Situated cognition and the role of multi-agent models in explaining language structure

Henry Brighton, Simon Kirby, Kenny Smith
Language Evolution and Computation Research Unit
Department of Theoretical and Applied Linguistics
The University of Edinburgh
Adam Ferguson Building
40 George Square
Edinburgh EH8 9LL
{henryb, simon, kenny}@ling.ed.ac.uk

August 12, 2002

Abstract

How and where are the universal features of language specified? We consider language users as situated agents acting as conduits for the cultural transmission of language. Using multi-agent computational models we show that certain hallmarks of language are adaptive in the context of cultural transmission. This observation requires us to reconsider the role of innateness in explaining the characteristic structure of language. The relationship between innate bias and the universal features of language becomes opaque when we consider that significant linguistic evolution can occur as result of cultural transmission.

1 Introduction

There must be a biological basis for language. Animals cannot be taught language. Now imagine having a thorough knowledge of this capacity: a detailed explanation of whatever cognitive processes are relevant to learning, understanding, and producing language. Would this understanding be sufficient for us to predict universal features of language? Human languages

exhibit only a limited degree of variation. Those aspects of language that do not vary are termed *language universals*. The assumption of contemporary linguistics and cognitive science is that these hallmarks can shed light on the cognitive processes underlying language. In the discussion that follows we reflect on the reverse implication, and argue that language universals cannot be fully explained by understanding biologically determined aspects of cognition. The relationship between the two is opaque, and mediated by a cultural dynamic in which some linguistic forms are adaptive (Kirby, 1999).

In addressing this question one must reconsider the traditional practice in cognitive science of, first, isolating a competence from its cultural context and then, secondly, attempting to understand that competence such that its behaviour can be fully explained. This practice is questioned by the proponents of embodied cognitive science (Dreyfus, 1972; Winograd & Flores, 1986; Clancy, 1997; Brooks, 1999; Pfeifer & Scheier, 1999). We examine the claims of embodied cognitive science, specifically the principle of situatedness, and relate this enterprise to recent work in the field of computational evolutionary linguistics. We note that these two approaches share a methodological assumption, one that singles out cultural context as being a theoretically significant consideration. In the discussion that follows we show how this notion of cultural context can be modeled using multi-agent computational models. In short, we aim to show how multi-agent systems can be used to shed light on some fundamental issues in linguistics, but also cognitive science in general.

First we discuss alternative standpoints in explaining why, as a cognitive process, language exhibits certain designs. We argue that situatedness must form part of any explanation – a thorough understanding of linguistic competence cannot lead to a thorough explanation for the universal aspects of language structure. To flesh this claim out we present work on an agent-based framework for studying the evolution of language: the *iterated learning model*. In particular, we focus on compositionality in language. Insights gained from these models suggest that language designs cannot be explained by understanding language in terms of a detached individual's knowledge of language. An argument for this stance is presented Section 4 where we make explicit the foundational principles that underly our approach to understanding the characteristic structure of language.

2 Explaining universal features of language

Take all the world's languages and note the structural features they have in common. On the basis of these universal features of language, we can propose

a *universal grammar*, a hypothesis that circumscribes the core features of all possible human languages (Chomsky, 1965). On accepting this hypothesis, we should ask: Why is linguistic form subject to this set of universal properties? More precisely, how and where are these restricted set of structures specified? The discussion that follows will address the manner in which this question is answered.

The hunt for an explanation of universal features is traditionally mounted by arguing that universal grammar is an innate biological predisposition that partially defines the manner in which language is learned by a child. The linguistic stimulus a child faces, be it Chinese or Spanish, through the process of learning, results in a knowledge of language. For Chomsky learning is "better understood as the growth of cognitive structures along an internally directed course under the triggering and partially shaping effect of the environment"(Chomsky, 1980, p34). So an innate basis for language, along with the ability to learn, permits the child to arrive at a knowledge of language. The degree to which language is specified innately is a matter of heated debate. At one extreme, we can imagine a highly specialised "language instinct" (Pinker, 1994) and at the other, we can imagine a domain general learning competence which serves language as well other cognitive tasks (Elman et al., 1996).

2.1 The object of study

For a moment, let us stand back from this debate and examine the vocabulary of explanation we have employed to answer the original question: How and where are universal features of language specified? We notice that an explanation of a population level phenomena – language – has been reduced to the problem of an individual's knowledge of language. Languages vary greatly, but we are specifically interested in the features common to all languages. Universal properties of language, to a greater or lesser extent, are specified innately in each human. This de-emphasis of context, culture and history is recurring theme in cognitive science, as Howard Gardner notes: "Though mainstream cognitive scientists do not necessarily bear any animus [...] against historical or cultural analyses, in practice they attempt to factor out these elements to the maximum extent possible." (Gardner, 1985, p41). Taking this standpoint helps in mounting a practical investigation into a possible answer to the question. The universal aspects of language we see in the world are strongly correlated with an individual's act of cognition, which is taken to be biologically determined. Now we have isolated the real object of study. Understanding the innate linguistic knowledge of humans will lead us to an understanding of why language is the way it is. For the purposes of

this study, let us characterise this position.

Principle 1 (Principle of detachment) *A total explanation of the innate basis for language, along with an explanation of the role played by the linguistic stimulus during the language acquisition process, would be sufficient for a thorough explanation for the universal properties of language.*

Now the problem is to account for a device that relates input (linguistic stimulus) to output (knowledge of language). For example, Chomsky discusses a language acquisition device (LAD) in which the output takes the form of a grammatical system of rules. He states that "An engineer faced with the problem of designing a device for meeting the given input-output conditions would naturally conclude that the basic properties of the output are a consequence of the design of the device. Nor is there any plausible alternative to this assumption, so far as I can see" (Chomsky, 1967). In other words, if we want to know how and where the universal design features of language are specified, we need look no further than an individual's competence derived from primary linguistic data via the LAD. This position, which we have termed the principle of detachment, runs right through cognitive science and amounts to a general approach to studying cognitive processes. For example, in his classic work on vision, Marr makes a convincing case for examining visual processing as a competence understood entirely by considering a series of transformations of visual stimulus (Marr, 1977, 1982). We will now consider two bodies of work that suggest that the principle of detachment is questionable[1].

2.1.1 Explanation via synthetic construction

One of the aims of cognitive science, and in particular, artificial intelligence (AI), is to explain human, animal, and alien cognition by building working computational models. Those working in the field of AI often isolate a single competence, such as reasoning, planning, learning, or natural language processing. This competence is then investigated in concordance with the principle of detachment, more often than not, in conjunction with a simplified model of the environment (a micro-world). These simplifying assumptions, given the difficulty of the task, are quite understandable. So the traditional approach is centred around the belief that investigating a competence with respect to a simplified micro-world will yield results that, by and large, hold true when that agent is placed in the real world. General

[1]There are other arguments for questioning the principle of detachment, for example, those presented by Winograd and Flores (1986), but we omit them for the sake of brevity.

theories that underly intelligent action can therefore be proposed by treating the agent as a detached entity operating with respect to an environment. Crucially, this environment is presumed to contain the intrinsic properties found in the environment that "real" agents encounter.

This is a very broad characterisation of cognitive science and AI. Nevertheless, many within cognitive science see this approach as misguided and divisive, for a number of reasons. For example, we could draw on the wealth of problems and lack of progress traditional AI is accused of (Pfeifer & Scheier, 1999, p59-78). Some within AI have drawn on this history of perceived failure to justify a new set of principles collectively termed *Embodied Cognitive Science* (Pfeifer & Scheier, 1999), and occasionally *New AI* (Brooks, 1999). Many of these principles can be traced back to Hubert Dreyfus' critique of AI, 20 years earlier (Dreyfus, 1972). The stance proposed by advocates of embodied cognitive science is important because they refine Dreyfus' stance, build on it, and crucially cite examples of successful engineering projects. This recasting of the problem proposes, among others, *situatedness* as a theoretical maxim (Clancy, 1997). Taking the principle of situatedness to its extreme, the exact nature of the environment is to be taken as primary and theoretically significant. For example, the environment may be partly constructed by the participation of other agents (Bullock & Todd, 1999). In other words, certain aspects of cognition can only be fully understood when viewed in the context of participation (Winograd & Flores, 1986; Brooks, 1999). It is important to note that this "new orientation" is seen by many as opposing the branches of mainstream AI, or at least the branches of AI that claim to explain cognition.

If, for a moment, we believe the advocates of embodied cognitive science, they are telling us that any explanation for a cognitive capacity must be tightly coupled with an understanding of the environment. What impact does this discussion have on our questions about language universals? First, it provides a source of insights into investigating cognition through building computational models. A theory faces a different set of constraints when implemented as a computational model. An explanation that is grounded by a synthetic artifact can act as a sanity check for theory. Second, this discussion admits the possibility that investigating cognition without assuming the principle of detachment can be fruitful. In the context of language and communication, the work of Luc Steels is good example of this approach. Steels investigates the construction of perceptual distinctions and signal lexicons in visually grounded communicating robots (Steels, 1997, 1998). In this work signals and the meanings associated with signals emerge as a result of self-organisation.

2.1.2 The evolutionary explanation

Only humans have language. How did language evolve? The communication systems used by animals do not even approach the sophistication of human language, so the question must concern the evolution of humans over the past 5 million years, since our last non-linguistic ancestor, *Australopithecus* (Jones, Martin, & Pilbeam, 1992). Unfortunately, there is no fossil evidence offering concrete insights into the evolution of language in humans. We can, for example, analyse the evolution of the vocal tract, or examine skulls and trace a path through the skeletal evolution of hominids, but the kind of conclusions we can draw from such evidence can only go so far (Lieberman, 1984; Wilkins & Wakefield, 1995).

Over the past 15 years computational evolutionary linguistics has emerged as a source of alternative answers. This approach uses computational models to try and shed light on the very complex problem of the evolution of language in humans (Hurford, 1989; Kirby, 2002). One source of complexity is the interaction between two substrates, each one operating on a different time-scale. More precisely, linguistic information is transmitted on two evolutionary substrates: the biological and the cultural. For example, you are born with some innate predisposition for language which evolved over millions of years. The linguistic forms you inherit from your culture have evolved over hundreds of years, and your linguistic competence emerges over tens of years.

Much of the work on the evolution of language, particularly in the context of computational modeling, has analysed this interaction. By modeling linguistic agents as learners and producers of language, and then investigating how communication systems evolve in the presence of both biological and cultural transmission, computational evolutionary linguistics attempts to shed light on how language could evolve from non-linguistic communities. This approach draws on disciplines such as cognitive science, artificial life, complexity, and theoretical biology. Recent work in this field has focussed on how certain hallmarks of human language can arise in the absence of biological change. This observation must lead us to consider how far a biological explanation for language can take us. For example, the very possibility of trademark features of language not being fully explained in terms of an individual's (biologically determined) cognitive capacity raises important questions.

We detail this work in the next section, but raise the issue here as it impacts on the current discussion. In explaining how and why language has its characteristic structure, the evolutionary approach is in line with the claims made by proponents of embodied cognitive science. A thorough

explanation for language universals may lie outside the traditional vocabulary of explanation, in which case the principle of detachment will need to be breached.

2.2 Summary

This discussion has outlined the basis for asking two questions. First, what kind of explanatory vocabulary should be invoked when explaining universal features of language? Secondly, can situatedness shed light on this problem?

Building multi-agent computational models allows us to analyse how cognitive agents interact, specifically, what role this interaction plays in explaining the behaviour we observe in nature. This approach serves an important purpose for cognitive science generally, which traditionally views the individual as the locus of study. For linguistics, being subfield of cognitive science, a multi-agent approach to understanding cognition, one which takes situatedness as theoretically significant, is an untapped resource.

We aim to fully investigate how relevant multi-agent systems are to the question of explaining universal features of language. This is a timely investigation. For example, on the validity of artificial intelligence Chomsky notes "in principle simulation certainly can provide much insight" (Chomsky, 1993, p30). Perhaps more relevant is the remark made by another prominent linguist, Ray Jackendoff: "If some aspects of linguistic behaviour can be predicted from more general considerations of the dynamics of communication in a community, rather than from the linguistic capacities of individual speakers, then they should be." (Jackendoff, 2002, p101). Taking these two observations together we should at least consider the role of situatedness in explaining the universal features of language. The next section presents recent work on exploring precisely this question.

3 Language Evolution and Iterated Learning

The iterated learning model (ILM) is a general framework for modeling the cultural transmission of language (Kirby, 2001; Brighton, 2002), and is based on Hurford's conception of the expression/induction model (Hurford, 1989, 1990). The basis of an iterated learning model is a series of generations. Each generation consists of a population of agents which learn language from utterances produced by the previous generation. Each agent represents a language user, and begins life as an infant observing the language of adult agents in the previous generation. The agent learns from these observations and induces a knowledge of language. After doing so, the infant becomes

an adult. Once an adult, an agent will be prompted to form utterances which infant agents, in the next generation, observe. This process, depicted in Figure 1, is repeated for some number of generations, typically in the thousands.

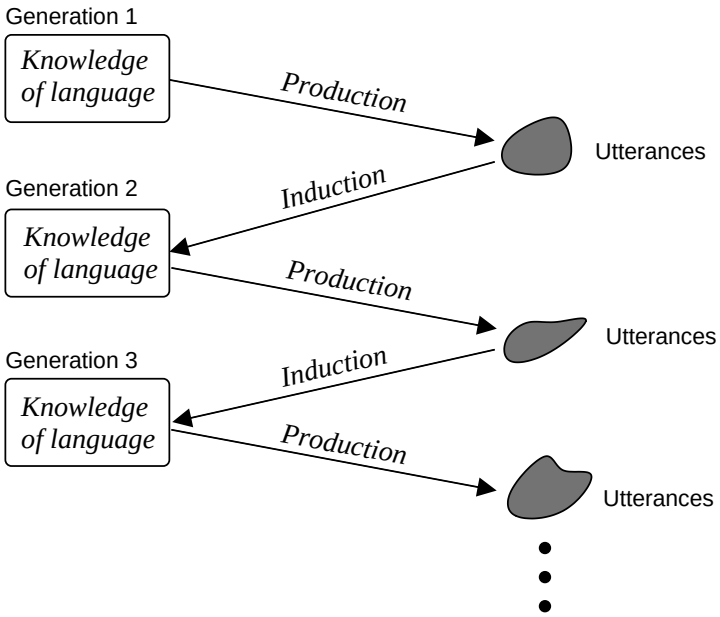

Figure 1: The agents in the ILM produce utterances. These utterances are used by the agents in the next generation to induce a knowledge of language. By repeating this process, the language evolves.

In this article we will concentrate on models which have one agent in each generation. This simplification is important, and is discussed later. In brief, the iterated learning model allows us to see how a language evolves over time, as it passes through a repeated cycle of induction and production. The agents themselves act as a conduit for language, with the bias inherent in the processes of learning and generalisation defining, in part, how language will evolve from one generation to the next.

In the ILM a language is defined as a mapping from meanings to signals. Meanings are regarded as abstract structured entities, and modeled here as feature vectors. Signals differ from meanings in that they are of variable length. Signals are built by concatenating abstract symbols drawn from some alphabet. These idealisations are consistent with Pinker and Bloom's characterisation of language as the "transmission of propositional structures over a serial channel" (Pinker & Bloom, 1990). One of the hallmarks of

human language, which we will be considering in detail, is the property of *compositionality* (Montague, 1974):

> The meaning of a signal is a function of the meaning of its parts, and how they are put together.

Compositional languages are those exhibiting the property of compositionality. We can contrast these with holistic languages, where parts of the meaning do not correspond to parts of the signal — the only association that exists is one that relates the whole meaning to the whole signal. Before going into the details of the ILM, it is worth considering three examples of communication systems found in nature:

1. The alarm calls of Vervet monkeys provide us with the classic example of a largely innate holistic communication system (Cheney & Seyfarth, 1990).

2. Bird song has learned signals with elaborate structure, but the meaning the song conveys is believed to be holistic – a structured song refers to the meaning as whole (Hauser, 1996).

3. Honey bees do have a compositional communication system, but it is innate (von Frisch, 1974).

Significantly, the only communication system that is learned and exhibits compositionality is human language. Both compositional and holistic utterances occur in human language. For example, the idiom "kicked the bucket" is a holistic utterance which means *died*. Contrast this utterance with "large green caterpillar" for which the meaning is a function of the meaning of its parts: "large", "green", and "caterpillar".

A simple[2] example of a holistic language, using the formalisation of language in the ILM, might be set of meaning signal pairs $L_{holistic}$:

$$\begin{aligned} L_{holistic} = \quad & \{\langle\{1,2,2\}, \text{sasf}\rangle, \langle\{1,1,1\}, \text{ac}\rangle, \langle\{2,2,2\}, \text{ccx}\rangle, \\ & \langle\{2,1,1\}, \text{q}\rangle, \langle\{1,2,1\}, \text{pols}\rangle, \langle\{1,1,2\}, \text{monkey}\rangle\} \end{aligned}$$

No relation exists between the signals and the meanings, other than the whole signal standing for the whole meaning. In contrast, an example of a compositional language is the set:

[2]The languages used in the simulations we discuss are usually larger than the examples presented here.

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

Notice that each signal is built from symbols that map directly onto feature values. Therefore, this is a compositional language; the meaning associated with each signal is a function of the meaning of the parts of that signal.

Now, at some point in evolutionary history, we presume that a transition from a holistic to a compositional communication system occurred (Wray, 1998). This transition formed part of what has been termed the eighth major transition in evolution – from an animal communication system to a full blown human language (Maynard Smith & Szathmary, 1995). Using the ILM, we can try and shed light on this transition. In other words, how and why might a holistic language such as $L_{holistic}$ spontaneously pass through a transition to a compositional language like $L_{compositional}$?

3.1 Technicalities of the ILM

Agents in the ILM learn a language on the basis of a set of observed meaning/signal pairs L'. This set L' is some random subset of the language which could have been spoken in the previous generation, denoted as L. That is, L' is the set of utterances of L that were produced. Humans are placed in precisely this position. First, we hear signals and then we somehow associate a meaning to that signal. Second, we suffer from the *the poverty of the stimulus* (Pullum & Scholz, 2002) – we learn language in light of remarkably little evidence. For example, there is no way any human language can ever be externalised as a set of utterances. Languages are just too large, in fact, they are ostensibly infinite. This restriction on the degree of linguistic stimulus available during the language learning process we term the *transmission bottleneck*. This process is illustrated in Figure 2.

Once an agent observes the set of utterances L', it forms a hypothesis, h, for this observed language using a learning mechanism. In our experiments we draw on a number of machine learning approaches to achieve this task. Once an appropriate hypothesis has been induced, the agent is considered an adult, and can now form utterances of its own. By interrogating the hypothesis, signals can be produced for any given meaning. Sometimes the agent will be called to produce for a meaning it has never observed in conjunction with a signal, and it therefore might not be able to postulate a signal by any principled means. In this situation some form of invention is required. Invention is a last resort, and introduces randomness into the language. However,

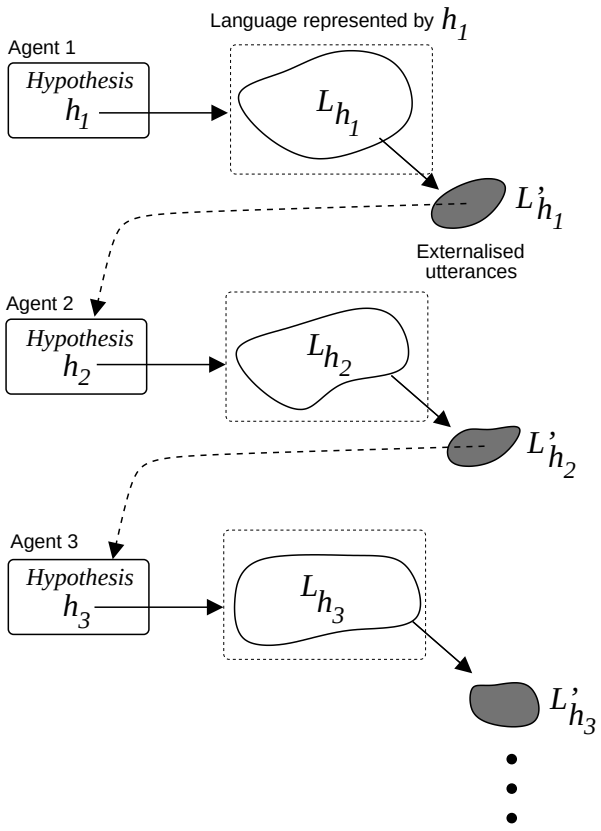

Figure 2: The hypothesis of agent 1, h_1, represents a mapping between meanings and signals, L_{h_1}. On the basis of some subset of this language, L'_{h_1}, the agent in the next generation induces a new hypothesis h_2. This process of utterance observation, hypothesis induction, and production, is repeated generation after generation.

if structure is present in the language, there is the possibility of generalisation. In such a situation, the hypothesis induced could lead to an ability to produce signals for all meanings, without recourse to invention, even though all the meaning/signal pairs have not been observed.

With a transmission bottleneck in place, a new dynamic is introduced into the ILM. Because learners are learning a mapping by only observing a subset of that mapping, through the process of invention, they might make "mistakes" when asked to convey parts of that mapping to the next generation. This means that the mapping will change from generation to generation. In other words, the language evolves. How the language evolves, and the possibility and nature of steady states, are the principle objects of study within

the ILM. We now consider these two questions.

3.2 The evolution of compositional structure

Recall that, from an initially holistic language, we are interested in the evolution of compositional language. Specifically, we would like to know which parameters lead to the evolution of compositional structure. The parameters we consider in the discussion that follows are:

1. The severity of the transmission bottleneck, b ($0 < b \leq 1$), which represents the proportion of the language utterable by the previous generation that is actually observed by the learner. The poverty of the stimulus corresponds to the situation when $b < 1.0$.

2. The structure of the meaning space. Meanings are feature vectors of length F. Each feature can take one of V values. The space from which meanings are drawn can be varied from unstructured (scalar) entities ($F = 1$) to highly structured entities with multiple dimensions.

3. The learning and production bias present in each agent. The learning bias defines a probability distribution over hypotheses, given some observed data. The production bias defines, given a hypothesis and a meaning, a probability distribution over signals.

To illustrate how compositional language can evolve from holistic language we present the results of two experiments. The first experiment is based on a mathematical model identifying steady states in the ILM (Brighton & Kirby, 2001; Brighton, 2002), and the second considers the dynamics of an ILM in which neural networks are used as a model of learning (Smith, 2002a). We refer the reader to these articles if they require a more detailed discussion.

3.2.1 Compositional structure is an attractor in language space

Using a mathematical model we show that, under certain conditions, compositional language structure is a steady state in the ILM. In these experiments the processes of learning and generalisation are modeled using the Minimum Description Length Principle (Li & Vitányi, 1997) with respect to a hypothesis space consisting of finite state transducers. These transducers map meanings to signals, and as a result of compression, can permit generalisation so that utterances can be produced for meanings which have never been observed.

Primarily we are interested in steady states. A steady state corresponds to a language which repeatedly survives the communication bottleneck: It is stable within the ILM. We can define language stability as the degree to which the hypotheses induced by subsequent agents agree on the mapping between meanings and signals. When a bottleneck is in place, the initial language, that produced by the first agent, will be unstable because it is holistic and therefore uncompressable. Note that when there is no communication bottleneck in place, all languages are stable because an agent will have observed the whole language, and could therefore just construct a lookup table that associates every meaning with a signal.

A stable language is one that can be compressed, and therefore pass through the communication bottleneck. Compression can only occur when structure is present in a language, so compression can be thought of as exploiting structure to yield a smaller description of the data. This is why holistic language cannot fit through the bottleneck – it has no structure.

Ultimately, we are interested in the degree of stability advantage conferred by compositional language over holistic language. Such a measure will reflect the probability of the system staying in a stable (compositional) region in language space. More formally, we define the expressivity, E of a language L as the number of meanings that the hypothesis induced on the basis of L, which we term h, can express without recourse to invention.

Given a compositional language L_c, and a holistic language L_h, we use a mathematical model to calculate the expected expressivity of the transducer induced for each of these language types. We denote these measures of expressivity E_c and E_h, respectively. These expressivity values tell us how likely the transducer is to be able to express an arbitrary meaning, and therefore, how stable that language will be in the context of the ILM. Finally, the value we are really interested in is that of *relative stability*, S:

$$S = \frac{E_c}{E_c + E_h}$$

This tells us how much more stable compositional language is than holistic language. In short, the model relates relative stability, S, to the parameters b (severity of the communication bottleneck), F, and V (the structure of the meaning space). Figure 3(a)-(d) illustrates how these three variables interact. Each surface represents, for a different bottleneck value, how the meaning space structure impacts on the relative stability, S, of compositional language over holistic language. We now analyse these results from two perspectives.

Tight bottleneck. The most striking result depicted in Figure 3 is that for low bottleneck values, where the linguistic stimulus is minimal, there is a

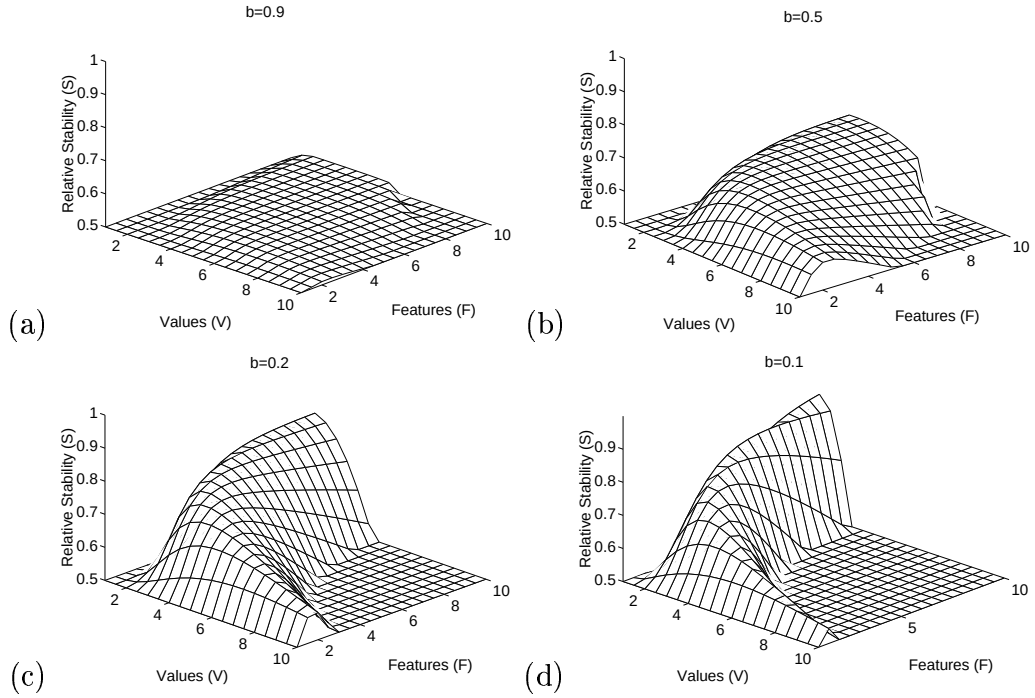

Figure 3: The bottleneck size has a strong impact on the relative stability of compositionality, S. In (a), $b = 0.9$ and little advantage is conferred by compositionality. In (b)-(d) the bottleneck is tightened to 0.5, 0.2, and 0.1, respectively. The tighter the bottleneck, the more stability advantage compositionality offers. For low bottleneck sizes, a sweet spot exists were highly structured meanings lead to increased stability.

high stability payoff for compositional language. For large bottleneck values (0.9), compositionality offers a negligible advantage. This makes sense, as we noted above, because without a bottleneck in place all language types are equally stable. But why exactly is compositional language so advantageous when a tight bottleneck is in place? When faced with a holistic language we cannot really talk of learning, but rather memorisation. Without any structure in the data, the best a learner can do is memorise: generalisation is not an option. For this reason, the expressivity of an agent faced with a holistic language is the number of distinct utterances observed.

Note that when agents are prompted to produce utterances, the meanings are drawn at random from the meaning space. A meaning can therefore be expressed more than once. Expressivity is precisely the number of distinct utterances observed. When there is structure in the language, expressivity is

no longer a function of the number of utterances observed, but rather some function, say f, of the number of distinct feature values observed, as these are the structural entities that generalisation exploits. Whenever a meaning is observed in conjunction with a signal, F features values are observed while only a single meaning is observed. The mathematical model we have developed proposes a function f on the basis of the MDL principle. Recall the parallel between the communication bottleneck and the the situation known as the poverty of the stimulus: all humans are placed in the situation where they have to learn a highly expressive language with relatively little linguistic stimulus. These results suggest that for compositionality to take hold the poverty of the stimulus is a requirement. Traditionally, poverty of stimulus, introduced in Section 3.1, is seen as evidence for the view that we have innate linguistic knowledge. Because a language learner is faced with an impoverished body of linguistic evidence, innate language specific knowledge is one way of explaining how language is learned so reliably (Chomsky, 1965; Pinker, 1994; Pullum & Scholz, 2002). The results presented here suggest an alternative viewpoint: stimulus poverty introduces an adaptive pressure for structured, learnable languages.

Structured meaning spaces Certain meaning spaces lead to a higher stability payoff for compositionality. Consider one extreme, where there is one dimension ($F = 1$). Here, only one feature value is observed when one meaning is observed. Compositionality is not an option in such a situation, as there is no structure in the meaning space. When we have a highly structured meaning space, the payoff in compositionality decreases. This is because feature values are likely to co-occur infrequently as the meaning space becomes vast. Somewhere in between these extremes sits a point of maximum stability payoff for compositionality.

3.2.2 An agent-based model

The results presented above tell us something fundamental about the relation between expressivity and learning. The model, stripped bare, relates language expressivity to two different learning models by considering the combinatorics of entity observation. We compare two extremes of language structure: fully structured compositional languages and structureless holistic languages. In this respect, the model is lacking because human language exhibits a mixture of both. Some utterances we use are holistic, some are compositional (Wray, 1998). We also skirt round the question of dynamics. The model is an analysis of Lyapounov stable states: places in language space that, if we start near, we stay near (Glendinning, 1994).

We now briefly discuss a second experiment that addresses both these issues. In this experiment, the dynamics of language evolution are modeled explicitly using an agent-based simulation. Agents in this experiment are associative neural networks. This model is an extension of a model of simple learned vocabulary (Smith, 2002b). Using an associative network in conjunction with learning rules based on Hebbian learning, the mapping between meanings and signals is coded using a meaning layer, two intermediate layers, and a signal layer. Languages exhibiting all degrees of compositionality, holistic to compositional, and all gradations in between, are learnable by this network (Smith, 2002a).

The first generation of the ILM starts with a network consisting of weighted connections, all of which are initialised to zero. The network is then called to express meanings drawn from an *environment* which we define as some subset of the meaning space. One dimension of variation over environments is *dense* to *sparse*. This means that the set of possible meanings to be communicated are drawn from a large proportion of the space (dense) or a small proportion of the space (sparse). The second dimension of variation concerns *structured* and *unstructured* environments. A structured environment is one where the average inter-meaning hamming distance is low, so that meanings in the environment are clustered. Unstructured environments have a high inter-meaning hamming distance.

Once again, the bottleneck parameter, the proportion of the environment used as learning data, is varied. First, let us consider the case where no bottleneck is present — a hypothesis is chosen on the basis of a complete exposure to the language of the previous generation. Figure 4(a) depicts, for 1000 independent ILM runs, the frequency of the resultant (stable) languages as a function of compositionality. Compositionality is measured as the degree of correlation between the distance between pairs of meanings and distance between the corresponding pairs of signals. We see that few compositional languages evolve. Contrast this behaviour with Figure 4(b), where a bottleneck of 0.4 is imposed. Compositional languages are now by far the most frequent end-states of the ILM. The presence of a bottleneck makes compositionality adaptive in the ILM. We also note that structured environments lead reliably to compositional language.

This experiment, when considered in more detail, illustrates the role of clustering in the meaning space, and the impact of different network learning mechanisms (Smith, 2002b). But for the purposes of this discussion, the key illustration is that the bottleneck plays an important role in the evolution of compositional languages. In short, these results validate those of the previous section.

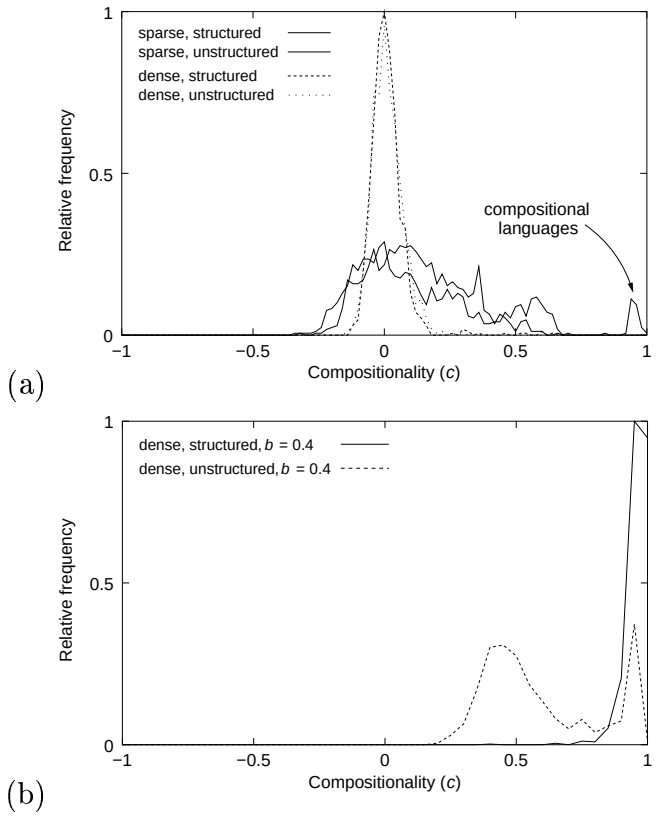

Figure 4: In (a) we see how the lack of a bottleneck results in little pressure for compositional languages. In (b), where a bottleneck of 0.4 is imposed, compositional languages reliably evolve, especially when the environment is structured.

3.3 Using the ILM to explain language structure

The learning bias and hypothesis space of each agent is taken to be innately specified. Each generation of the ILM results in the transfer of examples of language use only. In the absence of a bottleneck, compositionality offers little advantage, but as soon as a bottleneck is imposed, compositional language becomes an attractor in language space. So even though agents have an innate ability to learn and produce compositional language, it is the dynamics of transmission that result in compositionality occurring in the ILM. We must reject the idea that an innate ability to carry out some particular behaviour necessarily implies its occurrence. We aim to strengthen this claim, and refine it.

Previous work investigating the ILM has shown that linguistic features

such as recursive syntax (Kirby, in press), and regular/irregular forms (Kirby, 2001) can also be framed in this context. The idea that we can map innate properties such as, for example, the learning and generalisation process, the coding of environmental factors, and the fidelity of utterance creation *directly* onto a properties of evolved languages is not wholly justifiable. This approach should be seen as building on the Kirby's analysis of language universals (Kirby, 1999) in which issues such as, for example, constraints on representation and processing are shown to bring about functional pressures that restrict language variation. Here, we also note that the relationship between innate bias and universal features of language is not transparent, but concentrate on the constraints introduced by cultural transmission. These constraints result in certain linguistic forms being adaptive; we can think of language evolving such that it maximises its chances of survival. For a linguistic feature to persist in culture, it must accommodate the constraints imposed by transmission pressures. Compositionality is one example of an adaptive feature of language.

If we want to set about explaining the characteristic structure of language, then an understanding of the biological machinery forms only part of the explanation. The details of these results, such as meaning space structure and the configuration of the environment, are not important in the argument that follows. Nevertheless, factors relating to the increase in semantic complexity have been cited as necessary for the evolution of syntactic language (Schoenemann, 1999). We believe that the scope of the ILM as a means to explain and shed light on language evolution is wider than we have suggested so far.

To summarise, by taking compositionality as an example, we argue that its existence in all the world's languages is due to the fact that compositional systems are learnable, generalisable, and therefore are adaptive in the context of human cultural transmission. This explanation cannot be arrived at when we see the individual as the sole source of explanation. Viewing individuals engaged in a cultural activity allows us to form explanations like these.

4 Underlying Principles

We began by considering explanations for the hallmarks of language. So far we have investigated an agent's role in the context of cultural transmission. In this section we aim to tie up the discussion by making explicit a set of underlying principles. We start by noting that any conclusions we draw will be contingent on an innateness hypothesis:

Principle 2 (Innateness hypothesis) *Humans must have an biologically determined predisposition to learn and produce language. The degree to which this capacity is language specific is not known.*

Here we are stating the obvious – the ability to process language must have a biological basis. However, the degree to which this basis is specific to language is unclear. We have no definitive answer to the question of innately specified features of language (Pullum & Scholz, 2002). Next, we must consider the innateness hypothesis with respect to two positions. First, assuming the principle of detachment, the innateness hypothesis must lead us to believe that there is a clear relation between patterns we observe in language and some biological correlate. If we extend the vocabulary of explanation by rejecting the principle of detachment, then the question of innateness is less clear cut. We can now talk of a biological basis for a feature of language, but with respect to a cultural dynamic. Here, a cultural process will mediate between a biological basis and the occurrence of that feature in language. This discussion centres around recasting the question of innateness. This observation leads us to accepting that situatedness plays a role.

Principle 3 (Situatedness hypothesis) *A thorough explanation of language competence would not amount to a total explanation of language structure. A thorough explanation of language competence in conjunction with an explanation of the trajectory of language adaption would amount to a total explanation of language structure.*

The degree of correlation between a biological basis and the observed language universal is hard to quantify. However, Figure 5 illustrates the general point. A biological basis will admit the possibility of some set of communication systems $C_{possible}$. A detached understanding of language can tell us little about which members of $C_{possible}$ will be adaptive and therefore observed. The situatedness hypothesis changes the state of play by considering which communication systems are adaptive, $C_{adaptive}$, on a cultural substrate.

Rejecting the situatedness hypothesis must lead us to consider the issue of representation. The only way a thorough knowledge of language universals can be arrived at, while at the same time accepting the principle of detachment, is that universal features are somehow "represented" explicitly. How else could we understand a universal feature of language by understanding a piece of biological machinery? An acceptance of the situatedness hypothesis allows us to explain a feature of language in terms of a biological trait realised as a bias which, in combination with the adaptive properties of this bias over repeated cultural transmission, leads to that feature being observed.

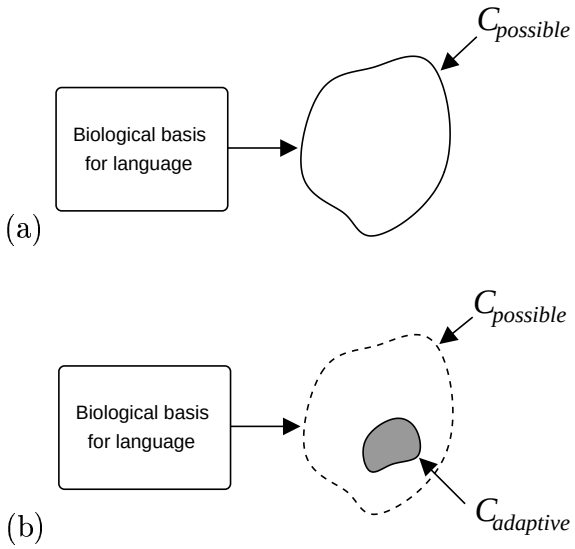

Figure 5: In (a), which assumes the principle of detachment, we can only make a claim about possible communication systems. In (b), assuming the situatedness hypothesis, an explanation accounts for the resulting communication systems which are adaptive over cultural transmission.

However, if one accepts cultural transmission as playing a pivotal role in determining language structure, then one must also consider the impact of other factors effecting adaptive properties. But as a first cut, we need to understand how much can be explained without resorting to any functional properties of language:

Principle 4 (Language function hypothesis) *Language structure can be explained independently of language function.*

A defence of this hypothesis is less clear cut. However, the models we have discussed make no claims about, nor explicitly model, any notion of language function. Agents simply observe the result of generalisation. The fact that compositional structure results without a model of language function suggests that this is a fruitful line of enquiry to pursue. The treatment of language in discussions on embodied cognitive science often assume language function is salient (Winograd & Flores, 1986), but we must initially assume it is not. The kind of cognitive processes that we consider are processes such as memory limitations and learning bias.

4.1 The role of modeling

The previous section we examined the basis for explaining language universals. The claims we made are partly informed by modeling. Is this methodology valid? Many issues relating to language processing are not modeled. For example, those involved in the study of language acquisition will note that our learners are highly implausible: the language acquisition process is an immensely complex and incremental activity (Elman, 1993). It must be stressed that our models of learning and generalisation should be seen as abstracting the learning process . We are interested in the justifiable kind of generalisations that can be made from data, not a plausible route detailing how these generalisations are arrived at. The output of a cognitively plausible models of learning is generalisation, just as it is in our models. Rather than modeling the language acquisition process, we are modeling the result (or output) of the language acquisition process. We make no claims about the state of learners during the act of learning. We also have not addressed the role of population dynamics. The models presented here represent a special case of the ILM, one where there is a single agent in each population. We regard this simplification as a necessary first step. Extending the models to contain multiple agents in each generation is a current research project.

5 Conclusions

Cognitive science has traditionally restricted the object of study by examining cognitive agents as detached individuals. For some aspects of cognition this emphasis might be justifiable. But this assumption has become less appealing, and many have taken to the idea that notions of situatedness, embeddedness, and embodiment should be regarded as theoretically significant and should play an active role in any investigation of cognition. Our aim is to consider this claim by building multi-agent models, where agents are learners and producers of language. Specifically, we aim to investigate how multi-agent models can shed light on the problem of explaining the characteristic structure of language.

When explaining universal features of language, the traditional standpoint, which we characterised in Principle 1, assumes that cultural context is not a theoretically significant consideration. We attempt to shed light on the question of how and where the universal features of language are specified. The approach we take is in line with the intuitions of embodied cognitive science. By examining the role of the cultural transmission of language over many generations, we show that certain features of language are adaptive:

significant evolution of language structure can occur on the cultural substrate.

Taking the example of compositionality in language, we illustrate this point using two models. The first model identifies compositionality as an Lyapounov stable attractor in language space when a transmission bottleneck is place. The second model offers additional evidence by demonstrating that compositionality evolves from holistic language. The upshot of these two experiments is that cultural transmission in populations of agents endowed with a general ability to learn and generalise can lead to the spontaneous evolution of compositional syntax. Related work has shown that recursive syntax and regular/irregular forms are also adaptive in the context of cultural transmission. The implications of this work lead us to reconsider how features of language should be explained. More precisely, the relationship between any innate (but not necessarily language-specific) basis for a language feature, and the resulting feature, is opaque.

We place the discussion in the context of three principles that need to be considered when explaining features of language. First, Principle 2 lays down an innateness hypothesis, which makes clear that language must have a biological basis. What form this biological basis takes is very much an open question. Secondly, we propose Principle 3, a situatedness hypothesis which makes explicit the claim that understanding the biological machinery behind language alone is not enough to explain universal features of language. This claim constitutes the core of the argument. Principle 4 identifies a hypothesis relating to the relationship between language function and language structure. The idea that language function, such as issues of communicability, has an impact on language universals is unclear.

By rejecting Principle 1 and pursuing a line of enquiry guided by Principles 2-4 we have shown that multi-agent models can provide important insights into some fundamental questions in linguistics and cognitive science. The work presented here should be seen as the first steps towards a more thorough explanation of the evolution of linguistic structure. We believe that multi-agent models will become an increasingly important tool in the study of language.

References

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

Brighton, H., & Kirby, S. (2001). The survival of the smallest: stability

conditions for the cultural evolution of compositional language. In J. Kelemen & P. Sosík (Eds.), *Advances in artificial life: Proceedings of the 6^{th} european conference on artifical life, prague, september 2001* (pp. 592–601). Springer-Verlag.

Brooks, R. A. (1999). *Cambrian intelligence.* Cambridge, MA: MIT Press.

Bullock, S., & Todd, P. M. (1999). Made to measure: Ecological rationality in structured environments. *Minds and Machines*, *9*(4), 497–541.

Cheney, D., & Seyfarth, R. (1990). *How monkeys see the world: Inside the mind of another species.* Chicago, IL: University of Chicago Press.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1967). Recent contributions to the theory of innate ideas. *Synthese*, *17*, 2-11.

Chomsky, N. (1980). *Rules and representations.* London: Basil Blackwell.

Chomsky, N. (1993). *Language and thought.* Moyer Bell.

Clancy, W. J. (1997). *Situated cognition.* Cambridge: Cambridge Univeristy Press.

Dreyfus, H. L. (1972). *What computers still can't do* (2nd ed.). Cambridge, MA: MIT Press.

Elman, J. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Gardner, H. (1985). *The minds's new science.* New York: Basic Books.

Glendinning, P. (1994). *Stability, instability, and chaos: An introduction to the theory of nonlinear differential equations.* Cambridge: Cambridge University Press.

Hauser, M. D. (1996). *The evolution of communication.* Cambridge, MA: MIT Press.

Hurford, J. R. (1989). Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua*, *77*, 187–222.

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

Jackendoff, R. (2002). *Foundations of language.* Oxford: Oxford Univeristy Press.

Jones, S., Martin, R., & Pilbeam, D. (Eds.). (1992). *The Cambridge encyclopedia of human evolution.* Cambridge: Cambridge University Press.

Kirby, S. (1999). *Function, selection and innateness: the emergence of language universals.* Oxford: Oxford University Press.

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

Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, *8*, 185-215.

Kirby, S. (in press). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic evolution through language acquisition: Formal and computational models.* Cambridge: Cambridge University Press.

Li, M., & Vitányi, P. (1997). *A introduction to Kolmogorov complexity and its applications.* New York: Springer-Verlag.

Lieberman, P. (1984). *The biology and evolution of language.* Cambridge. MA: The University of Harvard Press.

Marr, D. (1977). Artificial intelligence: A personal view. *Artificial Intelligence*, *9*, 37–48.

Marr, D. (1982). *Vision.* Freeman.

Maynard Smith, J., & Szathmary, E. (1995). *The major transitions in evolution.* Oxford University Press.

Montague, R. (1974). *Formal philosophy: Selected papers of Richard Montague.* Newhaven: Yale University Press.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

Pinker, S. (1994). *The language instinct.* Penguin.

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

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

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

Smith, K. (2002a). *Compositionality from culture: the role of environment structure and learning bias* (Tech. Rep.). Department of Theoretical and Applied Linguistics, The University of Edinburgh.

Smith, K. (2002b). The cultural evolution of communication in a population of neural networks. *Connection Science*, *14*(1), 1–21.

Steels, L. (1997). Constructing and sharing perceptual distinctions. In M. van Someren & G. Widmer (Eds.), *Proceedings of the european conference on machine learning.* Berlin: Springer-Verlag.

Steels, L. (1998). The origins of syntax in visually grounded robotic agents. *Artificial Intelligence*, *103*(1,2), 133–156.

von Frisch, K. (1974). Decoding the language of the bee. *Science*, *185*, 663–668.

Wilkins, W. K., & Wakefield, J. (1995). Brain evolution and neurolinguistic preconditions. *Behavioral and Brain Sciences*, *18*, 161–226.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition.* Adison-Wesley.

Wray, A. (1998). Protolanguage as a holistic system for social interaction. *Language and Communication*, *18*, 47–67.