Learning biases for the evolution of linguistic structure: an associative network model

Kenny Smith

Language Evolution and Computation Research Unit,
School of Philosophy, Psychology and Language Sciences,
The University of Edinburgh,
Adam Ferguson Building, 40 George Square, Edinburgh EH8 9LL
kenny@ling.ed.ac.uk

Abstract. Structural hallmarks of language can be explained in terms of adaptation, by language, to pressures arising during its cultural transmission. Here I present a model which explains the compositional structure of language as an adaptation in response to pressures arising from the poverty of the stimulus available to language learners and the biases of language learners themselves.

1 Introduction

The goal of evolutionary linguistics is to explain the origins and development of human language — how did language come to be structured as it is? Recent research, much of which uses A-Life techniques to develop its argument, attempts to answer this question by appealing to cultural evolution (see [4] for review). Language is culturally transmitted to the extent that language learners acquire their linguistic competence on the basis of the observed linguistic behaviour of others. A key contribution of those working within the cultural framework is to show that the cultural transmission of language leads to an adaptive dynamic — the adaptation, by language itself, to pressures acting during its cultural transmission. This cultural evolution can lead to the emergence of at least some of the characteristic structure of language.

This process of cultural adaptation must be dependent on some biological endowment. What is not clear is what form this endowment takes, or to what extent it is language-specific. In this paper I will present a series of experiments, using a computational model of the cultural transmission of language, which allow us to refine our understanding of the necessary biological basis for a particular structural characteristic of language, compositionality. In this model a learner's biological endowment consists of a particular way of learning, with an associated learning bias.

2 Elements of the Model

I will present an Iterated Learning Model (ILM) which allows us to investigate the role of stimulus poverty and learning bias in the evolution of compositional language. The ILM is based around a simple treatment of languages as a mapping between meanings and signals (see Section 2.1). Linguistic agents are modelled using associative networks

(see Section 2.2). These agents are slotted into a minimal population model to yield the ILM. For the purposes of this paper, we will consider an ILM with the simplest possible population dynamic[1]— the population consists of a set of discrete generations, with each generation consisting of a single agent. The agent at generation n produces some observable behaviour (in this model a set of meaning-signal pairs), which is then learned from by the agent at generation $n + 1$.

2.1 Compositionality and a Model of Languages

Compositionality relates semantic structure to signal structure — in a compositional system the meaning of an utterance is dependent on the meaning of its parts. For example, the utterance "John walked" consists of two words, a noun ("John") and a verb ("walked"), which further consists of a stem ("walk-") and a suffix (-"ed"). The meaning of the utterance as a whole is dependent on the meaning of these individual parts. In contrast, in a non-compositional or *holistic* system the signal as a whole stands for the meaning as a whole. For example, the meaning of the English idiom "bought the farm" (meaning died) is not a function of the meaning of its parts.

The simplest way to capture this is to treat a language as a mapping between a space of meanings and a space of signals. In a compositional language, this mapping will be neighbourhood-preserving. Neighbouring meanings will share structure, and this shared structure will result in shared signal structure — neighbouring meanings in the meaning space will map to neighbouring signals in signal space. Holistic mappings are not neighbourhood-preserving — since the signal associated with a meaning does not depend on the structure of that meaning, shared structure in meaning space will not map to shared signal structure, unless by chance.

For the purposes of this model, meanings are treated as vectors and signals are strings of characters. Meanings are vectors in some F-dimensional space, where each dimension takes V possible values. F and V therefore define a meaning space $\mathcal{M}$.[2] The world, which provides communicatively relevant situations for agents in the model, consists of a set of objects, where each object is labelled with a meaning drawn from the meaning space $\mathcal{M}$.[3] Signals are strings of characters of length 1 to l_{max}, where characters are drawn from the character alphabet Σ.[4] l_{max} and Σ therefore define a signal space $\mathcal{S}$.

Given these representations of meanings and signals, we can now define a measure of compositionality. This measure is designed to capture the notion given above, that compositional languages are neighbourhood-preserving mappings between meanings and signals, and is based on a measure introduced in [1]. Compositionality (c) is

[1] More complex population dynamics have consequences for the cultural evolution of communication systems, as discussed in Smith & Hurford (this volume).

[2] The structure of this meaning space has been shown to have consequences for the cultural evolution of compositional structure [2]. However, I will not vary this parameter. All results reported here are for the case where $F = 3$ and $V = 5$.

[3] All results presented here are for the case where the world contains 31 objects, each object is labelled with a distinct meaning, and those meanings are drawn from a hypercube subspace of the space of possible meanings.

[4] For the results reported here, $l_{max} = 3$ and $\Sigma = \{a, b, c, d, e, f, g, h, i, j\}$.

based on the meaning-signal pairs that an agent produces, and is the Pearson's Product-Moment correlation coefficient of the pairwise distances between all the meanings and the pairwise distances between their corresponding signals.[5] $c = 1$ for a perfectly compositional system and $c \approx 0$ for a holistic system.

2.2 A Model of a Linguistic Agent

We now require a model of a linguistic agent capable of manipulating such systems of meaning-signal mappings. I will describe an associative network model of a linguistic agent. This model is based upon a simpler model of a linguistic agent, used to investigate the cultural evolution of vocabulary systems [6]. The main advantage of this model is that it allows the biases of language learners to be manipulated and investigated. For full details of the network model, the reader is referred to [5].[6]

Representation Agents are modelled using networks consisting of two sets of nodes $\mathcal{N}_M$ and $\mathcal{N}_S$ and a set of bidirectional connections $\mathcal{W}$ connecting every node in $\mathcal{N}_M$ with every node in $\mathcal{N}_S$. Nodes in $\mathcal{N}_M$ represent meanings and partial specifications of meanings, while nodes in $\mathcal{N}_S$ represent partial and complete specifications of signals.

As summarised above, each meaning is a vector in F-dimensional space where each dimension has V values. *Components* of a meaning are (possibly partially specified) vectors, with each feature of the component either having the same value as the given meaning, or a wildcard. Similarly, components of a signal of length l are (possibly partially specified) strings of length l. Each node in $\mathcal{N}_M$ represents a component of a meaning, and there is a single node in $\mathcal{N}_M$ for each component of every possible meaning. Similarly, each node in $\mathcal{N}_S$ represents a component of a signal and there is a single node in $\mathcal{N}_S$ for each component of every possible signal.

Learning During a learning event, a learner observes a meaning-signal pair $\langle m, s \rangle$. The activations of the nodes corresponding to all possible components of m and all possible components of s are set to 1. The activations of all other nodes are set to 0. The weights of the connections in $\mathcal{W}$ are adjusted according to some weight-update rule. In Section 4 this weight-update procedure will be a parameter of variation. However, initially, we will consider the rule

$$\Delta W_{xy} = \begin{cases} +1 \text{ if } a_x = a_y = 1 \\ -1 \text{ if } a_x \neq a_y \\ \ 0 \ \text{ if } a_x = a_y = 0 \end{cases} \quad (1)$$

where $W_{xy} \in \mathcal{W}$ gives the weight of the connection between nodes x and y and a_x gives the activation of node x. The learning procedure is illustrated in Fig. 1 (a).

[5] Distance in the meaning space is measured using Hamming distance. Distance in the signal space is measured using Levenstein (string edit) distance.

[6] Available for download at http://www.ling.ed.ac.uk/∼kenny/publications.html

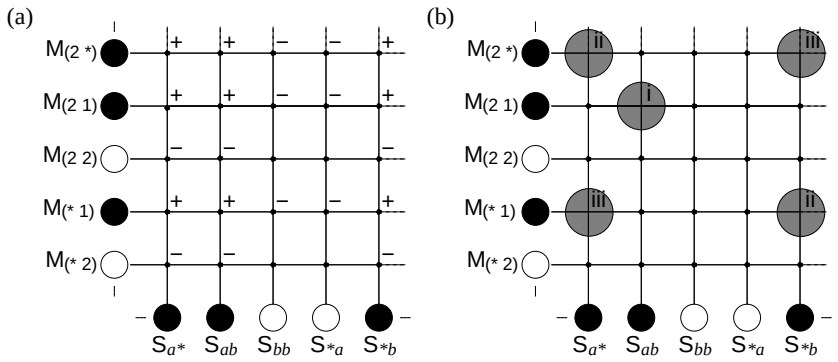

Fig. 1. (a) Storage of the meaning-signal pair $\langle (2\ 1)\ , ab \rangle$. Nodes are represented by large circles and are labelled with the component they represent. For example, $M_{(2\ *)}$ is the node which represents the meaning component $(2\ *)$, where $*$ is an unspecified feature value. Nodes with an activation of 1 are represented by large filled circles. Small filled circles represent weighted connections. During the learning process, nodes representing components of $(2\ 1)$ and ab have their activations set to 1. Connection weights are then either incremented $(+)$, decremented $(-)$ or left unchanged. (b) Retrieval of three possible analyses of $\langle (2\ 1)\ , ab \rangle$. The relevant connection weights are highlighted in grey. The weight for the one-component analysis $\langle \{(2\ 1)\}\ , \{ab\} \rangle$ depends on the weight of the connection between the nodes representing the components $(2\ 1)$ and ab, marked as i. The weight for the two-component analysis $\langle \{(2\ *)\ , (*\ 1)\}\ , \{a*, *b\} \rangle$ depends on the weighted sum of two connections, marked as ii. The weight of the alternative two-component analysis $\langle \{(2\ *)\ , (*\ 1)\}\ , \{*b, a*\} \rangle$ is given by the weighted sum of the two connections marked iii.

Production An *analysis* of a meaning or signal is an ordered set of components which fully specifies that meaning or signal. During the process of producing utterances, agents are prompted with a meaning and required to produce a meaning-signal pair. Production proceeds via a winner-take-all process. In order to produce a signal based on a given meaning $m_i \in \mathcal{M}$, every possible signal $s_j \in \mathcal{S}$ is evaluated with respect to m_i. For each of these possible meaning-signal pairs $\langle m_i, s_j \rangle$, every possible analysis of m_i is evaluated with respect to every possible analysis of s_j. The evaluation of a meaning analysis-signal analysis pair depends on the weighted sum of the connections between the relevant nodes. The meaning-signal pair which yields the analysis pair with the highest weighted sum is returned as the network's production for the given meaning. The production process is illustrated in Fig. 1 (b).

3 A Familiar Result

I will begin by replicating a familiar result: the emergence of compositional structure through cultural processes depends on the presence of a *transmission bottleneck* [2, 3]. Recall that a learner in the model acquires their linguistic competence on the basis of a set of observed meaning-signal pairs. That set of meaning-signal pairs is drawn from the linguistic behaviour of some other individual, which is a consequence of that individual's linguistic competence. I will investigate two possible conditions. In the *no*

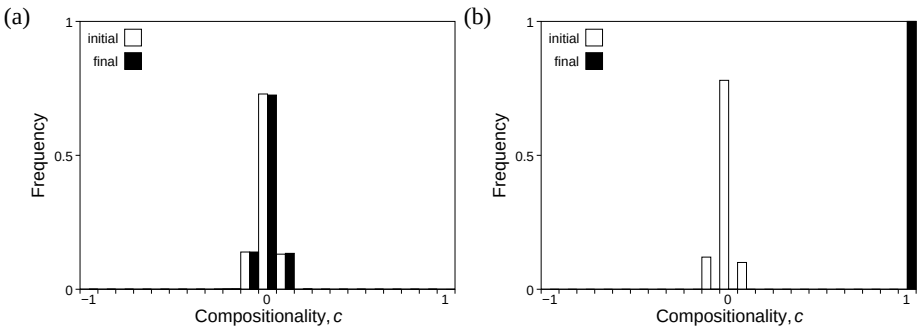

Fig. 2. The impact of the transmission bottleneck. (a) gives frequency by compositionality for runs in the no bottleneck condition. (b) gives frequency by compositionality for runs where there is a bottleneck on transmission.

transmission bottleneck condition, this set of meaning-signal pairs contains examples of the signal associated with every possible meaning, and each learner is therefore presented with the complete language of the agent at the previous generation. In the *transmission bottleneck* condition, the set of observed behaviour does not contain examples of the signal associated with every meaning, therefore each learner is presented with a subset of the language of the agent at the previous generation.[7] The transmission bottleneck constitutes one aspect of the poverty of the stimulus problem faced by language learners — they must acquire knowledge of a large (or infinite) language on the basis of exposure to a subset of that language.

In both conditions, the initial agents in each simulation run have all their connections weights set to 0, and therefore produce every meaning-signal pair with equal probability. Subsequent agents have connection weights of 0 prior to learning. Runs were allowed to progress for a fixed number of generations (200). Figs. 2 (a) and (b) plot compositionality by frequency for the initial and final languages, for the no bottleneck and bottleneck conditions respectively.[8]

As can be seen from the figure, in the absence of a bottleneck on transmission, there is no cultural evolution and compositional languages do not emerge. In contrast, in the bottleneck condition highly compositional systems emerge with high frequency — cultural evolution leads to the emergence of compositional language from initially holistic systems. This confirms, using a rather different model of a language learner, previously established results [2, 3].

In the absence of a transmission bottleneck, the initial, random assignment of signals to meanings can simply be memorised. Consequently, there is no pressure for compositionality and the holistic mapping embodied in the initial system persists. However, holistic systems cannot survive in the presence of a bottleneck. The meaning-signal pairs of a holistic language have to be observed to be reproduced. If a learner only ob-

[7] For all simulations involving a transmission bottleneck described in this paper, learners observed approximately 60% of the language of the previous agent.

[8] The results for the no bottleneck condition are based on 1000 independent runs of the ILM. The results for the bottleneck condition are based on 100 runs — fewer runs are required as there is less sensitivity to initial conditions.

serves a subset of the holistic language of the previous generation then certain meaning-signal pairs will not be preserved — the learner, when called upon to produce, will produce some other signal for that meaning, resulting in a change in the language. In contrast, compositional languages are generalisable, due to their structure, and remain relatively stable even when the learner observes a subset of the language of the previous generation. Over time, language adapts to this pressure to be generalisable. Eventually, the language becomes highly compositional, highly generalisable and consequently highly stable.

4 Exploring Learning Biases

To what extent is this fundamental result, that the transmission bottleneck leads to a pressure for compositional language, dependent on the model of a language learner? There is indirect evidence that this result is to some extent independent of the model of a language learner — a wide range of learning models all produce this fundamental result (see [7] for review). However, do these models share a common element? Is there some *learner bias* common across all these models which is required for compositional language to evolve culturally?

In order to investigate this question, further experiments were carried out, in which the parameter of interest is the weight-update rule used to adjust network connection weights during learning. The general form of the weight-update rule is as follows:

$$\Delta W_{xy} = \begin{cases} \alpha \text{ if } a_x = a_y = 1 \\ \beta \text{ if } a_x = 1 \wedge a_y = 0 \\ \gamma \text{ if } a_x = 0 \wedge a_y = 1 \\ \delta \text{ if } a_x = a_y = 0 \end{cases} \quad (2)$$

For the results described in the previous Section, $\alpha = 1$, $\beta = \gamma = -1$, $\delta = 0$. I will now consider a wider range of weight-update rules, restricting myself to rules where $\alpha, \beta, \gamma, \delta \in \{-1, 0, 1\}$. This yields a set of $3^4 = 81$ possible weight-update rules. In order to ascertain the biases of the different weight-update rules, each weight-update rule is subjected to three tests:[9]

Acquisition test: Can an isolated agent using the weight-update rule acquire a perfectly compositional language, based on full exposure to that language? To evaluate this, an agent using the weight-update rule was trained on a predefined perfectly compositional ($c = 1$) language, being exposed once to every meaning-signal pair included in that language. The agent was judged to have successfully acquired that language if it could reproduce the meaning-signal pairs of the language in production and reception.

Maintenance test: Can a population of agents using the weight-update rule maintain a perfectly compositional language over time in an ILM, when there is a bottleneck on transmission? To evaluate this, 10 runs of the ILM were carried out for the weight-update rule, with the agent in the initial generation having their initial connection

[9] A similar technique has been applied to the investigation of learning biases required for the cultural transmission of vocabulary systems [6].

Test result			Number of rules
Acquire?	Maintain?	Construct?	
no	no	no	63
yes	no	no	16
yes	yes	yes	2

Table 1. Summary of the results of the three tests. The table gives the three types of performance exhibited, and the number (out of 81) of weight-update rules fitting that pattern of performance.

weights set so as to produce a perfectly compositional language. Populations were defined as having maintained a compositional system if c remained above 0.95 for every generation of ten 200 generation runs.

Construction test: Can a population of agents using the weight-update rule construct a highly compositional language from an initially random language, when there is a bottleneck on transmission (as happened in the results outlined in the previous Section)? To evaluate this, 10 runs of the ILM were carried out for the weight-update rule, with the agent in the initial generation having initial connection weights of 0 and therefore producing a random set of meaning-signal pairs. Populations were defined as having constructed a compositional system if c rose above 0.95 in each of ten 200 generation runs.

The results of these sets of experiments are summarised in Table 1. Only a limited number of weight-update rules (two of 81) support the evolution of compositional language through cultural processes. Why? What is it about the assignment of values to the variables α, β, γ and δ in these rules that make them capable of acquiring, maintaining and constructing a compositional system?

A full analysis is somewhat involved, and I will simply summarise the key point here — for full details the reader is referred to [5]. The two weight-update rules which pass the acquisition, maintenance and construction tests satisfy three conditions: 1) $\alpha > \beta$; 2) $\delta > \gamma$; 3) $\alpha > \delta$. These two rules[10] are the only weight-update rules which satisfy these conditions. By returning to the network and examining the way in which connection weights change on the basis of exposure to individual meaning-signal pairs, we can identify the consequences of these restrictions.

1. $\alpha > \beta$ ensures that, if an agent is exposed to the meaning-signal pair $\langle m_i, s_j \rangle$, they will in future tend to prefer produce s_j when presented with m_i, rather than $s_{k \neq j}$.
2. $\delta > \gamma$ ensures that, if an agent is exposed to $\langle m_i, s_j \rangle$, they will prefer *not* to produce s_j when presented with $m_{k \neq i}$.
3. $\alpha > \delta$ ensures that, if an agent is exposed to $\langle m_i, s_j \rangle$, they will tend to reproduce this meaning-signal pair in a manner which involves the maximum number of components.

[10] To be explicit, the two rules are:
$\alpha = 1$, $\beta = -1$, $\gamma = -1$, $\delta = 0$
and $\alpha = 1$, $\beta = 0$, $\gamma = -1$, $\delta = 0$.

Points 1 and 2 in combination lead to a preference for *one-to-one* mappings between meanings and signals — agents with the appropriate weight-update rules are biased in favour of learning languages which map each meaning to a constant signal (one-to-many mappings are avoided, see Point 1), and which map each distinct meaning onto a distinct signal (many-to-one mappings from meanings to signals are avoided, see Point 2). Point 3 corresponds to a bias in favour of memorising associations between elements of meaning and elements of signal, rather than between whole meanings and whole signals.

5 Conclusions

I have presented an Iterated Learning Model of the cultural evolution of compositional structure. This model has been used to replicate a familiar result — the poverty of the stimulus available to language learners (as imposed by the transmission bottleneck) leads to the emergence of compositional structure. However, novelly, this cultural evolution has been shown to be dependent on language learners possessing two biases:

1. a bias in favour of one-to-one mappings between meanings and signals.
2. a bias in favour of exploiting regularities in the input data, by acquiring associations between parts of meanings and parts of signals.

Both these biases are present in most computational models of the evolution of linguistic structure and, significantly, there is also evidence to suggest that human language learners bring these biases to the language acquisition task [7]. Compositionality, a fundamental structural property of language, can therefore be explained in terms of cultural evolution in response to two pressures — a pressure arising from the poverty of the stimulus, and a pressure arising from the biases of language learners. The source of this learning bias in humans is a topic for further research — is the bias a consequence of some general cognitive strategy, or a specific biological adaptation for the acquisition of language?

References

1. H. Brighton. Experiments in iterated instance-based learning. Technical report, Language Evolution and Computation Research Unit, 2000.
2. H. Brighton. Compositional syntax from cultural transmission. *Artificial Life*, 8(1):25–54, 2002.
3. S. Kirby. Syntax out of learning: the cultural evolution of structured communication in a population of induction algorithms. In D. Floreano, J. D. Nicoud, and F. Mondada, editors, *Advances in Artificial Life: Proceedings of the 5th European Conference on Artificial Life*. Springer, Berlin, 1999.
4. S. Kirby. Natural Language from Artificial Life. *Artificial Life*, 8(2):185–215, 2002.
5. K. Smith. Compositionality from culture: the role of environment structure and learning bias. Technical report, Language Evolution and Computation Research Unit, 2002.
6. K. Smith. The cultural evolution of communication in a population of neural networks. *Connection Science*, 14(1):65–84, 2002.
7. K. Smith. Learning biases and language evolution. In *Proceedings of the 15th European Summer School on Logic, Language and Information*. forthcoming.