Compositional Syntax From Cultural Transmission

Henry Brighton
Language Evolution and Computation Research Unit
Department of Theoretical and Applied Linguistics
University of Edinburgh
Adam Fergusen Building, George Square
Edinburgh, EH8 9LL
UK
henryb@ling.ed.ac.uk

Abstract A growing body of work demonstrates that syntactic structure can evolve in populations of genetically identical agents. Traditional explanations for the emergence of syntactic structure employ an argument based on genetic evolution: Syntactic structure is specified by an innate language acquisition device (LAD). Knowledge of language is complex, yet the data available to the language learner are sparse. This incongruous situation, termed the "poverty of the stimulus," is accounted for by placing much of the specification of language in the LAD. The assumption is that the characteristic structure of language is somehow coded genetically. The effect of language evolution on the cultural substrate, in the absence of genetic change, is not addressed by this explanation. We show that the poverty of the stimulus introduces a pressure for compositional language structure when we consider language evolution resulting from iterated observational learning. We use a mathematical model to map the space of parameters that result in compositional syntax. Our hypothesis is that compositional syntax cannot be explained by understanding the LAD alone: Compositionality is an emergent property of the dynamics resulting from sparse language exposure.

Keywords
language, evolution, syntax, learning, compression, culture

1 Introduction

Children master complex features of language on the basis of surprisingly little evidence. This phenomenon, termed the poverty of the stimulus, has been taken as evidence for innate linguistic knowledge [6]. How can such a rich body of linguistic knowledge be induced from an impoverished body of linguistic evidence? The traditional explanation appeals to an innate language acquisition device (LAD) [5]. The LAD is a language-specific module encapsulating the knowledge of language that cannot be induced from primary linguistic data. This explanation attributes the bulk of the specification of language to the human biological endowment, rather than the linguistic stimulus. In short, when considering the origins of language, humans, and possibly other hominids, developed language directly as a result of the biological evolution of the LAD [17, 18]. If the syntactic structure of language can be explained in terms of the LAD alone, the fundamentals of the story of how language emerged, and why language has the structure it does, are in place.

An important linguistic dynamic is missing from this explanation: Language itself can evolve, on a cultural substrate, among genetically homogeneous language users. By investigating the degree to which this dynamic can result in language evolution, recent agent-based simulations have questioned the primacy of innatist accounts of linguistic structure. For example, Kirby [9] and Batali [1] demonstrate that recursive and compo-

sitional syntax can emerge in the absence of genetic change. These experiments raise important questions. Can we explain why language has its characteristic structure by analyzing the LAD alone? How much of the structure of language is determined by the dynamics resulting from evolution on a cultural substrate? The situation characterized by the poverty of the stimulus poses a design problem. One solution is an innate, structure-specifying LAD. Using a mathematical model, we propose an alternative to this solution. The poverty of the stimulus introduces a strong pressure for compositional structure when we take into account cultural evolution: Our hypothesis is that much of the characteristic structure of language, of which compositionality is one distinctive feature, emerges as a result of pressures on transmission. It is important to stress that to some degree, an innate LAD is required for language, but the syntactic structure of language need not be explicitly specified by this LAD.

Previous work can be characterized as establishing that compositional and recursive syntax, two of the hallmarks of language, can emerge relative to a population of genetically homogeneous language users [1, 9, 10]. We build on this work, first by mapping the space of model parameters that lead to linguistic structure, and second by introducing a model that accounts for statistical effects during the language acquisition process. Instead of aiming to demonstrate that linguistic structure can emerge in the absence of genetic change, our methodology focuses on establishing the range of conditions for emergence. Our results show that instead of characterizing the poverty of the stimulus as a constraint on transmission, overcome by the LAD, it is best conceived of as a determinant in the evolution of compositional syntax. We argue that compositional structure is best explained in terms of the dynamics of language evolution, rather than the internals of the LAD.

To model language evolution we use the *iterated learning model*, an agent-based model of cultural evolution where language change results from repeated observational learning [3, 11]. In Section 2 the details of the iterated learning model are presented. In this model, each agent has the ability to generalize from observed examples of language. The generalization bias of each agent determines how language changes from generation to generation. In Section 3 we introduce a generalization procedure based on the minimum description length principle. An abstracted form of this generalization procedure is developed in Section 4 and is used to form part of a mathematical model. This model, covered in Section 5, is used to map the parameter space. The results are discussed in Section 6.

2 The Iterated Learning Model

The iterated learning model (ILM) provides a framework for modeling the cultural evolution of language [3, 11]. Language is transmitted from one generation to the next by one agent forming utterances, and the next agent observing these utterances. This process is repeated generation after generation, and in linguistic terms characterizes the translation from language performance (language use) to language competence (language knowledge) via observational learning: Agents induce their knowledge of language through observation. The ILM captures an important characteristic of language. Language is learned by language users, and the input to the learning process is based on the output of learning itself. This is an important phenomenon. The space of possible languages is restricted to contain only those languages that can be produced by an agent. The effects of any bias on either learning or production will be amplified at each generation.

The ILM is an agent-based model where each agent represents a language user. Given the most basic population model, where each generation comprises one agent, the ILM proceeds as follows. The first generation agent, A_1, forms utterances. The

second generation agent A_2 observes these utterances and forms a hypothesis to account for them. The next generation proceeds by A_2 producing utterances that the next agent A_3 observes. This process is repeated, often for thousands of generations. How, for example, does the language of A_{233} differ to that of A_{1047}? Does the system reach a steady state? What kind of languages emerge and characterize steady states? The ILM allows us to answer questions like these. By exploring the properties of an agent, along with certain pressures on information transmission, we can try to attack the issue of how certain language structures evolve. The key idea is that language itself evolves; each agent in the model is initially identical. Agents arrive at different hypotheses only when they are exposed to different bodies of linguistic evidence.

2.1 Simplifying Assumptions

Before examining the ILM in more detail, it is worth making explicit the simplifying assumptions we have made in the construction of the model.

1. Agents have the ability to "mind read." When an agent observes a signal, the intended meaning of that signal is also given. This simplification avoids the problem of modeling the ascription of a meaning to a signal. An agent must associate a signal with a meaning somehow, but we regard this as a separate, nontrivial problem [24–26]. In short, we assume meaning transmission occurs over a noiseless channel.
2. Issues of communication are not considered. For example, communicative accuracy, the agents' intentions, or any model of success or failure in language use is not considered. How much of the structure of language is due to pressures on learning? To begin to answer this question, we must strip the model of any assumptions about the functional aspects of communication. Part of our hypothesis is that issues relating to communication are not the principle determinants of language structure.
3. Population effects are not considered. Each agent learns from the output of only one other agent. Similarly, utterances produced by an agent are only observed by a single infant agent.

From a modeling perspective, these simplifications are crucial. Ultimately we seek the minimal set of assumptions and hypotheses with which linguistic structure can be explained. By employing such a minimal model, the degree of similarity between the resulting languages and natural language will be negligible. Properties of natural language, such as parts of speech and tense, will not occur. However, by introducing additional details to the model, certain aspects of natural language can emerge. For example, the occurrence of regular/irregular forms emerge in the ILM when we introduce a pressure for small signals in conjunction with a nonuniform distribution over meanings [11].

It is important to note that although elements of language structure can evolve through iterated learning, biological evolution must form the backbone of any story of language evolution (see, for example, [7]). The structure of the agents, how they learn, and how they conceptualize communicatively relevant situations, is taken to be biologically determined.

2.2 Iterated Learning in Detail

Each agent senses an external environment that contains n objects $\{\omega_1, \omega_2, \ldots, \omega_n\}$. Each object ω_i represents some communicatively relevant situation. We do not attempt

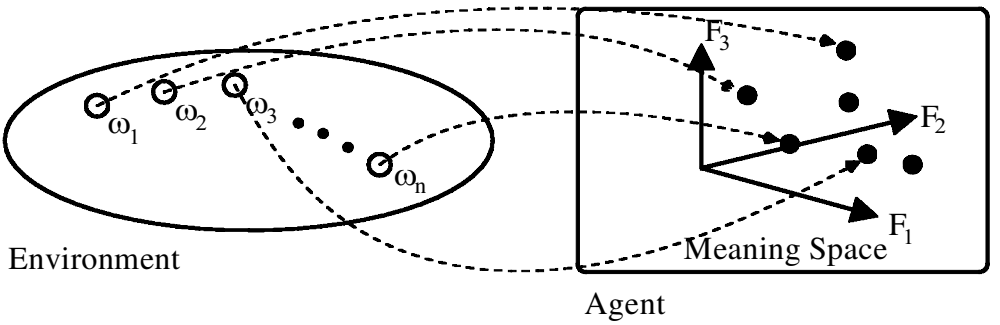


Figure 1. The relationship between the environment and an agent. The environment contains a set of n objects $\{\omega_1, \omega_2, \ldots, \omega_n\}$ that represent communicatively relevant situations. These objects are perceived by the agent in terms of a semantic space. In this example, a three-dimensional meaning space is used to represent the objects internally.

to specify the details of these situations. The important point is that objects are conceptualized in terms of a meaning space internal to the agent. To an agent, an object corresponds to a point in its meaning space. The distinction between an object and a meaning is important. In previous work, every meaning in a meaning space is an object; during a simulation, all meanings represent communicatively relevant situations [9, 11]. We introduce the object/meaning distinction as it allows different meaning space structures to be used for a given environment. For example, given an environment containing 100 objects, we can use one from any number of meaning spaces to conceptualize (label) the objects.

Figure 1 illustrates the relationship between the external environment and the internal meaning space. More formally, a meaning is a vector drawn from a space defined by two parameters: F, the number of features, and V, the cardinality of the set from which feature values are drawn. A simplifying assumption is made that each feature has the same number of possible values, V. We make this assumption as it simplifies the mathematics of the model. In short, meanings are points in an F-dimensional space, with each dimension having V discrete values. The environment remains constant over time, and we assume that different agents sense the same object in the same way: The correspondence between object and meaning is identical for different agents. The mapping from objects to meanings is random. From the point of view of the agent some objects will be indistinguishable if the same meaning is used to label more than one object.

2.2.1 Language as a Mapping Between Meanings and Signals

Agents, on the basis of their linguistic competence, utter signals that stand for meanings. A language defines the relationship between meanings and signals. More formally, a language is a mapping between meanings and signals. We define it as follows. First, signals are defined as strings of symbols drawn from some alphabet Σ. The maximum length of a signal is $l_{\max}$. Given a meaning m and signal s, a meaning/signal pair is denoted by $\langle m, s \rangle$ such that m is drawn from the meaning space $\mathcal{M}$:

$$\mathcal{M} = \{(f_1, f_2, \ldots, f_F): \ 1 \leq f_i \leq V \quad \text{and} \quad 1 \leq i \leq F\}$$

and the signal s is drawn from the signal space $\mathcal{S}$:

$$\mathcal{S} = \{w_1 w_2 \ldots w_l: \ w_i \in \Sigma \quad \text{and} \quad 1 \leq l \leq l_{\max}\}$$

A language L is defined as a set of meaning/signal, $\langle m, s \rangle$, pairs. From the space of all languages, two classes of language structure will be considered: *holistic languages* and *compositional languages*. Central to this distinction is the relationship between meaning structure and signal structure. A holistic language exhibits no structural relationship between meanings and signals: The whole signal stands for the whole meaning, rather than parts of the signal referring to parts of the meaning. A compositional language is one where a relationship does exist between meaning and signal: The meaning of a signal is a function of the meaning of its parts, and how they are assembled [15]. Holistic languages and compositional languages are constructed as follows:

- To construct a holistic language L_{holistic} given a set of meanings M, a random signal is assigned to each meaning. Each signal is of some random length l ($1 \leq l \leq l_{\text{max}}$) and is composed of symbols drawn randomly from Σ.
- To construct a compositional language L_{comp} given a set of meanings M, we use a dictionary of subsignals. Each subsignal is used to represent a feature value. For simplicity, the assumption is made that each feature value, for each feature, has a unique entry in the dictionary, that is, a subsignal refers to one and only one feature value, for one feature. For each meaning in M a signal is constructed on the basis of the dictionary: A signal is formed by concatenating the corresponding subsignal for each feature value in the meaning. For simplicity, the order of subsignals occurring in the complete signal reflects the order of feature values in the meaning. This simplification is made for language construction only and does not restrict the class of compositional languages we account for in the model.

In a compositional language similar meanings will map to similar signals. Similar meanings, by definition, must have elements in common and so their corresponding signals will also have elements in common. The mapping from meanings to signals will be neighborhood preserving: Groups of similar meanings always map to groups of similar signals. This kind of regularity does not occur, unless by chance, when constructing holistic languages.

2.2.2 Learning, Production, and Language Change

After observing some language L an agent selects a hypothesis H that best describes L. When called upon to express objects, which appear as meanings, the agent then uses the hypothesis H to find an appropriate signal for each of these meanings. The linguistic competence of the agent is defined as the ability to express signals for meanings. The set of possible hypotheses, the process by which the hypothesis is selected, and the manner in which appropriate signals are chosen for meanings are collectively termed the *generalization process*. Consider the case when the observed language L is a subset of some larger language $\hat{L}$. We say the agent observes $\hat{L}$ subject to a *transmission bottleneck*. This situation resembles what Chomsky [6] termed the poverty of the stimulus. The language L is an impoverished version of $\hat{L}$. Depending on the effectiveness of the generalization process, and the structure in L, it is possible for an agent to reproduce all the meaning/signal pairs in $\hat{L}$ after only observing L. Learners of natural language are placed in exactly this situation: Nobody learns English by observing all English sentences.

The behavior of an agent can be thought of as a function that maps the language of generation t, L_t, to the language of generation L_{t+1}:

$$f\colon\ L_t \mapsto L_{t+1}$$

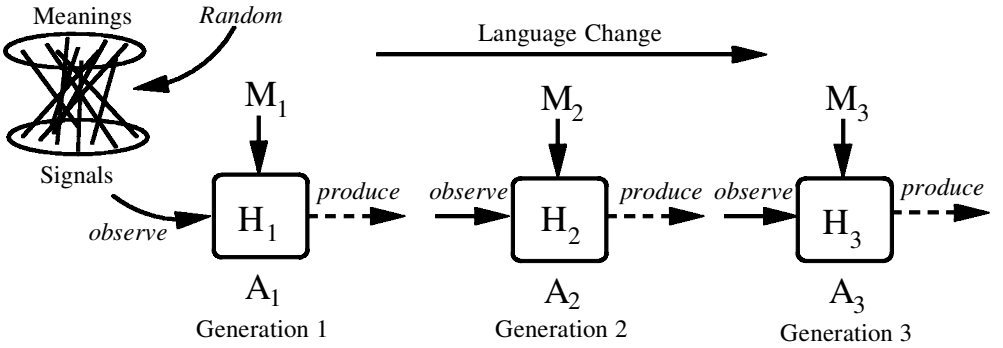


Figure 2. The first three generations of the iterated learning model. The first agent in the chain, A_1, is presented with example utterances (meaning/signal pairs) from some holistic language. A hypothesis H_1 is then chosen to account for this linguistic evidence. A_1 is given a set of meanings M_1 that correspond to objects drawn randomly from the set of objects. For each of these meanings, an appropriate signal is deduced and uttered by A_1. A_2 observes these utterances and forms a hypothesis H_2 to explain the set of meaning/signal pairs. The process is repeated: A_2 utters a signal for each of those meanings in M_2 for agent A_3 to observe. A hypothesis corresponds to an agent's linguistic knowledge. Utterances (meaning/signal pairs) represent the agents' linguistic performance.

Under certain conditions $L_t = L_{t+1}$, in which case L_t is a stable language. The iterated learning process rests on the fact that the input to an agent is the output of another agent. In linguistic terms, an infant learns language by observing the linguistic behavior of an adult. In this model, we show that language evolution will only occur when agents suffer from the poverty of the stimulus. In the iterated learning model the poverty of the stimulus occurs as a result of the transmission bottleneck: Language learners never learn from a complete exposure to the language of the previous generation.

Figure 2 depicts this process in more detail. The first agent A_1 observes a language best described by the hypothesis H_1. Objects in the environment are then presented to A_1, at random, some number of times. The agent must utter a signal for each of these objects. This process is termed production. The next agent A_2 then observes these meaning/signal pairs. This process is repeated; each agent is presented with a random series of object observations. Random meanings, along with the induced hypothesis, are then used to produce the language for the next generation to observe. The first agent in the model observes utterances drawn from a holistic language.

2.3 Language Evolution

In this model, language evolves when an agent attempts to reconstruct the language of the previous generation on the basis of sparse language exposure. Inconsistencies between the two languages are introduced because either the structure of language makes generalization impossible, or the generalization procedure is inadequate. In this article we are principally interested in steady states, rather than the transitions a language passes through before stability results. Before focusing on steady states, this section will clarify the factors that drive language change.

To arrive at and maintain a steady state requires specific conditions. In the absence of a transmission bottleneck all language structures are stable because all objects are observed in conjunction with a signal. Providing the hypothesis is consistent with the data, no uncertainty can arise when an object needs to be expressed. As soon as a learner forms a hypothesis on the basis of a subset of the whole language, instability can result; some of the objects may not have been observed during the lifetime of the agent. In this situation the agent must use knowledge of what it has observed to postulate solutions to what it has not observed. For this reason, the set of language structures that result in stability will depend on the size of the transmission bottleneck.

2.3.1 Unstable States and Language Evolution

How exactly does the language of one generation get transformed into the language of the next generation? Consider the meaning/signal pair $\langle m, s \rangle$, that occurs in some language L_n. One of three mechanisms will be responsible for the association of m with some signal s' in the next generation, L_{n+1}:

1. *Memorization.* Here, $s = s'$ as the learner uses the signal that accompanied the meaning when it was observed. The mapping between m and s does not change. Memorization can only occur when $\langle m, s \rangle$ is observed.
2. *Generalization.* In the case that the learner has never observed $\langle m, s \rangle$, but enough is known about the relationship between the observed meanings and their signals, an appropriate signal can be arrived at by induction. In this situation, either $s = s'$, and the association remains stable, or via a progressive induction decision, $s \neq s'$. The latter change is progressive in that it will reinforce any structure occurring in the mapping between meanings and signals.
3. *Invention.* The learner has not observed $\langle m, s \rangle$, and the learner cannot discern any structure in the observed language that will help in finding a signal for m. In this situation, an inventive production decision is required. Some random element must be used in the production of m. Alternatively, the learner can choose not to produce a signal for m at all. Either way, we assume $s \neq s'$, and any existing association between m and s is broken. The case when $s = s'$ will only occur by chance.

These three mechanisms define when an association between meaning and signal will change from one generation to the next. Only some combinations of these processes can maintain a stable language. The language of one generation is transformed into the language of the next generation by using one of four combinations of production mechanisms (invention/memory, invention/generalization/memory, generalization/memory, memory). Figure 3 illustrates the relationship between the production mechanisms and the occurrence of divergence. Divergence is shown in Figure 3c, d, where invention is required. Stability can only result when invention is not used; Figure 3a, b illustrates this fact.

Invention, which only occurs in the presence of a transmission bottleneck, serves an important purpose. When the language is unstructured, invention will occur more frequently. By chance, an invented signal could introduce some structure to the language of the next generation where this structure was absent in the language of the previous generation. Critically, a structured relation between meanings and signals is more likely to survive the transmission bottleneck than an unstructured relation [10]. Consider some region of the mapping between meanings and signals that is structured. Not all the parts of this structured region need to be observed for the structured relation to survive the transmission bottleneck. Contrast this with an unstructured (random) region. For this region to be represented in the language of the next generation requires that all the meaning/signal pairs representing that region be observed. Structure is compressible.

It is these stochastic inventions, which introduce structure where it was previously absent, that drive the language evolution toward regions of stability. The more structure the mapping contains, the less frequently invention will occur. Rather like simulated annealing [12], iterated learning can be seen as a search strategy. Providing the language is unstructured and a transmission bottleneck is in place, the initial temperature will be high; invention will occur frequently. In time the temperature decreases, and invention will occur less frequently. The search follows a trajectory toward language that best fits the combined biases of hypothesis selection and production.

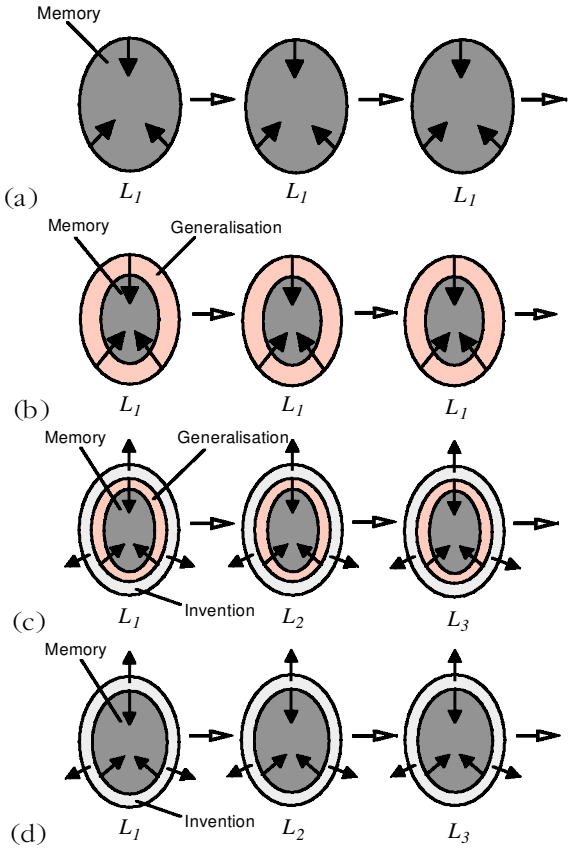


Figure 3. The four ways in which a language L_1 can evolve. These diagrams show how the mechanism used to produce a signal can effect a change in the language from generation to generation. Inward arrows signify the reinforcement of the language structure. Outward arrows indicate divergence from the current language structure. First, a language can only persist when one of two combinations of production mechanism is in play: (a) memory, and (b) memory and generalization. Divergence will occur when either (c) memory, generalization, and invention occur, or (d) when memory and invention occur. Invention always causes divergence.

So, iterated learning can effect a cumulative evolution toward structure. This process is the linguistic equivalent of what Tomasello [28] terms the *ratchet effect* in cultural evolution. However, only when certain conditions are met will iterated learning reliably converge to a stable state. We aim to understand and formalize these conditions.

2.4 The Parameter Space

Six parameter groups define the behavior of the iterated learning model. In the context of the model developed here, they are as follows:

1. *The meaning space.* The space of possible meanings. The meaning space is defined by the number of features, F, and the cardinality of the set from which feature values are drawn, V.
2. *The signal space.* Signals are constructed by drawing symbols from some alphabet Σ. The maximum length of a signal is $l_{\max}$.

3. *The transmission bottleneck.* The number of random object observations, R. The value of R is related to the degree of language exposure. Some probability distribution over objects specifies how likely an object is to be observed. We assume this distribution is uniform, unless otherwise stated.
4. *The perceptual bias.* Utterances may not be perceived accurately. For example, the transmission channel might be noisy, or there may be limits on the working memory of an agent, thereby restricting the set of perceivable utterances [8]. In the model developed here, we assume all utterances are transmitted noise free; no restrictions on perception are modeled.
5. *The learning bias.* The learning bias defines the space of possible hypotheses, and which hypothesis is chosen given some data.
6. *The production bias.* Given a meaning and a hypothesis, the production bias defines which signal is chosen to express the meaning.

Stable language only occurs for certain parameter combinations. Recall that without a transmission bottleneck, all languages are stable. As soon as a transmission bottleneck is in place, only certain languages are stable, and these rely on certain combinations of learning and production bias. Before mapping this parameter space, we consider the role of learning and production.

3 Compression, Learning, and Generalization

Agents act as a conduit for language. An agent observes a subset of the language of the previous generation. The process of learning from this subset, and then producing utterances for the next generation, is a complex one. At one level an agent simply maps one language onto another. At a more detailed level, the function that defines this mapping is composed of a learning mechanism and a production mechanism. Learning, in this context, is the process of arriving at a hypothesis that explains the observed language. Production is the process that, given a hypothesis, completes the mapping from meanings to signals. The production mechanism defines how the hypothesis is interrogated to yield signals. Because the chosen hypothesis might reflect some regular structure existing in the observed language, unobserved regions of the language could be recovered by the production mechanism exploiting the structure in the hypothesis. The combination of learning and production is termed generalization. In this section we propose a candidate computational model of generalization based on the minimum description length (MDL) principle.

3.1 Minimum Description Length Learning

Ranking potential hypotheses by minimum description length is a principled and elegant approach to hypothesis selection [13, 21]. The MDL principle can be derived from Bayes's rule, and in short states that the best hypothesis for some observed data is the one that minimizes the sum of (a) the encoding length of the hypothesis, and (b) the encoding length of the data, when represented in terms of the hypothesis. A trade-off then exists between small hypotheses with a large data encoding length and large hypotheses with a small data encoding length. When the observed data contains no regularity, the best hypothesis is one that represents the data verbatim, as this minimizes the data encoding length. However, when regularity does exist in the data, a smaller hypothesis is possible that describes the regularity, making it explicit, and as result the hypothesis describes more than just the observed data. For this reason, the cost of encoding the data increases. MDL tells us the ideal trade-off between the length of

the hypothesis encoding and the length of the data encoding described relative to the hypothesis. More formally, given some observed data D and a hypothesis space $\mathcal{H}$ the best hypothesis h_{MDL} is defined as

$$h_{\text{MDL}} = \min_{h \in \mathcal{H}} \{L_{C_1}(h) + L_{C_2}(D|h)\} \quad (1)$$

where $L_{C_1}(h)$ is the length in bits of the hypothesis h when using an optimal coding scheme over hypotheses. Similarly, $L_{C_2}(D|h)$ is the length, in bits, of the encoding of the observed data *using* the hypothesis h. We use the MDL principle to find the most likely hypothesis for an observed set of meaning/signal pairs passed to an agent. When regularity exists in the observed language, the hypothesis will capture this regularity, when justified, and allow for generalization beyond what was observed. By employing MDL we have a theoretically solid justification for generalization. The next section will clarify the MDL principle—we introduce the hypothesis space and coding schemes.

3.1.1 The Hypothesis Space

We introduce a novel model for mapping strings of symbols to meanings, which we term a finite state unification transducer (FSUT). This model extends the scheme used by Teal and Taylor [27] to include variable length signals and, more importantly, meanings. Given some observed data, the hypothesis space consists of all FSUTs that are consistent with the observed data. Both compositional and noncompositional languages can be represented using the FSUT model.

A FSUT is specified by a 7-tuple $(Q, \Sigma, F, V, \delta, q_0, q_F)$ where Q is the set of states used by the transducer, and Σ is the alphabet from which symbols are drawn. F and V define the structure of the meaning space. The transition function δ maps state/symbol pairs to a new state, along with the (possibly underspecified) meaning corresponding to that part of the transducer. Two states, q_0 and q_F need to be specified; they are the initial and final state, respectively. Consider an agent A that receives a set of meaning/signal pairs during language acquisition. For example, an observed language might be the set

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

This language is compositional. It was constructed using the following dictionary:

	Value 1	Value 2
Feature 1	a	b
Feature 2	c	d
Feature 3	e	f

So, for example, the subsignal corresponding to feature value 2 for the first feature is "b". Figure 4a depicts a FSUT that models L. We term this transducer the *prefix tree transducer*—the observed language and only the observed language is represented by the prefix tree transducer. The power of the FSUT model only becomes apparent when we consider possible generalizations made by merging states and edges:

1. *State merge.* Two states q_1 and q_2 can be merged to form a new state if the transducer remains consistent. All edges that mention q_1 or q_2 now mention the new state.

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

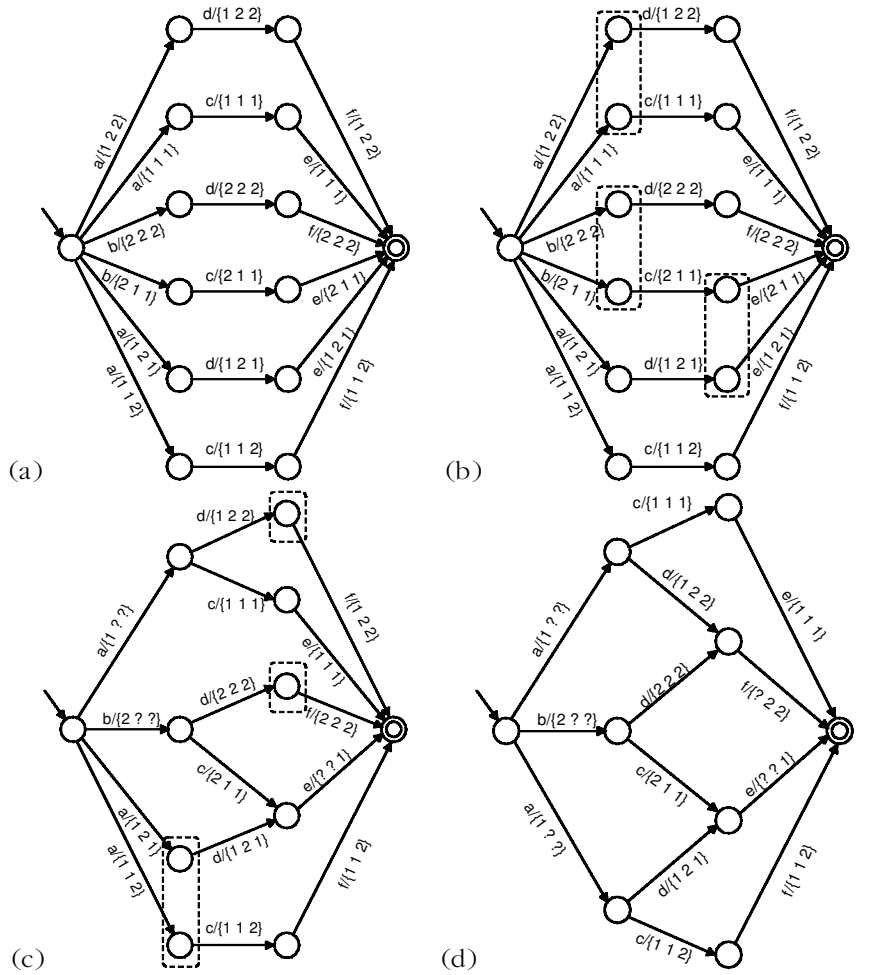
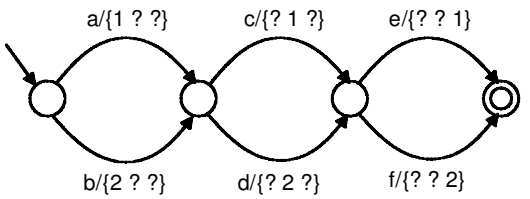


Figure 4. Given the compositional language *L*, the prefix tree transducer shown in (a) is constructed. By performing edge and state merge operations, outlined in (b) and (c), the transducer can be compressed. The transducer shown in (d) is compressed but does not lead to any generalizations.

2. *Edge merge.* Two edges e_1 and e_2 can be merged if they share the same source and target states and accept the same symbol. The result of merging the two edges is a new edge with a new meaning label. Meanings are merged by finding the intersection of the two component meanings. Those features that do not have values in common take the value "?"—a wild card that matches all values. As fragments of the meanings may be lost, a check for transducer consistency is also required. Without this consistency check, some observed meaning/signal pairs will not be accounted for by the resulting transducer.

Figure 4b and c illustrates some possible state and edge merge operations. The transducer resulting from these merge operations is shown in in Figure 4d. Figure 5 depicts

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



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

Figure 5. Given the compositional language L a series of state and edge merge operations, beginning with those shown in Figure 4, result in the compressed transducer shown here. As a result of compression, the transducer can express more meanings than those contained in L. All members of language L^+ can be expressed.

a fully compressed transducer, which is found by performing additional state and edge merge operations. The fully compressed transducer can express meanings that are not present in L. The language L^+, shown in Figure 5, contains all the meaning/signal pairs that can be expressed by the fully compressed transducer. By compressing the prefix tree transducer, the structure in the compositional language has been made explicit, and as result, generalization can occur.

3.1.2 Encoding Lengths

To apply the MDL principle we need an appropriate coding scheme for (a) the hypotheses, and (b) the data using the given hypothesis. These schemes correspond to C_1 and C_2 introduced in Equation 1. The requirement for the coding scheme C_1 is that some machine can take the encoding of the hypothesis and decode it in such a way that a unique transducer results. Similarly, the coding of the data with respect to the transducer must describe the data uniquely. To encode a transducer $T = (Q, \Sigma, F, V, \delta, q_0, q_F)$ containing n states and e edges we must calculate the space required, in bits, of encoding a state ($S_{\text{state}} = \log_2(n)$), a symbol ($S_{\text{symbol}} = \log_2(|\Sigma|)$), and a feature value ($S_{\text{fvalue}} = \log_2(V)$). The number of bits required to encode a meaning relies not only on S_{fvalue}, but also the cost of encoding the wild-card specifier. The number of bits used to encode an arbitrary meaning $m = \{f_1, \ldots, f_F\}$ is given by

$$S_{\text{meaning}}(m) = \sum_{i=1}^{F} Z(f_i)$$

where f_i denotes the value of the ith feature, and

$$Z(f_i) = \begin{cases} 1 & : \text{ when} f_i = ? \\ 1 + S_{\text{fvalue}} & : \text{ otherwise} \end{cases}$$

That is, $Z(f_i)$ represents the number of bits required to encode either a feature value or the wild-card specifier. The initial bit is used to differentiate between these two

possibilities. Denoting the meaning associated with the ith edge by m_i, the encoding length of the transducer is then

$$S_{\mathrm{T}} = \sum_{i=1}^{e} \{2S_{\mathrm{state}} + S_{\mathrm{symbol}} + S_{\mathrm{meaning}}(m_i)\} + S_{\mathrm{state}}$$

which corresponds to encoding the transition function δ along with the identity of the accepting state. For the transducer T to be uniquely decoded, we must also specify the lengths of constituent parts of the transducer. We term this part of the encoding the *prefix block*:

$$S_{\mathrm{prefix}} = S_{\mathrm{state}} + 1 + S_{\mathrm{symbol}} + 1 + S_{\mathrm{fvalue}} + 1 + S_{\mathrm{F}} + 1$$

Where S_{F} is the encoding length, in bits, required to define the number of features in a meaning: $S_{\mathrm{F}} = \log_2(F)$. To calculate $L_{C_1}(h)$ we then use the expression:

$$L_{C_1}(h) = S_{\mathrm{prefix}} + S_{\mathrm{T}} \tag{2}$$

Recall that $L_{C_1}(h)$ defines the length of the encoding of the hypothesis h using the coding scheme C_1. This quantity is termed the grammar encoding length (GEL) [27]. Similarly, the length of the encoding of the data, in terms of the hypothesis h, $L_{C_2}(D|h)$, is termed the data encoding length (DEL). The DEL is far simpler to calculate than the GEL. For some string s composed of symbols $w_1 w_2 \ldots w_{|s|}$ we need to detail the transition we choose after accepting each symbol with respect to the given transducer. The list of choices made describes a unique path through the transducer. Additional information is required when the transducer enters an accepting state as the transducer could either accept the string or continue parsing characters, as the accepting state might contain a loop transition. Given some data D composed of p meaning/signal pairs, $L_{C_2}(D|h)$ is calculated by

$$L_{C_2}(D|h) = \sum_{i=1}^{p} \sum_{j=1}^{|s_i|} \{\log_2 z_{ij} + F(s_{ij})\} \tag{3}$$

where s_i is the signal of the ith meaning/signal pair, and z_{ij} is the number of outward transitions from the state reached after parsing j symbols of the signal of the ith meaning/signal pair. The state reached after parsing j symbols of the signal of the ith meaning/signal pair is denoted by s_{ij}. The function F handles the extra information for accepting states:

$$F(s_{ij}) = \begin{cases} 1 & : \text{ when the transducer is in } q_{\mathrm{F}} \\ 0 & : \text{ otherwise} \end{cases}$$

Prefix tree transducers are compressed by applying the merge operators described above. We use a beam search [14]. The merge operators are chosen at random and applied to a random point in the transducer. The resulting transducer must be consistent with the observed data to enter the beam. Transducers with the smallest encoding lengths are most likely to remain in the beam, as new transducers, formed by applying the merge operations, replace the transducer with the largest encoding length. Smaller transducers are therefore more likely to be used in further exploration of the hypothesis space. When no operator can be applied, the search stops, and the transducer with the smallest encoding length is chosen.

3.2 MDL, Hypothesis Selection, and Generalization

With respect to some observed data, MDL provides a ranking over candidate hypotheses. Identifying the hypothesis that yields the smallest encoding length is a practical problem that can be addressed by employing a search strategy. But the selection of the hypothesis solves only half the problem of mapping a meaning to a signal. To complete the mapping a production mechanism is required that interrogates the hypothesis. With respect to our implementation of MDL, a procedure is required that takes a FSUT and a meaning and produces a signal for that meaning. One such mechanism proceeds by performing a depth-first search for a path through the FSUT such that no transition is taken that has a meaning label that is inconsistent with the target meaning. Providing the set of transitions result in the correct meaning being parsed, and the final transition leads to the accepting state, the resulting signal is formed by concatenating the symbols found on each transition. Using this procedure, generalization can occur. This model of generalization was used in an agent-based simulation reported by Brighton and Kirby [3].

Here, we only consider two FSUT structures: prefix tree machines and compressed machines. These machines correspond to the hypotheses chosen when the observed language is either holistic or compositional, respectively. The FSUT model defines a space of hypotheses. The application of the MDL principle over this space of hypotheses, in conjunction with the iterated learning model, can account for the evolution of all gradations of language type: from holistic language to fully compositional language, and mixtures of both structures. In the model developed here, we simplify the issue and only consider stability conditions for holistic and fully compositional language.

4 Optimal Generalization

Assume an agent knows, before observing any utterances, that the observed language will have compositional structure. Recall that a compositional language is totally defined by the dictionary used to construct it. Given the expectation of a compositional language, what degree of exposure to the language is required before the dictionary can be derived? The earliest point at which the dictionary could be constructed is when all feature values have been observed. Disregarding the details of the procedure for reconstructing the dictionary, this is the minimum degree of exposure before reconstruction is possible at all. We term the ability to express all the meanings for which all the feature values have been observed the *optimal generalization bias*.

In this section we aim to formalize the notion of the optimal generalization bias. We show that this degree of bias is paralleled in the MDL model of generalization discussed above. In short, we aim to show that given a compositional language, in all but a few circumstances, the compressed transducer outlined in the previous section will have the smallest encoding length. Relating the optimal generalization bias to the MDL model allows us to model language stability without the need to perform lengthy agent-based simulations.

The strongest possible generalization bias is one that results in some fixed compositional language L_C being expressed, whatever the input. But any such scheme would be inconsistent with all observed languages other than L_C. We consider only hypotheses that are consistent with the observed data. Compositional languages, for our purposes, are those where the feature values appearing in the meaning are associated with unique subsignals. A dictionary relating every feature value to a subsignal totally defines the compositional language. This is a slight simplification of the problem as the dictionary does not define the order in which subsignals are assembled. We assume the ordering of the feature values in the meaning is reflected in the construction of the signal. To express a meaning, the subsignal corresponding to each feature value in the meaning

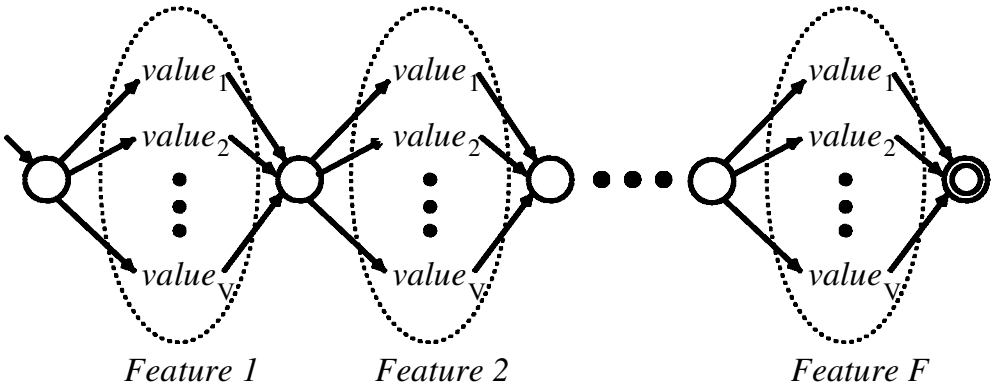


Figure 6. The general structure of a compressed transducer. Each feature is represented by a section of the transducer. Within a section, individual feature values are represented by a unique path.

is located in the dictionary. The signal is formed by concatenating these subsignals. Now, if not enough evidence to build the whole dictionary is observed, expressivity will be suboptimal. Some objects that need to be expressed will be represented by meanings that contain unobserved feature values: The entry in the dictionary will be missing. The optimal generalization bias is the ability to express all meanings that are built from observed feature values.

Definition 1 (Optimal generalization): *Given a compositional language L_C, and a learner with the optimal generalization bias, a meaning $m \in L_C$:*

$$m = (v_1, v_2, \ldots, v_F)$$

can be expressed providing each value v_i has been observed at least once. Given that some v_i has been observed, the optimal generalization bias makes the assumption that the subsignal corresponding to v_i can always be deduced.

The optimal generalization bias serves as an upper bound on the degree of inductive bias over compositional language.

4.1 MDL and the Optimal Generalization Bias

Imagine that the dictionary used to construct a compositional language can also act as a hypothesis for that language. Under which circumstances will the MDL hypothesis selection result in such in hypothesis? The compressed FSUT illustrated in Figure 5 corresponds to the notion of a dictionary. Each feature used to construct the meaning space is represented as a separate region in the transducer. Within these regions, each feature value is also represented. Meanings constructed from all combinations of the feature values are expressible given such a transducer structure. Figure 6 depicts the general structure of a compressed transducer.

For a compositional language, under what circumstances will MDL choose a compressed transducer? We show that given certain assumptions about the distribution of meanings, compressed machines are always chosen by MDL, given a compositional language as input. Bear in mind this is a statement concerning the ranking of hypotheses with respect to the the FSUT model and the MDL principle. How one performs the search through the space of hypotheses to verify this fact is a practical issue. We assume an exhaustive search. Our claim is that applying the MDL principle over the

space of FSUTs results in a hypothesis equivalent to that characterized by the optimal generalization bias. To demonstrate this fact, first we present an analytic argument, then we illustrate the argument using data extracted from a simulation.

4.1.1 The Relationship Between DEL and GEL

The compressed machine yields the smallest grammar encoding length possible. Smaller machines do exist, but they will not be consistent with the observed data. We therefore rule them out. If hypotheses were selected solely on the basis of their size, then compressed machines would always be chosen for compositional language, and our analytic demonstration would be complete. This policy is known as Occam's razor (see, for example, [14]). However, as we are using the MDL principle, we must also consider the size of the data represented in terms of the hypothesis. The following analysis demonstrates that under two assumptions, a uniform distribution over meanings and the presence of a transmission bottleneck, the impact of the data encoding length is negligible. As a result MDL hypothesis selection can be simplified to the policy of Occam's razor: Always pick the smallest hypothesis. The hypothesis chosen by MDL will only differ from that picked using Occam's razor when nonuniform statistical effects are present in the data.

A series of meaning/signal pairs are presented to an agent. This series can be represented as a set of p distinct meaning/signal pairs. The number of times some arbitrary meaning/signal pair i is observed is denoted as $K(i)$. It is useful to specify the size, or severity, of the transmission bottleneck in terms of the expected number of distinct objects observed, rather than the number of object observations. The expected object coverage, denoted as c, after R random observations of N objects is defined as

$$c = \frac{\log_2(1 - R)}{\log_2(1 - \frac{1}{N})} \tag{4}$$

The value of c represents the proportion of the objects that we expect to observe.

For a compositional language, we know that $p \leq V^F$ as V^F is the maximum number of unique meaning/signal pairs. Consider using a prefix tree transducer to encode this language: It will require $\log_2(p)$ bits to encode a single meaning/signal pair, because, in a prefix tree transducer, each meaning/signal pair is represented by a unique nonbranching path through the transducer. In total, for some compositional language L,

$$\mathrm{DEL}_{\mathrm{prefix}}(L) = \sum_{i=1}^{p} \{K(i) \cdot \log_2(p)\}$$

bits are required to encode all the meaning/signal pairs in L. Now consider how many bits are required to encode a single meaning/signal pair, given that a compressed transducer is used. Assuming all the feature values have been observed, this will be

$$\log_2(V^F)$$

We arrive at this expression as follows. Each feature in the meaning is represented by a different section of the transducer. A path through one of these sections can take one of V possible routes. Specifying a single feature value therefore requires $\log_2(V)$ bits. As there are F features, to encode a whole meaning/signal pair requires

$$F \cdot \log_2(V) = \log_2(V^F)$$

bits. In total a compressed machine, given a compositional language L, will require

$$\mathrm{DEL}_{\mathrm{comp}}(L) = \sum_{i=1}^{p} \{K(i) \cdot \log_2(V^F)\}$$

bits to encode the p meaning/signal pairs. The difference between $\mathrm{DEL}_{\mathrm{comp}}(L)$ and $\mathrm{DEL}_{\mathrm{prefix}}(L)$ is important. Only when this difference is greater than the size difference between the grammar encoding lengths will prefix tree machines be chosen. More formally, the encoding length of a prefix tree transducer, $\mathrm{EL}_{\mathrm{prefix}}$, is defined as:

$$\mathrm{EL}_{\mathrm{prefix}}(L) = \mathrm{GEL}_{\mathrm{prefix}}(L) + \mathrm{DEL}_{\mathrm{prefix}}(L)$$

and similarly for a compressed transducer

$$\mathrm{EL}_{\mathrm{comp}}(L) = \mathrm{GEL}_{\mathrm{comp}}(L) + \mathrm{DEL}_{\mathrm{comp}}(L)$$

Now, only when

$$\mathrm{EL}_{\mathrm{prefix}}(L) < \mathrm{EL}_{\mathrm{comp}}(L) \tag{5}$$

will prefix tree transducers be selected over a compressed transducer, given a compositional language as input. This situation requires that the difference in the grammar encoding lengths is *less* than the difference in the data encoding lengths. This situation is discussed below.

4.1.2 A Nonuniform Distribution Over Meanings

The difference between $\mathrm{DEL}_{\mathrm{prefix}}(L)$ and $\mathrm{DEL}_{\mathrm{comp}}(L)$ is usually much smaller than the difference between $\mathrm{GEL}_{\mathrm{prefix}}(L)$ and $\mathrm{GEL}_{\mathrm{comp}}(L)$. This is because transducer compression results in the removal of many transducer states, but the degree to which this loss of states increases the cost of encoding the data is usually small. The upshot of this disparity is that, given a compositional language, for a prefix tree transducer to be preferred over a compressed transducer, the number of occurrences of each meaning/signal pair, $K(i)$, for $1 \leq i \leq p$, must be large.

Assuming a uniform distribution over meanings, this situation will not occur unless each meaning/signal pair is observed many times. For this to happen $c \approx 1.0$, that is, the whole language is observed. This situation is not modeled here—we assume there is a transmission bottleneck. In terms of MDL, the justification for this relationship rests on the assumption that the more evidence we have for a set of observations, the less likely novel observations are to occur.

Figure 7 illustrates how, for a compositional language, hypothesis selection depends on the probability distribution over meanings (and hence objects). First, given a transmission bottleneck ($c < 1.0$) and a uniform distribution over meanings, compressed machines are always chosen as $\mathrm{EL}_{\mathrm{prefix}} - \mathrm{EL}_{\mathrm{comp}} > 0$ holds irrespective of coverage. Contrast this with a situation where we fix $K(i) = 100$ for $1 \leq i \leq p$, also shown in Figure 7. For low coverage values, the inequality $\mathrm{EL}_{\mathrm{prefix}} - \mathrm{EL}_{\mathrm{comp}} > 0$ no longer holds—prefix tree transducers are preferred for low coverage values. In practice, given a uniform distribution over meanings, this situation cannot occur unless $c \approx 1.0$.

However, if we consider a distribution such as that resulting from Zipf's law, then this situation can occur when $c < 1.0$. A Zipfian distribution, in this context, would mean that the frequencies of meanings decay as a power function of their rank [11, 30]. In this case, as c increases, $K(i)$, for some values of i, will become huge. In this situation,

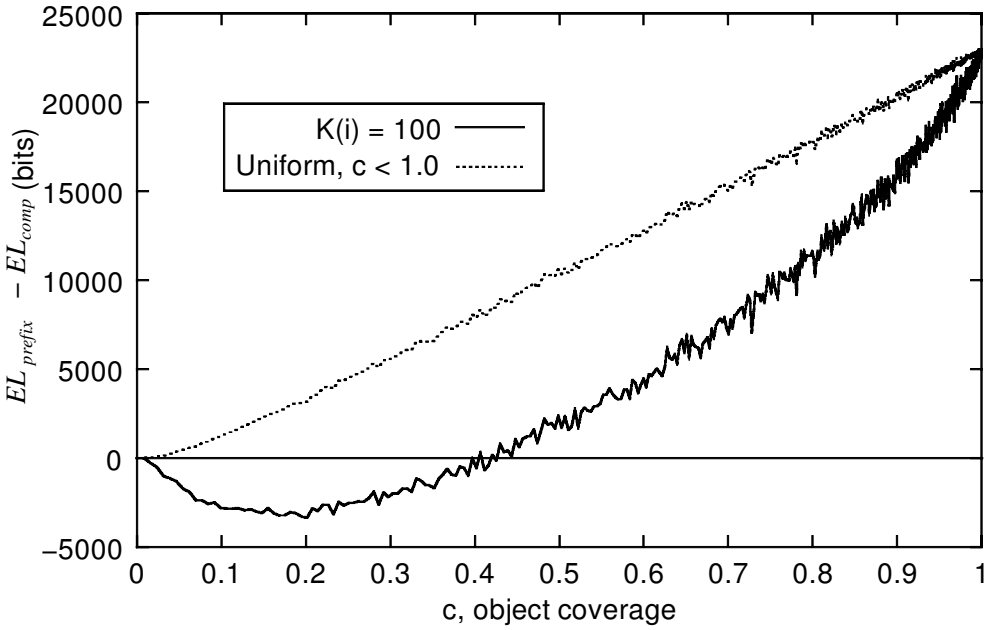


Figure 7. When $EL_{prefix} - EL_{comp} < 0$, prefix transducers are chosen by MDL for compositional language. With a uniform distribution and a transmission bottleneck, this does not occur. Only when, for example, we fix $K(i) = 100$ will prefix tree transducers be chosen over compressed machines.

the rule that compressed transducers are always selected for compositional language no longer holds. By introducing strong statistical effects into the data, hypothesis selection by MDL begins to diverge from that of Occam's razor. The occurrence of communicatively relevant situations are unlikely to conform to a uniform distribution [4]. The assumption of a uniform distribution made in the model analysis therefore restricts the structure of languages we consider. The introduction of a nonuniform distribution would be an important extension to the model: We are currently investigating Zipfian distributions.

4.2 Summary

Short of knowing the language in advance, the optimal generalization bias describes the least amount of evidence required before generalization can occur. We have demonstrated that, assuming a uniform distribution over meanings, compressed transducers are preferable to prefix tree transducers for all compositional languages. As compressed transducers are a model of optimal generalization, we reduce the problem of generalization over compositional language to that of optimal generalization. In the next section, a mathematical model is developed that relates the stability advantage conferred by compositional language to the parameters of the iterated learning model. The model relies on the notion of the optimal generalization bias.

5 Modeling the Conditions for Stability

Understanding the conditions for stability requires an analysis of the interaction between generalization bias, the transmission bottleneck, and the structure of the meaning space. The following analysis employs a mathematical model, based on the agent-based simulation reported by Brighton and Kirby [3]. Modeling by simulation has disadvantages

when the parameter space is large. Parts of the parameter space that were intractable to map using the Monte Carlo simulations outlined in previous work are no longer intractable. Mathematical models of cultural evolution have been proposed before [2], but mathematical models of language evolution typically focus on arguments appealing to natural selection [16, 17].

The model developed here estimates the stability advantage offered by compositional language. By varying the severity of the transmission bottleneck and the degree of structure in the meaning space, the conditions for which compositional language offers the greatest stability advantage can be established. We conjecture that it is precisely these regions of the parameter space from which compositional language is most likely to emerge. In contrasting one language structure with another, we are pitting one type of hypothesis, prefix tree transducers in the case of holistic language, against another, compressed transducers in the case of compositional language. The model estimates the likelihood of a language type emerging on the basis of the expressivity of the corresponding hypothesis. For a given language type, the appropriate hypothesis structure was found using the MDL principle.

5.1 The Expressivity of a Prefix Tree Transducer

Using a prefix tree transducer, only meanings that have been observed in conjunction with a signal can be expressed. Given a holistic language, a prefix tree transducer is the best we can do: There is no principled way of expressing a novel meaning on the basis of previously observed meanings. To calculate the expected expressivity of an agent using a prefix tree transducer we need to calculate the probability of observing some arbitrary meaning m drawn from the meaning space. For a meaning m to be observed by an agent it must first be used to label an object, as there is no guarantee that m will be used at all (recall Figure 1). Second, the object(s) labeled with m need to be observed at least once by the agent. The likelihood of observing a meaning/signal pair (representing an object) depends on the severity of the transmission bottleneck.

After observing a series of objects, each of which is represented by a meaning, the agent will have accumulated a set of observed meanings. We denote this set as O. The probability of observing some arbitrary meaning m, that is, $\Pr(m \in O)$, is determined by the number of objects in the environment (N), the number of meanings in the meaning space (M, where $M = V^F$), and the number of random object observations during an agent's lifetime (R). The probability of observing the meaning m is defined as

$$\Pr(m \in O) = \sum_{x=0}^{N} \left\{ \frac{x}{N} \cdot \left(\sum_{r=1}^{R} \left(\frac{N-x}{N} \right)^{r-1} \right) \cdot \left(\frac{(M-1)^{N-x}}{M^N} \right) \cdot \binom{N}{x} \right\} \tag{6}$$

This expression takes into account that first m must be used to label an object, and second, that after R observations of the objects, m is observed at least once. The Appendix provides an explanation for Equation 6. To estimate the number of distinct meanings observed, and therefore the number of distinct meanings that can be expressed, we simply multiply this probability by the number of possible meanings, V^F. The expressivity of an agent using a prefix tree transducer as a hypothesis is denoted as E_{prefix}:

$$E_{\text{prefix}} = \Pr(m \in O) \cdot V^F \tag{7}$$

To summarize, we first calculate the probability of the meaning being used to label an object *and* that object being observed. This probability, multiplied by the number of possible meanings, yields the expected number of meanings observed by an agent. As

no generalization can occur, the expressivity of the agent using a prefix tree transducer is equivalent to the number of meanings observed.

5.2 The Expressivity of a Compressed Transducer

Given a compositional language the compressed transducer structure maximizes the expressivity of an agent: Expressivity becomes a function of the number of feature values observed, rather than a function of the number of meanings observed. Recall that given the optimal generalization bias, to express a meaning m requires that all the feature values in m have been observed. The rate at which feature values are observed is at least the rate at which whole meanings are observed. As a result, the expressivity achieved by a compressed transducer is always at least the degree of expressivity achieved when using a prefix tree transducer.

To estimate the expressivity of an agent using optimal generalization we need to calculate the probability of observing some arbitrary feature value. Recall that this probability will be greater than the probability of observing some arbitrary meaning, because fewer entities need to be seen, given the same number of observations. More formally, for some arbitrary feature f_i $(1 \leq i \leq F)$, f_i can take on a value drawn from the set of possible values $\{1, 2, \ldots, V\}$. We require the probability that some arbitrary feature value drawn from this set is observed after R object observations.

As before, this is a two-stage process. First, we must take into account that the arbitrary feature value must be used in at least one of the meanings chosen to label the objects. The feature value may be used once or more, or not at all. Given that for some arbitrary feature there are V possible values, N objects are labeled, and R random samples of these objects are taken, we can use an expression similar to Equation 6 to estimate the probability of observing some arbitrary feature value v one or more times. After R observations, for the arbitrary feature f_i, the set O_{f_i} of feature values is observed. We denote the probability of observing an arbitrary feature value as $\Pr(v \in O_{f_i})$ and define it as

$$\Pr(v \in O_{f_i}) = \sum_{x=0}^{N} \left\{ \frac{x}{N} \cdot \left(\sum_{r=1}^{R} \left(\frac{N-x}{N} \right)^{r-1} \right) \cdot \left(\frac{(V-1)^{N-x}}{V^N} \right) \cdot \binom{N}{x} \right\} \tag{8}$$

Now, for some meaning $m = (v_1, v_2, \ldots, v_F)$ the ability to express m requires that each feature value v_i has been observed. We can therefore express the probability of expressing m as the combined probability of expressing each of the feature values of m. We denote this probability as $\Pr(v_1 \in O_{f_i} \wedge \cdots \wedge v_F \in O_{f_i})$ and define it as

$$\Pr(v_1 \in O_{f_i} \wedge \cdots \wedge v_F \in O_{f_i}) = \Pr(v \in O_{v_i})^F \tag{9}$$

This is the probability of being able to express some arbitrary meaning m. Contrast this probability with that represented in Equation 6. For some meaning m, Equation 9 tells us the probability of being able to express m. Equation 6 tells us the probability of *observing* an arbitrary m. This is an important distinction, as to estimate the expressivity of an agent using a compressed transducer we need to multiply the probability of expressing a meaning m with the expected number of expressible meanings. The expected number of distinct meanings used when labeling N objects is

$$N_{\text{used}} = \left(1 - \left(1 - \frac{1}{V^F} \right)^N \right) \cdot V^F \tag{10}$$

To find the expected number of objects for which a signal can be derived we then multiply the probability of expressing a meaning by the number of expressible meanings:

$$E_{\text{generalization}} = \Pr(v \in O_{f_i})^F \cdot N_{\text{used}} \tag{11}$$

Compressed transducers are derived from prefix tree transducers by means of state and edge merging operations. Compression can only increase expressivity because consistency with the observed data is always maintained. The minimum expressivity for a compressed transducer is therefore E_{prefix}. We can now define the expressivity of a compressed transducer as

$$E_{\text{compressed}} = \max\{E_{\text{generalization}}, E_{\text{prefix}}\} \tag{12}$$

5.3 Language Stability

The above analysis relates transducer structure to expressivity. Given a compositional language, hypothesis selection using the MDL principle results in a compressed transducer. The expressivity in this case is defined by the expression for $E_{\text{compressed}}$. A holistic language results in a prefix tree transducer being chosen. The expressivity in this case is defined by E_{prefix}.

Next, we relate expressivity to stability. In the ILM, language stability is the degree to which the hypotheses induced by subsequent agents maintain the mapping between meanings and signals. The entire set of associations between meanings and signals may not be externalized as utterances. They may exist by virtue of the hypothesis alone. The stability of a language is therefore related to the proportion of the objects that can be expressed without resorting to invention. Invention either introduces or maintains unstructured areas of the mapping between meanings and signals. Stability is therefore related to expressivity. The notion of stability used here corresponds to the probability of a language being transmitted from one generation to the next without change. The stability value of a language is the proportion of the objects that can be expressed through generalization. We can now characterize the degree of stability for a language type. The degree of stability of a compositional language, $S_{\text{compositional}}$, and of a holistic language S_{holistic} can be defined as

$$S_{\text{compositional}} \propto \frac{E_{\text{compressed}}}{N} \tag{13}$$

$$S_{\text{holistic}} \propto \frac{E_{\text{prefix}}}{N} \tag{14}$$

An important measure we employ is that of *relative stability*. We denote relative stability as S and define it as

$$S = \frac{S_{\text{compositional}}}{S_{\text{compositional}} + S_{\text{holistic}}} \tag{15}$$

For a coverage value and a meaning space structure (defined by c, F, and V, respectively), the S value tells us the degree to which compositional language is more stable than holistic language. S reflects how much larger $S_{\text{compositional}}$ is than S_{holistic}, $(0.0 \leq S \leq 1)$. When $S > 0.5$ compositional language is more stable than holistic language. The case when $S < 0.5$, which does not occur in these experiments, corresponds to the situation when holistic language is more stable than compositional language. There is no stability advantage to either language type when $S = 0.5$.

5.4 Model Analysis

The key measurement used in the model is the relative stability, S, of compositional language over holistic language. The value of S is dependent on four variables:

1. The number of features used to construct the meaning space, F.
2. The number of values used by each feature in the meaning space, V.
3. The bottleneck size, represented as the expected object coverage, c (see Equation 4).
4. The number of objects in the environment, N.

The value of S indicates more than just the relative stability of compositional language over holistic language; S also reflects the likelihood of compositional language emerging. When, as a result of invention, compositional structure is introduced, the relative stability of compositional language tells us how likely this compositional structure is to persist (recall Section 2.3.1). Figures 8 and 9 show the relationship, for different degrees of coverage, between S and the structure of the meaning space. The parameter N is not of any real consequence, as changing it will just shift the landscape away from the origin: Larger meaning spaces are required to represent more objects. In contrast, the model reveals that several important relationships exist between F, V, c, and S:

1. S is at a maximum for small bottleneck sizes. In Figure 8a S nears the maximum value of 1 for certain meaning space structures. This result demonstrates that compositional language can still be learned even though exposure is limited. In contrast, holistic language will not persist over generations when only a small subset is observed. This result is important, as it demonstrates that the poverty of the stimulus is an important determinant of compositional language.
2. High S values only occur once a certain degree complexity in the meaning space has been reached. This means that the conceptual space of the agent must be broken up into multiple features and values for compositionality to become an option. For meaning spaces with only a few features, S cannot reach a high value. In short, there must be a certain degree of feature structure before compositional language can become advantageous.
3. Most clearly illustrated in Figure 8a, the largest meaning spaces, those that contain more than approximately 2 million meanings, lead to small values of S. In all the examples (Figures 8 and 9), a very high degree of structure leads to low S values. Beyond a certain point, the more highly structured the meaning space, the less likely different meanings are to share feature values. Consider the most extreme case where each object will be labeled with a meaning containing feature values not present in labeling of any other object.
4. When labeling N objects there is a trade-off between the number of features used and the number of values per feature used. The more features used, the fewer the number of feature values required to represent N objects. When a meaning is observed, F feature values are observed. This means that all feature values will be observed sooner if the N objects are conceptualized more in terms of features than feature values. This is a relationship most strikingly illustrated by the surface shown in Figure 8a. Here, compositional language is far more stable when many features are used.

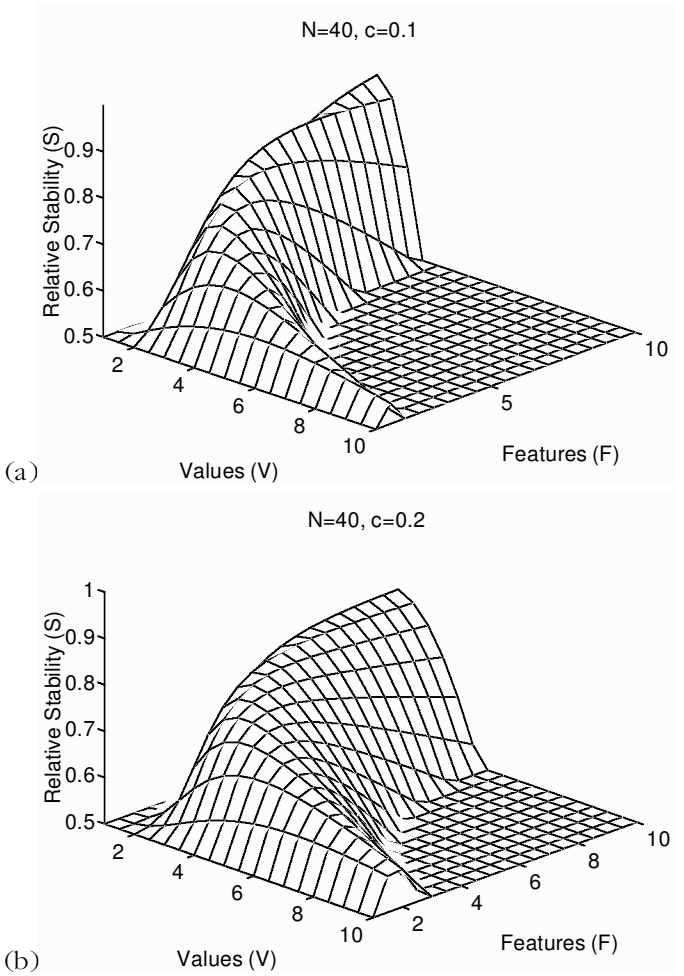


Figure 8. The relationship between meaning space structure, low coverage, and relative stability. In these examples, the number of objects, *N*, is 40. The two surfaces demonstrate the relation between meaning space structure and *S* for low coverage values: (a) $c = 0.1$, (b) $c = 0.2$. The highest *S* values ($S \approx 1.0$) occur for low coverage ($c \approx 0.1$) and medium complexity, illustrated in (a). The number of features plays a more significant role than the number of values per feature.

Before relating these results to language and its evolution, it is worth considering the interactions between F, V, c, and S in more detail.

Figure 8a would indicate that a meaning space $\mathcal{M}_1$ with $F = 10$ and $V = 2$ results in higher relative stability than a meaning space $\mathcal{M}_2$ with $F = 2$ and $V = 10$. Why is the number of features so important? First, there is an order of magnitude more meanings in $\mathcal{M}_2$ that in $\mathcal{M}_1$. From looking at the surfaces shown in Figures 8 and 9, it is not easy to discern the size difference between the meaning spaces. The number of meanings found in $\mathcal{M}_1$ fits the other parameters well for high S; the number of meanings found in $\mathcal{M}_2$, on the other hand, is large enough that the likelihood of objects sharing feature values is relatively low.

Recall that F feature values are observed when a single object is observed. As more objects are observed, the likelihood of full feature value coverage increases rapidly

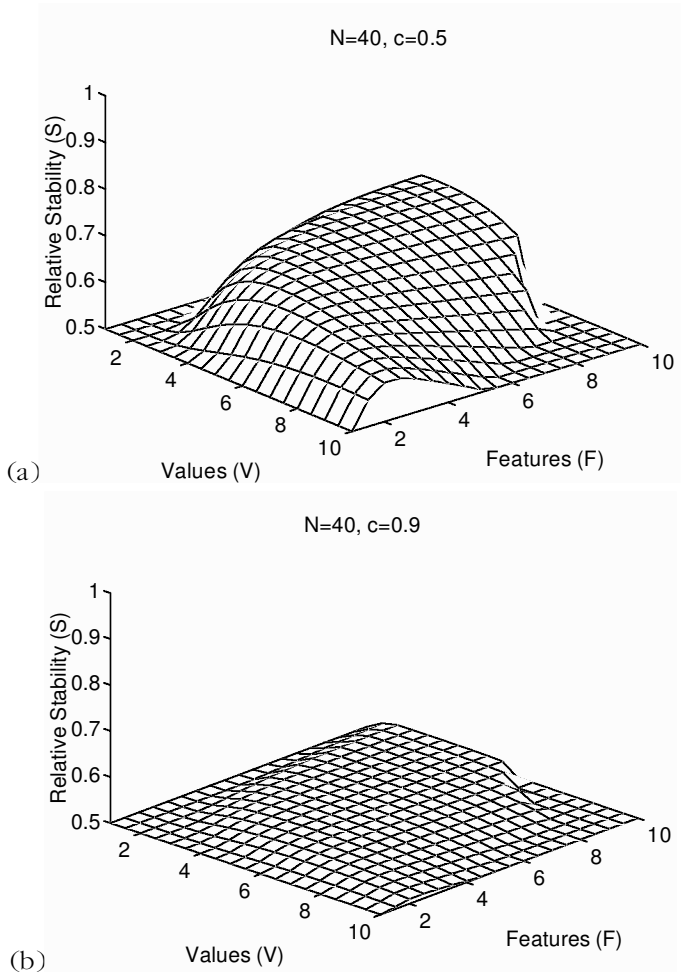


Figure 9. The relationship between meaning space structure, mid and high coverage, and relative stability. Each surface demonstrates the relation between meaning space structure and S for mid and high coverage values: (a) $c = 0.5$, and (b) $c = 0.9$. The maximum S value decreases rapidly when coverage is increased. The size of the region in which $S > 0.5$ grows as the coverage increases. This indicates that for smaller values of S, more meaning space structures lead to an advantage for compositionality.

in comparison to the degree of object coverage. This discrepancy in feature value coverage and object coverage is greatest for small c (bottleneck) values. This is why compositional language is relatively more stable in comparison to holistic language in this region of the parameter space. Expressivity is a function of object coverage for holistic language. Expressivity is a function of feature value coverage for compositional language. Figure 10 illustrates this relationship: The expressivity achieved, represented as a proportion of the objects, reaches a maximum for much lower coverage values when a compressed transducer is chosen.

As the degree of object coverage increases the set of meaning spaces for which $S > 0.5$ increases. The higher the coverage the greater the number of objects observed. This relationship allows larger meaning spaces to be used, as enough observations are occurring such that previously rare co-occurrences of feature values now become more likely.

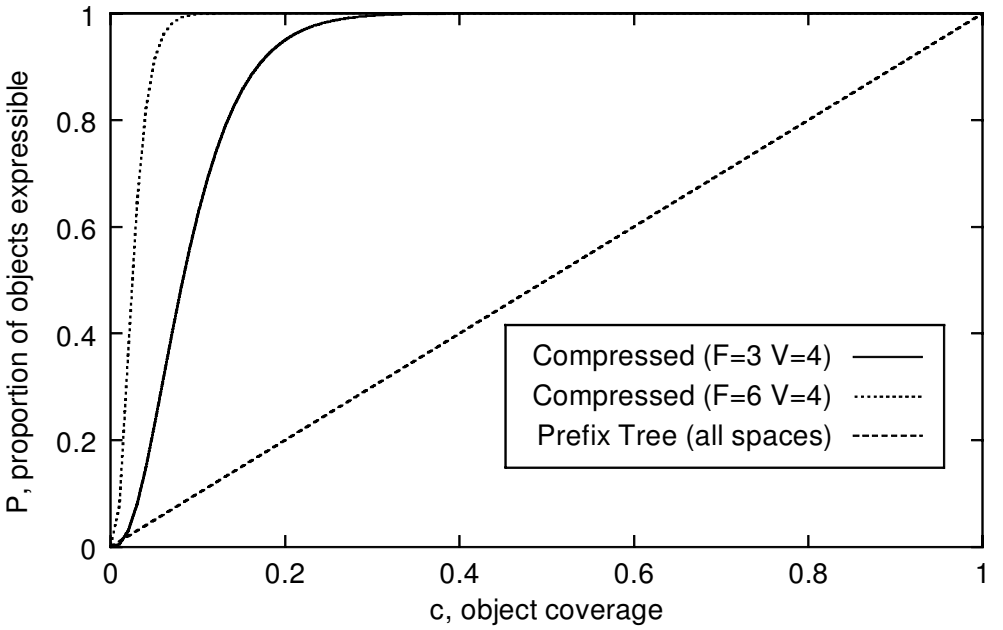


Figure 10. The rate at which expressivity increases as a function of object coverage. For compressed transducers the increase in expressivity is much faster than that found with prefix tree transducers. The proportion of the objects which can be expressed, P, is plotted against the degree of coverage of the object space, c. Compressed transducers quickly reach high expressivity. Prefix tree transducers can only express those objects they have observed, hence the linear relationship between c and E. The value of E is shown for two different meaning spaces ($F = 3$, $V = 4$ and $F = 6$, $V = 4$). The more features used, the fewer the number of observations required before all feature values are observed.

5.5 Summary

By pairing compositional language with the compressed transducer structure, and holistic language with the prefix tree transducer structure, language stability can be reduced to the issue of transducer expressivity. The model does exactly this. The relative stability measure reflects the degree of stability advantage conferred by compositional language. Using the model, the parameter combinations that yield high S were found to occupy the part of the parameter space characterized by a combination of low object coverage and high, but not outrageously high, meaning space complexity.

6 Discussion

How much of the characteristic structure of language is explicitly coded in the LAD? If we neglect the pressures of language evolution on the cultural substrate, and take seriously the claim of the poverty of the stimulus, the LAD must explicitly code much of language structure. Rather than neglecting the role of language evolution on the cultural substrate, we treat it as the fundamental determinant of language evolution. Our hypothesis is that compositional structure is a function of the dynamics of cultural evolution.

In the model presented here, agents have the innate ability to produce compositional language. One could criticize this model, and other models of the cultural evolution of language, on the grounds that compositional structure is built in: The emergence of structure is not surprising. Any evolutionary model operates within a space of possible solutions: a space of genomes in the case of genetic evolution, and a space of languages in the case presented here. The possibility of a design does not imply its occurrence.

What makes the case of language evolution so problematic is the interaction between the two evolutionary substrates. When examining this interaction, any explanation for language structure must appeal, to some degree, to a genetic component. It is far from clear how much of the structure of language is specified innately: The "argument from the poverty of the stimulus" is a prima facie explanation for linguistic nativism [19]. Consider the other extreme, where we entertain the possibility that the innate basis for language is not entirely language specific. For example, a bias toward the compression of representations is clearly not language specific. A domain-specific evolutionary argument is not required to account for all uses of learning. In short, the work presented here should be seen as exploring the possibility of a weak form of linguistic nativism: The interaction between any innate basis for language and a general ability to learn is complex, and certainly not well understood.

We argue that compositional structure is an emergent property of iterated observational learning. Several parameters control the behavior of the iterated learning model. Importantly, the role of learning has been framed as a search for the hypothesis with the smallest encoding length, according to the MDL principle. This approach to learning is motivated by the need for organisms to compress observed data [13, 20, 22], specifically, linguistic data [29]. Using a mathematical model of language stability, in conjunction with the view of learning as compression, we have mapped a large part of the parameter space of the iterated learning model.

The most important parameter is the severity of the transmission bottleneck, denoted by c in the model. This parameter controls the degree of exposure an agent has to the language of the previous generation. Without a transmission bottleneck—when c approaches the maximum value of 1—all languages are equally stable. Introducing a bottleneck, by decreasing the value of c, leads to the set of stable languages being restricted to contain only those that are learnable from limited exposure. These are compositional languages. We draw a parallel between the presence of a transmission bottleneck and the situation of the poverty of the stimulus. The poverty of the stimulus, when we take into account iterated observational learning, results in compositional language structure constituting a steady state.

The structure of the meaning space, defined by F and V, is another determinant of compositional structure in our model. There is a trade-off. Low structural complexity in the meaning space means that compositionality offers little advantage over holistic language. This occurs when few feature values are used, and objects are discriminated primarily in terms of feature values. With low structural complexity, the components (feature values) of the meanings co-occur too infrequently. The biological evolution of semantic complexity, which is assumed in our model, has been proposed by Schoenemann [23] as a necessary determinant in the emergence of syntax. The model supports this conclusion. Furthermore, we found that too much complexity in the meaning space is counterproductive. There is a limit to the stability payoff gained from increased structural complexity, as the feature values used in constructing the meanings will co-occur too infrequently for generalization to function.

The findings presented here strengthen the compelling argument that iterated learning, the process of information transmission via observational learning, is a candidate explanatory mechanism for the emergence of syntactic structure. We have taken the foundational work of Kirby [10] and Batali [1], which establishes the ability to evolve structured language, and have built on this work by identifying the key requisite conditions. We focused on the transmission bottleneck, the most salient model parameter, and drew a parallel with the the poverty of the stimulus. The poverty of the stimulus is traditionally characterized as a problem, overcome by innately specified compositional syntax, but we argue it is in fact a fundamental determinant of the emergence of compositional syntax on a cultural substrate.

Acknowledgments

The author would like to thank Simon Kirby, T. Mark Ellison, Caroline Round, Kenny Smith, and all the members of the LEC research unit. The input provided by the anonymous reviewers is greatly appreciated.

References

1. Batali, J. (in press). The negotiation and acquisition of recursive communication systems as a result of competition among exemplars. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge, UK: Cambridge University Press.
2. Boyd, R., & Richerson, P. (1985). *Culture and the evolutionary process*. Chicago: University of Chicago Press.
3. Brighton, H., & Kirby, S. (2001). The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In J. Kelemen & P. Sosík (Eds.), *Advances in artificial life* (Vol. 1704 of *Lecture Notes in Artificial Intelligence*, pp. 592–601). New York: Springer.
4. Bullock, S., & Todd, P. M. (1999). Made to measure: Ecological rationality in structured environments. *Minds and Machines, 9*, 497–541.
5. Chomsky, N. (1976). *Reflections on language*. London: Temple Smith.
6. Chomsky, N. (1980). *Rules and representations*. London: Basil Blackwell.
7. Dunbar, R. I. M. (1993). Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*, 681–735.
8. Kirby, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford: Oxford University Press.
9. Kirby, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, & J. R. Hurford (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form* (pp. 303–323). Cambridge, UK: Cambridge University Press.
10. Kirby, S. (2001). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models*. Cambridge, UK: Cambridge University Press.
11. Kirby, S. (2001). Spontaneous evolution of linguistic structure: An iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5*.
12. Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*, 671–680.
13. Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. New York: Springer.
14. Mitchell, T. M. (1997). *Machine learning*. New York: McGraw-Hill.
15. Montague, R. (1974). *Formal philosophy: Selected papers of Richard Montague*. New Haven, CT: Yale University Press.
16. Nowak, M. A., Komarova, N. L., & Niyogi, P. (2001). Evolution of universal grammar. *Science, 291*, 114–118.
17. Nowak, M. A., Plotkin, J. B., & Jansen, V. A. A. (2000). The evolution of syntactic communication. *Nature, 404*, 495–498.
18. Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*, 707–784.
19. Pullum, G. K., & Scholz, B. C. (in press). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19*(1–2).

20. Reznikova, Z. I., & Ryabko, B. Y. (1986). Analysis of the language of ants by information-theoretical methods. *Problems of Information Transmission, 22*, 245–249.

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

22. Ryabko, B., & Reznikova, Z. (1996). Using Shannon entropy and Kolmogorov complexity to study the communicative system and cognitive capacities in ants. *Complexity, 2*, 37–42.

23. Schoenemann, P. T. (1999). Syntax as an emergent property of the evolution of semantic complexity. *Minds and Machines, 9*, 309–346.

24. Smith, A. D. M. (2001). Establishing communication systems without explicit meaning transmission. In J. Kelemen & P. Sosík (Eds.), *Advances in artificial life* (Vol. 1704 of *Lecture notes in Artificial Intelligence*, pp. 381–390). New York: Springer.

25. Steels, L. (1997). Constructing and sharing perceptual distinctions. In M. van Someren & G. Widmer (Eds.), *Proceedings of the European Conference on Machine Learning*. Berlin: Springer.

26. Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence, 103*, 133–156.

27. Teal, T. K., & Taylor, C. E. (2000). Effects of acquisition on language evolution. *Artificial Life, 6*, 129–143.

28. Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge, MA: Harvard University Press.

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

30. Zipf, G. K. (1936). *The psycho-biology of language*. London: Routledge.

Appendix: Objects, Meanings, Feature Values, and the Transmission Bottleneck

Part of the model developed in Section 5 requires a calculation that estimates the likelihood of observing an entity. The calculation appears in two guises: first, where the entities are meanings, and second, where entities are feature values. We gloss the problem in terms of the first interpretation:

> Given N objects and M meanings, a random meaning is assigned to each object. After R random object observations, what is the probability of observing some arbitrary meaning?

Abstracting from the details of the model, the problem can be generalized as follows:

> Given N balls and M colors, first, all the balls are assigned a color. For each ball, a color is chosen at random, with every color being equi-probable. Second, balls are sampled at random with replacement, R times. The question is then, for some particular N, M, and R, what is the probability of observing some arbitrary color at least once?

First, given M colors, we assign a random color to each of the N balls. What is the probability that a single ball is colored A?

$$\Pr(\text{colored } A) = \frac{1}{M}$$

The probability that the ball is not colored A is then

$$\Pr(\text{not colored } A) = \frac{M-1}{M}$$

We denote the event that p balls are colored A as $\mathcal{X}(p)$, where A is an arbitrary color. Now, after allocating colors to all N balls we can say

$$\Pr(\mathcal{X}(0)) = \left(\frac{M-1}{M}\right)^N \binom{N}{0}$$

$$\Pr(\mathcal{X}(1)) = \frac{1}{M}\left(\frac{M-1}{M}\right)^{N-1} \binom{N}{1}$$

In general,

$$\begin{aligned}\Pr(\mathcal{X}(x)) &= \left(\frac{1}{M}\right)^x \left(\frac{M-1}{M}\right)^{N-x} \binom{N}{x} \\ &= \left(\frac{M-1}{M^N}\right)^{N-x} \binom{N}{x} \end{aligned} \tag{16}$$

Now we consider sampling R balls at random, with replacement. First, we observe that

$$\Pr(\text{1st ball sampled has color } A \mid N \text{ balls colored } A) = \frac{X}{N}$$

which means that

$$\Pr(\text{1st ball sampled is not colored } A \mid N \text{ balls colored } A) = \frac{N-X}{N}$$

Let the event that the nth ball sampled has color C, given that x balls are colored C, be denoted by the term $\text{color}(n) = C$. To simplify matters, we also write $\text{color}(n_1, n_2, \ldots) = C$ to denote the event that each of $n_1, n_2, \ldots$ balls sampled are colored C. We are interested in the event that *at least* one ball sampled has color C and denote this event as $\mathcal{O}$. We can represent $\Pr(\mathcal{O}|\mathcal{X}(x))$ as follows:

$$\begin{aligned}\Pr(\mathcal{O}|\mathcal{X}(x)) = {} & \Pr(\text{color}(1) = C) \\ & + \Pr(\text{color}(2) = C \mid \text{color}(1) \neq C) \\ & + \Pr(\text{color}(3) = C \mid \text{color}(1, 2) \neq C) \\ & + \Pr(\text{color}(4) = C \mid \text{color}(1, 2, 3) \neq C) \\ & + \cdots \\ & + \Pr(\text{color}(R) = C \mid \text{color}(1, 2, \ldots, (R-1)) \neq C)\end{aligned}$$

More formally, in terms of x, N, and R,

$$\Pr(\mathcal{O}|\mathcal{X}(x)) = \frac{x}{N} + \frac{x}{N}\left(\frac{N-x}{N}\right) + \frac{x}{N}\left(\frac{N-x}{N}\right)^2 + \cdots + \frac{x}{N}\left(\frac{N-x}{N}\right)^{(R-1)}$$

$$= \frac{x}{N} \cdot \sum_{r=1}^{R} \left(\frac{N-x}{N} \right)^{(r-1)}$$

We can express $\Pr(\mathcal{O})$ in the following manner:

$$\begin{aligned} \Pr(\mathcal{O}) = {} & \Pr(\mathcal{O}|\mathcal{X}(0)) \cdot \Pr(\mathcal{X}(0)) \\ & + \Pr(\mathcal{O}|\mathcal{X}(1)) \cdot \Pr(\mathcal{X}(1)) + \cdots + \Pr(\mathcal{O}|\mathcal{X}(N)) \cdot \Pr(\mathcal{X}(N)) \end{aligned}$$

And then, using Equation 16, we can express $\Pr(\mathcal{O})$ as follows:

$$\begin{aligned} \Pr(\mathcal{O}) &= \sum_{x=0}^{N} \Pr(\mathcal{O}|\mathcal{X}(x)) \cdot \Pr(\mathcal{X}(x)) \\ &= \sum_{x=0}^{N} \left\{ \frac{x}{N} \cdot \left(\sum_{r=1}^{R} \left(\frac{N-x}{N} \right)^{r-1} \right) \cdot \left(\frac{(M-1)^{N-x}}{M^N} \right) \cdot \binom{N}{x} \right\} \end{aligned} \tag{17}$$

To summarize, given N balls, and for some arbitrary color A, the probability of observing at least one ball colored A after R observations is given by Equation 17.