxford: Oxford Univ. Press; 1999.

[33] Dawkins R. The selfish gene. Cambridge: Cambridge Univ. Press; 1989.

[34] Deacon T. The symbolic species. London: Penguin; 1997.

[35] Doherty MJ. Children's difficulty in learning homonyms. Journal of Child Language 2004;31(1):203–14.

[36] Durham WH. Coevolution: Genes, culture and human diversity. Stanford, CA: Stanford Univ. Press; 1991.

[37] Elman JE, Bates EA, Johnson MH, Karmiloff-Smith A, Parisi D, Plunkett K, editors. Rethinking innateness: A connectionist perspective on development. Cambridge, MA: MIT Press; 1996.

[38] Elman JL. Generalization from sparse input. In: Proceedings of the 38th annual meeting of the Chicago Linguistic Society. Chicago Linguistic Society; 2003.

[39] Feyerabend P. Against method. London: Verso; 1978.

[40] Glendinning P. Stability, instability, and chaos: An introduction to the theory of nonlinear differential equations. Cambridge: Cambridge Univ. Press; 1994.

[41] Grünwald P. Model selection based on minimum description length. Journal of Mathematical Psychology 2000;44:133–52.

[42] Hare M, Elman JL. Learning and morphological change. Cognition 1995;56(1):61–98.

[43] Hopcroft JE, Ullman JD. Introduction to automata theory, languages, and computation. Reading, MA: Addison–Wesley; 1979.

[44] Hurford JR. Biological evolution of the Saussurean sign as a component of the language acquisition device. Lingua 1989;77(2):187–222.

[45] Hurford JR. Nativist and functional explanations in language acquisition. In: Roca IM, editor. Logical issues in language acquisition. Dordrecht: Foris; 1990. p. 85–136.

[46] Jackendoff R. Foundations of language: Brain, meaning, grammar, evolution. Oxford: Oxford Univ. Press; 2002.

[47] Kagan J. The second year. Cambridge, MA: Harvard Univ. Press; 1981.

[48] Kimball JP. The formal theory of grammar. Englewood Cliffs, NJ: Prentice-Hall; 1973.

[49] Kirby S. Function, selection and innateness: The emergence of language universals. Oxford: Oxford Univ. Press; 1999.

[50] Kirby S. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In: Knight C, Studdert-Kennedy M, Hurford J, editors. The evolutionary emergence of language: Social function and the origins of linguistic form. Cambridge: Cambridge Univ. Press; 2000. p. 303–23.

[51] Kirby S. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. IEEE Transactions on Evolutionary Computation 2001;5(2):102–10.

[52] Kirby S. Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe E, editor. Linguistic evolution through language acquisition: Formal and computational models. Cambridge: Cambridge Univ. Press; 2002. p. 173–203.

[53] Krifka M. Compositionality. In: Wilson RA, Keil F, editors. The MIT encyclopaedia of the cognitive sciences. Cambridge, MA: MIT Press; 2001.

[54] LePage RB, Tabouret-Keller A. Acts of identity. Cambridge: Cambridge Univ. Press; 1985.

[55] Li M, Vitányi P. An introduction to Kolmogorov complexity and its applications. New York: Springer-Verlag; 1997.

[56] Lidz J, Waxman S. Reaffirming the poverty of the stimulus argument: A reply to the replies. Cognition 2004;93:157–65.

[57] Lidz J, Waxman S, Freedman J. What infants know about syntax but couldn't have learned: Evidence for syntactic structure at 18-months. Cognition 2003;89:65–73.

[58] Lightfoot D. The development of language: Acquisition, change, and evolution. Oxford: Blackwell; 1999.

[59] Mańczak W. Laws of analogy. In: Fisiak J, editor. Historical morphology. The Hague: Mouton; 1980. p. 283–8.

[60] Markman EM. Categorization and naming in children: Problems of induction. Cambridge, MA: MIT Press; 1989.

[61] Markman EM. Constraints on word learning: Speculations about their nature, origins, and domain specificity. In: Gunnar M, Maratsos M, editors. Modularity and constraints in language and cognition: The Minnesota symposium on child psychology, vol. 25. Hillsdale, NJ: Erlbaum; 1992. p. 59–101.

[62] Markman EM, Wachtel GF. Children's use of mutual exclusivity to constrain the meaning of words. Cognitive Psychology 1988;20:121–57.

[63] Maynard Smith J, Szathmáry E. The major transitions in evolution. Oxford: Oxford Univ. Press; 1995.

[64] Maynard Smith J, Szathmáry E. The origins of life: From the birth of life to the origins of language. Oxford: Oxford Univ. Press; 1999.

[65] Mazzocco MM. Children's interpretations of homonyms: A developmental study. Journal of Child Language 1997;24(2):441–67.

[66] Mesoudi A, Whiten A, Laland KN. Is human cultural evolution Darwinian? Evidence reviewed from the perspective of *the origin of species*. Evolution 2004;58(1):1–11.

[67] Newmeyer FJ. Uniformitarian assumptions and language evolution research. In: Wray A, editor. The transition to language. Oxford: Oxford Univ. Press; 2002.

[68] Niyogi P, Berwick RC. Evolutionary consequences of language learning. Linguistics and Philosophy 1997;20:697–719.

[69] Nowak MA, Komarova NL. Towards an evolutionary theory of language. Trends in Cognitive Sciences 2001;5(7):288–95.

[70] O'Grady W, Dobrovolsky M, Katamba F. Contemporary linguistics: An introduction. 3rd ed. London: Longman; 1996.

[71] Oliphant M. The learning barrier: Moving from innate to learned systems of communication. Adaptive Behavior 1999;7(3/4):371–84.

[72] Osherton D, Stob M, Weinstein S. Systems that learn. Cambridge, MA: MIT Press; 1986.

[73] Pinker S, Bloom P. Natural language and natural selection. Behavioral and Brain Sciences 1990;13(4):707–84.

[74] Pitt MA, Myung IJ, Zhang S. Toward a method of selecting among computational models of cognition. Psychological Review 2002;109(3):472–91.

[75] Pullum GK. Learnability, hyperlearning, and the poverty of the stimulus. In: Johnson J, Juge ML, Moxley JL, editors. Proceeding of the 22nd annual meeting: General session and parasession on the role of learnability in grammatical theory. Berkeley, CA: Berkeley Linguistic Society; 1996. p. 498–513.

[76] Pullum GK, Scholz BC. Empirical assessment of stimulus poverty arguments. The Linguistic Review 2002;19(1–2):9–50.

[77] Rissanen J. Modeling by shortest data description. Automatica 1978;14:465–71.

[78] Rissanen J. Stochastic complexity in learning. Journal of Computer and System Sciences 1997;55:89–95.

[79] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 1958;65:368–408.

[80] Slobin DI. Cognitive prerequisites for the development of grammar. In: Freguson CA, Slobin DI, editors. Studies of child language development. New York, NY: Holt, Rinehart and Winston; 1973.

[81] Slobin DI. Language change in childhood and history. In: Macnamara J, editor. Language learning and thought. London: Academic Press; 1977. p. 185–221.

[82] Slobin DI. Crosslinguistic evidence for the language-making capacity. In: Slobin DI, editor. The crosslinguistic study of language acquisition, vol. 2. Hillsdale: Lawrence Earlbaum Associates; 1985. p. 1157–249.

[83] Smith ADM. Intelligent meaning creation in a clumpy world helps communication. Artificial Life 2003;9(2):175–90.

[84] Smith K. The cultural evolution of communication in a population of neural networks. Connection Science 2002;14(1):65–84.

[85] Smith K. Compositionality from culture: The role of environment structure and learning bias. Tech. Rep., Language Evolution and Computation Research Unit, University of Edinburgh; 2003.

[86] Smith K. The evolution of vocabulary. Journal of Theoretical Biology 2004;228(1):127–42.

[87] Smith K, Brighton H, Kirby S. Complex systems in language evolution: The cultural emergence of compositional structure. Advances in Complex Systems 2003;6(4):537–58.

[88] Smith K, Hurford JR. Language evolution in populations: Extending the iterated learning model. In: Banzhaf W, Christaller T, Dittrich P, Kim JT, Ziegler J, editors. Advances in artificial life (Proceedings of the 7th European conference on artificial life). Berlin: Springer-Verlag; 2003. p. 507–16.

[89] Smith K, Kirby S, Brighton H. Iterated learning: A framework for the emergence of language. Artificial Life 2003;9(4):371–86.

[90] Steels L. Constructing and sharing perceptual distinctions. In: van Someren M, Widmer G, editors. Proceedings of the European conference on machine learning. ECML '97. Berlin: Springer-Verlag; 1997. p. 4–13.

[91] Teal TK, Taylor CE. Effects of compression on language evolution. Artificial Life 2000;6(2):129–43.

[92] Tomasello M. The cultural origins of human cognition. Harvard: Harvard Univ. Press; 1999.

[93] Trask RL. Historical linguistics. London: Arnold; 1996.

[94] Trudgill P. Sex, covert prestige, and linguistic change in the urban British English of Norwich. Language in Society 1972;1:179–96.

[95] Uriagereka J. Rhyme and reason: An introduction to minimalist syntax. Cambridge, MA: MIT Press; 1998.

[96] von Frisch K. Decoding the language of the bee. Science 1974;185:663–8.

[97] Wagner L. Defending nativism in language acquisition. Trends in Cognitive Sciences 2001;6(7):283–4.

[98] Wexler K. On the argument from the poverty of the stimulus. In: Kasher A, editor. The Chomskyan turn. Oxford: Blackwell; 1991.

[99] Wolff JG. Language acquisition, data compression, and generalization. Language and Communication 1982;2(1):57–89.

[100] Zadrozny W. From compositional to systematic semantics. Linguistics and Philosophy 1994;17:329–42.