

Compositionality from culture: the role of environment structure and learning bias

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

<http://www.ling.ed.ac.uk/~kenny>

Abstract

This technical report is based on Chapter 5 of Smith (2003).

1 An Iterated Learning Model

The ILM outlined in this report is an extension of the associative network ILM described in Smith (2002a) and is designed to allow investigation into the importance and interaction of three factors in the cultural evolution of compositional structure: transmission bottleneck, environment structure, and learning bias.

1.1 Languages and communication

A language L consists of a production function $p(m)$, mapping from meanings m to signals s , and a reception function $r(s)$, mapping from signals s to meanings m . m and s are selected such that $m \in \mathcal{M}$ and $s \in \mathcal{S}$ where $\mathcal{M} = \{m_1, m_2 \dots m_{|\mathcal{M}|}\}$ and $\mathcal{S} = \{s_1, s_2 \dots s_{|\mathcal{S}|}\}$.

In the model of unstructured communication systems each $m \in \mathcal{M}$ and each $s \in \mathcal{S}$ is a distinct atomic unit. In the model of structured languages, each $m \in \mathcal{M}$ is a vector drawn from an F -dimensional space, where each dimension has V possible values. More formally, F and V define a meaning-space \mathcal{M} :

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

We can therefore define a distance measure between two meanings m_i and m_j , $HD(m_i, m_j)$. This is simply the Hamming distance between the two meanings, the number of features for which m_i and m_j have different values.

I will introduce a new term, the *environment*, \mathcal{E} , appealing to the notion that the agents' external environment provides the situations which they are required to produce signals for. The meanings in \mathcal{E} constitute a *subset* of \mathcal{M} .

*The author is supported by ESRC Research Grant No. R000223969.

Each signal $s \in \mathcal{S}$ is a string of characters of length 1 to l , $l \leq l_{max}$, where the characters are drawn from the alphabet Σ . l_{max} and Σ define a signal space \mathcal{S} :

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

We can define a distance measure between two signals s_i and s_j , $LD(s_i, s_j)$. LD is the Levenstein (string edit) distance between those two signals and gives the minimum number of deletions, insertions or substitutions required to convert s_i into s_j .

1.2 Linguistic agents

Linguistic agents in the model must be capable of representing such communication systems, modelling production and reception functions of the type outlined above and modifying their behaviour based on observations of systems of the type outlined above. The associative network model outlined in Smith (2002a) is used as a basis for the model of a communicative agent. The basic associative network model is altered to allow the manipulation of structured meanings and signals. The most fundamental changes are in the processes of production and reception, while the process of learning remains largely unchanged.

1.2.1 Representation

Agents are modelled using networks consisting of two sets of nodes \mathcal{N}_M and \mathcal{N}_S and a set of weighted bidirectional connections \mathcal{W} , which connect every node in \mathcal{N}_M with every node in \mathcal{N}_S .

What do these nodes represent? In the associative network outlined in Smith (2002a), meanings and signals are discrete atomic items, therefore each node represented a particular meaning or signal. However, in the case of structured languages each distinct meaning or signal has structure, some of which may be shared with other meanings or signals. In order to represent systems of this type, it is necessary to introduce two new pieces of terminology — the notion of *components* of meanings and signals, and *analyses* of meanings and signals. I will define components now, and postpone the definition of analyses until the section on production and reception.

As summarised above, each meaning is a vector in F -dimensional space where each dimension has V values. *Components* of meanings are (possibly partially specified) vectors, with each feature of the component either having the same value as the meaning, or a wildcard. More formally, if c_m is a component of meaning m , then the value of the j th feature of c_m is:

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

where $*$ represents a wildcard. Similarly, components of signals of length l are (possibly partially specified) strings of length l . I impose the additional constraint that a component must have a minimum of one specified position. For example, the components of the meaning represented by the vector (1 2) are (1 2), (1 *) and (* 2), but not (1 3) (value of feature 2 doesn't match) or (* *) (no specified features). Similarly, the components of the signal represented by the string bd are bd , $b*$ and $*d$, but not $e*$ (first character doesn't match), $**$ (no specified characters) or a (not of correct length).

Each node M_i 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 S_i 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. In order to represent the meaning component c_m the activation, $a_{M c_m}$, of node $M c_m$ is set to one. In order to represent the signal component c_s , $a_{S c_s}$ is set to 1. This representational scheme is illustrated in Figure 1.

1.2.2 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 W where W is specified as a 4-tuple $(\alpha \beta \gamma \delta)$, where α, β, γ and δ take integer values in the range $[-1, 1]$. The value in α specifies how the weight of connection $w_{i,j}$ should be adjusted when $a_i = a_j = 1$, the value in β specifies how $w_{i,j}$ should be adjusted when $a_i = 1$ and $a_j = 0$, the value in γ specifies how $w_{i,j}$ should be adjusted when $a_i = 0$ and $a_j = 1$ and the value in δ specifies how $w_{i,j}$ should be adjusted when $a_i = a_j = 0$. The storage process is illustrated in Figure 2.

1.2.3 Production and reception

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

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

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

During the process of producing utterances, agents are prompted with a meaning and required to produce a meaning-signal pair. Production proceeds via a winner-take-all process. In order to retrieve a signal $s_i \in \mathcal{S}$ based on an input 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 yields a score g . The meaning-signal pair which yields the analysis pair with the highest g is returned as the network's production for the given meaning. The score for a meaning analysis (which consists of a set of meaning components) paired with a signal analysis (a set of signal components) is given by:

$$g \left(\left\{ c_m^1, c_m^2 \dots c_m^n \right\}, \left\{ c_s^1, c_s^2 \dots c_s^n \right\} \right) = \sum_{i=1}^n \omega \left(c_m^i \right) \cdot w_{c_m^i, c_s^i}$$

where n is the number of components in the analysis of meaning and signal, $w_{c_m^i, c_s^i}$ gives the weight of the connection between the nodes representing the i th component of the meaning analysis and the i th component of the signal analysis and $\omega(x)$ is a weighting function which gives the non-wildcard proportion of x . The production process is illustrated in Figure 3.

How is this process to be interpreted? A meaning analysis-signal analysis pair can be interpreted as a parse tree where each terminal node of the tree is labelled with both a component of meaning and a component of signal. The i th node of the tree is labelled by the i th component of the meaning analysis and the i th component of the signal analysis. This yields the fairly natural interpretation that

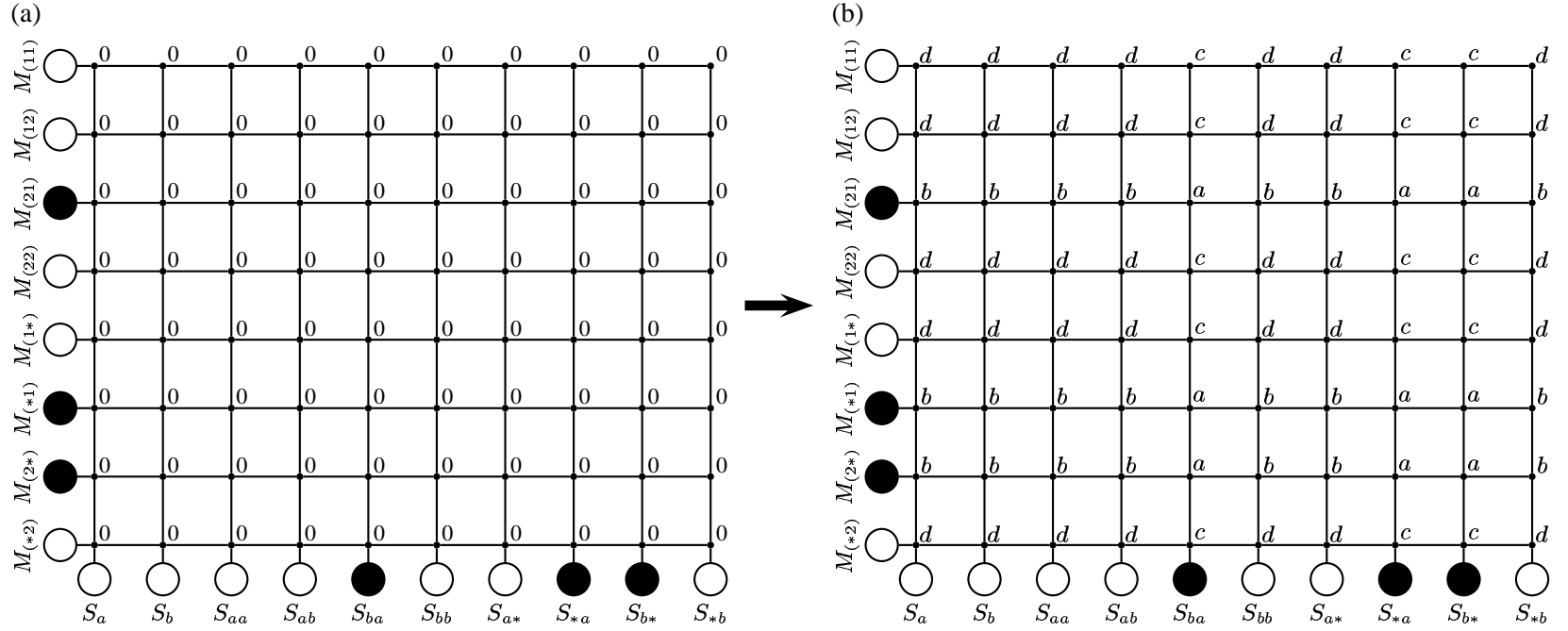


Figure 2: Learning of the meaning-signal pair $\langle (2\ 1), ba \rangle$ using the weight-update rule $W = (a\ b\ c\ d)$. In (a), the nodes in \mathcal{N}_M and \mathcal{N}_S have been set to the patterns of activation representing the components of $(2\ 1)$ and ba . All connections have weight 0. In (b) the result of the application of the learning process is shown — all connections now have weights of a , b , c or d , depending on the activations of the nodes they connect.

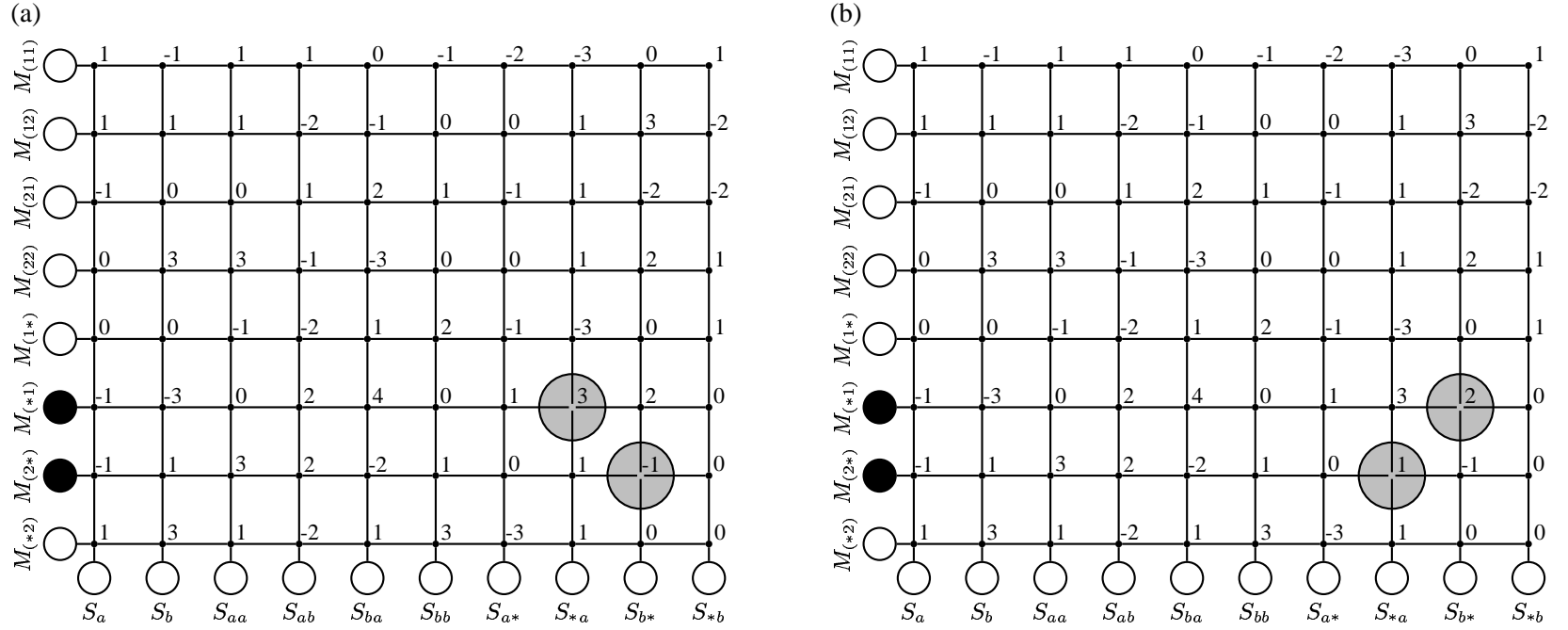


Figure 3: Evaluation of two meaning analysis-signal analysis pairs, encoded as patterns of activation over \mathcal{N}_M . In (a) the pair $\langle\langle(2*), (*1)\rangle\rangle, \langle\langle b*, *a\rangle\rangle$ is evaluated by taking the connection weights $w_{M(2*), S_{b*}}$ and $w_{M(*1), S_{*a}}$ (highlighted in grey) and calculating g . $g(\langle\langle(2*), (*1)\rangle\rangle, \langle\langle b*, *a\rangle\rangle) = 1$. In (b) this process is repeated for the pair $\langle\langle(2*), (*1)\rangle\rangle, \langle\langle *a, b*\rangle\rangle$, yielding $g(\langle\langle(2*), (*1)\rangle\rangle, \langle\langle *a, b*\rangle\rangle) = 1.5$.

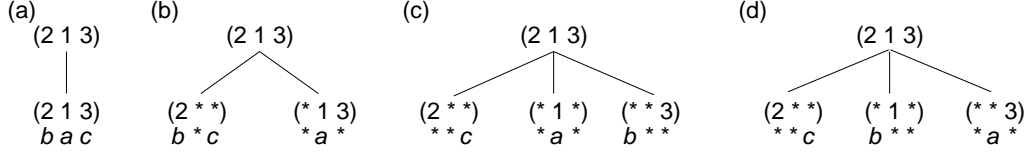


Figure 4: Parse trees corresponding to four of the possible 169 analyses pairs of the meaning-signal pair $\langle (2\ 1\ 3), bac \rangle$. (a) gives the parse tree for the analysis pair $\langle \{(2\ 1\ 3)\}, \{bac\} \rangle$. There is a single node in the tree, which is labelled with the single component from both meaning and signal analysis. (b) gives the parse tree for the analysis pair $\langle \{(2\ **), (*\ 1\ 3)\}, \{b*c, *a*\} \rangle$. The left daughter node is labelled with the 1st component of both analyses, and the right daughter node is labelled with the 2nd component of both analyses. (c) and (d) give the parse trees for the analyses pairs $\langle \{(2\ **), (*\ 1\ *), (*\ * 3)\}, \{**c, *a*, b**\} \rangle$ and $\langle \{(2\ **), (*\ 1\ *), (*\ * 3)\}, \{**c, b**, *a*\} \rangle$ respectively. These differ in the order of the second and third components of the signal, which leads to different interpretations of the semantics of string-initial b and medial a .

the i th component of the meaning analysis is conveyed by the i th component of the signal analysis. This is illustrated in Figure 4.

Given this interpretation, we can justify the simplifying assumption implicit in the g measure above that meaning analysis-signal analysis pairs consist of a meaning analysis and a signal analysis with the same number of components. We can make a further simplifying assumption during production and reception that, in the case where two or more meaning analysis-signal analysis pairs would produce equivalent trees, only one is evaluated. For example, $\langle \{(1\ *), (*\ 2)\}, \{a*, *b\} \rangle$ and $\langle \{(*\ 2), (1\ *)\}, \{*b, a*\} \rangle$ produce equivalent trees, therefore there is no need to evaluate both.

1.3 The Iterated Learning Model

I will consider the simplest possible ILM, where the population consists of a single individual at any one time. That individual produces some observable behaviour and then is removed. This observable behaviour is then observed and learned from by a new individual, and the process iterates.

Initialisation Create a population of one agent using the weight-update rule W and possessing communication system L .

Iteration

1. Generate a set of meaning-signal pairs for the single agent in the population by applying the network production process to every meaning $m \in \mathcal{E}$.
2. Remove the current agent.
3. Create a new population consisting of a single agent with connection weights of 0 who uses weight-update rule W .
4. The new agent receives e exposures to the observable behaviour produced by the preceding agent. During each of these e exposures the new agent observes a *single meaning-signal pair* and updates their connection weights according to the observed meaning-signal pair and their weight-update rule W .

5. Return to 1.

In this ILM, each of the e exposures consists of the single observation of a single meaning-signal pair. There is therefore a potential bottleneck on cultural transmission — learners are not guaranteed to make observations of every meaning in the environment, and therefore may subsequently be required to produce a signal for a meaning which they themselves have not observed paired with a signal. We can calculate the expected *coverage* for a given \mathcal{E} and e , $c(\mathcal{E}, e)$ (Equation taken from Brighton (2002)):

$$c(\mathcal{E}, e) = 1 - \left(1 - \frac{1}{|\mathcal{E}|}\right)^e$$

$c(\mathcal{E}, e)$ gives the expected proportion of the meanings in \mathcal{E} that will be observed given e random selections of meanings from \mathcal{E} , and therefore the severity of the bottleneck on cultural transmission. As $e \rightarrow \infty$, $c \rightarrow 1$ and the bottleneck virtually disappears. However, there will still be a (possibly remote) chance that an individual will be called upon to produce a meaning for a signal that they themselves have not observed. It is impossible to remove the bottleneck completely by increasing e . I will therefore replace step 4 in the iteration algorithm given above with one of two options:

4 (no bottleneck) The new agent receives $e = |\mathcal{E}|$ exposures to the observable behaviour produced by the preceding agent. During each of these exposures the new agent observes a single meaning-signal pair and updates their connection weights according to the observed meaning-signal pair and their weight-update rule W . Each $m \in \mathcal{E}$ is selected in turn, therefore the learner observes the full set of observable behaviour produced by the preceding agent.

4 (bottleneck) The new agent receives e exposures to the observable behaviour produced by the preceding agent. During each of these exposures the new agent observes a single, randomly selected, meaning-signal pair and updates their connection weights according to the observed meaning-signal pair and their weight-update rule W . The agent will therefore observe approximately $|\mathcal{E}| \cdot c(\mathcal{E}, e)$ distinct meanings, paired with their corresponding signals.

1.4 Environments

As discussed in Section 1.1, a distinction has been made between the meaning space \mathcal{M} and the environment \mathcal{E} , the set of meanings for which agents are required to produce signals. In Brighton’s model (Brighton 2002) meanings in the environment are selected at random from the meaning space — \mathcal{E} is random subset of \mathcal{M} . However, it is not necessarily the case that meanings in the environment should be selected at random from the space of possible meanings. I will introduce a notion of *environment structure*, in contrast to Brighton’s notion of meaning space structure. In an *unstructured* environment, meanings in \mathcal{E} are selected at random from \mathcal{M} . In a *structured* environment, meanings in the environment are drawn from a hypercube subset of the space of possible meanings.

In addition to this notion of environment structure, we can define a measure of environment *density*. This is simply the proportion of the space of possible meanings which are contained in \mathcal{E} , and can be defined as $p(\mathcal{E})$:

$$p(\mathcal{E}) = \frac{|\mathcal{E}|}{|\mathcal{M}|} = \frac{|\mathcal{E}|}{V^F}$$

Low p corresponds to low density, and high p corresponds to high density.

Figure 5 shows three unstructured environments, of various densities. Figure 6 shows three structured environments, of various densities¹.

¹Recall that values within a feature are unorganised. Therefore, structured environments are not simply *smaller* than

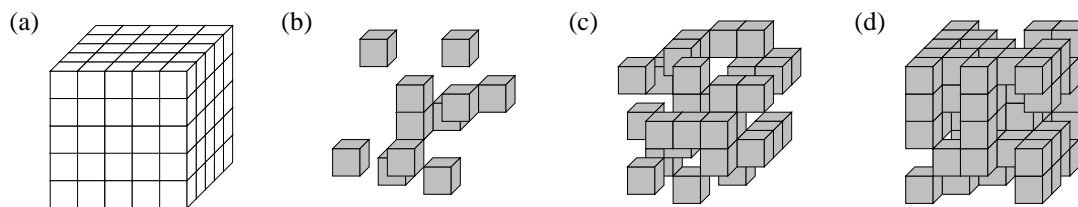


Figure 5: Unstructured environments. F and V define a meaning space \mathcal{M} . For $F = 3$ and $V = 5$ the meaning space can be visualised as a cube, as in (a). The three dimensions of the cube each correspond to the three feature values. The five subdivisions on each dimension correspond to the five values for each feature. Each point in this cube corresponds to a particular meaning. (b) is a sparse, unstructured environment ($p = 0.096$), where the meanings in \mathcal{E} are highlighted in grey. (c) is a medium density, unstructured environment ($p = 0.248$). (d) is a dense, unstructured environment ($p = 0.504$).

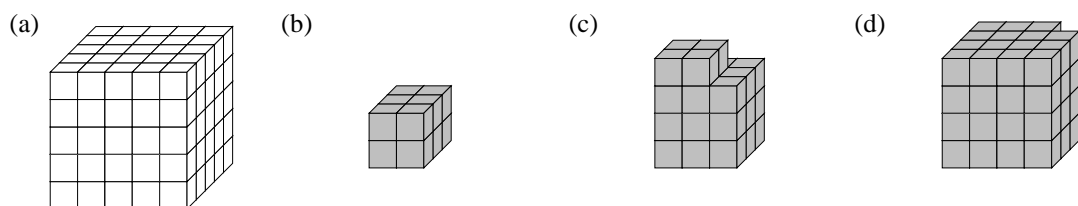


Figure 6: Structured environments. (a) is the meaning space. (b) is a sparse, structured environment ($p = 0.096$). (c) is a medium density, structured environment ($p = 0.248$). (d) is a dense, structured environment ($p = 0.504$).

1.5 Measuring compositionality

In a compositional language the meaning of an expression is a function of the meanings of its parts and the way in which they are combined. In contrast, in a non-compositional (or *holistic*) system, the meaning of an expression is not dependent on the meaning of its parts. I will define two measures of compositionality here, both drawing on slightly different interpretations of this informal definition.

Compositionality is a property that can be observed in externalized language, without knowing the language-user internal manipulations which lead to the produced language. In this sense, a compositional language is a mapping between meanings and signals which preserves neighbourhood relationships — neighbouring meanings will share structure, and that shared structure in meaning space will map to shared structure in the signal space. For example, the sentences *John walked* and *Mary walked* have parts of an underlying semantic representation in common (the notion of someone having carried out the act of walking at some point in the past) and will be near one another in semantic representational space. This shared semantic structure leads to shared signal structure (the inflected verb *walked*) — the relationship between the two sentences in semantic and signal space is preserved by the compositional mapping from meanings to signals. A holistic language is one which does not preserve such relationships — as the structure of signals does not reflect the structure of the underlying meaning, shared structure in meaning space will not necessarily result in shared signal structure.

The external compositionality measure which I describe here captures this notion, and is based on the measure developed in Brighton (2000) for Euclidean meaning and signal spaces. The measure of external compositionality is simply the degree of correlation between the distance between pairs of meanings and the distance between the corresponding pairs of signals. If shared meaning structure leads to shared signal structure then there will be a positive correlation between the distance between pairs of meanings and the distance between the corresponding pairs of signals. If shared structure does not necessarily lead to shared signal structure then there will be no correlation. This measurement will be referred to as *external compositionality*, or e-compositionality, given that it refers to the agent-external, observable behaviour resulting from production.

In order to evaluate the e-compositionality of an agent’s communication system, the production process is applied to every $m \in \mathcal{E}$ to produce the set \mathcal{O} , the observable meaning-signal pairs produced by that agent. In order to measure the degree of external compositionality we measure the degree to which the distances between all the possible pairs of meanings correlates with the distances between their associated pairs of signals. More formally, we first take all possible pairs of meanings $\langle m_i, m_{j \neq i} \rangle$, where $m_i \in \mathcal{E}$ and $m_j \in \mathcal{E}$. We then find the signals these meanings map to in the set of observable meaning-signal pairs \mathcal{O} , $\langle s_i, s_j \rangle$. This will give us a set of n meaning-meaning pairs and a set of n signal-signal pairs. Let $\Delta m_n = HD(m_i, m_j)$ be the Hamming distance between the two meanings in the n th pair of meanings and $\Delta s_n = LD(s_i, s_j)$ be the Levenstein distance between the n th pair of signals. Furthermore, let $\overline{\Delta m} = \frac{\sum_{i=1}^n \Delta m_n}{n}$ be the average inter-meaning Hamming distance and $\overline{\Delta s} = \frac{\sum_{i=1}^n \Delta s_n}{n}$ be the average inter-signal Levenstein distance. We can then compute the Pearson correlation coefficient for the distance pairs $\langle m_n, s_n \rangle$, which gives the e-compositionality of a set of observable behaviour, $E(\mathcal{O})$:

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

unstructured environments — each structured environment could be rearranged so as to appear to more fully fill the meaning space. It is the degree of sharing of feature values which defines environment structure, not apparent closeness.

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

The e-compositionality measure makes no reference to the agent-internal representations which lead to the observable behaviour that an agent produces. This is something of a shortcoming, as can be illustrated by two simple thought experiments. Firstly, imagine a speaker of a language, say English, who produces what appear to be entirely grammatical, normal sentences of English. However, under a suitably sophisticated set of experimental conditions it is revealed that this speaker of English has in fact simply memorised hundreds of thousands of unanalysed sentences of English, paired with their meanings, in a massive lexicon. While from their external behaviour we might conclude that the meaning of an expression for this speaker was a function of the meaning of its parts, this would not be the case — for this speaker, complete meanings are stored paired with complete sentences, and those sentences as a whole stand for the whole meaning. The imaginary speaker is in fact producing English as a holistic system, with no real, internal compositional knowledge. It has been suggested (Wray & Perkins 2000) that much of language used for social interaction is in fact processed in this way. However, it would be undesirable to say that our imaginary speaker possesses a compositional knowledge of English, even if their external behaviour appears to comply to a compositional analysis.

As a second thought experiment, imagine a speaker who has a thorough knowledge of several hundred thousand languages. Each time this individual speaks, they choose a language from their massive arsenal at random, and produce their utterance in that language. Internally, this imaginary speaker is behaving compositionality for each utterance they produce — just like a native speaker of whichever language they happen to be using, the meaning of their utterance is a function of the meaning of the parts of that utterance and the way those parts are combined. However, to an external observer, their language would appear to be non-compositional — even closely related meanings would be communicated by radically different expressions, with no obvious structure-preserving mapping between meanings and signals.

Our second measure of compositionality, which I will term *internal compositionality*, or i-compositionality, addresses these deficiencies of the e-compositionality measure by quantifying the degree to which utterances are constructed by the combination of agent-internal representations.

During production or reception the set of possible meaning analysis-signal analysis pairs are evaluated, with the meaning-signal pair which yields the analysis pair with the highest g being returned as the network’s production or reception behaviour. In order to evaluate the i-compositionality of an agent’s communication system, the production process is applied to every $m \in \mathcal{E}$ to produce the set \mathcal{A} , the set of meaning analysis-signal analysis pairs which yield the highest g for each meaning. The i-compositionality of a set of meaning analysis-signal analysis pairs \mathcal{A} , $I(\mathcal{A})$, is:

$$I(\mathcal{A}) = \sum_{k=1}^{k=|\mathcal{A}|} \frac{1}{|\mathcal{A}|} i(A_k)$$

where $i(A_k)$ is the i-compositionality of the k th meaning analysis-signal analysis pair $\langle a_m, a_s \rangle$, and is given by:

$$i(\langle a_m, a_s \rangle) = \frac{|a_m| - 1}{\min(l_{max}, F) - 1}$$

$i(\langle a_m, a_s \rangle) = 0$ when the meaning analysis and signal analysis consist of a single component, and $i(\langle a_m, a_s \rangle) = 1$ where each analysis consists of the maximum number of components, which is constrained by the smaller of the maximum string length and the dimensionality of the meaning space. $I(\mathcal{A}) = 0$ for an i-holistic mapping, and 1 for a perfectly i-compositional language.

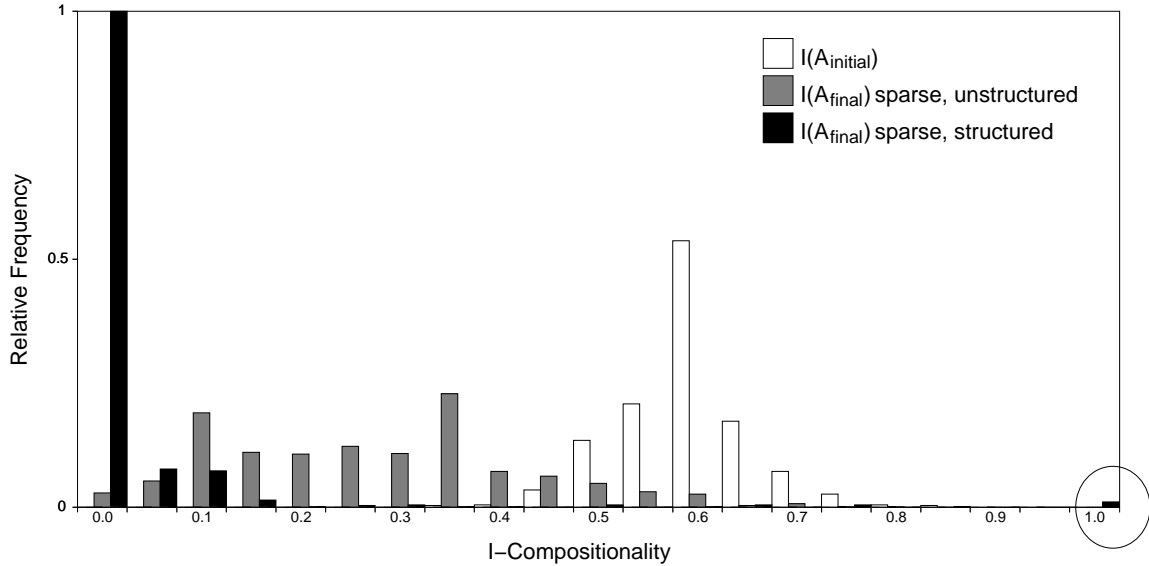


Figure 7: I-compositionality of initial and final, stable systems in sparse environments, when there is no bottleneck on cultural transmission. The initial systems are partially i-compositional. The final systems are less i-compositional, although highly i-compositional systems (circled) do occur with very low frequency when the environment is structured.

2 The impact of transmission bottleneck and environment structure

I will begin by presenting results for an ILM where every agent uses the weight-update rule $W = (1 \ -1 \ -1 \ 0)$. The agents in the initial generation of each ILM have connection weights of 0, and therefore use the maximum entropy system where every meaning analysis-signal analysis pair occurs with equal probability — the initial language L is random. The meaning space ($F = 3$ and $V = 5$) and six environments illustrated in Figures 5 and 6 were used. The signal space is given by $l_{max} = 3$, $\sigma = \{a, b, c, d, e, f, g, h, i, j\}$.

2.1 Linguistic evolution in the absence of a bottleneck

Runs of the ILM were carried out, using the no-bottleneck variant of step 4 — each individual observes the full set of observable behaviour produced by the preceding agent. 1000 runs were carried out for each of the six environments. Each run was allowed to proceed to a stable state, where parent and child produce identical observable behaviour. At this point, in the absence of a bottleneck on cultural transmission, further change is impossible. Figures 7 and 8 give the distributions of systems with respect to $I(\mathcal{A})$ and $E(\mathcal{O})$, for the 1000 runs of the ILM with the sparse unstructured and sparse structured environments².

²The plotting style requires some justification. The measures of i-compositionality and e-compositionality are real numbers. We are interested in the frequency of systems exhibiting a given degree of compositionality. Such information is typically conveyed using a histogram or frequency polygon. I have chosen to use a histogram, given the problems with edge effects arising from using a frequency polygon. Bins of width 0.05 are used for all results plotted here. The y-axis gives relative, rather than absolute frequency — the relative frequency is simply absolute frequency divided by the absolute frequency of the most frequent value. The most frequent value therefore has a relative frequency of 1.

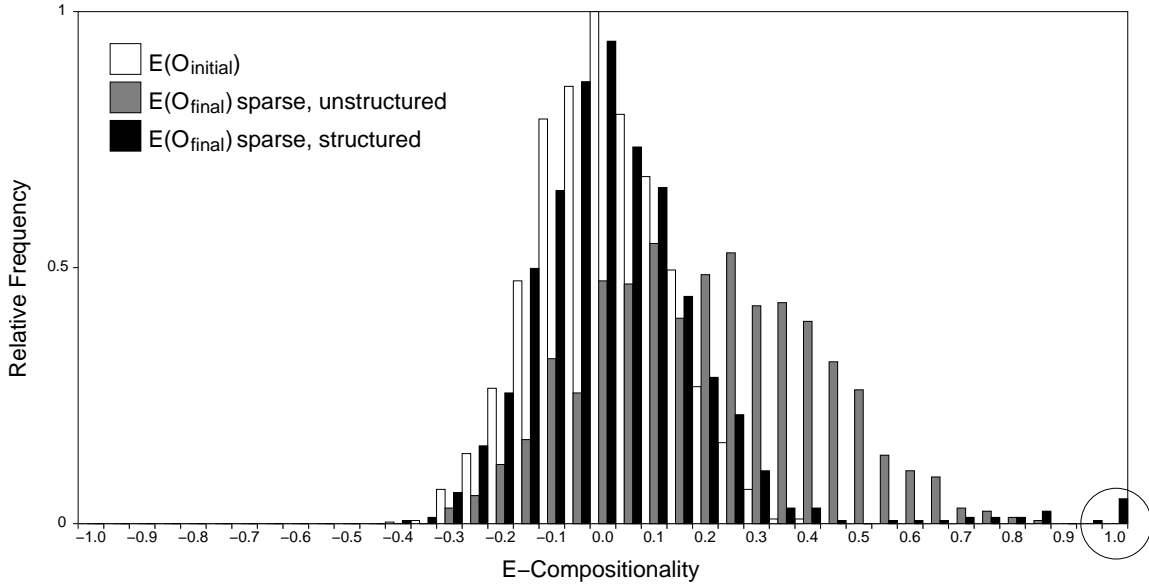


Figure 8: E-compositionality of initial and final, stable systems in sparse environments, when there is no bottleneck on transmission. The initial systems have low e-compositionality. The final systems are also of low or medium e-compositionality. Highly e-compositional systems (circled) occur infrequently, and only when the environment is structured.

In Figure 7, values of $I(\mathcal{A}_{initial})$ are distributed around 0.6, while values for $I(\mathcal{A}_{final})$ are typically lower. This is due to the random behaviour of the initial agents — each meaning analysis-signal analysis pair occurs with equal probability, and given that there are more multi-component analyses pairs than single-component analyses pairs, the initial random behaviour scores highly in terms of internal compositionality. The final stable systems tend to have lower internal compositionality. In structured environments, the i-compositionality of the final systems tends to be around 0. In unstructured environments, final i-compositionality tends to be somewhat higher. Highly i-compositional systems occur very infrequently and only where the environment is structured.

In Figure 8, values of $E(\mathcal{O}_{initial})$ are distributed around 0, indicating that the initial, random systems are not highly e-compositional. As with the internal compositionality measure, the final stable systems tend not to be highly compositional according to the external measure, with unstructured environments leading to a slightly higher level of compositionality. Highly e-compositional systems occurring infrequently and only when the environment is structured.

The i-compositionality measure has the somewhat undesirable property of treating the random initial behaviour as partially compositional. However, for the stable states the internal and external measures are equivalent — for the data in Figures 7 and 8, there is a high degree of correlation between $I(\mathcal{A}_{final})$ and $E(\mathcal{O}_{final})$ ($r = 0.842, p < 0.001$). This reflects the fact that agents produce an e-compositional language in an i-compositional manner, and similarly produce an e-holistic system in an i-holistic manner — i-compositional internal representation leads to e-compositional language. For the remaining results I will focus on the external compositionality measure, and unless otherwise indicated, “compositional” will mean “e-compositional”. The internal measure will be returned to below in Section 3.

Figures 9 and 10 give the distributions of systems with respect to $E(\mathcal{O})$, for 1000 runs of the ILM

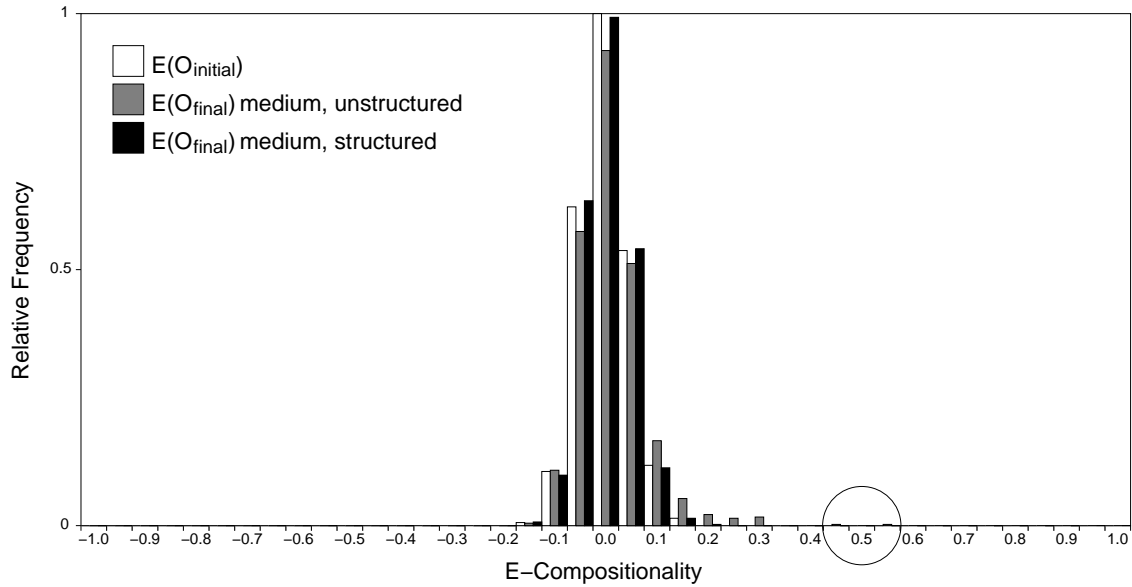


Figure 9: E-compositionality of initial and final, stable systems in medium density environments, when there is no bottleneck on transmission. The initial systems and the vast majority of the final systems have low e-compositionality. Partially e-compositional final systems (circled) occur with very low frequency, and only when the environment is unstructured.

with the medium and dense environments.

Comparison of Figures 8, 9 and 10 shows that, as environment density increases the frequency of final non-compositional systems increases. In the sparse environments, partially compositional systems do occur, and are more frequent in the unstructured environment. Highly compositional systems occur with very low frequency in the sparse, structured environment. Partially e-compositional systems occur with very low frequency in the medium, unstructured environment. In the dense environments all systems are non-compositional, regardless of the degree of environment structure.

These results suggest three questions. Firstly, why are highly compositional systems so infrequent? Previous results (e.g. Kirby (2002), Brighton (2002)) lead us to expect that, in the absence of a bottleneck on cultural transmission, compositional and holistic systems will be equally stable. Given that the initial random systems are holistic we would expect these systems to remain stable over time. This is what happens in dense environments, or medium density, structured environments. The emergence of partially or highly compositional systems in low density environments, or medium density unstructured environments, is therefore somewhat surprising, which leads us on to the second and third questions.

Secondly, why does environment density impact on the compositionality of the emergent systems? Figure 11 plots the values of $E(O_{initial})$ against $E(O_{final})$ for the simulation runs in the sparse, structured environment. As can be seen from the Figure, the runs can be split into three groups:

- runs where $E(O_{initial}) = E(O_{final})$ (group (a) in the Figure).
- runs where $E(O_{initial}) \neq E(O_{final})$, where $E(O_{final})$ is below 0.9 (group (b) in the Figure).
- runs where $E(O_{initial}) \neq E(O_{final})$, where $E(O_{final})$ is close to 1 (group (c) in the Figure).

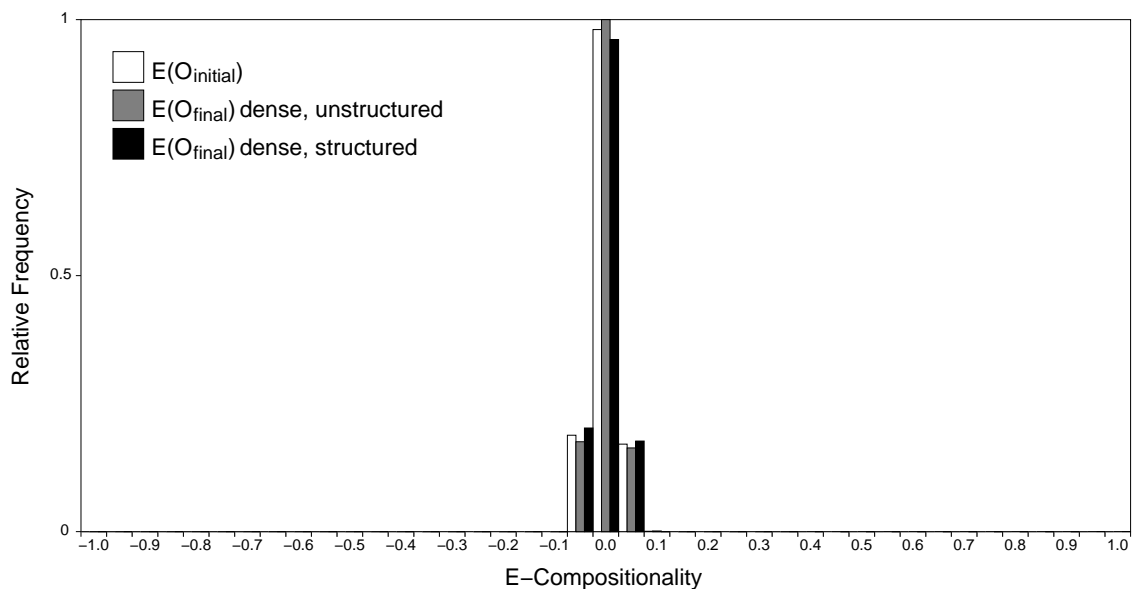


Figure 10: E-compositionality of initial and final, stable systems in dense environments, when there is no bottleneck on transmission. Both the initial and final systems have low e-compositionality.

All environments exhibit runs falling into groups (a) and (b). Only when the environment is sparse and structured do group (c) points occur, representing runs which converge on highly compositional languages. Is $E(\mathcal{O}_{final})$ related to $E(\mathcal{O}_{initial})$? Table 1 gives the mean and standard deviations of the initial values of $E(\mathcal{O}_{initial})$, categorised according to which group they fall into.

As can be seen from the first column of the Table, runs in all environments have a mean value of $E(\mathcal{O}_{initial})$ of approximately 0. However, these initial values are much more tightly distributed around the mean in the more densely filled environments. The second column gives the mean $E(\mathcal{O}_{initial})$ for simulation runs where $E(\mathcal{O}_{initial}) = E(\mathcal{O}_{final})$. These are somewhat lower than the overall mean, and are lower than the mean $E(\mathcal{O}_{initial})$ for simulation runs which move away from initial value (excluding the values for the dense environments, which buck the overall trend). Also, for the group b runs in sparse and medium environments, the mean value of $E(\mathcal{O}_{initial})$ is lower for unstructured environments than for structured environments. Finally, the mean $E(\mathcal{O}_{initial})$ for simulation runs which converge on highly compositional languages is higher still.

These results suggest that there is a degree of sensitivity to the compositionality of the initial, random system. Where this initial mapping exhibits compositional tendencies, yielding $E(\mathcal{O}_{initial})$ above the mean, there is an increased likelihood of the system moving, over iterated learning events, towards more compositional languages. The compositional tendencies of the initial system spread to other parts of the system over time, resulting in an increase in compositionality. However, this progression is not guaranteed — not all simulation runs where $E(\mathcal{O}_{initial})$ is above the mean eventually converge on more compositional systems. For the more densely-filled environments, partially or highly compositional systems emerge infrequently due to the fact that the initial systems tend to be clustered more tightly around the non-compositional mean. When the environment contains few meanings the initial system may, by chance, exhibit some compositional tendencies. However, when the environment contains a large number of meanings such tendencies are likely to be drowned out by the majority non-compositional mapping.

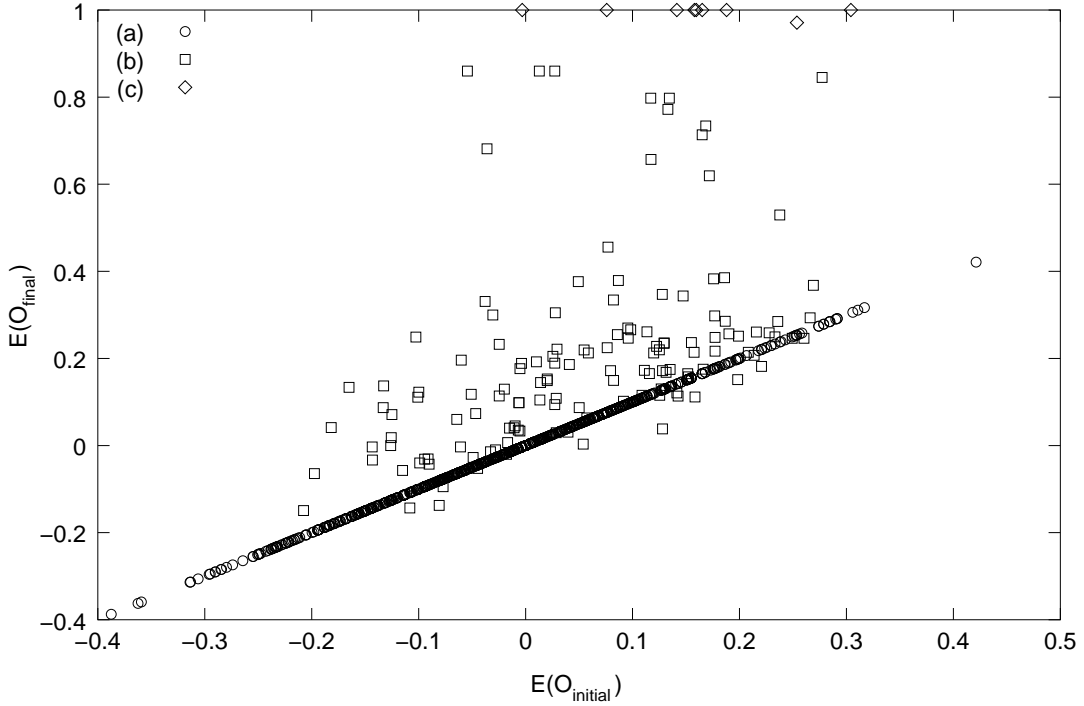


Figure 11: Initial e-compositionality against final e-compositionality for runs in the sparse structured environment. Each point represents a single run. The runs can be separated into three groups. Points labelled as (a) have the same initial and final e-compositionality. For points labelled as (b) the e-compositionality of the initial and final systems is different, but the final system is not highly e-compositional. Points labelled (c) represent runs where the final language is highly e-compositional.

Environment	Initial e-compositionality by group			
	all	a	b	c
sp, us	$\mu = 0.0013, \sigma = 0.1246$	$\mu = -0.0512$	$\mu = 0.0160$	NA
sp, s	$\mu = -0.0029, \sigma = 0.1246$	$\mu = -0.0136$	$\mu = 0.0536$	$\mu = 0.1603$
m, us	$\mu = -0.0004, \sigma = 0.0470$	$\mu = -0.0047$	$\mu = 0.0209$	NA
m, s	$\mu = -0.0011, \sigma = 0.0457$	$\mu = -0.0017$	$\mu = 0.0306$	NA
d, us	$\mu = 0.0002, \sigma = 0.0221$	$\mu = 0.0002$	$\mu = -0.0009$	NA
d, s	$\mu = -0.0011, \sigma = 0.0232$	$\mu = -0.0010$	$\mu = -0.0040$	NA

Table 1: Sensitivity to initial conditions. The table gives the mean (μ) and standard deviation (σ) of the e-compositionality of the initial systems in the various environments (sp = sparse, m = medium, d = dense, us = unstructured, s = structured), broken down by the three groups identified in Figure 11. Standard deviation is given once only, as σ for each subgroup is approximately the same as that for all groups combined. The mean for group c points is higher than that for group b points, which is generally higher than that for group a points. As environment density increases, the initial values are clustered more tightly around the mean.

Thirdly, why does environment structure impact on the e-compositionality of systems at the lower densities? This is related to the previous question. At lower densities, as discussed above, compositional tendencies in the initial system spread, over time, to other parts of the system. In structured environments, distinct meanings tend to have feature values in common with a large number of other meanings. In unstructured environments distinct meanings have feature values in common with few other meanings. If the initial random system has a tendency to express a given feature value with a certain substring then this can spread to cover all meanings involving that feature value — the system becomes consistent with respect to that feature value, which can have knock-on consequences for other values at that feature and other features. In structured environments the potential for spread of the substring associated with a particular feature value is wider than is the case in unstructured environments, given that more meanings will share that feature value. Any initial compositional tendency will therefore spread more widely in structured environments, with more possible follow-on consequences, resulting in the more frequent emergence of highly compositional languages.

However, while shared feature values allow the possibility of the spread of compositionality, they also inhibit it — in a structured environment, any compositional tendency in the initial random mapping has to cover a large number of meanings which share feature values. If only some of these meanings share a character for a particular feature value, then the other meanings, which do not share the character, are likely to outweigh the slight compositional tendency. In contrast, in unstructured environments fewer meanings share feature values, therefore the initial random system has to be less ‘lucky’ in the assignment of characters to feature values. This is reflected in the fact that the mean $E(\mathcal{O}_{initial})$ has to be higher in structured environments before $E(\mathcal{O}_{final})$ moves away from $E(\mathcal{O}_{initial})$, and also in the fact that the average $E(\mathcal{O}_{final})$ in unstructured environments is higher (see Figure 8). In structured environments, the initial compositional tendency has to be strong to escape the attraction of the overall non-compositional mapping, but once this attraction has been escaped highly compositional systems can emerge. In contrast, in unstructured environments the attraction of the initial non-compositional mapping is weaker, due to the reduced degree of feature-value sharing, but the potential spread of compositionality is reduced.

2.2 Linguistic evolution in the presence of a bottleneck

The simulation results outlined in the previous Section show that, in the absence of a bottleneck on cultural transmission, highly compositional languages emerge infrequently. Their emergence is dependent on the density and structure of the environment, and there is a degree of sensitivity to the compositionality of the original, random system of meaning-signal mappings. It is now time to investigate how a transmission bottleneck impacts on the compositionality of the emergent systems.

To this end, runs of the ILM were carried out, with step 4 of the iteration algorithm being replaced by the bottleneck condition — each individual observes e meaning-signal pairs, randomly selected from the set of observable behaviour produced by the preceding agent. e will be experimentally varied. Selection of a value of e depends on the desired degree of coverage $c(\mathcal{E}, e)$, but also on the number of distinct feature values included in the meanings in \mathcal{E} and the rate of seeing distinct feature values with respect to the rate of seeing distinct meanings. For example, in the sparse unstructured environment (see Figure 5 a) there are 12 distinct meanings, and 15 distinct feature values (values 1,2,3,4 and 5 for each feature). In the case where $e < 12$ in this environment it is therefore impossible for a learner to observe all possible feature values paired with a (sub)signal. This will have consequences for the stability of the communication systems through the bottleneck. Consequently, as a simplifying rule of thumb we will not consider the case where $e < 12$. This rules out simulation runs in the sparse environments where $c(\mathcal{E}, e) < 0.65$.

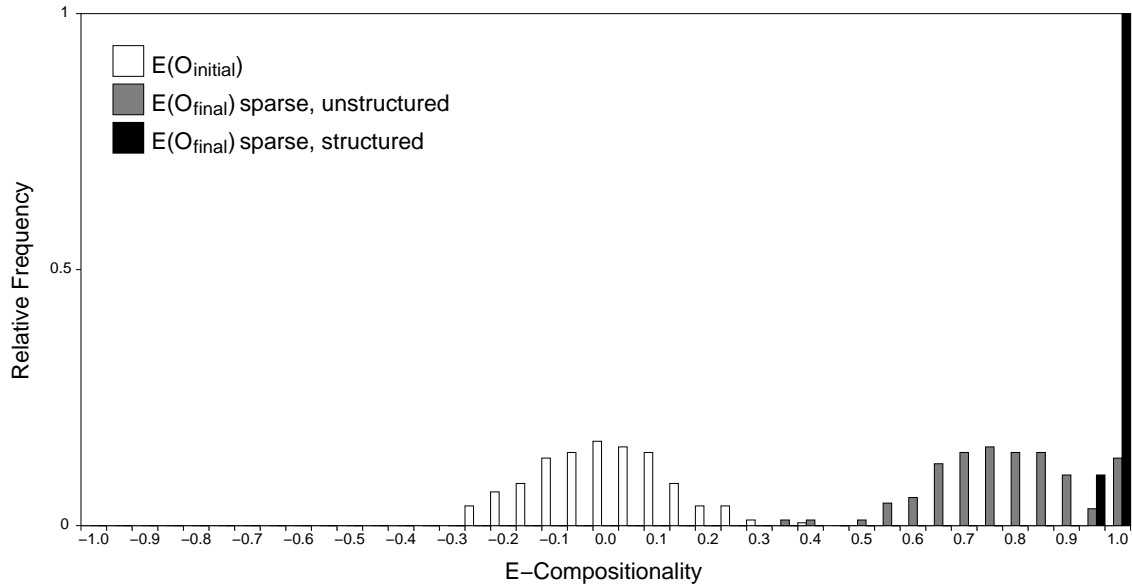


Figure 12: Compositionality of initial and final languages in sparse environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.8$). Highly compositional languages are highly frequent, and are most frequent when the environment is structured.

100 runs of the ILM were carried out for:

- the sparse unstructured and structured environments given in Figures 5 (b) and 6 (b) with $c = 0.8$ ($e = 19$).
- the medium unstructured and structured environments in Figures 5 (c) and 6 (c) with $c = 0.4$ ($e = 16$), $c = 0.6$ ($e = 28$) and $c = 0.8$ ($e = 49$).
- the dense unstructured and structured environments in Figures 5 (d) and 6 (d) with $c = 0.25$ ($e = 18$), $c = 0.4$ ($e = 32$), $c = 0.6$ ($e = 57$) and $c = 0.8$ ($e = 101$)

The distribution of the initial and final systems for these runs in terms of e-compositionality is given in Figures 12–19.

When there is a bottleneck on cultural transmission, the compositionality of the emergent languages is far less sensitive to the compositionality of the initial meaning-signal mappings. Consequently, 100 runs, rather than 1000 runs, were sufficient. While in the absence of a bottleneck runs were allowed to proceed until a stable state was reached, in the bottleneck condition runs were terminated after a fixed number of generations (200). The random selection of meanings from the environment for which to produce utterances means that, as with any stochastic system, a skewed distribution of meanings could lead to the loss of structure. The results reported here accurately reflect the behaviour of the system — allowing the runs to proceed for several hundred more generations gives a similar distribution of languages. In other words, the *distribution* of systems is stable, while individual languages may oscillate between varying levels of compositionality.

The main result apparent from these Figures is that, in the presence of a bottleneck, highly compositional languages emerge with high frequency, and emerge most frequently when the environment

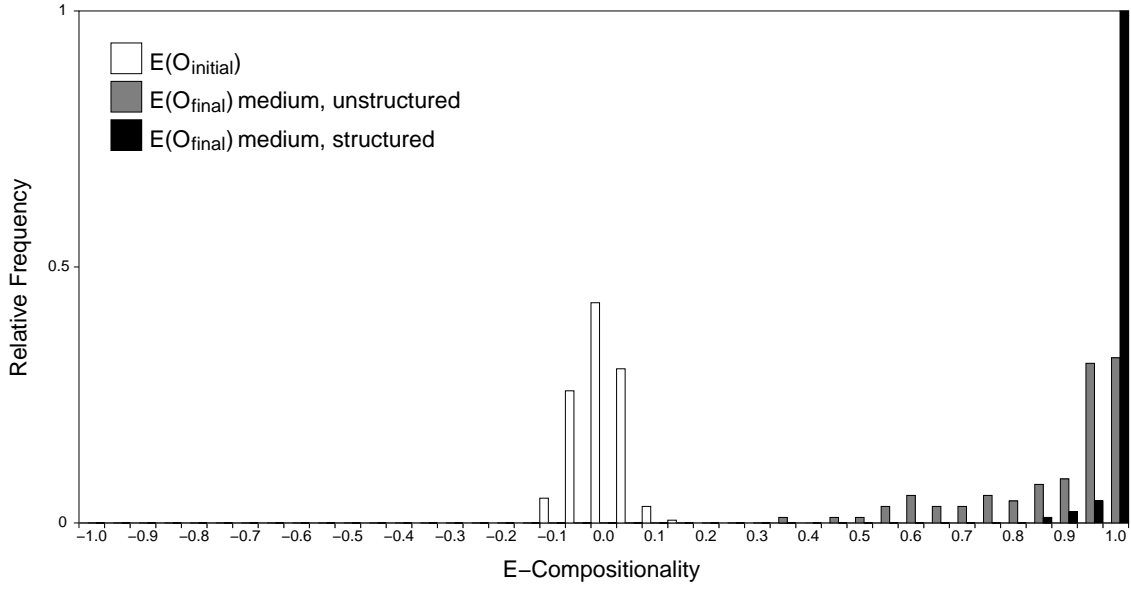


Figure 13: Compositionality of initial and final languages in medium density environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.4$). As in Figure 12, highly compositional languages are highly frequent, and are most frequent when the environment is structured.

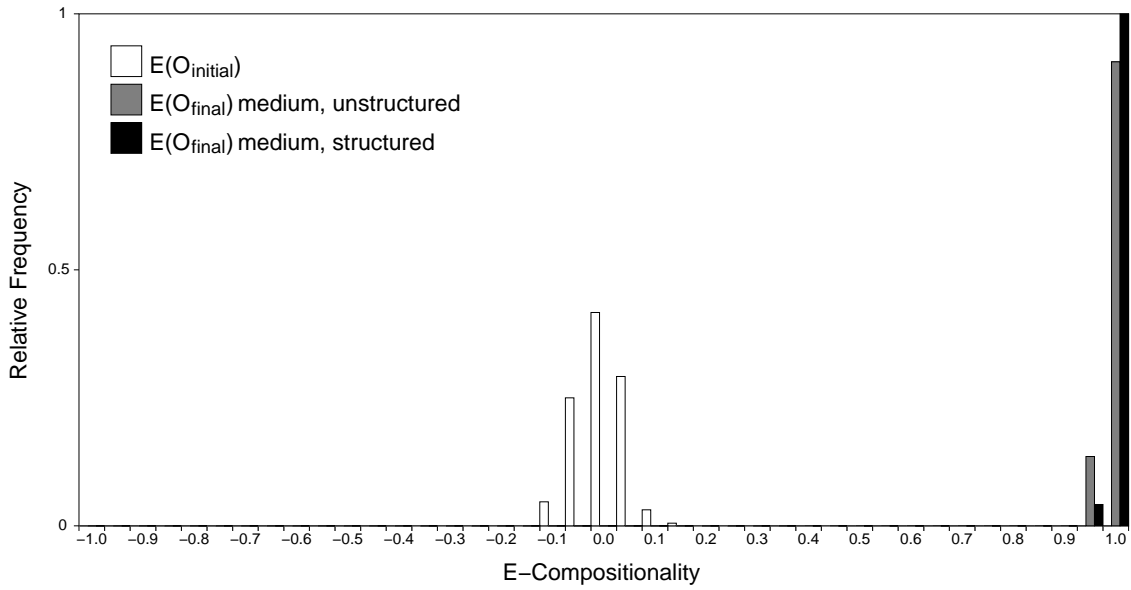


Figure 14: Compositionality of initial and final languages in medium density environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.6$). There is less disparity between unstructured and structured environments.

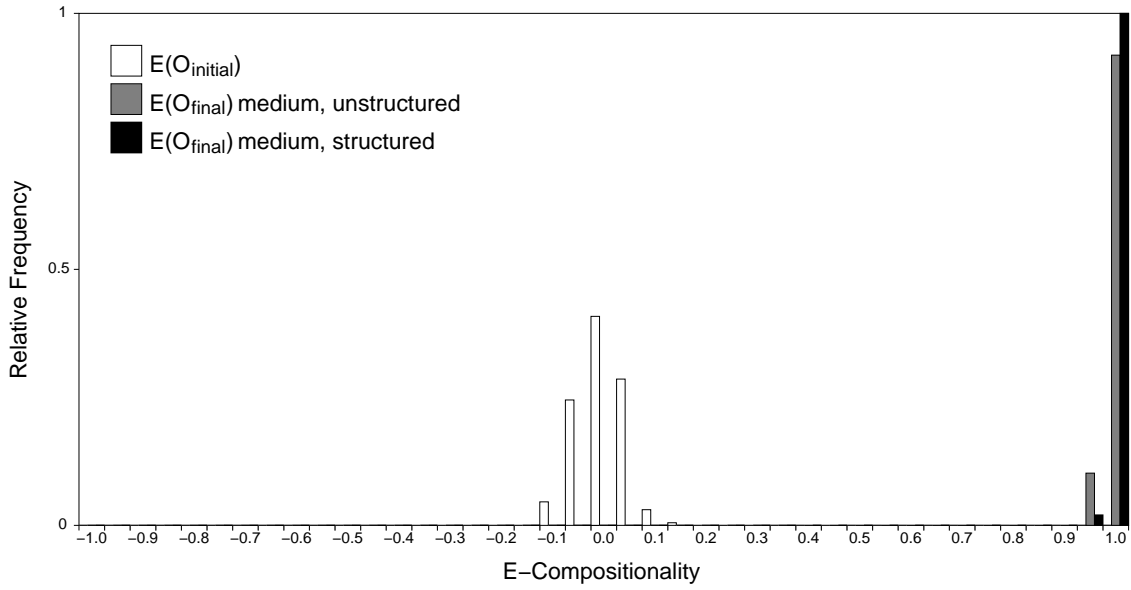


Figure 15: Compositionality of initial and final languages in medium density environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.8$).

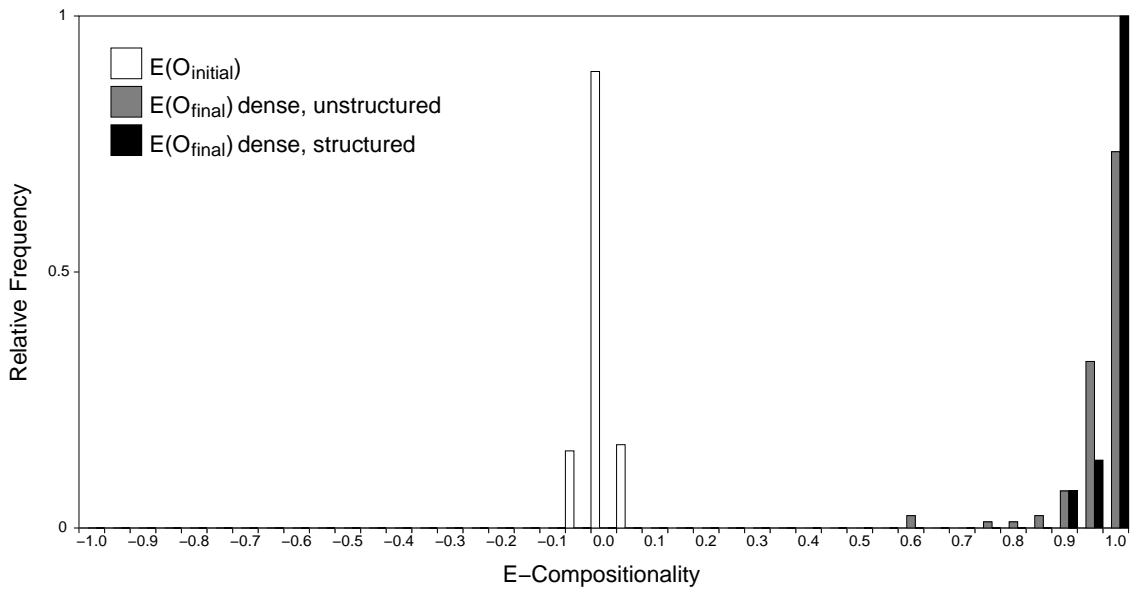


Figure 16: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.25$). While the majority of final languages are highly compositional, some partially compositional systems do exist when the environment is unstructured.

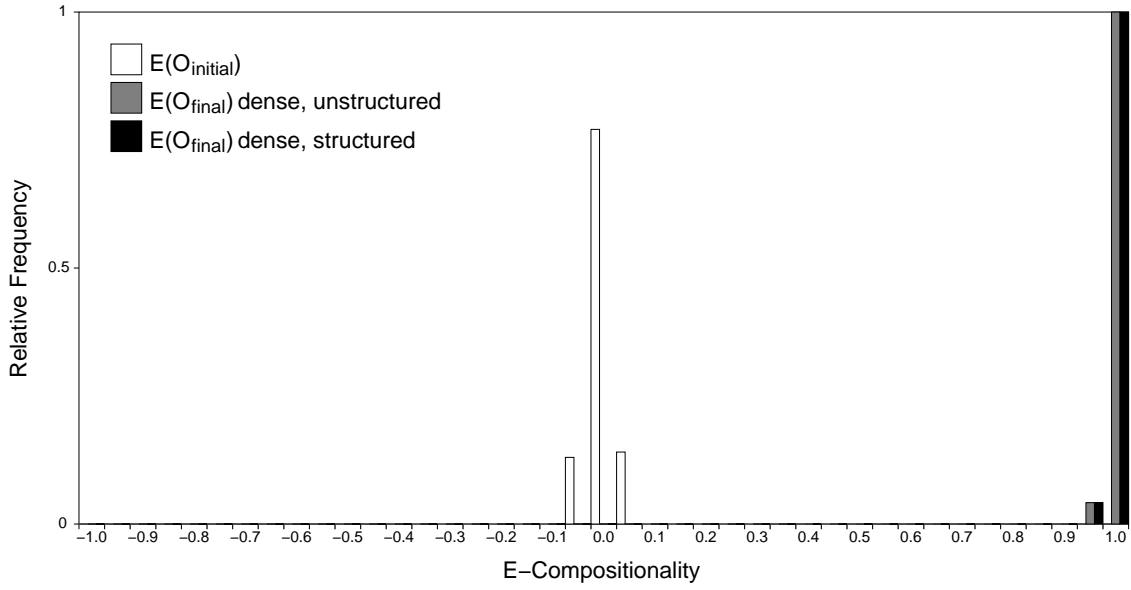


Figure 17: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.4$).

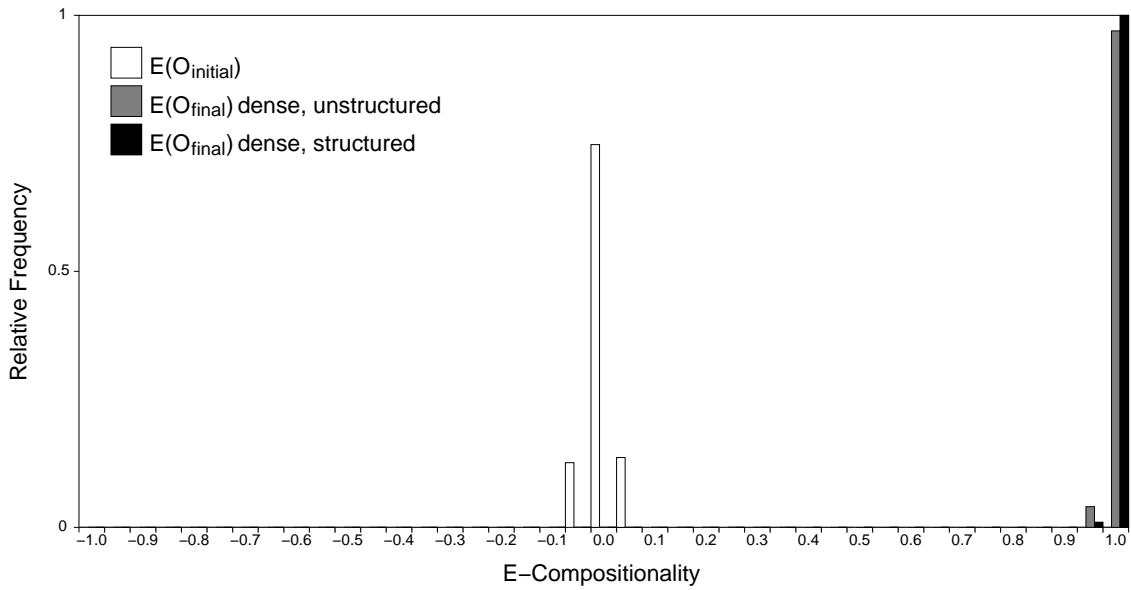


Figure 18: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.6$).

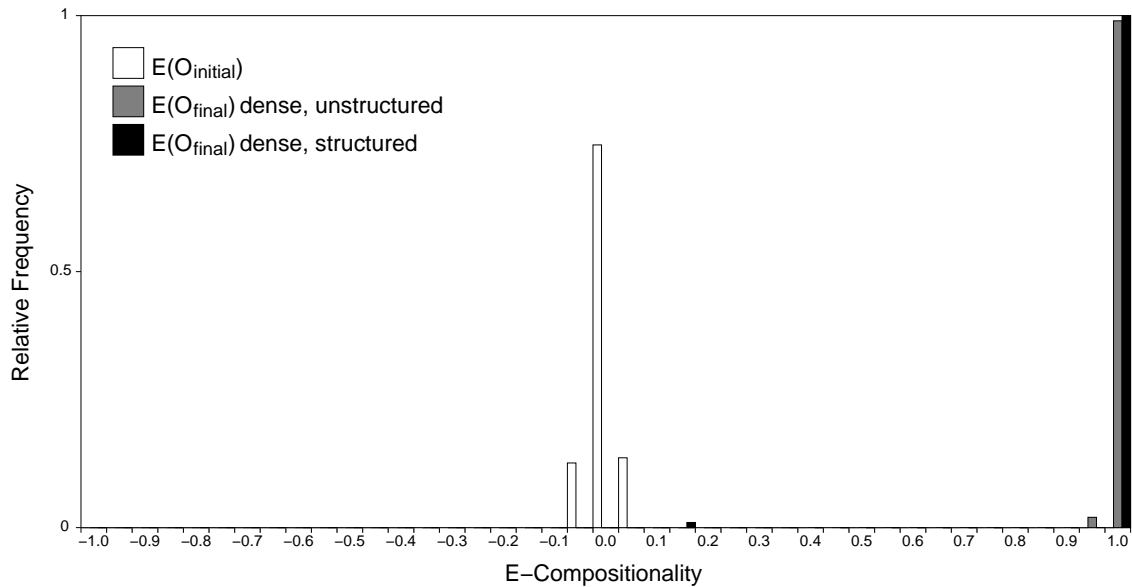


Figure 19: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.8$).

is structured. The nuances of this general result will be returned to below. First, however, the main result must be explained.

Brighton’s (2002) mathematical model predicts that, in the presence of a bottleneck on cultural transmission, compositional language will be more stable than holistic language. The results from the computational model bear this out, but also show that it is possible to move from an initially holistic system to a highly compositional system over time — compositional languages can emerge from initially holistic communication through purely cultural processes, provided there is a bottleneck on cultural transmission.

Why are compositional languages so strongly preferred when there is a bottleneck on transmission? Holistic languages cannot persist in the presence of a bottleneck. The meaning-signal pairs of a holistic language have to be observed to be reproduced. When a learner only observes 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 small subset of the language of the previous generation. Over time, language adapts to the pressure to be generalisable. Eventually, particularly when the environment is structured, the language becomes highly compositional, highly generalisable and consequently highly stable.

In a structured environment the advantage of compositionality is at a maximum. In such environments, meanings share feature values with several other meanings. A language mapping these feature values to a signal substring is highly generalisable. When the environment is unstructured, meanings share feature values with few or no other meanings. In the most extreme case, a meaning may have a value for a particular feature which no other meaning has. The signal associated with that meaning cannot then be deduced from observations of the signals associated with other meanings, and has to be observed to be learned. Consequently, compositional language in an unstructured environment is

less stable through the transmission bottleneck. The resultant languages are a compromise between the push towards compositionality introduced by the bottleneck and the pull back towards randomness resulting from the possibility of not observing a particular feature value paired with a subsignal.

The severity of transmission bottleneck, $c(\mathcal{E}, e)$, does appear to have an impact on the compositionality of the emergent languages — for example, comparison of Figure 13 with Figures 14 and 15 suggests that the difference between structured and unstructured environments is most pronounced when c is low. However, consideration of all the results taken together suggests that variation in c alone may not explain the observed patterns of behaviour. The results presented here can be split into two groups on gross qualitative grounds:

1. Situations where highly compositional systems are highly frequent for both structured and unstructured environments, with emergent languages in structured environments tending to be slightly more compositional. This occurs in the medium density environments for $c(\mathcal{E}, e) = 0.6$ or 0.8 (Figures 14 and 15), and in dense environments for $c(\mathcal{E}, e) = 0.4, 0.6$ or 0.8 (Figures 17–19)
2. Situations where highly compositional systems emerge with high frequency in structured environments, and the emergent systems in unstructured environments exhibit a range of compositionality, from partial to high. The sparse environment runs where $c(\mathcal{E}, e) = 0.8$ (Figure 12) match this description, as do the results for the medium density environment where $c(\mathcal{E}, e) = 0.4$ (Figure 13) and the dense environment where $c(\mathcal{E}, e) = 0.25$ (Figure 16).

The general trend is that the difference between structured and unstructured environments is at a maximum when $c(\mathcal{E}, e)$ is low, and decreases as $c(\mathcal{E}, e)$ increases. In the presence of *any* bottleneck on cultural transmission ($c < 1$), there will be a pressure for compositional language. There is pressure acting on languages to be generalisable from a subset, discussed above. For high values of c , there will be little difference between structured and unstructured environments. However, for low c a difference will emerge. When c is low a learner will only see a small subset of the language of the previous generation. Provided c is not too low, this is not a problem when the environment is structured — a learner need only observe a few meanings to get an idea of the substring each feature value should map to. However, in unstructured environments low c is more of a problem — some meanings have feature values which are shared with few other meanings and as a consequence the substrings associated with these feature values in a compositional language are prone to being lost during transmission. The pressure for compositionality is counteracted to some extent by randomness reintroduced at each generation to cover feature values which have not been observed. This results in the emergence of partially stable, partially compositional systems. When c gets very low this problem begins to affect structured environments too. Eventually, when c gets low enough, no stable language is possible, regardless of the degree of structure in the environment.

This explanation, based purely on $c(\mathcal{E}, e)$, breaks down when confronted with the results for sparse environments, with high c (Figure 12). While the theory predicts little difference between the compositionality of the languages in structured and unstructured environments, the results show a large difference. The importance of environment structure is *greater* than the theory predicts for that level of c . The theory also fails somewhat to account for the results for dense environments with $c(\mathcal{E}, e) = 0.25$. In this case the theory predicts a wider disparity between structured and unstructured environments than observed in the medium density environment where $c(\mathcal{E}, e) = 0.4$, given the lower value of c . However, the structure of the environment in fact has slightly *less* impact.

A more satisfying analysis can be gained by adapting Brighton’s (2002) equations for calculating the probability of seeing a particular feature value given a particular number of exposures. After e

observations a learner will have accumulated a set of observations of values for the particular i th feature f_i . Let us call this set of observations O_{f_i} . Brighton gives the probability of a particular value v being in this set of observations, $Pr(v \in O_{f_i})$, as:

$$Pr(v \in O_{f_i}) = \sum_{x=1}^N \left\{ \frac{x}{N} \cdot \left(\sum_{\epsilon=1}^e \left(\frac{N-x}{N} \right)^{\epsilon-1} \right) \cdot \left(\frac{(V-1)^{N-x}}{V^N} \right) \cdot \binom{N}{x} \right\}$$

where there are N meanings ($N = |\mathcal{E}|$) and V distinct values for the feature f_i . x in this equation represents the number of objects labelled with v , the feature value of interest. The first two terms of the product

$$\dots \frac{x}{N} \cdot \left(\sum_{\epsilon=1}^e \left(\frac{N-x}{N} \right)^{\epsilon-1} \right) \dots$$

give the probability of seeing at least 1 occurrence of v , given that there are x objects labelled with v . The remainder of the equation simply sums over the probabilities of labelling x out of N objects with v . This part of the equation is not required for our analysis, given that the number of objects labelled with a particular feature value is given by the predefined environment. We can therefore simplify Brighton's equation to:

$$Pr(v \in O_{f_i}) = \frac{x}{N} \cdot \left(\sum_{\epsilon=1}^e \left(\frac{N-x}{N} \right)^{\epsilon-1} \right)$$

where x is simply the number of meanings in the environment \mathcal{E} which have value v for feature f_i . We can then calculate the probability of being able to express a particular meaning $m = (v_1 v_2 v_3)$, where v_1 is the value for f_1 and so on, $Pr(m|O_{f_1}, O_{f_2}, O_{f_3})$:

$$Pr(m|O_{f_1}, O_{f_2}, O_{f_3}) = Pr(v_1 \in O_{f_1}) \cdot Pr(v_2 \in O_{f_2}) \cdot Pr(v_3 \in O_{f_3})$$

In other words, the probability of being able to express a given meaning compositionally is the product of the probabilities of having seen each feature value paired with a subsignal. We can then average $Pr(m|O_{f_1}, O_{f_2}, O_{f_3})$ for all $m \in \mathcal{E}$, to give the average probability of being able to produce an utterance compositionally. Table 2 compares the values of $\overline{Pr}(m)$ (the value of $Pr(m|O_{f_1}, O_{f_2}, O_{f_3})$, averaged over all meanings in \mathcal{E}) for structured and unstructured environments, for various values of e .

The difference between values of $\overline{Pr}(m)$ for structured and unstructured environments provide a useful measure of the relative stability advantage of compositional language in structured environments over compositional language in unstructured environments. This size of this value corresponds fairly well to the differences between the final systems in structured and unstructured environments observable in Figures 12–19. For example, the observable difference is greatest in Figure 12, followed by Figure 13, and these two settings of environment density and e yield the largest and second-largest differences in the Table.

2.3 Summary

To summarise the results presented so far, it has been shown that environment density, environment structure, and bottleneck impact on the cultural evolution of compositionality. In the absence of a bottleneck, highly compositional language is unlikely to evolve. Highly compositional languages only evolve in structured environments, due to the increased potential for spread of compositionality

Density	e	c	Difference
sparse	19	0.8	6×10^{-2}
medium	16	0.4	9×10^{-2}
medium	28	0.6	9×10^{-3}
medium	49	0.8	2×10^{-4}
dense	18	0.25	4×10^{-2}
dense	32	0.4	2×10^{-3}
dense	57	0.6	1×10^{-5}
dense	101	0.8	2×10^{-9}

Table 2: Comparison of $\overline{Pr}(m)$ for structured and unstructured environments of various densities, for various values of e . The Difference column gives $\overline{Pr}(m)$ for structured environments minus $\overline{Pr}(m)$ for unstructured environments, and is a measure of the relative stability advantage of compositional systems in the structured environment — the greater the difference, the greater the stability advantage of compositionality in structured environments.

arising from the large number of shared feature values between meanings. The emergence of such systems is highly sensitive to the initial, random assignment of signals to meanings.

In the presence of a bottleneck on cultural transmission, highly compositional languages reliably emerge from initially random, holistic mappings, in both structured and unstructured environments. This is due to the pressure on the evolving languages to be generalisable, introduced by the transmission bottleneck. Compositional languages emerge most frequently in structured environments, as generalisations in such environments have a higher yield than in unstructured environments. Finally, the advantage of compositional language in structured environments over compositional language in unstructured environments can be quantified by applying Brighton’s equations for the probability of observing feature values given a certain number of exposures. In structured environments, the high degree of shared structure between meanings increases the probability that the subsignals paired with each feature value will have been seen after a small number of observations.

3 Exploring the impact of learning bias

In the previous section the impact of environment structure and bottleneck on the cultural evolution of compositional language was investigated. The learning bias of the associative network was kept constant for all these experiments — all results reported were for the case where every network used the weight-update rule $(1 \ - \ 1 \ - \ 1 \ 0)$. In this Section, I will investigate the impact of using different weight-update rules, with different associated learning biases, while keeping the severity of the transmission bottleneck and the degree of environment density and structure constant.

The investigations outlined in this Section will roughly follow the format of the experiments outlined in Smith (2002a) — the ability of networks with various weight-update rules to acquire, maintain and construct an optimal system will be explored in Sections 3.1–3.3. In Section 4 I will describe the key properties of the learning bias required to acquire, maintain and construct optimal compositional languages.

Value	Feature		
	1	2	3
1	j	e	h
2	h	i	f
3	a	c	e
4	b	a	d
5	e	d	b

Table 3: A feature value lookup table for a compositional language. The signal characters are concatenated in the order of the feature values — for example, (1 1 1) would be expressed as *je h*.

3.1 Acquisition of a compositional system

The first issue is to ascertain whether individual agents, in isolation, can acquire a perfectly compositional, unambiguous communication system. Such a language, \mathcal{L} , was constructed according to the feature value–character mapping given in Table 3. As in the previous section, $F = 3$, $V = 5$, $l_{max} = 3$ and $\Sigma = \{a, b, c, d, e, f, g, h, i, j\}$. The medium density, structured environment illustrated in Figure 6 (c) provided the set of meanings \mathcal{E} . With respect to this environment (or indeed any other), L is perfectly e-compositional — $E(\mathcal{L}) = 1$.

Agents using each of the 81 possible weight-update rules ($\alpha, \beta, \gamma, \delta \in [-1, 1]$) were then trained on \mathcal{L} , by storing each meaning-signal pair in \mathcal{L} in their network. The agents were then evaluated to see if they a) successfully acquired the meaning-signal mapping in \mathcal{L} and b) would reproduce \mathcal{L} in an i-compositional or i-holistic manner.

Agents were judged to have acquired the meaning-signal mapping if, for every $\langle m_i, s_j \rangle \in \mathcal{L}$, both:

- Production of the signal associated with m_i always³ resulted in s_j being produced, i.e. $\langle m_i, s_j \rangle$ can be reproduced in production.
- **and** reception of s_j always resulted in the interpretation m_i , i.e. $\langle m_i, s_j \rangle$ can be reproduced in reception, meaning that the agent would communicate optimally with itself or another agent using the same weight-update rule exposed to \mathcal{L} .

18 of the 81 possible rules succeeded in the acquisition task, with the remaining 63 failing to reproduce the observed system. These 18 successful rules were classified as [+maintainer, \pm constructor]⁴ in the simple associative network tests outlined in Smith (2002a).

Note that there is both a degree of continuity with the previous classification — the same 18 rules are separated out in both classifications — and a discontinuity — 31 rules were able to acquire unambiguous systems in the previous classification, whereas in the new classification only 18 can. Rules which were [+learner, –maintainer] in the earlier categorisation⁵ are incapable of acquiring \mathcal{L} . This is due to the fact that the learning procedure for the structured communication systems requires that learners be able to handle multiple active input nodes when learning — all components

³As before, the term “always” is reduced to “for every one of 1000 trials”.

⁴I will avoid where possible specifying redundant features of weight-update rules — [+maintainer] implies [+learner].

⁵Bear in mind that the terms +learner, +maintainer and so on are based on the classification given in Smith (2002a), and are shorthands for a set of restrictions on the relationship between the values of α, β, γ and δ in the weight-update rules. While the mnemonic of +learner was quite transparent in Smith (2002a), it is less so here — certain weight-update rules which fit the pattern glossed as +learner cannot learn a compositional system. However, these terms will be preserved, for reasons which will become obvious.

of a meaning and signal are presented simultaneously to the learner. As discussed in Smith (2002a), [+learner, –maintainer] weight-update rules are biased in favour of many-to-one mappings, and this bias, given the multiple active units present in each exposure, leads to an inability to acquire \mathcal{L} .

Of those rules which were classified as capable of acquiring \mathcal{L} , a further evaluation was made as to whether they reproduced their acquired mapping in an i-compositional or non-i-compositional manner. Agents using the 18 [+maintainer] rules were trained on \mathcal{L} , as before. They were then called upon to produce for each $m \in \mathcal{E}$ to give a set of meaning-signal pairs, with an associated underlying set of winning meaning analysis–signal analysis pairs \mathcal{A}_p . Similarly, they were prompted with each s used in \mathcal{L} , to yield a set of meaning analysis-signal analysis pairs \mathcal{A}_r . \mathcal{A}_p and \mathcal{A}_r were evaluated according to the internal compositionality measure given in Section 1.5. Weight-update rules which yielded sets of analysis pairs where $I(\mathcal{A}_p) = I(\mathcal{A}_r) = 1$ were classified as [+ic-preserver], otherwise they were classified as [–ic-preserver].

7 weight-update rules were classified as [+ic-preserver], of which 2 were classified as [+constructor] and 5 were classified as [+maintainer, –constructor] in the previous classification. The remaining 11 rules (7 [+constructor] and 4 [+maintainer, –constructor]) were classified as [–ic-preserver].

3.2 Maintenance through a bottleneck

Next, maintenance tests, similar to those outlined for the associative network model in Smith (2002a), were carried out to assess the ability of the various weight-update rules to maintain an optimal system. In Smith (2002a) the maintenance tests assessed whether populations of agents using the weight-update rules were able to acquire an optimal system in the presence of noise, without a bottleneck. The maintenance test here measures the ability of populations of agents using these weight-update rules to preserve the optimal system through a bottleneck on cultural transmission.

Recall from the description of the ILM given in Section 1.3 above that the agents in the initial population use some predefined communication system L . For the experiments outlined in this section, the initial population’s set of weights \mathcal{W} were constructed such that the $p(m)$ of the initial L generates the unambiguous, perfectly e-compositional meaning-signal pairs encoded in \mathcal{L} , described above — in other words, the initial language in these simulations is predefined and perfectly compositional. ILMs were run with each of the 81 possible learning rules, with each learning receiving 28 exposures to the communication system of the previous generation ($e = 28$, $c(\mathcal{E}, e) = 0.6$). Populations were defined as having maintained a compositional system if $E(\mathcal{O})$ and $I(\mathcal{A})$ remained above 0.95 for every generation of ten 100 generation runs.

No [–maintainer] rules succeeded in this task. All [+maintainer, –ic-preserver] weight-update rules exhibited behaviour similar to run (a) in Figure 20, and failed to maintain the perfectly compositional system. Of the seven [+maintainer, +ic-preserver] rules, five behaved in a similar fashion to run (b) in Figure 20. Only two [+maintainer, +ic-preserver] weight-update rules succeeded in maintaining a compositional system. Populations using these two weight-update rules behaved like run (c) in Figure 20.

The five [+ic-preserver] rules which failed to maintain the perfectly compositional system were of the [+maintainer, –constructor] classification, whereas the two [+ic-preserver] rules which maintained the perfectly compositional system were of the [+constructor] classification. The one-to-one bias associated with [+constructor] rules is clearly crucial in maintaining a perfectly compositional system.

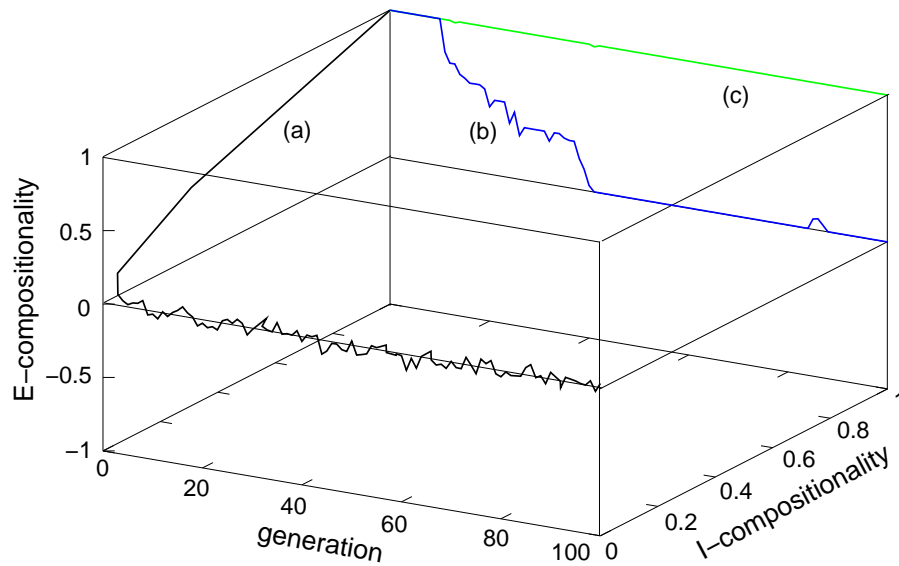


Figure 20: Three characteristic patterns of behaviour in populations using the 81 weight-update rules when attempting to maintain a perfectly compositional language. The population in run (a) rapidly collapses from using the initial system to an i- and e-holistic system. This behaviour characterises [+maintainer, -ic-preserver] rules. The population in run (b), which is using a weight-update rule classified as [+maintainer, -constructor, +ic-preserver], loses the initial language, although its final language is i-compositional. The population in run (c), which characterises [+constructor, +ic-preserver] weight-update rules, maintains the perfectly compositional initial language.

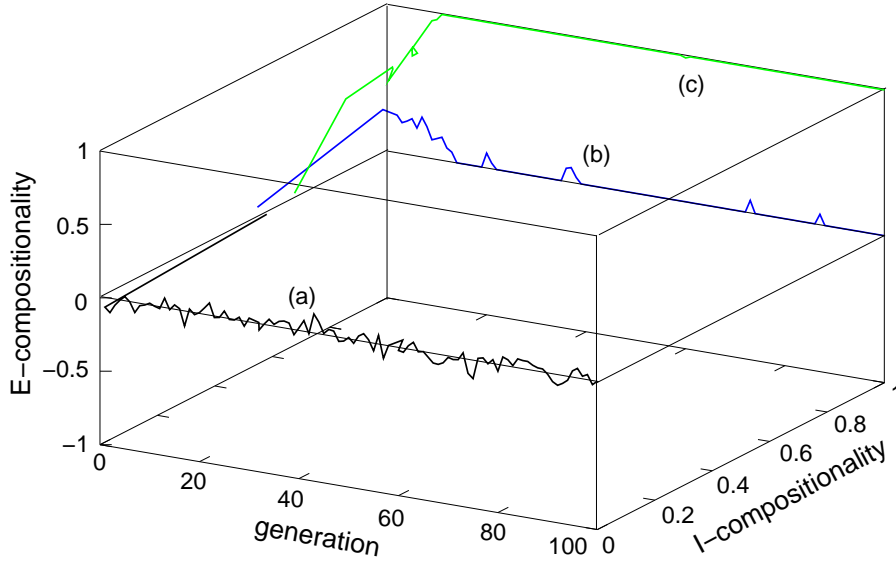


Figure 21: Three characteristic patterns of behaviour in populations using the 81 weight-update rules when attempting to construct a compositional language. The population in run (a) converges to an i- and e-holistic system. This behaviour characterises [+maintainer, -ic-preserver] rules. The population in run (b), which is using a weight-update rule classified as [+maintainer, -constructor, +ic-preserver] converges on a language which is i-compositional but e-holistic. The population in run (c), which characterises the two [+constructor, +ic-preserver] weight-update rules, constructs a language which is highly i- and e-compositional.

3.3 Construction through a bottleneck

Finally, the 81 weight-update rules were tested to see whether they could construct a compositional system from an initially random, holistic system, in the presence of a bottleneck on transmission.

In the previous section the initial population’s communication system, L , was perfectly compositional. In the ILMs outlined in this section the connection weights of every individual in the initial population are set to 0, resulting in an initial L with maximum entropy. As with the maintenance simulations outlined in the previous Section, each learner receives 28 exposures to the communication system of the previous generation ($e = 28$, $c(\mathcal{E}, e) = 0.6$). Populations were defined as having constructed a compositional system if $E(\mathcal{O})$ and $I(\mathcal{A})$ rose above 0.95 in every one of ten 100 generation runs.

Perhaps unsurprisingly, only the two weight-update rules which were capable of maintaining a preexisting optimal system were capable of constructing a compositional system from random initial behaviour. As mentioned earlier, these rules have the [+constructor, +ic-preserver] classification. Populations using these weight-update rules behaved in a similar fashion to population (c) in Figure 21. The [+maintainer, -ic-preserver] and [+maintainer, -constructor, +ic-preserver] weight-update rules were incapable of constructing an optimal system, and behaved similarly to population (a) and (b) respectively in Figure 21.

3.4 The classification hierarchy

The numbers of weight-update rules with the various complete classifications are given in Table 4.

Classification	Number	Acquire?	Acquire i-compositionally?	Maintain?	Construct?
[−learner, −maintainer, −constructor, −ic-preserver]	50	no	no	no	no
[+learner, −maintainer, −constructor, −ic-preserver]	13	no	no	no	no
[+learner, +maintainer, −constructor, −ic-preserver]	4	yes	no	no	no
[+learner, +maintainer, −constructor, +ic-preserver]	5	yes	yes	no	no
[+learner, +maintainer, +constructor, −ic-preserver]	7	yes	no	no	no
[+learner, +maintainer, +constructor, +ic-preserver]	2	yes	yes	yes	yes

Table 4: The number of weight-update rules of each particular complete classification, from the sample of 81, and a summary of their properties with respect to compositional languages. “Acquire?” indicates whether agents using rules with this classification can acquire a perfectly compositional language. “Acquire i-compositionally?” indicates whether they can reproduce an acquired system in a perfectly i-compositional manner. “Maintain?” indicates whether agents using a weight-update rule with this classification can maintain a perfectly compositional language through a bottleneck, in the context of the ILM. “Construct?” indicates whether such agents can construct a highly compositional system from random initial behaviour in the presence of a bottleneck, in the context of the ILM.

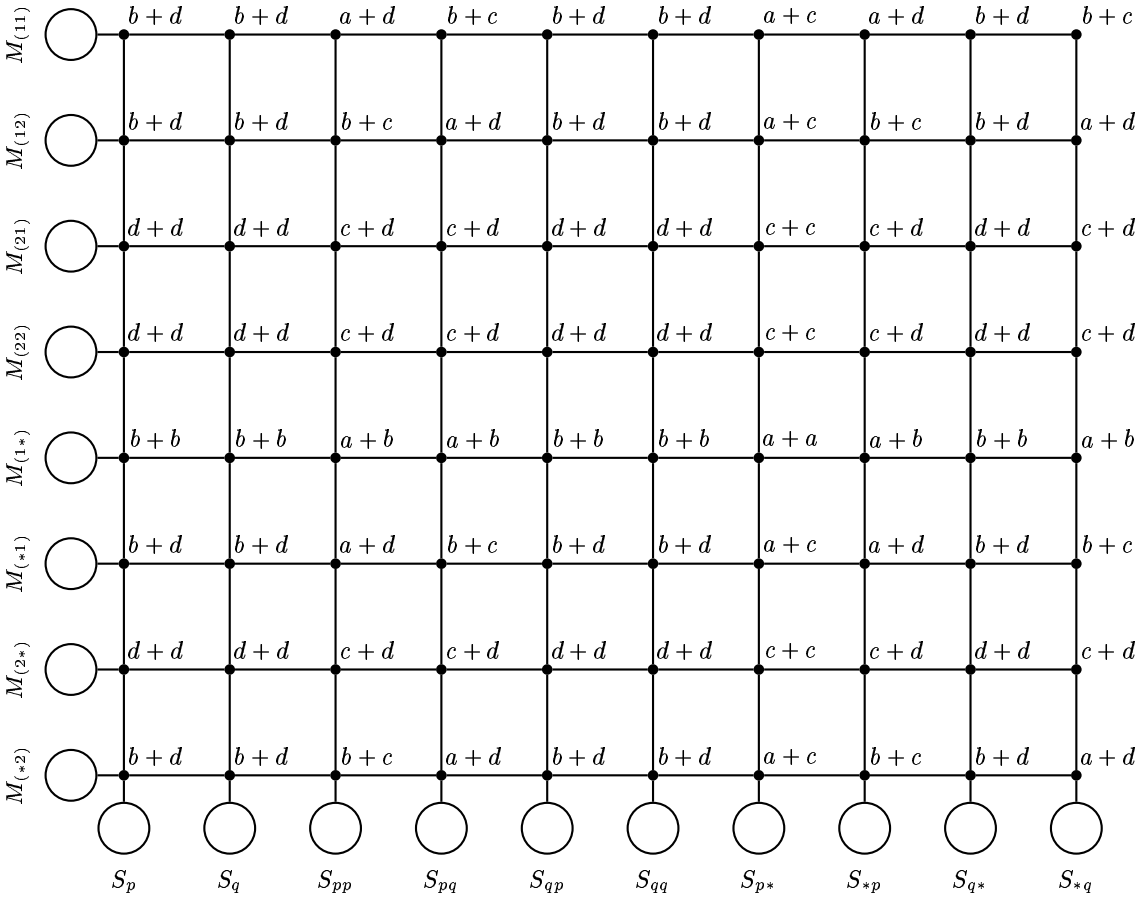


Figure 22: Connection weights after observing and learning the meaning-signal pairs $\langle (1\ 1), pp \rangle$ and $\langle (1\ 2), pq \rangle$ using weight-update rule $(a\ b\ c\ d)$.

4 The key bias

What pattern of assignment of values to α , β , γ and δ results in some weight-update rules being able to construct compositional languages from scratch, whereas other weight-update rules cannot maintain or learn such a system? In Smith (2002a) this type of question was tackled by considering the connection weights in a small network before and after exposure to a single meaning-signal pair. A similar strategy is pursued here. I will consider the simple case where $F = V = 2$, $l_{max} = 2$, $\Sigma = \{p, q\}$. A network of appropriate dimensions was trained on two meaning-signal pairs, $\langle (1\ 1), pp \rangle$ and $\langle (1\ 2), pq \rangle$, using the weight-update rule $(a\ b\ c\ d)$. The connection weights in this network after observing these two meaning-signal pairs are given in Figure 22.

4.1 An overview of the learning biases

In this Section I will focus on the g values for the various possible analyses in the network as a whole, in fairly broad terms, returning in more detail to smaller parts of the network in Section 4.2.

The g for a particular meaning analysis–signal analysis pair depends on one or two connection weights from the network shown in Figure 22. Table 5 gives the g values for various meaning analysis–signal analysis pairs. To simplify matters, only one possible ordering of components in the signal is given in the Table. For example, there are three possible analyses of the signal pp — $\{pp\}$, $\{p*, *p\}$ and $\{*p, p*\}$, but only the first two are included in Table 5.

4.1.1 [+maintainer, –constructor, ±ic-preserver] rules

It has already been established that [+maintainer] rules are characterised by the restriction:

A weight-update rule is [+maintainer] if $\alpha > \beta \wedge \delta \geq \gamma$

Such rules are neutral with respect to one-to-one mappings between meanings and signals (if $\delta = \gamma$) or biased in favour of one-to-one mappings (if $\delta > \gamma$, which yields the classification [+maintainer,+constructor]). Is there any pattern of assignment of values to α , β , γ and δ which distinguishes these rules on the [\pm ic-preserver] feature? Yes.

A weight-update rule is [+maintainer,+ic-preserver] if
 $\alpha > \beta \wedge \delta \geq \gamma \wedge \alpha > \delta$

Why does this pattern of weight changes lead to the network exploiting the internal compositional representations? I will focus first on weight-update rules which are characterised as [+maintainer,–constructor]. For these rules, $\delta = \gamma$. Table 6 gives the g values for various analyses for [+maintainer, –constructor, ±ic-preserver] rules.

Tables 6 (a) and (b) highlight the relevant values of g for [+maintainer, –constructor, –ic-preserver] rules. For these rules, $\alpha < \delta$ (as is the case in Table 6 (a)) or $\alpha = \delta$ (as is the case in Table 6 (b)). In the former case, the networks are strongly biased against compositional systems — the compositional representational capacities of the network are not exploited due to the fact that δ dominates α . In the latter case, the networks are neutral with respect to compositionality — both one- and two-component analyses are possible, due to the fact that $\alpha = \delta$. Finally, Table 6 (c) highlights the relevant values of g for [+maintainer,–constructor,+ic-preserver] rules. For these rules, α dominates δ . Consequently, analyses involving multiple components are preferred to those involving single components, indicating a bias in favour of i-compositional systems.

These weight-update rules, as we might expect, are neutral with respect to the one-to-one nature of the mapping between meanings and signals. For the meanings (2 1) and (2 2), which have not been observed paired with any signal, there are several possible candidate signals, including pp and pq , which have already been observed paired with a meaning.

4.1.2 [+constructor, ±ic-preserver] rules

[+constructor] rules are characterised by the restriction:

A weight-update rule is [+constructor] if $\alpha > \beta \wedge \delta > \gamma$

Such rules are biased in favour of one-to-one mappings between meanings and signals. [+constructor, +ic-preserver] rules are characterised as:

Meaning	Signal analysis									
	$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p*, *p\}$	$\{p*, *q\}$	$\{q*, *p\}$	$\{q*, *q\}$
(1 1)	$b+d$	$b+d$	$a+d$	$b+c$	$b+d$	$b+d$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+b+b+d)$	$\frac{1}{2}(b+b+b+c)$
(1 2)	$b+d$	$b+d$	$b+c$	$a+d$	$b+d$	$b+d$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(b+b+b+c)$	$\frac{1}{2}(a+b+b+d)$
(2 1)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(a+c+d+d)$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+d+d+d)$	$\frac{1}{2}(b+c+d+d)$
(2 2)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+c+c+d)$	$\frac{1}{2}(b+c+d+d)$	$\frac{1}{2}(a+d+d+d)$

Table 5: The g values of various meaning analysis -signal analysis pairs in the network after storing $\langle(1\ 1), pp\rangle$ and $\langle(1\ 2), pq\rangle$ using learning rule $(a\ b\ c\ d)$. Meanings which have not been observed paired with a signal are given in italics. Note that only one ordering of two-component signal analyses is given — there are in fact another 4 two-component signal analyses ($\{*p, p*\}$, $\{*q, p*\}$, $\{*p, q*\}$, $\{*q, q*\}$), but these other analyses can be safely ignored for the purposes of this analysis.

Meaning	Signal analysis									
	$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p*,*p\}$	$\{p*,*q\}$	$\{q*,*p\}$	$\{q*,*q\}$
(1 1)	$b + D$	$b + D$	$a + D$	$b + D$	$b + D$	$b + D$	$\frac{1}{2}(a + a + a + D)$	$\frac{1}{2}(a + a + b + D)$	$\frac{1}{2}(a + b + b + D)$	$\frac{1}{2}(b + b + b + D)$
(1 2)	$b + D$	$b + D$	$b + D$	$a + D$	$b + D$	$b + D$	$\frac{1}{2}(a + a + b + D)$	$\frac{1}{2}(a + a + a + D)$	$\frac{1}{2}(b + b + b + D)$	$\frac{1}{2}(a + b + b + D)$
(2 1)	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$
(2 2)	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$

Meaning	Signal analysis									
	$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p*,*p\}$	$\{p*,*q\}$	$\{q*,*p\}$	$\{q*,*q\}$
(1 1)	$b + D$	$b + D$	$a + D$	$b + D$	$b + D$	$b + D$	$\frac{1}{2}(a + a + a + D)$	$\frac{1}{2}(a + a + b + D)$	$\frac{1}{2}(a + b + b + D)$	$\frac{1}{2}(b + b + b + D)$
(1 2)	$b + D$	$b + D$	$b + D$	$a + D$	$b + D$	$b + D$	$\frac{1}{2}(a + a + b + D)$	$\frac{1}{2}(a + a + a + D)$	$\frac{1}{2}(b + b + b + D)$	$\frac{1}{2}(a + b + b + D)$
(2 1)	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$
(2 2)	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$

Meaning	Signal analysis									
	$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p*,*p\}$	$\{p*,*q\}$	$\{q*,*p\}$	$\{q*,*q\}$
(1 1)	$b + D$	$b + D$	$a + D$	$b + D$	$b + D$	$b + D$	$\frac{1}{2}(a + a + a + D)$	$\frac{1}{2}(a + a + b + D)$	$\frac{1}{2}(a + b + b + D)$	$\frac{1}{2}(b + b + b + D)$
(1 2)	$b + D$	$b + D$	$b + D$	$a + D$	$b + D$	$b + D$	$\frac{1}{2}(a + a + b + D)$	$\frac{1}{2}(a + a + a + D)$	$\frac{1}{2}(b + b + b + D)$	$\frac{1}{2}(a + b + b + D)$
(2 1)	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$
(2 2)	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$D + D$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$	$\frac{1}{2}(b + D + D + D)$	$\frac{1}{2}(a + D + D + D)$

Table 6: The g values for meaning analysis–signal analysis pairs in the network after training on two meaning-signal pairs using the weight-update rule ($a b c d$). This table focuses on [+maintainer, –constructor, \pm ic-preserver] rules, for which $c = d = D$ and $a > b$. The highest g values in each row are highlighted in grey. Meanings which have not been observed paired with a signal are given in italics. (a) The [+maintainer, –constructor, –ic-preserver] rule where $D > a$. The observed meaning-signal mappings can be reproduced. Only the one-component analyses are used, indicating a bias against i-compositionality. Meanings which have not been observed map to any single-component analysis with equal probability, indicating no bias against many-to-one meaning-signal mappings. (b) The [+maintainer, –constructor, –ic-preserver] rule where $a = D$. The observed meaning-signal mappings can be reproduced, with both one-component and two-component analyses being equally probable. This indicates neutrality with respect to i-compositionality. The non-observed meanings map to both one- and two-component analyses, again with no bias against many-to-one mappings. (c) The [+maintainer, –constructor, +ic-preserver] rule where $a > D$. The observed meaning-signal pairs are reproduced using two-component analyses, indicating a bias in favour of i-compositionality. Signals for the unobserved meanings are also produced using two-component analyses, but once again many-to-one mappings are not avoided.

A weight-update rule is [+constructor,+ic-preserver] if
 $\alpha > \beta \wedge \delta > \gamma \wedge \alpha > \delta$

Why does this pattern of weight changes lead to the network exploiting the internal compositional representations? Tables 7 (a) and (b) highlight the relevant values of g for [+constructor,−ic-preserver] rules. For these rules, $\alpha < \delta$ (as is the case in Table 7 (a)) or $\alpha = \delta$ (as is the case in Table 7 (b)). The parallels with the [+maintainer, −constructor, −ic-preserver] rules are clear. In the case where $\alpha < \delta$, the networks are strongly biased against i-compositional systems — the compositional representational capacities of the network are not exploited due to the fact that δ dominates α . In the case where $\alpha = \delta$, the networks are neutral with respect to i-compositionality — both one- and two-component analyses are possible. Finally, Table 7 (c) highlights the relevant values of g for [+constructor,+ic-preserver] rules. For these rules, α dominates δ . Consequently, analyses involving multiple components are preferred to those involving single components, indicating a bias in favour of i-compositional systems.

These weight-update rules are also biased in favour of one-to-one mappings between meanings and signals. This is reflected in the possible productions for the meanings (2 1) and (2 2), which have not been observed paired with any signal. For the [+constructor,−ic-preserver] rules there are several candidate signals. However, the signals pp and pq are ruled out, indicating a bias against many-to-one mappings between meanings and signals. For the [+constructor,+ic-preserver] weight-update rules, there is a single candidate signal for each of the non-observed meanings — an unambiguous, compositional system is constructed based on the exposure to the two meaning-signal pairs. This arises from the network’s one-to-one bias. This highlights the need for the alphabet to be larger than the number of values for each feature ($|\Sigma| > V$), as is the case in all the simulation results reported in this report. If this is not the case, then the one-to-one bias of [+constructor, +ic-preserver] agents allows them to reliably reconstruct the signal character associated with feature values they have not actually seen, provided that they have seen all other values for that feature. When $|\Sigma| > V$ this cannot be done reliably.

4.2 The two parts of the bias

The maintenance or construction of a compositional language through a bottleneck in a population of networks requires two elements. Firstly, the networks must be able to make generalisations from observed meaning-signal pairs to meanings which have not been observed. This requires that the compositional representational capacity of the networks is exploited — signals must be produced using multi-component analyses, in an internally-compositional fashion. If the networks consistently produce using single-component analyses, in an internally-holistic fashion, then generalisation from seen to unseen meanings is impossible.

In addition to this, there must be a principled system of mapping from feature values of meanings to signal characters. If, for example, a network produces utterances using multi-component analyses, but every distinct feature value maps onto the same signal character then an e-compositional system will not be constructed or maintained — the resulting, highly ambiguous system will not preserve neighbourhood relationships when mapping between meanings and signals.

The learning biases of the weight-update rules can be characterised along these two distinct dimensions — a bias in favour of (or against) exploiting internally compositional representations, and a bias in favour of (or against) a principled system of mappings from feature values to signal characters. Only when the correct biases for both aspects of the problem are in place will compositional languages be maintained or constructed.

(a)	Meaning	Signal analysis									
		$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p^*, *p\}$	$\{p^*, *q\}$	$\{q^*, *p\}$	$\{q^*, *q\}$
(1 1)	$b+d$	$b+d$	$a+d$	$b+c$	$b+d$	$b+d$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+b+b+d)$	$\frac{1}{2}(b+b+b+c)$	
(1 2)	$b+d$	$b+d$	$b+c$	$a+d$	$b+d$	$b+d$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(b+b+b+c)$	$\frac{1}{2}(a+b+b+d)$	
(2 1)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(a+c+d+d)$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+d+d+d)$	$\frac{1}{2}(b+c+d+d)$	
(2 2)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+c+c+d)$	$\frac{1}{2}(b+c+d+d)$	$\frac{1}{2}(a+d+d+d)$	

(b)	Meaning	Signal analysis									
		$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p^*, *p\}$	$\{p^*, *q\}$	$\{q^*, *p\}$	$\{q^*, *q\}$
(1 1)	$b+d$	$b+d$	$a+d$	$b+c$	$b+d$	$b+d$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+b+b+d)$	$\frac{1}{2}(b+b+b+c)$	
(1 2)	$b+d$	$b+d$	$b+c$	$a+d$	$b+d$	$b+d$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(b+b+b+c)$	$\frac{1}{2}(a+b+b+d)$	
(2 1)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(a+c+d+d)$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+d+d+d)$	$\frac{1}{2}(b+c+d+d)$	
(2 2)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+c+c+d)$	$\frac{1}{2}(b+c+d+d)$	$\frac{1}{2}(a+d+d+d)$	

(c)	Meaning	Signal analysis									
		$\{p\}$	$\{q\}$	$\{pp\}$	$\{pq\}$	$\{qp\}$	$\{qq\}$	$\{p^*, *p\}$	$\{p^*, *q\}$	$\{q^*, *p\}$	$\{q^*, *q\}$
(1 1)	$b+d$	$b+d$	$a+d$	$b+c$	$b+d$	$b+d$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+b+b+d)$	$\frac{1}{2}(b+b+b+c)$	
(1 2)	$b+d$	$b+d$	$b+c$	$a+d$	$b+d$	$b+d$	$\frac{1}{2}(a+a+b+c)$	$\frac{1}{2}(a+a+a+d)$	$\frac{1}{2}(b+b+b+c)$	$\frac{1}{2}(a+b+b+d)$	
(2 1)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(a+c+d+d)$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+d+d+d)$	$\frac{1}{2}(b+c+d+d)$	
(2 2)	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$\frac{1}{2}(b+c+c+c)$	$\frac{1}{2}(a+c+c+d)$	$\frac{1}{2}(b+c+d+d)$	$\frac{1}{2}(a+d+d+d)$	

Table 7: The g values for meaning analysis–signal analysis pairs in the network after training on two meaning–signal pairs using the weight-update rule $(a b c d)$. This table focuses on $[+constructor, \pm ic\text{-preserver}]$ rules, for which $a > b$ and $d > c$. (a) The $[+constructor, -ic\text{-preserver}]$ rule where $d > a$. The observed meaning–signal mappings can be reproduced. Only the one-component analyses are used, indicating a bias against i-compositionality. Meanings which have not been observed map to signals which have not been observed paired with any meaning, indicating a bias in favour of one-to-one mappings. (b) The $[+constructor, -ic\text{-preserver}]$ rule where $d = a$. The observed meaning signal mappings can be reproduced, with both one-component and two-component analyses being equally probable. This indicates neutrality with respect to i-compositionality. As with (a), non-observed meanings map to non-observed signals, indicating a one-to-one bias. However, both one- and two-component analyses are possible. (c) The $[+constructor, +ic\text{-preserver}]$ rule where $a > d$. The observed meaning–signal pairs are reproduced using two-component analyses, indicating a bias in favour of i-compositionality. As before, non-observed meanings map to non-observed signals, with only the two component analyses being used. The mapping is in fact perfectly one-to-one.

Relationship	Average $I(\mathcal{A})$	Average $I(\mathcal{A}_p)$	Average $I(\mathcal{A}_r)$
$\alpha > \delta$	0.98	0.98	0.98
$\alpha = \delta$	0.47	0.48	0.47
$\alpha < \delta$	0.16	0.17	0.14

Table 8: Average internal compositionality of production and reception behaviour combined ($I(\mathcal{A})$), production behaviour alone ($I(\mathcal{A}_p)$) and reception behaviour alone ($I(\mathcal{A}_r)$). Results are for all 81 weight-update rules, organised according to the relationship between α and δ

4.2.1 The internal compositionality bias

Comparison of Sections 4.1.1 and 4.1.2 reveals a common element to the rules which are [+maintainer, \pm constructor, +ic-preserver]:

A weight-update rule is [+maintainer, +ic-preserver] if
 $\alpha > \beta \wedge \delta \geq \gamma \wedge \alpha > \delta$

The $\alpha > \beta \wedge \delta \geq \gamma \dots$ part of this constraint relates to the [+maintainer] bias, which is a bias regarding the one-to-one nature of the mapping between meanings and signals. The $\dots \alpha > \delta$ part dictates that the i-compositional representational capacity of the network will be exploited. Both components of the bias are required in order to be classified as [+maintainer, +ic-preserver], as the [+ic-preserver] classification requires that the networks can reproduce an observed, unambiguous language. However, it is possible to abstract away from the actual meaning-signal mapping acquired, and investigate whether weight-update rules exploit the compositional representational capacity. This will reveal whether, as hypothesised, $\alpha > \delta$ leads to the use of multi-component analyses.

The acquisition tests outlined in Section 3.1 were repeated for all 81 weight-update rules. However, on this occasion the ability to reproduce the observed mapping was ignored, and the internal compositionality of their production ($I(\mathcal{A}_p)$) and reception ($I(\mathcal{A}_r)$) behaviour was measured. The results are summarised in Table 8.

This Table shows that the relationship between α and δ largely determines i-compositionality during production and reception. When $\alpha > \delta$ internal compositionality is high, indicating that the compositional representational capacities of the network are being exploited. When $\alpha = \delta$ internal compositionality is at an intermediate level, indicating that both the holistic and compositional representational capacities are used with approximately equal frequency. When $\alpha < \delta$ internal compositionality is low, indicating that holistic representations are preferred. The results for $\alpha > \delta$ and $\alpha < \delta$ are less clear cut than when the analysis is restricted to [+maintainer] rules, due to some of the more esoteric weight-update rules. However, the main point still stands:

Networks will tend to behave in an i-compositional manner if $\alpha > \delta$

4.2.2 The one-to-one bias

As discussed above in Sections 4.1.1 and 4.1.2, the biases of weight-update rules identified in Smith (2002a) carry over into the more complex model — [+maintainer, $-$ constructor, \pm ic-preserver] agents are neutral with respect to the one-to-one nature of the meaning-signal mapping, whereas [+constructor, \pm ic-preserver] agents are biased in favour of acquiring one-to-one mappings.

The relative values of α and δ determine whether an agent produces or receives in an internally-compositional manner. Parallel to this are the (α, β) and (δ, γ) relationships, which determines the

Meaning	Seen?	Signal					
		p	q	pp	pq	qp	qq
(1 1)	yes	b	b	a	b	b	b
(1 2)	no	d	d	c	d	d	d
(2 1)	no	d	d	c	d	d	d
(2 2)	no	d	d	c	d	d	d

Meaning	Seen?	Signal					
		p	q	pp	pq	qp	qq
(1 1)	yes	b	b	a	b	b	b
(1 2)	no	D	D	D	D	D	D
(2 1)	no	D	D	D	D	D	D
(2 2)	no	D	D	D	D	D	D

Meaning	Seen?	Signal					
		p	q	pp	pq	qp	qq
(1 1)	yes	b	b	a	b	b	b
(1 2)	no	d	d	c	d	d	d
(2 1)	no	d	d	c	d	d	d
(2 2)	no	d	d	c	d	d	d

Table 9: (a) The g values for i-holistic meaning analysis–signal analysis pairs after learning the meaning–signal pair $\langle (1\ 1), pp \rangle$ using the weight-update rule $(a\ b\ c\ d)$. (b) The g values for a network using a [+maintainer, –constructor, \pm ic-preserver] weight-update rule. In such rules $a > b$ and $c = d = D$. The highest g value in each row is highlighted in grey. The observed meaning–signal pair can be reproduced, but there is no bias against many-to-one mappings from meanings to signals. (c) The g values for a network using a [+constructor, \pm ic-preserver] weight-update rule. In such rules $a > b$ and $d > c$. The observed meaning–signal pair can be reproduced, and there is a bias against many-to-one mappings from meanings to signals — signal pp is avoided for all meanings apart from $(1\ 1)$.

bias with respect to the one-to-one quality of mappings. In other words, the (α, β) and (δ, γ) relationships determine whether internally-holistic and internally-compositional mappings are one-to-one or not, then the (α, δ) relationship determines which of the internally-holistic or internally-compositional analyses is actually used.

The learning biases with respect to holistic analyses and compositional analyses can therefore be looked at separately. In the previous Sections connection weights in a network exposed to two meaning-signal pairs were examined. In this Section it is sufficient to look at a network, identical in structure to the network given in Figure 22, which has been trained on the meaning-signal pair $\langle (1\ 1), aa \rangle$. The g values of interest for the i-holistic analyses are summarised in Table 9 (a).

Table 9 (b) highlights the dominant connection weights for [+maintainer, –constructor, \pm ic-preserver] weight-update rules. The observed meaning-signal pair can be reproduced holistically. The other meanings map to every possible holistic analysis with equal probability, including the already-observed signal pp . This indicates neutrality with respect to the one-to-one nature of the meaning-signal mapping, and is an identical result to that reported in Smith (2002a) for the simpler associative

network model.

Table 9 (c) highlights the dominant connection weights in a network using a [+constructor, ±ic-preserver] weight-update rule. The observed meaning-signal pair $\langle (1\ 1), aa \rangle$ can be reproduced. The other meanings map to p , q , pq , qp or qq with equal probability — pp is avoided. The results for the associative network model outlined in Smith (2002a) still hold in the more complex model — [+constructor] agents are biased in favour of acquiring one-to-one mappings.

[−maintainer] weight-update rules, as discussed in Smith (2002a), either cannot acquire observed holistic meaning-signal mappings (in the case of [−learner, −maintainer] rules) or can acquire such systems but are biased in favour of many-to-one mappings between meanings and signals (in the case of [+learner, −maintainer] rules). In the context of the i-holistic analyses part of structured networks, this has the consequence that networks using such rules either:

- cannot reliably reproduce the mapping from $(1\ 1)$ to pp or
- can reliably reproduce this mapping, but prefer to produce pp for all unobserved meanings.

In either case, an e-compositional system cannot be maintained or constructed.

Table 10 (a) gives the weights of the connections in the network between partially-specified components of meaning (organised according to the feature value which is specified) and partially-specified signal components (once again, different orderings are ignored — the subsignals given in the Table refer to signal components where the specified character is in the same position in the signal as the specified feature value — for example, row 1 column 1 of the Table gives the connection weight between $\{(1\ *)\}$ and $\{p*\}$).

Table 10 (b) highlights the dominant connection weights in the relevant portion of the network after learning using a [+maintainer, −constructor, ±ic-preserver] weight-update rule. The observed pairings of $(1\ *)$ with $p*$ and $(*\ 1)$ with $*p$ can be reproduced. Meaning components which have not been observed paired with any signal map to both possible signal substrings with equal probability — there is no bias against having a many-to-one mappings from feature values to signal substrings.

Table 10 (c) shows the dominant connection weights in a network trained on the meaning-signal pair $\langle (1\ 1), aa \rangle$ using a [+constructor, ±ic-preserver] weight-update rule. As with the [+maintainer, −constructor, ±ic-preserver] rule outlined above, the observed feature value-subsignal pairs can be reproduced. Unlike the [+maintainer, −constructor, ±ic-preserver] rules, use of a [+constructor, ±ic-preserver] weight-update rule results in a one-to-one mapping between feature values and subsignals — the subsignal q is preferred to p for unseen feature values. This bias results in the preferential acquisition of perfectly compositional, perfectly one-to-one mappings for networks using [+constructor, ±ic-preserver] weight-update rules.

[−maintainer] weight-update rules, as discussed above, either cannot acquire observed holistic meaning-signal mappings (in the case of [−learner, −maintainer] rules) or can acquire such systems but are biased in favour of many-to-one mappings between meanings and signals (in the case of [+learner, −maintainer] rules). In the context of compositional analyses, this has the consequence that networks using such rules either:

- cannot reliably reproduce the mapping from $(1\ *)$ to $p*$ and $(*\ 1)$ to $*p$ or
- can reliably reproduce this mapping, but prefer to produce $p*$ and $*p$ for $(2\ *)$ and $(*\ 2)$.

In either case, an e-compositional language cannot be maintained or constructed — if observed meaning-signal mappings cannot be reproduced then no stable language is possible, and if many-to-one mappings are preferred then the only stable language is highly ambiguous and therefore not e-compositional.

(a)

Feature	Value	Seen?	Subsignal	
			p	q
1	1	yes	a	b
1	2	no	c	d
2	1	yes	a	b
2	2	no	c	d

(b)

Feature	Value	Seen?	Subsignal	
			p	q
1	1	yes	a	b
1	2	no	D	D
2	1	yes	a	b
2	2	no	D	D

(c)

Feature	Value	Seen?	Subsignal	
			p	q
1	1	yes	a	b
1	2	no	c	d
2	1	yes	a	b
2	2	no	c	d

Table 10: (a) The g values for meaning component–signal component pairs after learning the meaning–signal pair $\langle (1\ 1), pp \rangle$ using the weight-update rule $(a\ b\ c\ d)$. As before, alternative orderings of signal components can be safely ignored — it is assumed that the value of the n th feature maps on to the n th character in the signal. (b) The g values for a network using a [+maintainer, –constructor, \pm ic-preserver] weight-update rule. In such rules $a > b$ and $c = d = D$. The observed feature value–signal character pairs can be reproduced, but there is no bias against many-to-one mappings from feature values to signal characters. (c) The g values for a network using a [+constructor, \pm ic-preserver] weight-update rule. In such rules $a > b$ and $d > c$. The observed feature value–signal character pairs can be reproduced, and there is a bias against many-to-one mappings from feature values to signal characters — signal character p is reserved for feature value 1.

4.3 The bias in other models

These two biases, in favour of exploiting compositional representations and in favour of one-to-one mappings between elements of meaning and elements of signals, are evident in other models of the cultural evolution of linguistic structure.

Learners in the ILM described in Kirby (2002) extract meaningful, recurring chunks from the utterances they observe wherever possible — they are biased in favour of acquiring internally compositional representations. They are also biased towards exploiting such meaningful chunks as much as possible during invention of signals — if an individual cannot express a whole meaning directly from its grammar then it invents a random signal for those subparts of the meaning which are not covered by the grammar, rather than inventing a random signal for the meaning as a whole.

In addition to this bias in favour of internally compositional representations, Kirby’s learners are biased against synonyms and homonyms. Unlike in the associative network model, these biases act as pre- and post-learning filters, rather than applying during the learning process itself. However, the net effect is the same. Kirby’s learners will not incorporate an observed utterance into their grammar if that utterance consists of a string which is already generable by the rules of their existing grammar — they have a global bias against acquiring homonymous utterances. The net effect of this bias, in combination with the chunking process, will be to prevent learners from acquiring homonymous lexical items.

The bias of Kirby’s agents against synonyms is rather less direct. During production, agents conduct a depth-first search through their grammars to find a combination of rules which allow them to express a given meaning. Consequently, when repeatedly called upon to express a meaning, they will reliably do so using the same signal — even if their grammar allows the possibility of utterance-level synonymy, their production behaviour will not be synonymous. This bias also leads to a bias against synonymy below the level of the whole utterance — if an individual has several ways of expressing the same atomic element of meaning they will, all other things being equal, express this atomic meaning consistently with a single element of the signal.

Batali’s (2002) exemplar-based learners are similarly biased. Recall that Batali’s learners induce a set of exemplars, where each exemplar has an associated cost. Exemplars which are used during learning have their costs reduced. This means that exemplars which encode small meaningful chunks will be used frequently during learning (as small elements of meaning and small parts of signals are more likely to recur in observed linguistic behaviour), will have their costs reduced rapidly and consequently will be even more likely to be used in future learning events. This has the effect of biasing learners to extract exemplars which associate small elements of meaning with parts of signals, and recombine these exemplars during learning, production and reception — Batali’s agents are biased in favour of acquiring internally compositional representations.

Batali’s scheme for manipulating exemplar costs also builds in biases against synonymy and homonymy. The bias against homonymy during learning is fairly explicit — after each learning event, learners search through their set of exemplars and increase the costs of exemplars which have a common signal but different meanings — homonymous exemplars are penalised, and therefore less likely to be used.

The bias against synonymy is less direct. Consider the case where there are two possible ways of expressing a given meaning, both with equal costs. During production of an utterance, the agent will be neutral with respect to these two variants, and will select one randomly. Another agent will learn from that production, and may reuse that exemplar when speaking to the agent who produced the utterance. The original agent may then learn based on that production (given the negotiation framework, agents can learn from individuals who they themselves have taught), in which case the

exemplar they used in the first place will have its cost reduced. This exemplar will then be used more frequently than its synonymous alternative. The reduction in costs leads to the reinforcement of one way of expressing each meaning, leading to the elimination of synonymy.

As a final example from the symbolic models of grammar induction, the learners used in Hurford (2000) are biased in favour of using internally compositional representations. Hurford's learners are strongly biased towards acquiring compositional rules — they can acquire such rules on the basis of a single observation, and also invent in a compositional manner. Hurford also includes a direct bias against synonymy, operating at the production level — if an individual has several ways of expressing a meaning, it uses the expression it acquired first. This production bias will lead to the rapid elimination of synonymy in the population's language. The bias of Hurford's learners with respect to homonymy is less clear. There is no obvious bias against homonymy built into the learning model — unlike Kirby and Batali's models, there is no prohibition on acquiring homonymous utterances. However, homonymy is perhaps less of a problem in Hurford's model due to the extremely large character alphabet available to his agents. While Kirby and Batali use the 26 letters of the alphabet, Hurford's agents have access to 2000 distinct 'syllables', which are combined to form utterances. The probability of homonymous utterances occurring by chance is therefore very low. I would anticipate that, assuming there is no hidden bias against homonymy in Hurford's learning model, a smaller syllable inventory would lead to the more frequent emergence of homonyms in his model, with a concomitant loss of e-compositionality in the emergent systems.

Finally, the learners in Brighton's (2002) model are biased to exploit internally compositional representations as much as possible — when acquiring a compositional system, they require a single observation of a feature value (paired with a signal substring) to be able to express that element of meaning. Biases with respect to homonymy and synonymy are not really relevant in Brighton's model — he assumes that such features are not present in the system presented to learners, and is not concerned with how they might be introduced due to misacquisition or invention.

Biases against homonymy and synonymy also determine the behaviour of ILMs involving neural network learners. Batali (1998) and Kirby & Hurford (2002) use feedforward networks with the obverter architecture. As discussed in Smith (2002b), this network architecture leads to a bias in favour of one-to-one mappings between whole meanings and whole signals. This bias also applies at the level of individual parts of meanings and parts of signals. Parts of meanings are represented by individual output nodes in these networks, and parts of signals are either represented by patterns of activation over the input nodes (in Batali's (1998) model) or as individual input nodes (in Kirby & Hurford's (2002) model). In either case, the obverter network architecture leads to a pressure for one-to-one mappings between individual input nodes, or patterns over groups of such nodes, and individual output nodes — many-to-one mappings are unstable, whereas one-to-many mappings are unlearnable.

In contrast, Hare & Elman (1995) use an imitator feedforward network architecture. This choice of network architecture leads to the loss, rather than emergence, of linguistic structure. This is a consequence of the many-to-one bias inherent in the imitator architecture (Smith 2002b), which applies at both the level of whole meanings and signals and at the level of subparts of meanings and signals.

The bias with respect to internally compositional representations of these networks is not clear — there is a continuum running from internally holistic to internally compositional representations. However, the well-established ability of feedforward networks to extract regularities from observed input-output mappings and generalise to unseen inputs using these regularities suggests a bias in favour of internally compositional representations. Indeed, this bias may form a general requirement for learning devices which can generalise.

4.4 Summary

In order to acquire, maintain and construct a compositional language, two components of learning bias must be in place. Firstly, learners must be biased towards using internally compositional representations. Secondly, they must be biased towards acquiring one-to-one mappings from feature values to signal substrings — they must prefer each distinct part of a meaning to be expressed by a distinct, unambiguous part of the signal. These two components of the necessary learning bias are found to be present in most other models where cultural evolution leads to the emergence of linguistic structure.

References

- BATALI, J. 1998. Computational simulations of the emergence of grammar. In *Approaches to the Evolution of Language: social and cognitive bases*, ed. by J. R. Hurford, M. Studdert-Kennedy, & C. Knight, 405–426. Cambridge: Cambridge University Press.
- 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Briscoe (2002), 111–172.
- BRIGHTON, H. 2000. Experiments in iterated instance-based learning. Technical report, Language Evolution and Computation Research Unit.
- 2002. Compositional syntax from cultural transmission. *Artificial Life* 8.25–54.
- BRISCOE, E. (ed.) 2002. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.
- HARE, M., & J. L. ELMAN. 1995. Learning and morphological change. *Cognition* 56.61–98.
- HURFORD, J. R. 2000. Social transmission favours linguistic generalization. In *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, ed. by C. Knight, M. Studdert-Kennedy, & J.R. Hurford, 324–352. Cambridge: Cambridge University Press.
- KIRBY, S. 2002. Learning, bottlenecks and the evolution of recursive syntax. In Briscoe (2002), 173–203.
- , & J. R. HURFORD. 2002. The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the Evolution of Language*, ed. by A. Cangelosi & D. Parisi, 121–147. Springer Verlag.
- SMITH, K. 2002a. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.
- 2002b. Natural selection and cultural selection in the evolution of communication. *Adaptive Behavior* 10.25–44.
- , 2003. *The Transmission of Language: models of biological and cultural evolution*. PhD Thesis, The University of Edinburgh.
- WRAY, A., & M. R. PERKINS. 2000. The functions of formulaic language: an integrated model. *Language and Communication* 20.1–28.