Compositionality, Linguistic Evolution, and Induction by Minimum Description Length

Henry Brighton

1 Introduction

The following question is crucial to both linguistics and cognitive science in general: How can we go about explaining why language has certain structural properties and not others? The dominant explanation proposes that constraints on linguistic variation – universal patterns found in language – are a direct reflection of properties of a genetically determined language faculty (eg., Chomsky, 1965, 2002). Compositionality is one such universal characteristic of language. The dominant explanation suffices if we regard the process of language acquisition to be in no way a process involving inductive generalizations. This is to say that the essential characteristics of language are not learned in any meaningful sense, as they are not the product of inductive generalizations made from linguistic data. This conjecture depends in large part on what is known as the *argument from the poverty of the stimulus* (APS) which states that the data required to make the appropriate inductive generalizations is simply not available to a child during language acquisition (eg., Wexler, 1991).

Despite the dominance of this theory, the assumption that linguistic universals are in no sense acquired as a result of inductive generalizations is controversial: the APS is only conjecture, and is in opposition to several alternative theoretical standpoints and empirical studies (eg., Cowie, 1999; Pullum and Scholz, 2002). In the discussion that follows, I consider how deviation from the extreme position suggested by the APS poses a problem when explaining linguistic universals such as compositionality. To address this deficiency, the following discussion considers an alternative explanation for the occurrence of compositionality in language. The validity of this alternative is tested using a computational model. In particular, I will argue that compositionality in language is not a direct reflection of our genetic endowment, but is instead fundamentally related to the way language is learned and culturally transmitted (see also Deacon, 1997; Christiansen and Kirby, 2003; Brighton et al., 2005). Cen-

tral to this explanation is the role of inductive inference. For this reason, the computational model discussed below employs a model of induction based on the minimum description length principle (Rissanen, 1978, 1989).

2 Issues in Explaining Linguistic Universals

Language is a particular system of relating sound 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 worlds languages. Why do all languages share these properties? Among those interested in language, a widespread hypothesis is that these intrinsic properties of language are, like the visual system, an expression of the genes (eg., Chomsky, 2002). To support this view, we can note that children master complex features of language on the basis of surprisingly little evidence. In fact, as we have seen, the APS is a conjecture stipulating that the knowledge of language children attain is surprising precisely because it *cannot* be derived solely from information made available by the linguistic environment (eg., Chomsky, 1965; Wexler, 1991; Cowie, 1999; Pullum and Scholz, 2002).

The modern debate on the innateness of language is dominated by the notion that the framework for linguistic development is innate, with the linguistic environment serving to supply information that steers an internally directed course of development. In this sense, languages themselves (eg., Spanish, Mandarin Chinese) 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? Linguistics, by conducting a thorough analysis of the world's languages, proposes a set of descriptive statements which capture these hallmarks of language. For example, the linguist may identify properties common to all languages they encounter, properties that occur according 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* (eg., Comrie, 1989; Croft, 1990; O'Grady et al., 1997). 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 aspect of its strength is that by coupling universal properties of language tightly to a theory of innate constraints our analysis of the structural hallmarks of language must center on a wholly psychological (ie., cognitive, mentalistic, or internal-

ist) explanation. As a consequence, by understanding those parts of the human cognitive system relevant to language we can understand why languages have certain structural characteristics and not others. With respect to understanding language, our object of study has been circumscribed to encompass a physical organ: the brain. Hence, the relationship between *descriptive statements* of universal properties of language and an *explanatory statements* of why language is this way are therefore largely transparent.

As we have seen, this position is largely substantiated by 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" (Chomsky, 2002). The claim that language is not learned causes a great deal of controversy and will impact heavily on the discussion to come. Nevertheless, to characterize 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 the same way, we should not claim that we learn speak.

Language Learning Under Innate Constraints

Linguistic nativism is far from accepted in the extreme form presented above (for insightful discussion, see Cowie, 1999; Jackendoff, 2002; Culicover, 1999). A less extreme alternative to this hypothesis is that the structure of language, to some extent, is learned by children: humans can arrive at complex knowledge of language without the need to have hard-wired (genetically determined) expectations of all dimensions of linguistic variation. This is the view I we will adopt throughout this article. I assume that to some degree language is learned through inductive generalizations from linguistic data, but to what degree it is learned is unclear. My position is therefore at odds with Chomsky's positional that knowledge of language goes "far beyond the presented primary linguistic data and is in no sense an 'inductive generalization' from these data." (Chomsky, 1965). What evidence can we draw on to resolve this debate? 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 (Pullum and Scholz, 2002; Culicover, 1999). This debate is often reduced to statements such as "linguistic structure is much more complex than the average empiricist supposes" (Wagner, 2001), and claims that "the attained grammar goes orders of magnitude beyond the information provided by the input data" (Wexler, 1991). These claims are backed up with specific exam-

ples designed to show how children's knowledge of language extends beyond what the data suggests (eg., Kimball, 1973; Baker, 1978; Crain, 1991; Lidz et al., 2003; Lidz and Waxman, 2004). Nevertheless, many still argue that the required information is in fact present in the linguistic data (Pullum, 1996; Pullum and Scholz, 2002), and to claim that it is not is "unfounded hyperbole" (Sampson, 1989). These rebuttles of the argument from the poverty of the stimulus are often based on the notion that "[l]earning is much more powerful than previously believed" (Bates and Elman, 1996). It should be noted that this stance in no way denies that fact 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 (Elman, 2003). In the light of this debate, I will 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 generalizations (ie., question the APS), then we must acknowledge that the linguistic environment must supply *information*. This information impacts on how universal properties of languages, like compositionality, are represented and processed within the cognitive system.

Towards an Evolutionary Explanation

The thrust of this discussion rests on the realization that the degree to which language is learned through a process of inductive generalization has a profound affect on the character of the explanatory framework we use to understanding why language has the structure that it does (Brighton et al., 2005). Why is this? 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. The environment must be supplying structural information in order for induction to occur. To achieve explanatory adequacy we must now explain why the environment is the way it is: How did this information come to exist? To address this issue I will argue for an evolutionary perspective, and seek to explain *how*, from a non-linguistic environment, compositional structure can develop through linguistic evolution. In short, this view casts doubt on the view that the hallmarks of language are, as Chomsky states, "coming from inside, not from outside." Necessarily, if inductive generalizations made from data contained in the environment determine the kind of linguistic structure we observe, then a wholly psychological theory of linguistic structure must be inadequate.

3 Linguistic Evolution Through Iterated Learning

Languages are 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 (eg., Andersen, 1973; Hurford, 1990). Focusing on this process of linguistic transmission, several computational models have demonstrated how phenomena of language change can be understood in these terms (Clark and Roberts, 1993; Hare and Elman, 1995; Niyogi and Berwick, 1997). The study of language change seeks to understand how full-blown human languages undergo structural change over time. For example, these models could an inform an enquiry into the morphological change that characterized the transition from, say, Latin to French. Of more importance to this discussion are studies that focus specifically on the linguistic evolution of linguistic complexity from non-linguistic communication systems (Batali, 2002; Kirby 2002, 2001; Brighton, 2002; Smith et al., 2003b). It should be noted, therefore, that linguistic evolution is a process that drives both *evolution* and *change* in language. Studies of language evolution are concerned with the origin of linguistic structure found in human languages while studies of language change are concerned with how such languages alter over time. Much of the work focusing on the evolution of language through linguistic evolution has been consolidated under a single computational modeling framework termed the *iterated learning model* (Kirby, 2001; Brighton, 2002; Smith et al., 2003a,b). In this article I will use an iterated learning model to demonstrate the evolution of compositionality. This model will be based on a contemporary theory of induction known as the minimum description length principle.

An iterated learning model is composed of a series of agents organized into generations. Language is transmitted through these agents: the agents act as a conduit for language. For a language to be transmitted from one agent to another, it must 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 the examples. 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).

A language is a particular mapping between a *meaning space* and a *signal*

space. Meanings are often modeled as compound structures such as feature vectors or logical expressions. Signals are usually serial structures, such as a string of symbols. Knowledge of language (a hypothesis) is a representation of this mapping. This representation could be modeled by any one of a number of computational models of inductive inference. The basic iterated learning model considers each individual agent in turn. Throughout this article I will consider the case when each generation contains only one agent. The first agent in the simulation, Agent 1, is initialized 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 mapping (language) L_{h_1}. These utterances are then passed to the next agent to learn from. Once the language has been transmitted from the first to the second agent, the first agent plays no further part in the simulation. The simulation proceeds by iteratively introducing a new agent to transmit the language. Each agent represent one generation, and the experiment is run for some number of generations. The important point is that, under certain conditions, the language will change from one generation to another; it will evolve and undergo adaptation. This process is illustrated in Figure 1.

One crucial driving force behind linguistic evolution is the *transmission bottleneck*, which imposes a constraint on how languages are transmitted. The transmission bottleneck reflects, as a constraint within the model, the fact that natural languages cannot be transmitted in totality from one individual to another. Linguistic data is never exhaustive; it is always sparse. For example, in the case of natural language, it is impossible for an infinitely large mapping between between meanings and signals to be externalized. In the iterated learning model this situation occurs too: Within each body of linguistic data only a subset of the set of possible meanings will be associated with a signal. This constraint will hit hard when we consider the process of language production. Production is the process by which agents find signals for meanings they are prompted to produce. The meanings the agent must produce signals for are, in the model that follows, always drawn at random from the meaning space. Production has to occur in the model, as an agent needs to create a set of utterances from which the next agent in the simulation is to learn from. If a meaning was not seen by an agent in conjunction with a signal during acquisition, then how is the agent to produce an appropriate signal? There are two courses of action. First, the agent can use the generative capacity of the induced hypothesis to yield an appropriate signal; in this case the agent will have *generalized*. Second, if generalization is not possible, then the agent will have to invent some signal for the meaning using some other means.

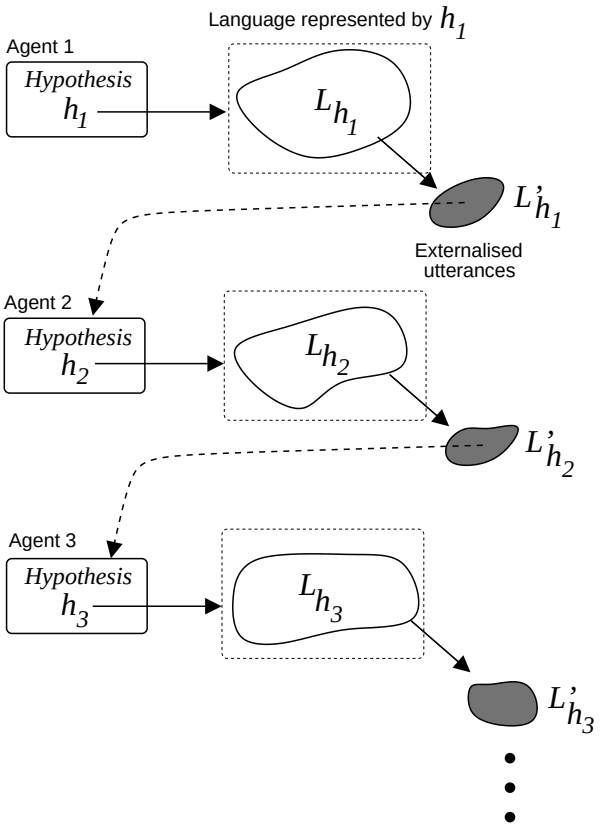


Figure 1: The iterated learning model. Agent 1 has knowledge of language represented by hypothesis h_1. This hypothesis represents a language L_{h_1}. Some subset of this mapping, L'_{h_1}, is externalized as linguistic performance by Agent 1 for the next agent, Agent 2, to learn from. On the basis of this performance, Agent 2 induces hypothesis h_2. The process is then repeated, generation after generation.

The impact of the transmission bottleneck has two interpretations within an iterated learning model. On the one hand, it represents a constraint on transmission. On the other, it represents a constraint on how much evidence is available to the learning algorithm used by each agent. By imposing sparsity in the available learning data a situation analogous to the poverty of stimulus, discussed above, is introduced. In order for an agent to acquire a generative capacity, the agent must generalize from the data it has have been given. In this sense, linguistic competence represents the ability to express meanings. To achieve a generative capacity requires that structure is present in the data. This is a crucial observation I will return to later.

Modeling Compositionality

A model of language needs to capture the fact that language is a particular relationship between sound and meaning. The level of abstraction used here will capture the fact that language is mapping from a "characteristic kind of semantic or pragmatic function onto a characteristic kind of symbol sequence" (Pinker and Bloom, 1990). When I refer to a model of language, I will be referring to set of possible relationships between, on the one hand, entities representing *meanings* and on the other, entities representing *signals*. Throughout this article I 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 number of features each meaning will have. The parameter V defines how many values each of these features can accommodate. A meaning space $\mathscr{M}$ specified by $F = 2$ and $V = 2$ would represent the set: $\mathscr{M} = \{(1,1),(1,2),(2,1),(2,2)\}$. Notice that meanings represent structured objects of a fixed length. Signals, in contrast, are represented as a variable length string of symbols drawn from some alphabet Σ. The length of signals within the model is bounded by the maximum length denoted by l_{max}. For example a signal space $\mathscr{S}$, defined by $\Sigma = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d}\}$ and $l_{max} = 4$, might be the set $\mathscr{S} = \{\mathsf{ba},\mathsf{ccad},\mathsf{acda},\mathsf{c},\ldots\}$.

We now have a precise formulation of the meanings and signals, but of greater importance to following discussion will be the kinds of structural relationships that exist between meanings and signals. It is the kind relationship between meanings and signals that makes human language so distinctive. As it stands, the model of language presented above can capture a key feature of language I will be focusing on: It can represent both compositional mappings and non-compositional mappings (for more exotic language models see Kirby 2002; Batali, 2002). Compositionality is a property of the mapping between meanings and signals. It is not a property of a set of meanings, nor a property of a set of signals. A compositional mapping is one where the meaning of a signal is some function of the meaning of its parts (eg., Krifka, 2001). Such a mapping is possible given the model of language developed so far. Consider the language L_{comp}:

$$\begin{aligned} L_{comp} = \; & \{\langle\{1,2,2\},\mathsf{adf}\rangle,\langle\{1,1,1\},\mathsf{ace}\rangle,\langle\{2,2,2\},\mathsf{bdf}\rangle, \\ & \langle\{2,1,1\},\mathsf{bce}\rangle,\langle\{1,2,1\},\mathsf{ade}\rangle,\langle\{1,1,2\},\mathsf{acf}\rangle\} \end{aligned}$$

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 symbol a, for example, represents

feature value 1 for the first feature. The precise relationship between meanings and signals can vary substantially. For example, one feature value can map to two separate parts of the signal, these parts of the signal can be of variable length, and some parts of the signal can correspond to no part of the meaning. But importantly, the property of compositionality is independent of such characteristics of the mapping. Compositionality is an abstract property capturing the fact that *some* function determines how parts of the signal correspond to parts of the meaning. The exact definition of a compositional relationship is subject to heated debate, and one I will sidestep in the interests of brevity (eg., Zadrozny, 1994). All human languages exhibit compositionality.

Instances of the model of language with no compositional structure whatsoever are also of interest. I will term such relationships *holistic* languages[1]: the whole signal maps to a whole meaning, such that no obvious relationship exists between parts of the signal and parts of the meaning. Here is an example of a holistic language $L_{holistic}$:

$$\begin{aligned} L_{holistic} = \; & \{\langle\{1,2,2\},\mathsf{sghs}\rangle,\langle\{1,1,1\},\mathsf{ppold}\rangle,\langle\{2,2,2\},\mathsf{monkey}\rangle, \\ & \langle\{2,1,1\},\mathsf{q}\rangle,\langle\{1,2,1\},\mathsf{rcd}\rangle,\langle\{1,1,2\},\mathsf{esox}\rangle\} \end{aligned}$$

A holistic language is usually constructed by associating a random signal to each meaning. For this reason, holistic languages may also be referred to as random languages in the discussion that follows. Given the model of language described above we can now consider in more depth how iterated learning models can be used to explore the linguistic evolution of compositionality.

4 Linguistic Evolution and Induction

A crucial component of any iterated learning model is the process of induction, as agents are required to induce hypotheses explaining the linguistic data they observe. Making an inductive inference involves choosing a hypothesis from a set of candidate hypotheses $\mathscr{H} = \{H_1, H_2, \ldots\}$ in the light of some data D. Such an inference, depending on the chosen hypothesis, can result in a general statement not only concerning the observed data, but also data yet been observed. The problem of induction is the problem of identifying the most appropriate hypothesis, and hence the most appropriate statement, that explains the given data D. Contemporary theories of induction regard this problem as one fundamentally resting on the issue of complexity (Rissanen, 1978; Li and Vitányi, 1997;

[1] 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 clarity.

Pitt et al., 2002). Complexity is the flexibility inherent in a class of hypotheses that allow them to fit diverse patterns of data. For example, choosing a hypothesis that is consistent with the observed data may describe the observed data but, due to the high degree of complexity of the hypothesis, may be woefully inadequate as an explanation of the data. The hypothesis, by virtue of its inherent complexity, may also describe an extremely diverse range of data. This makes the hypothesis less likely to be an appropriate model of the underlying data generating machinery. Similarly, a hypothesis with insufficient complexity will not possess the complexity required to explain the data. Accordingly, the inductive process is fundamentally an issue of identifying a trade-off in the complexity of hypotheses.

One approach to tackling this issue is the minimum description length (MDL) principle (Rissanen, 1978; Li and Vitányi, 1997). The MDL principle provides a means of judging, given a hypothesis space $\mathscr{H}$ and some data D, which member of $\mathscr{H}$ represents the most likely hypothesis given that D was observed. The key idea behind MDL is that the more we are able to compress the observed data, the more we have learned about it: any kind of pattern of regularity present in the data can potentially allow us to compress the data. Once we have identified such a pattern, we can re-describe the observed data using fewer symbols than a literal description of the data. This is philosophy behind the principle. MDL can be deployed in a practical sense by drawing on a theoretically solid and formally well understood body of techniques. Basing a model of induction on the MDL principle has the advantage that hypothesis selection is determined by the complexity of the hypotheses under consideration. In recent years, the MDL principle has become increasingly influential in the analysis of learning (Rissanen, 1997), model selection (Grünwald, 2002; Pitt et al., 2002), and many aspects of the cognitive system (Chater, 1999; Chater and Vitányi, 2003) including language acquisition (Wolff, 1982; Chater and Vitányi, 2004).

More formally, the MDL principle states that the most likely hypothesis is the one which minimizes 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 formalize 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 $\mathscr{H}$, and the observed data, D, the MDL principle selects a member of $\mathscr{H}$, H_{MDL}, as follows:

$$H_{MDL} = \min_{H \in \mathscr{H}} \{L_{C_1}(H) + L_{C_2}(D|H)\} \tag{1}$$

This expression states that the best hypothesis 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. Here we see how the selected hypothesis represents a point in a trade-off between high and low complexity explanations. The MDL principle tells us how to judge competing hypotheses with respect to this trade-off by exploiting the relationship between coding and probability (Cover and Thomas, 1991).

Learning Based on Minimum Description Length

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 $\mathscr{H}$, and coding schemes over these hypotheses. Recall that data in this discission are collections of utterances whose form is determined by the language model introduced earlier. One example is the following set of utterances, L_{comp}:

$$\begin{aligned} L_{comp} = \; & \{\langle\{1,2,2\},\text{adf}\rangle, \langle\{1,1,1\},\text{ace}\rangle, \langle\{2,2,2\},\text{bdf}\rangle, \\ & \langle\{2,1,1\},\text{bce}\rangle, \langle\{1,2,1\},\text{ade}\rangle, \langle\{1,1,2\},\text{acf}\rangle\} \end{aligned}$$

In order to apply the MDL principle to the selection of hypotheses given some arbitrary series of utterances, I will consider a hypothesis space composed of finite state unification transducers[2] (FSUTs) (Brighton, 2002). These transducers relate meanings to signals using 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 Figure 2(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 $\mathscr{H}$ 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 L, the space of candidate hypotheses will consist of all FSUTs consistent with the observed utterances. By consistent,

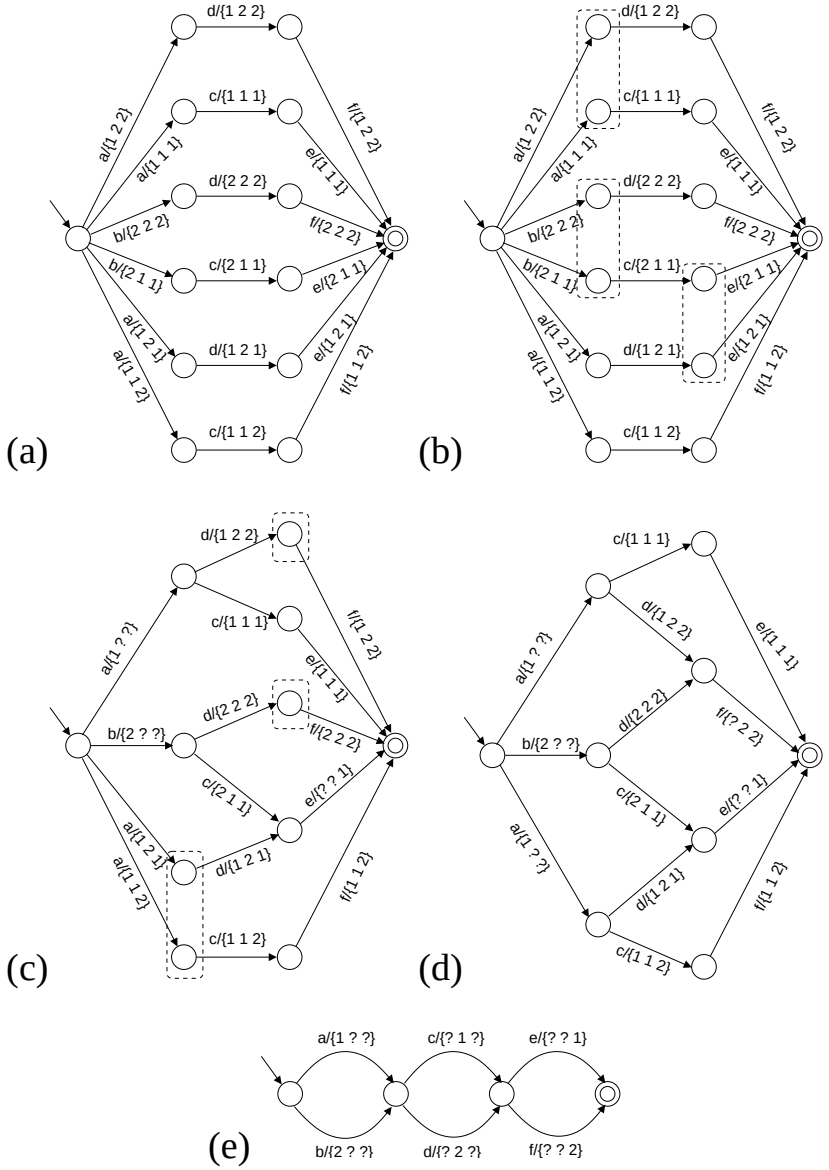
[2] A FSUT is a variation on the basic notion of a finite state transducer (Hopcroft and Ullman, 1979). The use of such transducers was inspired by and extends the work of Teal and Taylor (2000).

I mean that observed examples of meaning/signal associations are never discarded, the candidate hypotheses are constrained to always be able to generate, at a minimum, all the observed utterances. We are interested in situations in which a transducer is capable of generating utterances for meanings it has never observed; in such a situation, the transducer will have generalized.

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 applicable, can introduce generalizations by by merging states and edges. Given the 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, the reader should refer to the work presented in Brighton (2002, 2003). The important feature of the FSUT model, in combination with the MDL principle, is that compression can lead to generalisation. For example, Figure 2(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 shown in in Figure 2(d). Figure 2(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 they lead the transducer to express meanings which are inconsistent with the observed language. Nevertheless, 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 that where not mentioned in the linguistic data. To express a novel meaning, a search through the transducer is sought such that an appropriate series of edge transitions are found. Some of the meanings on these edge transitions, as a result of the application of the compression operators, may contain wildcard feature values that represent unbound feature values. These free variables are introduced when two transitions are merged which contain conflicting values for a particular feature. To express a novel meaning, the unification of the set of meanings occurring on the transitions must yield the meaning to be expressed. The resulting signal is formed by concatenating the symbols mentioned on each edge; the ordering of the symbols in the signal therefore reflects the ordering of the edge traversals when passing through the transducer. For example, a close inspection of the compressed transducer shown in Figure 2(e) reveals that meanings which are

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



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

Figure 2: Given the compositional language L_{comp}, the prefix tree transducer shown in (a) is constructed. By performing edge and state merge operations, the result of which are shown in (b), (c), and (d), the transducer can be compressed. The transducer shown in (e) represents a fully compressed transducer. It can generalize to the language L^{+}_{comp}.

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 in Figure 2, contains all the meaning/signal pairs which can be expressed by the fully compressed transducer in the above example. 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 has been made explicit, and as result, generalisation occurred. Generalisation will not be possible when structure is lacking in the observed data, and the result will be that some meanings cannot be expressed.

We now have a hypothesis space over which we can apply the MDL principle. The hypothesis chosen 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 (2003). With these model components in place, we are now in a position to assess the impact of induction based on the MDL principle within the iterated learning model.

5 The Evolutionary Consequences of the Simplicity Principle

The previous section summarised a model of learning based on compression constrained by the MDL principle. In this section I will describe how this model of learning leads to the cultural adaptation of the language as it is transmitted from one generation to the next within the iterated learning model. In order to specify this process in sufficient detail for simulation, several parameters need to defined. The meaning space is defined by F, the number of feature in each meaning, V, the number of values each feature can accommodate. The signal space is defined by Σ, the alphabet of symbols, and l_{max}, the maximum string length for randomly generated signals. A transmission bottleneck is imposed by restricting the number random utterances observed, R, to 32. These parameters are used to define the initial state of the system, including the initial language.

Figure 3 details these parameter values and depicts an example MDL FSUT induced from an initial random language constructed with the given parameter values. Negligible compression occurs. The language represented by the transducer is holistic; the compositional structure we seek to explain is lacking, and this is the situation we are interested in: how can a compositional mapping with maximum expressivity evolve through cultural adaptation?

Next, I will consider a crucial aspect of the model side-stepped so far: the issue of invention. Invention occurs when an agent is presented with a meaning it cannot express. That is, the meaning was not observed in conjunction with a signal during learning, and cannot be expressed as a result of generalisation. For example, the transducer in Figure 3 can only express the meanings which were

$$F = 3, V = 2, |\Sigma| = 20, l_{max} = 15, R = 32$$

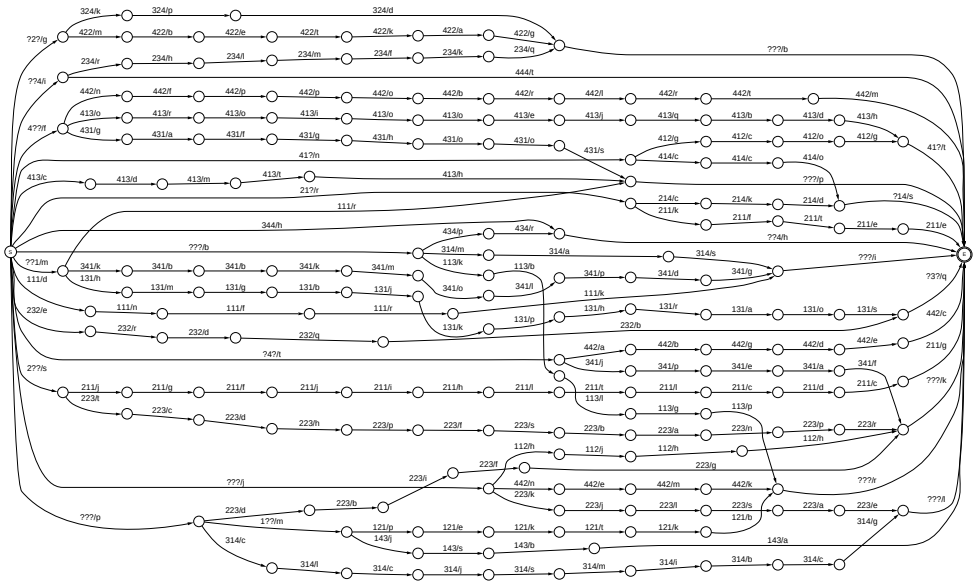


Figure 3: Given an initial random language defined by the above parameter values, H_{MDL} is an example of an induced FSUT. Negligible compression has occurred, and as a result the transducer does not generalise to novel meanings: 32 utterances were given as input, and each of these utterances is encoded by a unique path through the transducer.

present in the observed language. Within the iterated learning model, transducers will be required to express meanings which were not in the observed set of utterances. To solve this problem, a policy of random invention can be deployed, where a random signal is generated for novel meanings. This policy will be investigated first. Initialized with a random language the simulation is run for 200 iterations. Figure 4(a-b) illustrates how the system develops from one generation to the next. First of all, Figure 4(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}: $\alpha = 1 - \frac{|H_{mdl}|}{|H_{prefix}|}$. A high compression rate means that the language is compressible. Figure 4(a) illustrates that the compressibility of the language, from one generation to the next, changes very little. The initial random language undergoes no significant adaptation and remains unstructured and therefore uncompressible ($\alpha \approx 0.06$). Figure 4(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, labeled *A*, the systems fol-

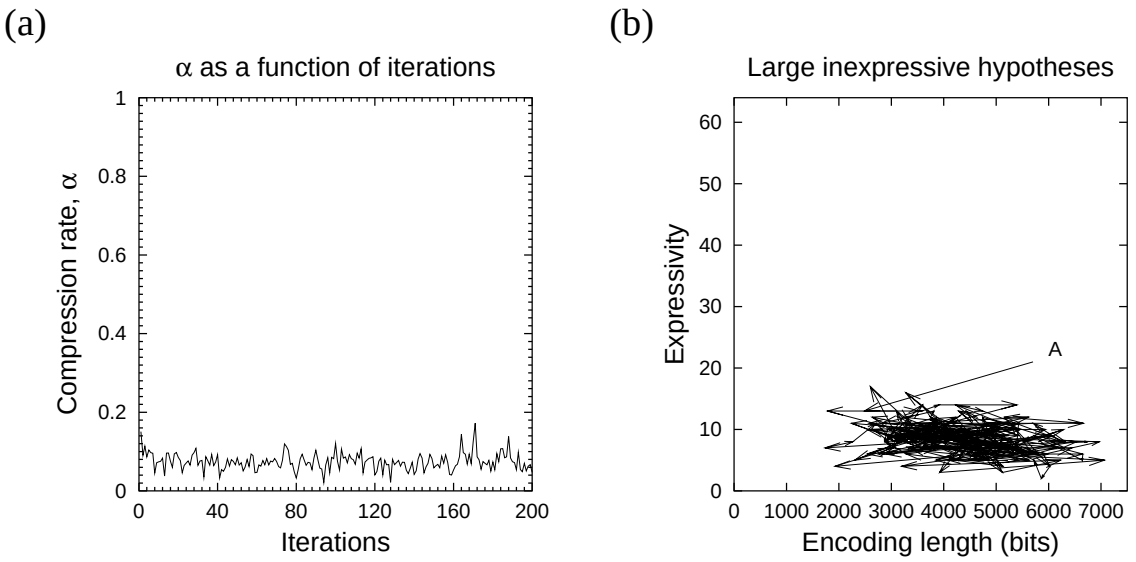


Figure 4: Linguistic evolution resulting from partially random invention.

lows 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. Why is this? The mechanisms supporting linguistic evolution – language learning and language production – are somehow inhibiting the cumulative evolution of structure. The source of this inhibition is the random invention procedure.

Invention Based on a Simplicity Principle

The MDL principle can tell nothing about the process of production. As a result, the process of interrogating the hypothesis with novel meanings to yield signals is not fully defined, and needs to be developed. To address this problem, a more principled invention mechanism is investigated where the invented signal is a derived using hypothesis itself rather than being constructed at random. The invented signal will be constrained by structure present in the hypothesis. The invention method I employ here 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 (2003), but the essentials of the process can be summarised as follows. The invented signal, if it where 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

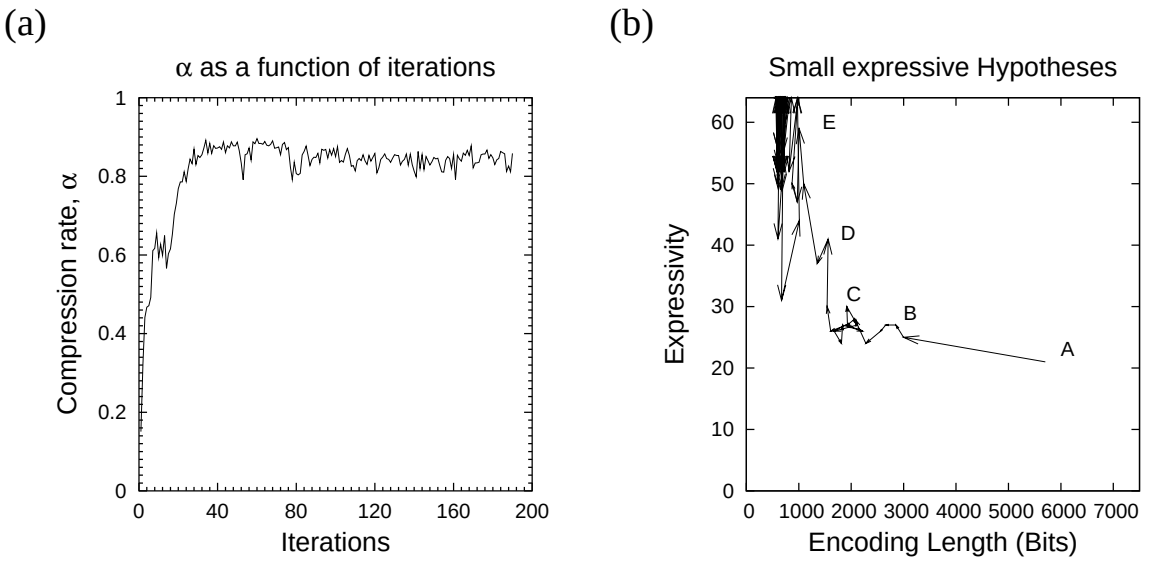


Figure 5: Linguistic evolution arising from simplicity based invention.

sense matches the structure of hypothesis. If such a signal cannot be found, then no signal is invented. In short, the invention procedure, rather than being random, now takes into account the structure present in the hypothesis.

Running the experiment with this invention procedure, Figure 5 illustrates exactly the same measurements as those shown in Figure 4. Strikingly, Figure 5 reveals that very different evolved states result as a consequence of the alternative invention procedure. Figure 5(a) illustrates an entirely different trajectory, 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 see small hypotheses capable of expressing all 64 meanings. The compressor scheme consistently directs linguistic evolution toward compositional systems. The first major 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 organizing 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 characterized by high expressivity. Figure 5(b) reflects these transitions. The compression rate rises in two stages corresponding to the points in the L-shaped trajectory.

Figure 6(a-d) depicts the transducers corresponding to the points *B*, *C*, *D*, and *E* in Figure 5(b). Figure 6(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 Figure 6(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. Figure 6(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. Although this transducer exhibits a large amount of redundancy, this redundancy does not effect its ability to be induced generation after generation. Initially, as the state approaches point *E*, some variation occurs across iterations before the steady state is arrived at. This suggests the stable regions of the state space are Liapounov stable: if the system were to start in this region, it would stay within this region (see, for example, Glendinning, 1994).

6 Linguistic Evolution of Compositionality

The previous section demonstrated how linguistic evolution can lead, from an initially holistic communication system, to the development of a compositional mapping between meanings and signals. It also indicated that certain conditions must be met if compositional structure is to develop at all: the invention mechanism, for example, cannot be random. This is one condition of many required in order for cumulative evolution to occur. It is not the case compositional structure is the inevitable outcome of an iterated learning model. In fact, the conditions required for cumulative evolution are strict. Brighton (2002) showed that the evolution of compositional structure requires that: (1) a transmission bottleneck imposing a sufficient degree of data sparsity is in place, and (2), that a sufficient degree of complexity is present in the meaning space. The parameters used in the previous section were chosen to maximize the likelihood of compositional systems according to the mathematical model reported by Brighton (2002). Although the example we have just considered represents

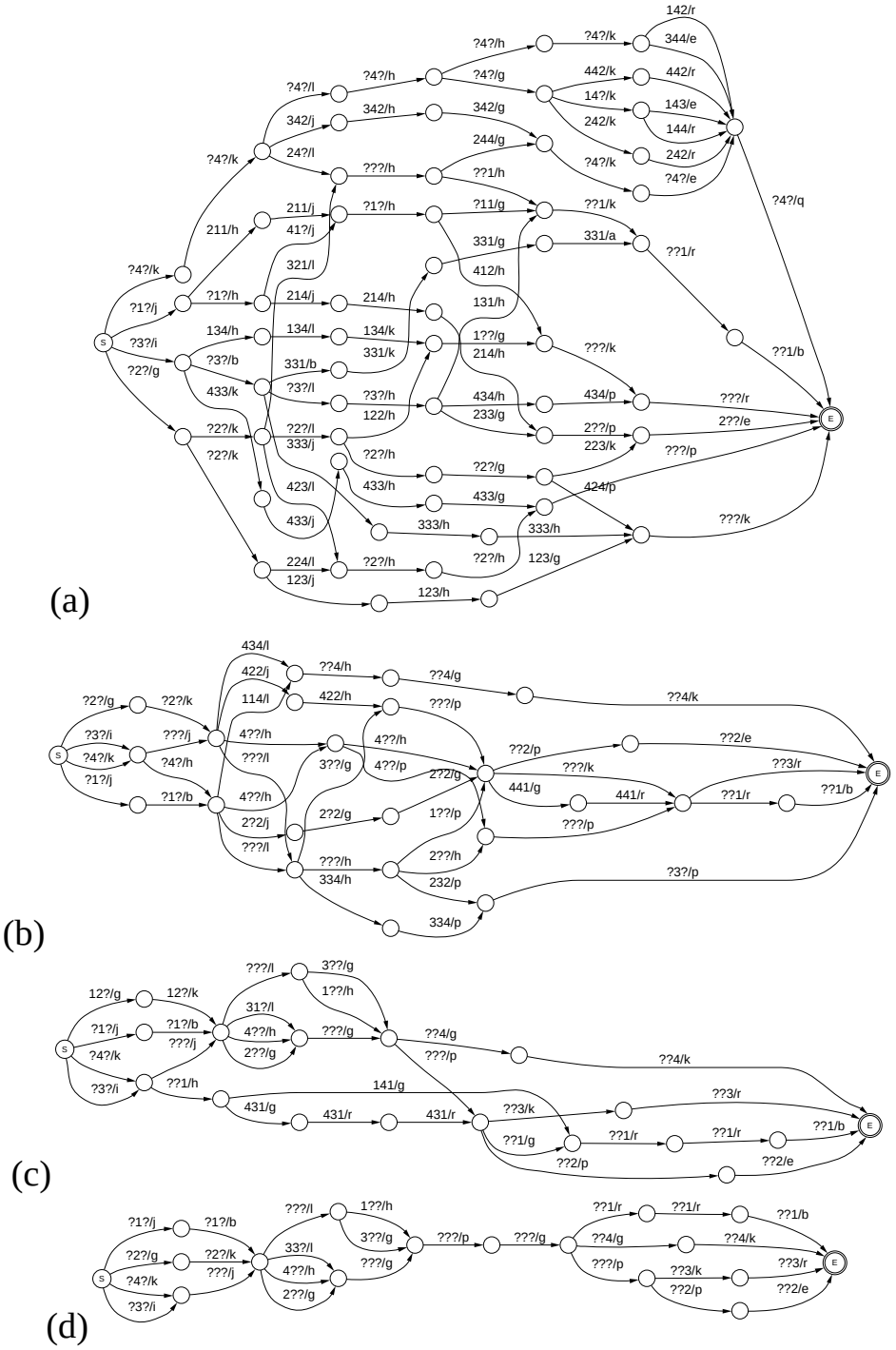
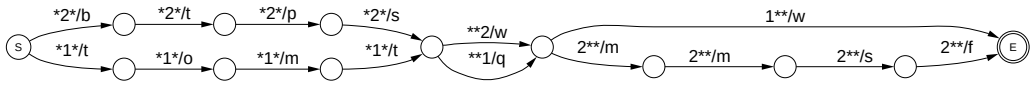


Figure 6: The transducers corresponding to positions *B*, *C*, *D*, and *E* highlighted in Figure 5.

the result of one simulation, this evolutionary trajectory is typical for the given parameter values. In short, it should be stressed that without a transmission bottleneck present languages will not change and compositional systems will not be observed. Similarly, without a sufficient degree of complexity in the meaning space, compositionality will not confer a stability advantage and therefore compositional languages are unlikely to be observed.

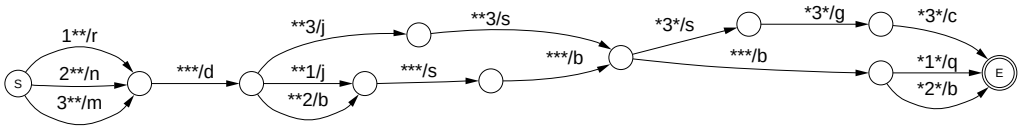
The sensitivity of the environmental conditions required for the evolution of compositional systems suggests that an explanation for why language exhibits compositionality cannot be framed exclusively in terms of the properties of the cognitive system. To fully appreciate this point, it is worth considering the nature of stable states in the model, as they provide an example of how the linguistic complexity observed is not trivially related to the properties of linguistic agents. Figure 7 shows two stable states result from the model. Figure 7(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, Figure 7(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.

(a)



G_1: S/x,y,z → A/x B/y C/z
A/1 → w
A/2 → mmsf
B/1 → tomt
B/2 → btps
C/1 → w
C/2 → q

(b)



G_2: S/x,y,z → A/x d B/y C/z
A/1 → r
A/2 → n
A/3 → m
B/y → D/y sb
B/3 → js
D/1 → j
D/2 → b
C/3 → sgc
C/z → b E/z
E/1 → q
E/2 → b

Figure 7: 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.

Optimal transducers, those with the lowest description length given the pa-

rameter 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 Figure 7 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 Figure 7, 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 are functional in the sense that they are stable, despite deviating from the "optimal" transducer, and this is why we observe them. The key point here is that given an understanding of the learning and production mechanisms of the linguistic agents, it is far from clear that such an understanding would by itself allow us to predict the outcome of the model. The process of linguistic evolution represents a complex adaptive system. This conclusion can be related to task of explaining why human languages have certain structural relationships and not others. If linguistic evolution plays a role in determining the structure of human language, then we must conclude that: (1) linguistic universals are not necessarily direct reflections of properties of the cognitive system, and (2), that an internalist or mentalistic explanation of linguistic universals is likely to be fundamentally lacking.

7 Conclusion

This discussion began with the following observation: If the universal features of language are in no sense acquired through a process of inductive generalizations, then we can rightfully circumscribe the cognitive system alone as the focus of an explanation for why language exhibits certain structural characteristics and not others. This must be the case, as the only remaining source of explanation has to appeal to characteristics of the cognitive system: characteristics of the environment are rendered irrelevant. The assumption that universal features of language are not acquired through inductive generalizations is controversial. The assumption has many critics. Importantly, deviating from this assumption necessarily creates a problem in explaining *why* language exhibits

these universal characteristics. If the acquisition of linguistic universals rely to some extent on properties of the linguistic data, then to retain explanatory adequacy requires us to explain *why* the environment contains certain structural characteristics and not others.

This discussion has focused on the process of linguistic evolution as a source explanation for why the linguistic data exhibits certain characteristics and not others. In particular, I have focused on the property of compositionality and explored the possibility that the process of linguistic evolution can explain how the linguistic environment came to contain compositional structure. To test this theory, I have used a computational model of linguistic evolution. The model predicts the cumulative evolution of compositional structure given certain environmental conditions. Importantly, the model also suggests that an understanding of the properties of linguistic agents cannot by itself satisfactorily explain why the evolved languages exhibit compositionality. This alternative standpoint places an explanation for why language exhibits certain hallmarks and not others fundamentally in the terms of an interaction between how language is acquired, how language is transmitted, and how the innate constraints on acquisition came to exist. In short, this work cast doubt on the utility of wholly psychological, cognitivistic, or internalist explanations for linguistic universals such as compositionality.

References

Andersen, H. (1973). Abductive and deductive change. *Language*, 40:765–793.

Baker, C. L. (1978). *Introduction to Generative-Transformational Syntax*. Prentice-Hall, Englewood Cliffs, NJ.

Batali, J. (2002). 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*, pages 111–172. Cambridge University Press, Cambridge.

Bates, E. and Elman, J. (1996). Learning rediscovered. *Science*, 274:1849–1850.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54.

Brighton, H. (2003). *Simplicity as a driving force in linguistic evolution*. PhD thesis, The University of Edinburgh.

Brighton, H., Kirby, S., and Smith, K. (2005). 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 University Press, Oxford.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology*, 52A:273–302.

Chater, N. and Vitányi, P. M. B. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences*, 7(1):19–22.

Chater, N. and Vitányi, P. M. B. (2004). A simplicity principle for language acquistion: Re-evaluating what can be learned from positive evidence. Manuscript under review.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Chomsky, N. (2002). *On Nature and Language*. Cambridge University Press, Cambridge.

Christiansen, M. H. and Kirby, S. (2003). Language evolution: consensus and controversies. *Trends in Cognitive Sciences*, 7(7):300–307.

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

Comrie, B. (1989). *Language Universals and Linguistic Typology*. Blackwell, Oxford, second edition.

Cover, T. and Thomas, J. (1991). *Elements of information theory*. Wiley Interscience, New York.

Cowie, F. (1999). *What's within? Nativism Reconsidered*. Oxford University Press, Oxford.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14:597–612.

Croft, W. (1990). *Typology and Universals*. Cambridge University Press, Cambridge.

Culicover, P. W. (1999). *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquistion*. Oxford University Press, Oxford.

Deacon, T. W. (1997). *The Symbolic Species*. W. W. Norton and Company.

Elman, J. L. (2003). Generalization from sparse input. In *Proceedings of the* 38^{th} *Annual Meeting of the Chicago Linguistic Society*.

Glendinning, P. (1994). *Stability, Instability, and Chaos: An Introduction to the Theory of Nonlinear Differential Equations*. Cambridge University Press, Cambridge.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44:133–152.

Hare, M. and Elman, J. L. (1995). Learning and morphological change. *Cognition*, 56:61–98.

Hopcroft, J. E. and Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Hurford, J. R. (1990). Nativist and functional explanations in language acquisition. In Roca, I. M., editor, *Logical issues in language acquisition*, pages 85–136. Foris Publications.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

Kimball, J. P. (1973). *The Formal Theory of Grammar*. Prentice-Hall, Englewood Cliffs, NJ.

Kirby, S. (2001). Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation*, 5(2):102–110.

Kirby, S. (2002). Learning, bottlenecks and the evolution of recursive syntax. In Briscoe, E., editor, *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, pages 173–203. Cambridge University Press, Cambridge.

Krifka, M. (2001). Compositionality. In Wilson, R. A. and Keil, F., editors, *The MIT Encyclopaedia of the Cognitive Sciences*. MIT Press, Cambridge, MA.

Li, M. and Vitányi, P. M. B. (1997). *A Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York.

Lidz, J. and Waxman, S. (2004). Reaffirming the poverty of the stimulus argument: a reply to the replies. *Cognition*, 93:157–165.

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

Niyogi, P. and Berwick, R. (1997). Evolutionary consequences of language learning. *Linguistics and Philosophy*, 20:697–719.

O'Grady, W., Dobrovolsky, M., and Katamba, F. (1997). *Contemporary Linguistics*. Longman, 3rd edition edition.

Pinker, S. and Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, 13:707–784.

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3):472–491.

Pullum, G. K. (1996). Learnability, hyperlearning, and the poverty of the stimulus. In Johnson, J., Juge, M. L., and Moxley, J. L., editors, *Proceeding of the 22nd Annual Meeting: General session and parasession on the role of learnability in grammatical theory*, pages 498–513. Berkeley Linguistic Society, Berkeley, CA.

Pullum, G. K. and Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review*, 19(1-2):9–50.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Rissanen, J. (1989). *Stochastic complexity and statistical inquiry*. World Scientific.

Rissanen, J. (1997). Stochastic complexity in learning. *Journal of Computer and System Sciences*, 55:89–95.

Sampson, G. (1989). Language acqusition: Growth or learning? *Philosophical Papers*, 18:203–240.

Smith, K., Brighton, H., and Kirby, S. (2003a). Complex systems in language evolution: The cultural emergence of compositonal structure. *Advances in Complex Systems*, 6(4):527–558.

Smith, K., Kirby, S., and Brighton, H. (2003b). Iterated learning: a framework for the emergence of language. *Artificial Life*, 9(4):371–386.

Teal, T. K. and Taylor, C. E. (2000). Effects of compression on language evolution. *Artificial Life*, 6(2):129–143.

Wagner, L. (2001). Defending nativism in language acqusition. *Trends in Cognitive Sciences*, 6(7):283–284.

Wexler, K. (1991). On the argument from the poverty of the stimulus. In Kasher, A., editor, *The Chomskyan Turn*, pages 253–270. Blackwell, Cambridge.

Wolff, J. G. (1982). Language acquisition, data compression, and generalization. *Language and Communication*, 2(1):57–89.

Zadrozny, W. (1994). From compositional to systematic semantics. *Linguistics and Philosophy*, 17:329–342.