Simplicity as a driving force in linguistic evolution

Henry Brighton
BSc. (Hons), MSc.

A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy

to

Dept of Linguistics,
University of Edinburgh

April 2003

Henry Brighton
BSc. (Hons), MSc.

Abstract

How did language come to have its characteristic structure? Many argue that by understanding those parts of our biological machinery relevant to language, we can explain why language is the way it is. If the hallmarks of language are simply properties of our biological machinery, elicited through the process of language acquisition, then such an explanatory route is adequate.

As soon as we admit the possibility that knowledge of language is learned, in the sense that language acquisition is a process involving inductive generalisations, then an explanatory inadequacy arises. Any thorough explanation of the characteristic structure of language must now explain why the input to the language acquisition process has certain properties and not others. This thesis builds on recent work that proposes that the linguistic stimulus has certain structural properties that arise as a result of linguistic evolution. Here, languages themselves adapt to fit the task of learning: they reflect an accumulated structural residue laid down by previous generations of language users.

Using computational models of linguistic evolution I explore the relationship between language induction and generalisation based on a simplicity principle, and the linguistic evolution of compositional structures. The two main contributions of this thesis are as follows. Firstly, using a model of induction based on the minimum description length principle, I address the question of linguistic evolution resulting from a bias towards compression. Secondly, I carry out a thorough examination of the parameter space affecting the cultural transmission of language, and note that the conditions for linguistic evolution towards compositional structure correspond to (1) specific levels of semantic complexity, and (2), induction based on sparse language exposure.

Ultimately, the story of the evolution of language in humans must depend on an account of the genetic evolution of the biological machinery underlying language. Rather than explicitly encoding the observed constraints on language, I argue that any explanation based on biological evolution should instead aim to explain how the conditions for linguistic evolution, outlined above, came about.

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Henry Brighton

Acknowledgements

I would like to thank my two supervisors, Simon Kirby and Jim Hurford. Simon Kirby, as result of his infatigable supervision and enthusiasm, has played a hugely influential role in the development of this thesis. Both Simon and Jim have demonstrated enormous patience in guiding my personal development from a non-linguist who initially doubted many of the views now articulated in this thesis.

I must also thank the members of the LEC, in particular Kenny Smith and Andrew Smith. Kenny and Andrew have fielded many naive questions concerning basic linguistics. Kenny, in particular, has helped greatly by reading earlier drafts of this thesis. External to the LEC, I must also thank T. Mark Ellison for providing information and insights on MDL and Kolmogorov complexity. Rob Clark has contributed some mathematical insight, as have Caroline Round and Korin Richmond. The computing support staff, in particular, Cedric McMartin and Mike Bennett, have showed patience in accommodating the significant computational demands of this research.

Outside Edinburgh, I must also thank the Santa Fe Institute, New Mexico, USA, and the Zentrum für interdisziplinäre Forschung at the Universität Bielefeld, Germany, for providing the opportunity to meet and work with other researchers.

All the software used in the development and presentation of this thesis has been open source. Thanks to projet Cristal at INRIA, Rocquencourt, for making Objective Caml such a pleasure to use. I must also acknowledge the contribution of GraphViz, which is developed and shared by AT&T Labs-Research.

Without the help of my parents, Roger Lindsay, Chris Mellish, Pete Whitelock, and Jim Hurford I would not have been able to embark on this PhD. Without the understanding of Richard Appignanesi, and his act of rearranging important deadlines, I may not have been able to finish it.

Contents

Chapter 1

Introduction

1.1 Language as an expression of the genes

Traditionally, knowledge of language is regarded as an innate biological predisposition. Languages may vary, but the fundamental structure of language is an expression of the genes. Accordingly, humans possess a faculty of language, or a *language organ*:

> It is hard to avoid the conclusion that a part of the human biological endowment is a specialized "language organ," the faculty of language (FL). Its initial state is an expression of the genes, comparable to the initial state of the human visual system, and it appears to be a common human possession to close approximation." (Chomsky 2002:85)

To support this view, we can note that children master complex features of language on the basis of surprisingly little evidence. In fact, what is often termed the *argument from the poverty of the stimulus*, is a conjecture stipulating that the knowledge of language children attain is surprising precisely because it *cannot* be derived solely from information made available by the environment (Chomsky 1965; Wexler 1991). This view can be traced back to Plato (427BC–347BC), who noted that humans come to know more than that suggested by the evidence they encounter, with language being just one example of this general phenomenon. John Locke (1632–1704) was perhaps the first to then ask the question: If knowledge of language is innate, then why does language exhibit so much variation? The modern debate on the innateness of language attempts to resolve this problem by noting

that the *framework* for linguistic development is innate, with the linguistic environment serving to supply information that steers an internally directed course of development:

> [the environment] provides primary linguistic data that enable the linguistic system to develop, just as it provides light and food that enable the visual and motor systems to develop. This sort of reasoning implies that the linguistic system is *profoundly abstract*. (Uriagereka 1998:523)

In this sense, languages themselves are not encoded entirely in the genes, but the fundamental, abstract properties of language are. How can we gain an understanding of these innately specified hallmarks of language? Linguistics, by proceeding to conduct a thorough analysis of the world's languages, proposes a set of fundamental properties common to all languages termed *language universals*. Linguistic universals define the dimensions of variation in language. Modern linguistic theory rests on the assertion that it is these dimensions of variation that are innately specified.

1.2 Linguistic evolution

The traditional view is based on an assumption: Children acquire knowledge that cannot be learned. One logical conclusion is to argue that language acquisition is therefore not a process of learning, but rather one of eliciting pre-defined, genetically determined properties of language. Now, is this conjecture entirely justified? Many argue that it is not. First of all, we can focus on the scientific status of the argument from the poverty of stimulus, and note that no rigorously attained body of evidence exists that lends strong support to it: the argument from the poverty of the stimulus is a prima facie explanation for linguistic nativism (Pullum & Scholz 2002). We can also conduct computational experiments to show that certain fundamental properties of language *can* be learned from the linguistic evidence available to the child. Here, techniques from *machine learning* are used to investigate the precise nature of the knowledge that can be induced from plausible bodies of linguistic evidence (Elman 1993; Elman *et al.* 1996; Christiansen & Devlin 1997).

The debate concerning the precise nature of the language acquisition process continues. But, if, for a moment, we accept a retreat from the position characterised by the argument from the poverty of stimulus, a fundamental problem arises for

linguistics. By questioning the assertion that the fundamental properties of language are solely an expression of the genes, then we require an explanation for *why* language exhibits certain hallmarks and not others. The only alternative is that they are somehow present in, or partly determined by, properties of the language learning environment. By re-addressing Plato's problem, the quandary of explaining why language is the way it is can be shifted from an explanation of properties of the child to an explanation of properties of the environment. Importantly, the problem still persists: Why does the language learning environment exhibit certain properties and not others? This question reveals a shortcoming in the explanatory practices of linguistics.

In this thesis I will investigate a solution to this explanatory shortcoming. The solution is reached by extending the current explanatory framework that we invoke when trying to understanding how and where the hallmarks of language are specified. The idea is simple: *languages themselves* adapt as a result of being repeatedly transmitted from one generation to another through the processes of production and induction (Christiansen 1994; Deacon 1997; Kirby 1999). The changes to a language that occur as result of this cultural transmission force language to have certain fundamental structural characteristics. It is these characteristics that we observed in the worlds languages. So rather than the hallmarks of language residing in biological structure that is an expression of the genes, they are taken to be *artifacts* reflecting the accumulated residue of language transmission. Accordingly, we can regard at least some of the universal properties of language as *frozen accidents* that persist because they are easily learned by children during the process of language acquisition.

1.2.1 Advances in the theory of linguistic evolution

The process by which languages themselves evolve is termed *linguistic evolution* (Briscoe 2002). This process should be contrasted with the process of biological evolution, which provides an explanation for the evolution of the biological machinery that supports the processing of language. In this thesis, I will investigate computational models of linguistic evolution. On the basis of these models, I will present several novel contributions to the theory of linguistic evolution:

Chapter 2. The problem of explaining why language has certain structural characteristics is a central concern of cognitive science. I will analyse Chomsky's view that language is not learned but specified by genetically determined cognitive structures.

I will note that this view is characterised by what I term the *principle of detachment*, which in turn has a far-reaching influence in the explanatory practices of the cognitive sciences. Alternatives to this view are considered, and I will relate the issue of explaining linguistic universals to recent advances in cognitive science that question the principle of detachment. In order to engage directly with current linguistic theory, I will outline three hypotheses that underly the view that language adapts to be learnable. The three hypotheses define an explanatory framework. This thesis will present evidence in support of the hypotheses. In short, I will seek to consolidate previous work by proposing a step towards a canonical set of hypotheses.

Chapter 3. To understand the process of cultural adaptation I will introduce the notion of *iterated learning*, where the learning process is situated by considering the impact that other learners have in constructing the environment for learning. Iterated learning *models* will be used throughout this thesis to test theories of linguistic evolution. On route to the development of more sophisticated models, Chapter 2 will outline two basic models that serve as illustrative examples. The second model I will develop addresses the issue of compositonality in language. The analysis of this model will reveal some foundation results.

Chapter 4. The major contribution made by this thesis will be the development of a model of linguistic evolution based on a simplicity principle. Iterated learning models revolve around the process of arriving at inductive generalisations from data. A rigorous model of induction can be approached by considering the theory of *Kolmogorov complexity* – a branch of information theory. I will present an argument for treating Kolmogorov complexity as a valid concept in investigating the role of simplicity principles in cognitive science. In particular, I will discuss the role of simplicity in Chomsky's minimalist program, as well as its role in the problem of language acquisition. This Chapter will close with the development of an iterated learning model driven by a computational approach to induction derived from the theory Kolmogorov complexity. This model will be analysed and developed in the next two Chapters.

Chapter 5. Rather than analysing the behaviour of iterated learning models through computational simulation, this Chapter will derive some foundation results. By employing mathematical and monte carlo models, I will conduct a wide-scale mapping of the parameter space of the iterated learning model developed in Chapter 4. In doing so, those parts of the parameters space that are computationally intractable

to map through simulation modelling can be understood such that general claims can be made about the behaviour of iterated learning models. These results relate directly to the issue of the argument from the poverty of the stimulus, which I will argue is a strong determinant of the linguistic evolution of structured language. In short, I will shed light on the *range* of conditions that must be met for linguistic evolution to converge on one of the hallmarks of language, *compositionality*.

Chapter 6. By finalising the model of linguistic evolution developed in Chapters 4 and 5, I will be in a position to conduct a set of full simulations. In doing so, induction from data based on a simplicity principle is modelled explicitly. I will show how linguistic evolution toward compositional language structure is possible given such a rigorous model of induction. These experiments will support and justify the results of previous models. Furthermore, by explicitly modelling the process of linguistic evolution, I will shown how structured languages exhibiting redundancy and sub-optimal properties are an inevitable result of linguistic evolution. These issues relate directly to Chomsky's minimalist program, where the imperfections in language are viewed as artifacts arising from *internal* constraints of the genetically determined language faculty.

Chapter 7. Finally, I will relate the conditions for linguistic evolution suggested by these models to the wider question of language evolution. In particular, by invoking an iterated learning model of linguistic evolution, the issue of which aspects of the cognitive system are (possibly innate) prerequisites for the emergence of language can be refined and recast.

In short, this thesis extends existing work suggesting that the universal properties of language need not be explained exclusively as genetically determined properties. Language is a complex system of relating sound to meaning that exhibits both structural elegance and structural redundancy. Considerations of the role of simplicity in linguistic evolution can shed light on explaining both these structural characteristics.

Chapter 2

Explaining linguistic structure

2.1 Introduction

Language universals are significant to the cognitive sciences. But what can the universal properties of language tell us about those parts of human cognitive system responsible for language? One possibility is that language universals offer an unparallelled insight into the mechanisms underlying those aspects of the cognitive system relevant to language – they can be taken as a reflection of a biologically determined linguistic competence (Chomsky 1965). Here, the implication is that the relation between linguistic universals and cognitive machinery is transparent. Using this insight, the question of how language evolved in humans is reduced to the problem of explaining how the biological machinery supporting the processing of language evolved. It is this reverse implication, that of explaining how the hallmarks of language came to be by explaining how the linguistic machinery came to be, that I will focus on in the discussion that follows.

The fundamental question is the following: Why does language have the structure that it does? At its root, this question concerns the degree to which an understanding of the cognitive structures underlying language amounts to a thorough explanation of the properties of language. If this explanation is to be adequate, then we have circumscribed the explanatory framework to include, firstly, a theory of the cognitive mechanisms relevant to processing language and, secondly, a theory of the role that the linguistic data available to a language learner plays in determining how these cognitive structures ultimately represent knowledge of language. Such an explanatory framework is characterised by what I will term the *principle of detachment*.

This explanatory framework accommodates substantial variation in the role played by the linguistic data. If the hallmarks of language are a direct consequence of the properties of the cognitive machinery alone, then the principle of detachment is wholly justifiable. But if the basic properties – universal linguistic characteristics – of the cognitive system are partly specified by the properties of the linguistic data, then a problem arises. The principle of detachment can offer no theory explaining why the linguistic stimulus has the form that it does.

The purpose of this chapter is threefold. Firstly, I will identify the assumptions made in the traditional approach to explaining universal properties of language. In line with the above argument, I will argue that this explanatory framework is inadequate. Furthermore, its assumptions are echoed in the traditional explanatory practices of the cognitive sciences. Secondly, I will sketch a candidate conceptual framework that can be used to combat this explanatory inadequacy. This approach, which I have termed *iterated learning*, has its roots in evolutionary linguistics. I draw parallels between this approach and that of situated cognitive science: both deem it necessary to explain the precise nature of the environment of cognition. These concepts need to be clarified, and this objective is the motivation for the third purpose of this chapter, and the most influential contribution to the progression of this thesis. I will outline three hypotheses that underly the theory that language adapts to be learnable. This theory offers a candidate explanation for why the linguistic stimulus has the form that it does.

Above, I noted the possibility that language universals offer an unparallelled insight into the workings of those aspects of the cognitive system relevant to language. This Chapter, then, considers an alternative: language universals are artifacts of a little-understood interaction between a cognitive system and an environment of adaptation. The relationship between language universals and the cognitive system is therefore opaque. This view has its roots in the work of, among others, Morten Christiansen, Terrence Deacon, and Simon Kirby (Christiansen 1994; Deacon 1997; Kirby 1999). Before investigating this position, I will place it in the context of the cognitive sciences, and make explicit its assumptions by postulating three hypotheses that underly the view that language adapts to be learnable. The fundamentals of this chapter can also be found in two forthcoming articles (Brighton *et al.* 2003b; Brighton *et al.* 2003a).

2.2 From description to explanation

Take all the world's languages and note the structural features they have in common. Now, on the basis of these universal features of language, we can propose a *universal grammar* (UG), a hypothesis circumscribing the core features of all human languages (Chomsky 1965). On its own, this hypothesis acts only as a description. But far from being an inert taxonomy, universal grammar sets the target for an explanatory theory. The kind of entities contained in UG that I will allude to consist of *absolute* and *statistical* language universals (Matthews 1997; O'Grady *et al.* 1997). Absolute universals are properties present in all languages. Statistical universals are properties present in a significant number of languages. Several further distinctions naturally arise when describing constraints in cross-linguistic variation, but in the interests of clarity I will restrict this discussion to one of absolute and statistical universals. In short, however we choose to describe the universal structural features of language, the impact of this descriptive theory is hugely important to the cognitive sciences, including linguistics. Indeed, for Chomsky:

> Real progress in linguistics consists in the discovery that certain features of given languages can be reduced to universal properties of language, and explained in terms of these deeper aspects of linguistic form. (Chomsky 1965:35)

But how can this progress in descriptive linguistics translate into real progress in those branches of the cognitive sciences concerned with explaining the cognitive structures underlying language? Precisely how UG is used to move from a purely descriptive theory to one offering an explanation for *why* languages are the way they are is worthy of some thought. Firstly, and quite naturally, UG can be employed as the backbone for a predictive theory of languages not just observed, but also for those languages that are possible. In addition, by adopting the uniformitarian assumption, we also postulate that past languages, for which we have no knowledge, also conform to the present conception of UG (Croft 1994; Newmeyer 2002). That is, the evidence that observed languages provide can be used to tell us about language in general. Such a theory, on the basis of induction, now becomes a putative theory of human language in the wider sense. Secondly, and more importantly in the discussion that follows, UG is taken as a plausible object to be explained by the cognitive sciences. For a cognitive scientist, this explanatory route naturally invites the question: Why is linguistic form subject to this set of universal properties? More

precisely, we seek an explanation for how and where this restricted set of linguistic features is specified. The outcome of such an enterprise would, ideally, be a theory detailing the cognitive structures responsible for the constraints on languages we observe.

In the following discussion I will analyse the possible routes we can take when forming such an explanation. UG can therefore be taken as shedding light on cognition. In essence, language is the result of a cognitive system; fundamental properties of language are then taken as reflecting fundamental properties of the cognitive system. This is the transparency that I referred to in my introductory remarks.

2.2.1 Chomsky and the cognitive sciences

An explanation for the universal features of language is traditionally reliant on an argument that universal grammar is an innate biological predisposition. So rather than a descriptive object, UG is invoked as a model of a state of the human cognitive system. More precisely, Chomsky states:

> the language organ is the *faculty of language* (FL); the theory of the initial state of FL, an expression of the genes, is *universal grammar* (UG); theories of states attained are *particular grammars*; the states themselves are *internal languages*, "languages" for short. (Chomsky 2002:64)

Universal grammar is therefore a model of a genetically determined state from which a transition occurs. On postulating UG as representing the features common to all languages, a theory of how the cognitive system proceeds through a series of transitions, ending in knowledge of a specific language, is required. Chomsky goes on to describe the processes that determine this transition:

> The initial state changes under the triggering and shaping effect of experience, and internally determined processes of maturation, yielding later stages that seem to stabilize at several stages, finally at about puberty. (Chomsky 2002:85)

So the linguistic stimulus a child faces, whichever language it is drawn from, in conjunction with "internally determined processes of maturation", is responsible

for an end-state corresponding to knowledge of language. With respect to an explanation for the universal features of language, the initial state – universal grammar – specifies these features of language. Those aspects of language that vary must therefore vary as a result of acquisition: the environment elicits internally specified structures. It is important to understand how exactly Chomsky imagines outside influences impacting on the yield of the acquisition process. Knowledge of language, for Chomsky, is taken to go "far beyond the presented primary linguistic data and is in no sense an 'inductive generalisation' from these data." (Chomsky 1965:33)

In line with this explanation, linguistic stimuli serve to provide a set of triggers that activate one of many possible states, or hypotheses. Chomsky has a tendency to avoid the term *learning* altogether. Language is not viewed as being learned in the sense that it is instructed by a teacher. Moreover, language acquisition is not generally regarded as a learning process at all, for example, language learning, or acquisition, for Chomsky, 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 1980:34)

Whether or not language acquisition can accurately be thought of as learning causes a great deal of controversy. Chomsky's position is an extreme; his observation that acquisition "in no sense an 'inductive generalization' " is questioned by many. There are two broadly defined perspectives from which this statement is questionable.

First, the use of the term *learning* in the cognitive sciences, and especially in computational learning theory, relates to hypothesis selection in light of observed data. Knowledge of a specific language is a hypothesis selected on the basis of primary linguistic data. Language acquisition, at this level, can therefore accurately be seen as the result of generalisation, if for no other reason than that knowledge of language describes an infinite number of utterances, and this knowledge was acquired in light of a finite set of experiences. However, it is likely that Chomsky's reticence in using the term is based on a wish to avoid confusion. Language is not learned in the everyday sense of the word:

> although you would normally say that an athlete learns how to throw a javelin, you wouldn't generally say that she learns how to grow her biceps. Language acquisition a process more like biceps growing than javelin throwing. (Uriagereka 1998:12)

The second perspective relating to language acquisition and generalisation is far more important. Taken to an extreme, the linguistic stimulus provides triggers that elicit biologically pre-configured linguistic forms (Gibson & Wexler 1994). Specialised acquisition processes spot these triggers, and the result is that the parameters of UG are set. In contrast, the role of the linguistic stimulus with respect to the opposite extreme is quite different. Here, relatively domain-general learning procedures induce the basic properties of language from the rich structure available to the learner. The intricacies of the structure of language are therefore induced by the learner, rather than specified as part of the machinery supporting acquisition. From a Chomskian perspective, this view of acquisition is controversial due to its failure to recognise the poverty of the linguistic stimulus: The chief complaint is that "linguistic structure is much more complex than the average empiricist supposes" (Wagner 2001). Nevertheless, before discussing the importance of this debate, I will briefly mention some convincing work in support of the view that generalisation does play a role in the acquisition of linguistic structure.

Christiansen & Devlin (1997) examine the learning of recursive structure in simple recurrent neural networks. After constructing 32 test grammars, which represent all combinations of head order, they found that grammars exhibiting consistent head ordering were easier to learn. Christiansen & Devlin go on to argue that word order universals can therefore be explained in terms of an acquisition procedure "without linguistic biases, demonstrating that recursive inconsistencies directly affect the learnability of a language" (Christiansen & Devlin 1997:113). Furthermore, these results are consistent with typological language data (Dryer 1992), suggesting that infrequently observed word ordering constraints are infrequent precisely because they are hard to learn. Christiansen & Devlin conclude that "innate linguistic knowledge may not be necessary to explain word order universals".

The acquisition of words themselves has also been shown to rely on non-linguistic learning biases. Saffran *et al.* (1996) examine the segmentation of words from fluent speech by infants. At precisely the point that infants are widely observed to acquire words – around 8 months old – Saffran *et al.* investigate the mechanisms infants use to isolate these words. They show that as a result of only 2 minutes of exposure, infants can exploit the statistical regularities in nonsense syllables they were presented with. Saffran *et al.* (1996) go on argue:

> that a fundamental task of language acquisition, segmentation of words from fluent speech, can be accomplished by 8-month-old infants based

> solely on the statistical relationships between neighbouring speech sounds. (Saffran *et al.* 1996)

If the infants are exploiting statistical regularities in the data, then we must question the view that generalisation plays no role in language acquisition.

How do these findings relate to the biological specification of a language-specific UG? The most extreme response is that language is learned without any language-specific constraints on the learning process, ie., there is no UG. I will interpret these results from a less extreme angle, and simply note that generalisation forms a central part of the language acquisition process. This position does not require a total rejection of the role of language-specific constraints. Indeed, it should be noted that nobody is proposing that language is learned from a *tabula rasa*. The fundamental point is this: Chomsky's argument – the argument from the poverty of the stimulus – is a prima facie explanation for innate linguistic knowledge (Pullum & Scholz 2002; Cowie 1999). The resolution of the debate is surely one of the most important outstanding questions in linguistics and the cognitive sciences in general. It is an empirical question for which no rigorously attained body of evidence has yet been presented. Just how influential the learning process is in arriving at knowledge of language remains frustratingly unclear. I will therefore leave this issue as an open question, and note that the only sensible position is one articulated by, for example, the prominent linguist Ray Jackendoff:

> I agree that learning which makes more effective use of the input certainly helps the child, and it certainly takes some of the load off the Universal Grammar. But I do not think it takes *all* the load off. It may allow Universal Grammar to be less rich, but it does not allow UG to be dispensed with altogether. (Jackendoff 2002:82)

2.2.2 The principle of detachment

In order to take stock, I will now set aside the issue of how much of language is innately specified and how much is learned. Whichever stance one takes, I will argue that the set of explanatory mechanisms adopted is essentially the same; further empirical findings will only narrow distinctions rather than open up alternatives.

It is important to reflect on the original question: How and where are the universal features of language specified? On accepting the above means of explanation it is vitally important to note that an explanation for the universal features of a

population level phenomenon – language – has been reduced to the problem of the knowledge of language acquired by individuals. Of course, languages vary greatly across populations, 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. But are we really justified in narrowing our explanatory vocabulary so quickly? Before doing so, it is surely useful to be aware of what kind of explanations are being discounted. The de-emphasis of context, culture and history is a recurring theme in the cognitive sciences, 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 1985:41)

Now, taking this standpoint is understandable and perhaps necessary when embarking on any practical investigation into cognition. The result of this line of explanation is that we consider universal features of language to be strongly correlated with an individual's act of cognition, which is taken to be biologically determined. The advantages of such an orientation are clear: we have isolated the object of study. Understanding the innate linguistic knowledge possessed by humans will lead us to an understanding of why universal features of language are the way they are. Chomsky is clearly confident that an explanation for language universals turns on an explanation of UG as the initial state of the language faculty:

> it must be that the basic structure of language is essentially uniform and is coming from inside, not from outside. (Chomsky 2002:93)

For the purposes of this study, I will characterise this position as the Principle of Detachment:

Definition 1 (Principle of detachment) *A thorough explanation of the cognitive processes relevant to language, coupled with an understanding of how these processes mediate between input (primary linguistic data) and output (knowledge of language), would be sufficient for a thorough explanation of the universal properties of language.*

In other words, an explanation for the universal features of language amounts to an explanation of the possible states of the biological machinery elicited by the

linguistic input. This explanatory orientation is characterised by circumscribing the objects of study: biological machinery, and structural changes to that machinery that arise as a result of the linguistic stimulus. To be absolutely clear on this point, Chomsky states that the language acquisition device (LAD) – a piece of biological machinery – on the basis of linguistic stimulus, maps this stimulus to a system of grammatical rules. He states:

> 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 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 I have termed the principle of detachment, runs right through the cognitive sciences and amounts to a general approach to studying competence as a cognitive process. 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; Marr 1982).

The important point here is that by highlighting the principle of detachment, the explanatory framework for explaining why language is the way it is is made explicit. This framework is tuned to the assumption that basic properties of language are a *direct* consequence of a piece of biological machinery. Now, if the result of language acquisition is simply to elicit properties of the machinery, then the principle of detachment is entirely justified. But if language acquisition is regarded as the process of learning – such that generalisation *does* play a role – then by definition properties of the state of the machinery and its output reside in the linguistic stimulus as well. This must be the case, as the possible states of the machinery are strongly determined by the properties contained in the linguistic stimulus. The first point to note is that the principle of detachment acknowledges this possibility: it requires an understanding of the primary linguistic data. Now here is the point at which the explanatory inadequacy creeps in: the principle of detachment cannot offer an explanation for *why* the linguistic stimulus contains the information it does. In short, if we are interested in a thorough explanation of linguistic structure, then there is a gaping hole in the explanatory vocabulary. All theories subscribing to

the principle of detachment can explain nothing about the *source* of the structure of the stimulus.

2.3 The nature of explanation in the cognitive sciences

The principle of detachment is questionable. An adequate explanation of why language exhibits the structure it does will require breaching this principle. The next step is to recognise that the principle of detachment, in a number of guises, has influenced the explanatory vocabulary of the cognitive sciences:

> It is probably no overstatement to suggest that much of cognitive science is still dominated by Chomsky's nativist view of the mind. (Quartz & Sejnowski 1997)

I will discuss two bodies of work that suggest how the current explanatory inadequacy can be resolved. First, I will focus on cognitive science, and in particular, artificial intelligence. I will note how both have historically adopted a conceptual framework influenced heavily by the Chomskian perspective. Recent work in artificial intelligence – and the cognitive sciences in general – has suggested an alternative orientation to address the deficiencies in artificial intelligence and cognitive modelling. Under this alternative approach, cognition is seen as a situated act, such that a full understanding of the environment of cognition is central to understanding its function and mechanisms. This orientation ultimately relates to how intelligent agents are built, but also informs the explanatory vocabulary used to understand intelligent action.

The second body of work, which will prove more relevant to the remainder of the thesis, is that of evolutionary linguistics. To understand how language evolved, we must be clear on *what* has evolved. The principle of detachment suggests that the biological machinery underlying the processing and understanding of language is the principal object of study. Biological evolution therefore lies at the root of the explanation of why language is the way it is. The principle of detachment is implicit in such an account: the evolution of the language faculty amounts to an explanation of the evolution of the characteristic structure of languages. Once again, this assumption has recently been questioned, and the insights motivating these questions will form the backbone of this thesis.

2.3.1 Artificial intelligence: Explaining by building

One of the aims of cognitive science, and in particular, Artificial Intelligence (AI), is to explain human and animal cognition by building working computational models. Some of the overarching principles that have guided AI research over the past half-century were laid down in an intellectual climate influenced by Chomsky's attack on behaviourism. The principle of detachment shares some of the assumptions made by what has now come to be termed *Good Old Fashioned AI* (GOFAI) (Haugeland 1997). Two characteristics of GOFAI are relevant to this discussion:

Disembodiment: Issues arising from the physical instantiation of the agent are not taken to be theoretically significant. A desirable theory of cognition is taken to be invariant over physical manifestation, and can therefore be investigated by designing abstract disembodied agents.

Detachment: Cognition is abstracted to a degree that makes the precise nature of the environment insignificant. In this sense, cognition is investigated out of context, and frequently detached from the issues of perception and action.

A conceptual framework adopting a disembodied and detached view of cognition makes building models a lot easier by focusing on abstracted and general explanations of cognition. It is worth noting that, traditionally, the goals of AI differ from the goals of this discussion. AI seeks to build models, and generally speaking, the model is taken as an explanation. In contrast, this discussion aims to draw up an acceptable explanatory framework for understanding the universal properties of language. These aims are quite separate, but related. AI is in line with the principle of detachment, as the machinery is seen as the focus of the explanation. Understand the cognitive machinery, and we can understand the workings of the cognitive system.

Disembodiment and detachment assume that the characteristic properties of the machinery are elicited, rather than partially determined by, the role of the environment. Therefore, by building the machinery, we assume that the basic properties of the device can be explained. Now, the key point I will make is that, over the course of AI's development, both these principles have been cast into doubt. Recent developments in the cognitive sciences provide a source of alternatives when questioning the principle of detachment. Can these developments shed light on an explanation for the universal properties of language? I will argue that they can: the conceptual framework for studying cognition has undergone significant revision. As Markman

and Dietrich note: "there is a revolution in the air in cognitive science" (Markman & Dietrich 2000:470).

Situated and embodied cognition

The fact is that after half a century of research, AI has failed to measure up to expectation, and as Jerry Fodor noted, "AI has walked into a game of 3-dimensional chess, thinking it was tic-tack-toe" (Dreyfus & Dreyfus 1990). For many, the hallmarks of this failure are: *lack of scalability*, where systems that operate with respect to micro-worlds are proven hopelessly inadequate when placed in more complex environments; *lack of robustness*, where systems fail to deal with unforeseen circumstances; *failure to operate in real-time*, where the information processing load of a system is so great as to be intractable as a valid theory (Pfeifer & Scheier 1999:59-78). Those who have sought to re-orientate AI, in one form or another, cite *situatedness* and *embodiment* as design maxims (Dreyfus 1972; Winograd & Flores 1986; Clancy 1997; Brooks 1999; Pfeifer & Scheier 1999).

Adopting the property of situatedness in an explanation implies that the precise nature of the environment is a theoretically significant consideration. In the limit, the property of situatedness means that an understanding of cognition cannot be separated from intricacies of the environment. For example, some environmental states may be persistent, in which case an agent need not represent them in the traditional sense, where a significant part of the processing load of the agent is to keep internal representations in synch with the outside world. Instead, the knowledge that the information will be there when it is required is sufficient. Without examining an act of cognition in this context, one might incorrectly attribute such persistent and detailed representations to the agent. In essence, as soon as one admits the possibility of a cognitive architecture exploiting the structure in its environment, then an abstract, detached treatment of cognition can become inaccurate. These inaccuracies relate directly to the perceived problems with GOFAI: systems may fail to operate in real time because exploitation of the environment is not possible if the precise nature of the environment is not considered; systems are not scalable precisely because cognitive structures are tuned to very specific environmental properties.

Embodiment refers to the importance of the nature of the physical instantiation of the agent. The sensory surfaces used by an agent define the nature and range of stimuli that can be exploited. Taking embodiment to its limit requires us to build physically instantiated biological agents. Without this physical grounding, and an

understanding of the perceptual and motor capacities that grounding requires, it is impossible to grasp an accurate reflection of the environment. Embodiment, then, is a property that is closely tied to the issue of situatedness, as an accurate picture of situatedness requires an accurate picture of the phenomenal world the agent inhabits.

The property of situatedness is most relevant to the present discussion. But how exactly does situatedness relate to an explanation of the universal properties of language? The explanatory inadequacy resulting from the principle of detachment – the lack of an explanation for the nature of the linguistic stimulus – occurs as a result of the assumption that situatedness is not a theoretically significant consideration. An explanation for the universal properties of languages requires an explanation of the situated nature of language acquisition. If language learning is possible as a result of the nature of the data – and learners are exploiting this input – then a large part of the explanation for the universal features of language will be an explanation for why the data is as it is. A fully situated account would explain this.

2.3.2 The evolutionary explanation

Only humans have language. The communication systems used by animals do not even approach the sophistication of human language, so the evolution of language must concern the evolution of the human line over the past 5 million years, since our last common ancestor with a non-linguistic species (Jones *et al.* 1992). Consequently, examining fossil evidence offers a source of insights into the evolution of language in humans. For example, we can analyse the evolution of the vocal tract, or examine skulls and trace a path through the skeletal evolution of hominids, but the kinds of conclusions we can draw from such evidence are severely limited (Lieberman 1984; Wilkins & Wakefield 1995).

One route to explaining the evolution of language in humans, which we can dub *functional nativism*, turns on the idea that language evolved in humans due to the functional advantages gained by linguistically competent humans. Language, therefore, was a trait selected for by biological evolution (Pinker & Bloom 1990; Nowak & Komarova 2001). Here, we can imagine an evolutionary trajectory starting from some biological predisposition present in proto-humans for using some set of communication systems C_{proto}. From this starting point, biological evolution led to the occurrence of the set of communication systems C_{UG}, which includes all human languages. The story of language evolution can then unfold by claiming that the

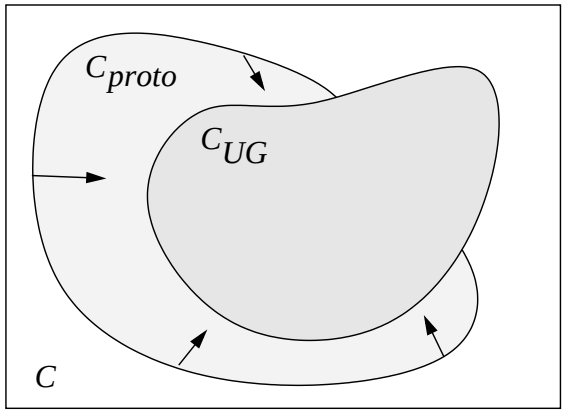


Figure 2.1: Functional nativism. From the set of all communication systems C, the communication systems of proto-humans, C_{proto}, evolved under some functional pressure towards C_{UG}.

biological machinery supporting C_{proto} evolved to support C_{UG} due to functional pressures (see Figure 2.1). Implicit in this account is the principle of detachment. The biological evolution of cognitive capacities supporting language are equated with the evolution of languages themselves.

Over the past 15 years computational evolutionary linguistics has emerged as a source for testing such hypotheses. This approach employs computational models to try and shed light on the problem of the evolution of language in humans (Hurford 1989; Kirby 2002b; Briscoe 2000). One source of complexity in understanding the evolution of language is the interaction between three complex adaptive systems, each one operating on a different time-scale. 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. In addition to these evolutionary systems, your linguistic competence emerges over tens of years. Much of the work in computational modelling has analysed this interaction. By modelling 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 can evolve in initially non-linguistic communities. This approach draws on disciplines such as cognitive science, artificial life, complexity, and theoretical biology. Recent work in this field has focused 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 neo-Darwinian 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.

I will detail this work in the next section, but I mention it here as it impacts on the current discussion. In explaining how and why language has its characteristic structure, the evolutionary approach, by investigating the interaction between biological and cultural substrates, is in line with the claims made by proponents of embodied cognitive science. Because languages themselves can adapt, independent of the biological substrate, certain features of language cannot be explained in terms of detached cognitive mechanisms alone.

2.4 Iterated learning

The previous section introduced the idea that linguistic evolution can occur on a cultural *substrate*. To clarify this point, it is worthwile noting that linguistic information is transmitted on two substrates. Firstly, biological evolution relies on a genetic substrate: information relevant to language is encoded in the genes. Secondly, cultural transmission and adaptation allows information relevant to language to be encoded within languages themselves. That is, language itself supports a substrate for the transmission of information relevant to language. In short, linguistic forms are determined not only by the underlying biological machinery, but also as a result of cultural adaptation. In this section I will flesh these ideas out, and relate them to an explanation of linguistic structure in more concrete terms.

An iterated learning model (ILM) is a framework for testing theories of linguistic evolution. Within an ILM agents act as a conduit for an evolving language – the language itself changes or evolves rather than the agents themselves. An ILM is a generational model: after members of one generation learn a language, their production becomes the input to learning in the next generation. Within this model of linguistic transmission, providing that the transfer of knowledge of language from one generation to the next is not entirely accurate or reliable, will result in diachronic change. Importantly, certain linguistic structure will survive transmission, while other forms may disappear. The precise nature of the information being transmitted depends on the theory in question. Here, I will discuss two broad categories of theory: those of language change and language evolution. The investigation of language change, although relying on an iterated learning framework, can only have a limited impact on our discussion of the principle of detachment: recall that the

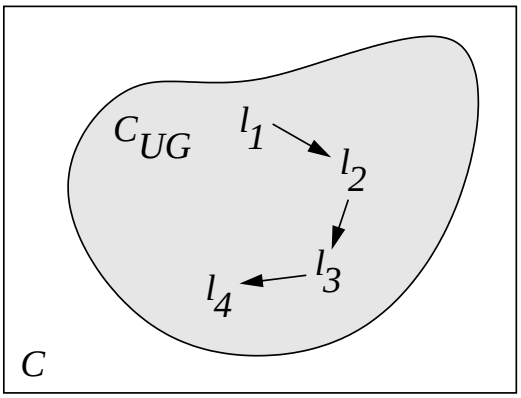


Figure 2.2: Language change. An example trajectory of language change through languages l_1, l_2, l_3, and l_4.

principle of detachment, according to its definition, requires an explanation of *absolute* language universals. Absolute universals are typically preconfigured in models of language change: an explanation of how these hallmarks came about is therefore outside the explanatory reach of such models. However, theories of language change, by explaining aspects of language which are observed to vary, can impact on an explanation of, for example, statistical universals.

2.4.1 Language change

In studying language change we often consider the trajectory of language through possible grammars. Any resulting explanation is therefore orientated neutrally with respect to explaining certain forms of absolute universal: the very possibility of a grammar presupposes compositional and recursive structure. From one grammar to the next, we presume many of the hallmarks of language to be ever-present (see Figure 2.2). Models of language change must invoke a situated component. A model must tackle the problem of language acquisition: a learner will deviate from the grammar of its teachers when the primary linguistic data fails to unambiguously represent the grammar from which it is derived. Knowledge of language is therefore not transmitted directly from mind to mind: some external correlate – linguistic performance – must stand proxy for knowledge of language (Hurford 1990). Modelling language change must therefore consider some environment allowing the transmission of language competence via language performance. This environment, importantly, is constructed by others.

Using iterated learning, we can construct computational models of language change. These studies are motivated by the observation that language change is driven by

considerations arising from language acquisition (Clark & Roberts 1993; Niyogi & Berwick 1997; Briscoe 2002). For example, using a principles and parameters approach to language description, Niyogi & Berwick (1997) develop a population model with which they investigate the dynamics of language change. In particular, they use a probabilistic model of grammar induction to focus on the loss of Verb-second position in the transition from Old French to Modern French, which results directly from misconvergences arising during language acquisition. In contrast, Hare and Elman address the problem of morphological change by looking at connectionist simulations of language learning, which, when placed in the context of iterated learning, can be used explain morphological changes such as verb inflection in Modern English arising from the inflectional system of Old English (Hare & Elman 1995). Importantly, the linguistic phenomena these models attempt to explain is relatively well documented: the historical accuracy of models of language change can be tested.

2.4.2 Language evolution

These studies of language change tell us that the learnability of languages, over the course of cultural transmission, has a bearing on the distribution of languages we observe. Now I will discuss extending the range of explanation offered by models of iterated learning to include the possibility of explaining hallmarks of language. The dynamics of iterated learning can make certain properties of communication systems ubiquitous. This must lead us to consider the fact that, just as the dimensions of variation can be explored via iterated learning, the undeviating features of language may also depend on issues of learnability.

The possibility that iterated learning models can shed light on an explanation of these properties will make a convincing case for questioning the principle of detachment. If the unvarying features of language can be explained in the same way as those that vary, then issues of innateness become problematic and less clear cut. For example, both Christiansen, Deacon, and Kirby have claimed previously that universals should, at least in part, be seen as arising from repeated transmission through learning:

> In short, my view amounts to the claim that most – if not all – linguistic universals will turn out to be terminological artifacts referring to mere side-effects of the processing and learning of language in humans (Christiansen 1994:127)

> Grammatical universals exist, but I want to suggest that their existence does not imply that they are prefigured in the brain like frozen evolutionary accidents. In fact, I suspect that universal rules or implicit axioms of grammar aren't really stored or located anywhere, and in an important sense, they are not *determined* at all. Instead, I want to suggest the radical possibility that they have emerged spontaneously and independently in each evolving language, in response to universal biases in the selection processes affecting language transmission. (Deacon 1997:115-116)

> The problem is that there are now two candidate explanations for the same observed fit between universals and processing — a glossogenetic one in which languages themselves adapt to the pressures of transmission through the arena of use, and a phylogenetic one in which the LAD adapts to the pressures of survival in an environment where successful communication is advantageous. (Kirby 1999:132)

These arguments place an explanation for the universal features of language well and truly outside the framework of explanation suggested by the principle of detachment. In the context of cultural transmission, I will term the process by which certain linguistic forms are adaptive and therefore evolve and persist *cultural selection for learnability.* More precisely:

Definition 2 (Cultural adaptation) *By cultural adaptation, I mean the occurrence of changes in the language due to the effects of cultural transmission.*

It is important to contrast the notion of cultural adaptation to that of genetic adaptation, where genetic changes occur as a result of natural selection. The notion of cultural adaptation refers to the *language* adapting, rather than the users of language. Next, I will define cultural selection for learnability:

Definition 3 (Cultural Selection for Learnability) *In order for linguistic forms to persist from one generation to the next, they must repeatedly survive the processes of expression and induction. That is, the output of one generation must be successfully learned by the next if these linguistic forms are to survive. Those forms that repeatedly survive cultural transmission are adaptive in the context of cultural transmission: they will be selected for due to the combined pressures of cultural transmission and learning.*

In this context, the terms *adaptive* and *selection* only loosely relate to the equivalent terms used in the theory of biological evolution. The idea that languages themselves adapt to be learnable, and in doing so organise themselves subject to a set of recurring structural properties, has been the subject of computational models that make explicit these assumptions. In particular, the experiments of Kirby (2002a) and Batali (2002) demonstrate that a collection of learners with the ability to perform grammar induction will, from an initially holistic communication system, spontaneously arrive at compositional and recursive communication systems. These experiments will be discussed more thoroughly in Chapter 3, but a general reading is as follows. Because language is ostensibly infinite, and cultural transmission can only result in the production of a finite series of utterances, generalisable forms will have a tendency to survive.

These experiments suggest that certain hallmarks of language are culturally adaptive: pressures arising from transmission from one agent to another cause these hallmarks to emerge and persist. For example, adaptive properties such as compositionality and recursion, which we can consider absolute language universals, are defining characteristics of stable systems. These models suggest that extending and enriching the explanatory framework suggested by the principle of detachment is a fruitful line of research. In particular, if the precise nature of the environment of language adaptation is to play a pivotal role, as suggested by situated theories of cognition, then further modelling may shed light on a wider range of linguistic forms. For example, Kirby (2001) demonstrates that by elaborating the environment by imposing a non-uniform distribution over the set of communicatively relevant situations, irregular forms emerge. By skewing the relative frequency of utterances, irregular forms can exist by virtue of the fact they are frequently used, and therefore are subject to a reduced pressure to be structured. Similarly, Smith *et al.* (forthcoming) show how clustering effects in the space of communicatively relevant situations leads to a stronger pressure for compositionality.

These studies demonstrate that the precise nature of the factors underlying language adaptation impacts on the resulting language structure. By understanding the import of environmental considerations on the evolved languages, in tandem with an investigation into plausible models of language acquisition, the aim is to shed further light on the relationship between cultural selection and the structure of evolved languages. In Chapter 5, I will conduct a large-scale exploration of such environmental considerations.

In this section I have discussed how models of language evolution and change based on a cultural, situated model of linguistic transmission can shed light on the occurrence of hallmarks of language. I will enter into more thorough discussions as the thesis unfolds. At this point, I am now in a position to present the theoretical foundations that will motivate and guide the remainder of the thesis.

2.5 Three hypotheses underlying the view that language adapts to be learnable

I began this discussion by considering the manner in which language universals should be explained. I now aim to make clear the principles that underly the view that language universals are, at least in part, the result of cultural selection for learnability. It is useful to distinguish two mutually supporting streams of discussion. First, the status of the notion of cultural selection in linguistic theory. For example, how does cultural selection for learnability impact on the theory of linguistic nativism? This stream of inquiry will, I hope, be clarified by the immediate discussion. The second stream of inquiry concerns the use of modelling to support and inform the theoretical claims. Before I focus on the latter, the theoretical foundations and claims need to be made explicit.

2.5.1 Innateness hypothesis

I will start by noting that any conclusions we draw will be contingent on an innateness hypothesis:

Hypothesis 1 (Innateness hypothesis) *Humans must have a biologically determined set of predispositions that impact on our ability to learn and produce language. The degree to which these capacities are language-specific is not known.*

Here I am 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. As I have noted, linguistics lacks a solid theory, based on empirical findings, that details which aspects of language can be learned, and which must be innate (Pullum & Scholz 2002). Next, I will consider the innateness hypothesis with respect to two positions. Firstly, 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. Secondly, if we extend the framework 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.

2.5.2 Situatedness hypothesis

This discussion now centres around recasting the question of innateness. Furthermore, this observation, because it relates to a cultural dynamic, leads us to accepting that situatedness plays a role:

Hypothesis 2 (Situatedness hypothesis) *A thorough understanding of the cognitive basis for language would not amount to a total explanation of universal language structure. However, a thorough understanding of the cognitive basis for language in conjunction with an understanding of the trajectory of language adaptation through cultural transmission would amount to a total explanation of language structure.*

Of course, the degree of correlation between a piece of biological machinery supporting some aspect of language and the resulting language universal is hard to quantify. But in general, some biological basis for language 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 culturally adaptive and therefore observed. The situatedness hypothesis changes the state of play by considering those communication systems that are adaptive, $C_{adaptive}$, on a cultural substrate, and therefore observed. In short, cultural selection for learnability occurs with respect to constraints on cultural transmission. These constraints determine which members of $C_{possible}$ are culturally adaptive, observed, and therefore become members of the set $C_{adaptive}$.

By conjecturing an opaque relationship between some biological basis for language and some observed language universal, the notion of UG becomes problematic. Universal grammar is often taken to mean one of two things. First, the term UG is sometimes used to refer to the set of features that all languages have in common[1]. Secondly, and perhaps more frequently, UG has been defined as the initial state of the language-learning child (Chomsky 1975). Figure 2.3 depicts how these two definitions relate to our discussion of the biological basis for language, the set of

[1]For a discussion on the historical use of the term *universal grammar*, see Jackendoff (2002:69-70).

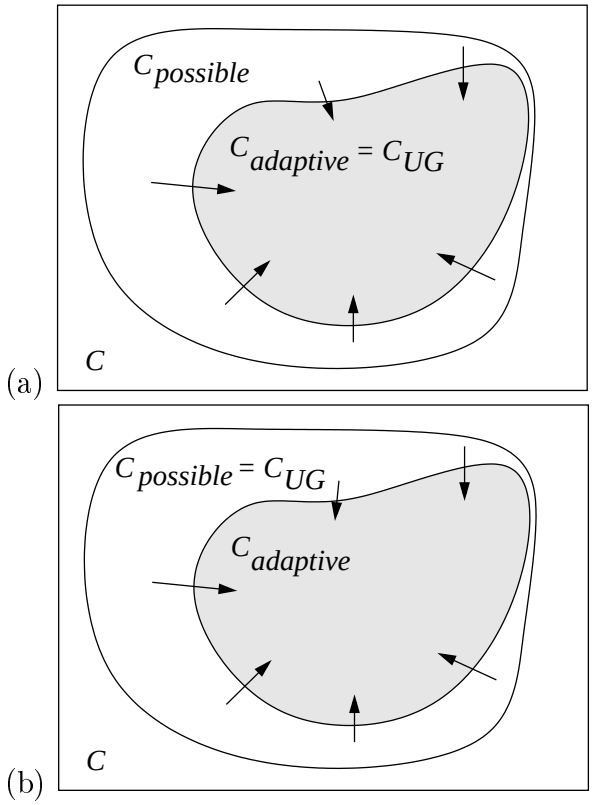


Figure 2.3: Of the set of possible human communication systems $C_{possible}$, some set $C_{adaptive}$ are adaptive in the context of cultural transmission, and therefore observed. Depending on how we define UG, the set of communication systems characterised by UG, C_{UG}, is either precisely those we observe ($C_{adaptive}$), or those that are possible, but not necessarily observed ($C_{possible}$).

possible communication systems, and the set of observed communication systems. The set of communication systems that conform to the definition of UG are denoted as C_{UG}. Depending on which definition of UG we adopt, this set will be equivalent to either $C_{possible}$ or $C_{adaptive}$. These two alternatives are now explored:

1. *UG as the set features common to all languages.* If we take UG as the set of features common to all observed languages, then C_{UG}, the set of communication systems conforming to UG, is identical to our set of culturally adaptive communication systems, $C_{adaptive}$. This must be the case, as only members of $C_{adaptive}$ are observed and can therefore contribute to a theory of UG under this reading. This position is represented in Figure 2.3(a).
2. *UG as the initial state of the language-learning child.* The alternative definition of UG, where UG defines the initial state of the learner, must encompass those communication systems which are possible, but not necessarily adaptive: $C_{possible}$. Because humans are equipped with the biological basis for using

> members of $C_{possible}$, their initial state must account for them. Hence, under this second reading of UG, $C_{possible} = C_{UG}$. As before, only some members of $C_{possible}$ will be culturally adaptive and therefore observed. Figure 2.3(b) reflects this relationship.

Irrespective of our definition of UG, 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. 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 resulting in adaptive properties emerging, for example, issues relating to communication and effective signalling.

2.5.3 Function independence hypothesis

As a first approximation, we need to understand how much can be explained without appealing to any functional properties of language. By functional, I mean relating to *communicative* function.

Hypothesis 3 (Function independence hypothesis) *Some aspects of language structure can be explained independently of language function.*

A defence of this hypothesis is less clear cut. Without doubt language is used for communication, but whether issues of communication determine *all* forms of language structure is by no means clear (Newmeyer 1998). The function independence hypothesis should be contasted with a functionalist stance, which

> analyzes grammatical structure [...] but it also analyzes the entire communicative situation: the purpose of the speech event, its participants, its discourse context. Functionalists maintain that the communicative situation motivates, constains, explains, or otherwise determines grammatical structure (Nichols 1984:97)

The orientation suggested by the function independence hypothesis is best summed up by Chomsky's statement that:

> language is not properly regarded as a system of communication. It is a system of expressing thought, something quite different. It can of

> course be used for communication, as can anything people do – manner of walking or style of clothes or hair, for example. But in any useful sense of the term, communication is not *the* function of language, and may even be of no unique significance for understanding the functions and nature of language. (Chomsky 2002:76)

The picture I am developing here suggests that constraints on learning and repeated cultural transmission play an important part in determining linguistic structure. The models I will discuss make no claims about, nor explicitly model, any notion of communicative function.

2.6 Chapter Summary

Absolute universals are hallmarks adhered to by every user of language. We might then take the individual as the locus of study when seeking an explanation for why absolute language universals take the form that they do. In line with this intuition, practitioners of cognitive science will often make the simplifying assumption that the behaviour of individuals can be understood by examining the internal cognitive processes of detached agents. The principle of detachment characterises this position.

In attempting to understand how and where language universals are specified, this discussion has focused on questioning the principle of detachment. I have explored two sources of insights that suggest that an explanation of the characteristic structure of language could benefit from breaching the principle of detachment. Firstly, advocates of situated cognitive science claim that the property of situatedness, a full understanding of the interaction between agent and environment, is theoretically significant. Secondly, recent work in the field of computational evolutionary linguistics suggests that cultural dynamics are fundamental to understanding why linguistic structure evolves and persists. I should stress here that in one respect languages are not stable, they are constantly changing. But in contrast, absolute language universals are entirely stable, or at least they have been over the duration of modern linguistic inquiry[2].

Taking these two sources as evidence, I introduced the notion of iterated learning as a means to explore the relation between absolute language universals and those

[2]See Newmeyer (2002) for discussion of this and other issues that relate to "uniformitarianism" in linguistics.

linguistic features that are adaptive in the context of cultural transmission. On the basis of this concept, I claim that cultural selection for learnability must form part of any explanation relating to how and where the hallmarks of language are specified. I also claim that, due to constraints on cultural transmission, languages adapt to reflect the biases present in language learners and producers. The relationship between these biases and the observed hallmarks of language is therefore opaque: a cultural dynamic mediates between the two.

Here is the message I wish to convey: Selection for learnability is an important determinant of language universals, and as such should be understood independently of any particular computational model. The aim is to outline the theoretical foundations of cultural selection for learnability. I have done this by proposing three hypotheses. First, the Innateness Hypothesis (Hypothesis 1) states that there must be a biological basis for our language-learning abilities, but the degree to which these abilities are language-specific is unclear. The second hypothesis, the Situatedness Hypothesis (Hypothesis 2), states that language universals cannot be explained through an understanding of the cognitive basis for language alone. I claim that certain properties of language are adaptive in the context of cultural transmission. The third hypothesis, the Function Independence Hypothesis (Hypothesis 3), makes clear that the communicative functions of language are not necessarily determinants of language structure. I note that an explanation for some absolute universals, such as compositional syntax, need not appeal to any notion of language function. In short, I seek an afunctional explanation for certain aspects of linguistic structure.

By questioning the principle of detachment and pursuing a line of enquiry guided by Hypotheses 1-3, I have argued that the concept of cultural selection for learnability can provide important insights into some fundamental questions in linguistics and cognitive science. This motivation for this Chapter is, first, to question the explanatory framework used in explaining the the universal characteristics of language, and second, to make explicit a set of hypotheses that can be used to engage with current linguistic theory. These hypotheses set the scene for exploring the role of learning in linguistic transmission. It is the role learning as a vehicle for information transmission that I will turn to next.

Chapter 3

Iterated Learning

3.1 Introduction

Iterated learning is a candidate explanatory framework for understanding why language exhibits certain hallmarks and not others. The following discussion will start by framing the process of iterated learning in the context of more general theories of cultural evolution. The main point of discussion, however, will be the development of iterated learning *models*. The foundational models of Batali (2002) and Kirby (2002a) are taken as the starting point in examining how iterated learning models can be used to explore the effects of repeated cultural transmission. These models make claims about the evolution of hallmarks of language – absolute universals – through linguistic evolution. A number of issues are raised by these experiments, and these questions suggest a programme for future research into iterated learning models. The motivation underlying this discussion is the need to introduce the key concepts in iterated learning. I will do this by way of examples. In Section 3.3, I will develop an illustrative example of iterated learning, where classification competence is subject to constrained cultural transmission. Then, in Section 3.4, I will extend this model and tackle the question of compositionality in language. In doing so, many of the fundamental issues and methods in iterated learning models will be set out. The contribution of this chapter is therefore twofold. First, it will act as an introduction to the key concepts in iterated learning as both a process and as a modelling framework. Second, foundation results concerning the emergence of compositional structure will be developed.

3.2 Foundations and context of iterated learning

Rather than simply eliciting pre-defined biologically determined linguistic properties, language acquisition may play a far more important role. If this is the case, then the linguistic stimulus may encode basic properties of language. As soon as we entertain the possibility that hallmarks of language are partly determined by the content of the linguistic stimulus, then, as I have argued, any acceptable explanation for these hallmarks must also explain the origin of the stimulus. It is clear that such an explanation requires a situated account of language learning; one that requires breaching the principle of detachment. In Figure 3.1(a) the detached account underlying the traditional explanation of knowledge of language is schematised: knowledge of language can be understood in terms of the causal properties of the linguistic stimulus in determining configurations of the biological machinery relevant to language. If the linguistic stimulus contains significant information, and this information determines what we trying to explain, then the source and nature of this information must be understood. An extension to the principle of detachment, such that this source is made explicit, is schematised in Figure 3.1(b): the linguistic stimulus is constructed by the behaviour of other agents. The situated account leads us to acknowledge that understanding knowledge of language requires an understanding of the role preceding language learners play in constructing the linguistic stimulus.

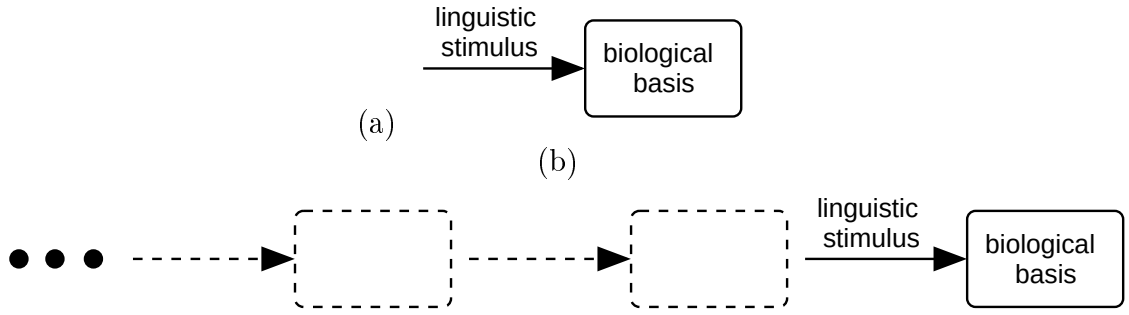


Figure 3.1: Detached versus situated accounts. In (a), an explanation consists of an understanding of the biological machinery in conjunction with the role of the linguistic stimulus. In (b), the explanatory range is widened to account for the origin of the linguistic stimulus by acknowledging the role of previous language users.

Iterated learning, where learners learn from the output of other learners, uncovers a set of issues that will impact on an explanation of why language is the way it is. Iterated learning leads to a form of information transmission in which linguistic information is continually translated between, using Chomsky's terminology,

I-Language and E-Language. I-language is the internal configuration of cognitive structures relevant to knowledge of language. E-Language is the externalised linguistic performance derived from internal linguistic competence, that is, I-language. Iterated learning therefore hinges on the observation that linguistic transmission is meditated by repeated translations between I-Language and E-Language. This idea is not new. Although it is implicit in the statements of Christiansen (1994), Deacon (1997), and Kirby (1999), cited in the previous Chapter, it has also been discussed, for example, by Andersen (1973) and Hurford (1990). The crux of the issue is that, across generations of language users, I-language is translated into E-language and then reverse-translated back into I-language. This process is repeated generation after generation, and, due to the imperfect nature of the processes of translation and reverse translation, results in linguistic change. Just how important this process is in determining hallmarks of language, however, is another issue entirely. The working assumption of this discussion is that, due to the effects of repeated language expression and language induction, the linguistic stimulus will contain structural properties introduced by language users many generations before: language reflects the accumulated residue of the effects of learning and production of preceding agents.

3.2.1 Cultural processes and models

In broad terms, iterated learning is an ongoing process in which learning agents take their input from the output of other learners. The example discussed above focuses on the specific case where learners are learning language. It is worth pointing out that iterated learning could be regarded as part of a more general information processing problem relevant to not just language, but other learned capacities found in the natural world. For the sake of terminological accuracy, we should regard iterated learning as a domain-general information processing problem. Language is a specific domain of learned behaviour, found in nature, for which the process of iterated learning is relevant.

Looking beyond language, iterated learning, broadly construed, is identical to the process of cultural transmission and evolution described by Tomasello (1999), and one half of the dual inheritance model discussed by Boyd & Richerson (1985). At a more general level, iterated learning can also be parallelled with Dawkins' notion of cultural evolution, but for the time being, it is far from clear to what extent iterated learning should be regarded as an evolutionary process in the usual sense (Dawkins 1982). In this context, culture is used to refer to those behavioural practices that

persist across generations as a result of being learned, rather than being innate. I will focus on Tomasello's discussion. Tomasello's argument addresses a paradox. A wide range of cognitive skills are exhibited by humans that are absent in other species. Furthermore, because humans share 99% of their genetic material with chimpanzees, biological evolution, Tomasello argues, cannot be invoked as the sole explanation for these differences. The time-scale over which these differences could have developed through biological evolution, Tomasello goes on to argue, is simply too small:

> The fact is, there simply has not been enough time for normal processes of biological evolution involving genetic variation and natural selection to have created, one by one, each of the cognitive skills necessary for modern humans to invent and maintain complex tool-use industries and technologies, complex forms of symbolic communication and representation, and complex social organizations and institutions. (Tomasello 1999:2)

Many species learn from observing the behaviour of others, and can therefore exploit the behaviour of conspecifics to make surviving in an environment less risky; rats, for example, learn from the feeding practices from their parents (Mundinger 1980). Importantly, as Boyd and Richerson point out, being the recipient of set of culturally inherited practices allows organisms to gain from the behaviour of genetically unrelated conspecifics; their maxim "inheritance as a shortcut to individual learning" therefore applies beyond inheritance from an organism's forebears (Boyd & Richerson 1985:14). But, as Tomasello argues, humans populations support a far more powerful notion of cultural evolution than any animals. This unique mode of cultural transmission supports *ratcheting*, where not only are cultural constructions successfully and repeatedly transmitted from generation to generation, but cultural *evolution* is possible due to variation and improvements. The ratchet effect is a property that gives rise to cumulative evolution:

> The process of cumulative cultural evolution requires not only creative invention but also, and just as importantly, faithful social transmission that can work as a ratchet to prevent slippage backward – so that the newly invented artifact or practice preserves its new and improved form at least somewhat faithfully until a further modification or improvement comes along. (Tomasello 1999:2)

Importantly, humans are unique in supporting such a cumulative evolution, and this uniqueness is a result of the human ability to "understand conspecifics as beings *like themselves* who have intentional and mental likes like their own." (Tomasello 1999). The notion of the ratchet effect is parallelled in Deacon's concept of linguistic evolution, where he discusses a "self-sustaining core, consisting of a grammar and syntax and a sufficient number of words to determine all critical word classes" (Deacon 1997:113). As I discussed in the previous Chapter, this core is sustained because it is easily learnable, with any variations arising from "transmission errors... [or] by the active creativity of its users" (Deacon 1997:114). It is clear that Tomasello's work on cultural evolution bears many parallels with the process of iterated learning. The view that significant explanatory purchase can be gained by considering cultural transmission is also mirrored in Tomasello's work:

> my central argument [...] is that it is these processes, not any specialized biological adaptations directly, that have done the actual work in creating many, if not all, of the most distinctive and important cognitive products and processes of the species *Homo sapiens*. (Tomasello 1999:11)

Now there is a danger of considerable overlap in terminology: iterated learning, in its widest sense, and cultural evolution refer to the same process. Nevertheless, those studies that invoke the term iterated learning have, so far, specifically concerned explanations of general linguistic properties. These properties are culturally transmitted and learned through observation rather than, for example, instruction. In this respect, iterated learning is a broadly defined *process* in which information is transmitted through observational learning. This discussion concerns *models* of iterated learning, and applies these models to explaining linguistic evolution.

3.2.2 Foundational models

The computational models of language change mentioned in Chapter 2 employ iterated learning models to investigate the impact of language acquisition on likely trajectories through the space of possible human languages (Clark & Roberts 1993; Hare & Elman 1995; Niyogi & Berwick 1997). Although these studies represent the first use of what we now term iterated learning models, they are of limited importance to the current question of the evolution of the hallmarks of language. The existence of absolute universals are an assumption in these models, rather than the target of explanation.

Of more importance to this discussion are the pioneering models of linguistic evolution through iterated learning (Batali 1998; Batali 2002; Kirby 2000; Kirby 2002a; Hurford 2000b). I will refer to four of these models which have been developed, independently, by John Batali and Simon Kirby. The early models of Batali (1998) and Kirby (2000) focus on the emergence of compositionality. Batali's model uses recurrent neural networks to map signals to meanings. In a fixed population without any turnover, Batali shows how repeated inter-agent production and learning of meaning/form pairs results in structured signals emerging, where signal structure reflects the structure present in the meanings. These pre-structured meanings are loosely interpreted by Batali as representing predicate-argument structure. Batali's conclusion is that such rudimentary syntactic structure, which must be a precursor to more complex syntactic forms, can be explained through repeated linguistic expression and induction. Kirby (2000) focuses more explicitly on the general property of compositionality, and presents a very different model, but obtains comparable results. Kirby's meanings are triplets representing predicate-argument relations. Unlike Batali's model, Kirby models a collection of agents subject to a turnover. After a number of rounds of communication, a random agent is removed and replaced by a new blank agent. In doing so, Kirby's model engages more directly with the question of language evolution through cultural transmission over generations of agents. Furthermore, this transmission is explicitly presented in terms of the translation between I-language and E-language. Once again, the outcome of the model is that, from an initially holistic relation between meanings and signals, a compositional relation emerges as the result of the evolution of the signals. Both these models therefore make a claim about the emergence of compositionality from initially holistic systems. The driving force for this linguistic evolution is iterated learning: learners learn from the output of other learners.

To strengthen the range of linguistic structure that iterated learning models can account for, both Batali (2002) and Kirby (2002a) present further models which claim not only to account for the emergence of compositionality, but also recursive structure. I will briefly focus on Kirby's model, as it has been subject to a further critical analysis which I will discuss shortly. Kirby's second model makes two important contributions. Firstly, the range of linguistic structure to be explained is extended to include recursive syntax. Agents in the model are grammar inducers. By enriching the structure of meanings in the model to include nested predicate-argument structure, agents are open to the possibility of processing structured signals that reflect this hierarchical structure. Secondly, the processes driving linguistic evolution are more clearly articulated. Agents are organised into generations. Each

generation contains only one agent. As the simulation proceeds, an agent will produce utterances that are observed by the agent in the next generation. Language is transmitted down the generations by the translation of I-language (grammar) into E-language (utterances). This model is therefore tuned more to an explanation of linguistic evolution across generations, rather than linguistic evolution within a single generation – which is the situation addressed in both of Batali's studies. For Kirby, the driving force for linguistic evolution is the *transmission bottleneck*:

> We can think of the transformations between I- and E-language as a bottleneck on the transmission of language over time... Since the number of meanings that the learners are exposed to is always lower than the total number of meanings, a totally idiosyncratic language *cannot* survive. (Kirby 2002a:194)

In short, the generalisable parts of the language are more likely to survive than idiosyncratic parts of the language. This dynamic leads the system towards fully generalisable languages, and these are syntactic languages. In effect, the bottleneck is restricting the set of stable languages to be those that are structured. These languages are those where the meaning structure is reflected in the signal structure: recursive and compositional syntax is the inevitable outcome of linguistic transmission under these conditions. These conclusions, reached using computational simulation, correspond to those proposed by Christiansen and Deacon. In this sense, models of iterated learning go some way to verifying their theoretical intuitions.

Outstanding questions

The computational simulations discussed above add weight to the view that hallmarks of language should, at least in part, be explained relative to a cultural dynamic. But rather than telling a complete story, these computational models demand an entirely new set of questions. For example, Kirby's model can be regarded as monolithic – it is hard to tease out the critical components. Similarly, the conditions that are critical to emergence are little understood. These concerns are raised by Tonkes & Wiles (2002):

> For computational models such as Kirby's, it is important to establish the features of the abstraction that lead to the observed results. That is,

> we should strive to understand the parts of the abstractions that are required, those which are superfluous, and those that must be constrained to a critical range of values. (Tonkes & Wiles 2002)

Tonkes and Wiles go on to single out the role of learning in Kirby's model as a source of doubt:

> It seems to us that the chosen induction algorithm is highly biased towards language-like, compositional structures, which is perhaps not surprising given that the algorithm was originally developed for computation linguistics. (Tonkes & Wiles 2002)

In order to fully understand the import of these foundational models, the questions raised by Tonkes and Wiles need to be explored. The issues raised by Tonkes and Wiles, in the most part, will be addressed by the next three chapters of this thesis. Firstly, the remainder of this chapter will proceed by incrementally building a model of the emergence of compositionality. In contrast to the models discussed above, I will describe several experimental models that fail to lead to the emergence of structure. The hope is to build assumptions layer upon layer before reaching a model of any explanatory weight. In doing so, we can gain a firmer grasp of the essential abstractions that Tonkes and Wiles mention. Secondly, the issue of learning is tackled in Chapters 4, 5, and 6, where I develop a model of learning based on a solid theory of induction. The impact of this model in the iterated learning framework is explored using mathematical and monte carlo models. In doing so, large regions of the parameter space of the iterated learning model can be explored. In short, the issues I will address are in line with the questions raised by Tonkes and Wiles. Before confronting these key issues, it is important to begin asking some basic questions about iterated learning. It is this issue I will now turn to.

3.3 Iterated learning: An illustrative example

With a view to shedding light on the three hypotheses proposed in Chapter 2, I will work towards developing an iterated learning model of linguistic evolution. This model is covered in Chapters 4, 5, and 6. Before introducing this model, several key concepts need to be established first. I will start by introducing a basic iterated learning model that will serve as an illustrative example. This model will set the scene for exploring a number of issues, and although the model is far removed from

a plausible model of linguistic transmission, it will highlight several fundamental issues related to modelling linguistic transmission.

This basic example of iterated learning concerns the transmission of classification competence. Each agent in the experiment will therefore be a *concept learner*, and each concept learner will learn concepts on the basis of the output of other concept learners. Here, then, is the first parallel with language that we can draw: knowledge of particular task – a competence – is transmitted down generations of agents by each agent externalising behaviour derived from its competence. This behaviour corresponds to the notion of performance in language.

3.3.1 Concept learning agents

In the context of machine learning, concept learning is the problem of forming a hypothesis in light of observed instances, where each instance is drawn from one of a number of classes. The success of a concept learner is measured by the accuracy with which it correctly classifies novel unclassified instances (Michalski 1992; Mitchell 1997). The classic example of concept learning concerns the classification of Iris plants found on the Gaspe peninsula. Here, three classes of Iris exist: *Setosa*, *Versicolor*, and *Virginica*. The problem is to formulate a decision process for determining the class of an Iris when the only information available is the height and width of the petal and stem (Fisher 1936). A concept learner is a computational description of such a decision process.

Typically, instances are represented by n-place feature vectors. Each feature represents some measurement, for example, colour or shape. Feature vectors have the general form $(v_1, v_2, \ldots, v_n)$ where each feature is represented by a value drawn from some set of values specific to that feature. Feature vectors have an optional class label drawn from some finite set C. The problem is then as follows. First, during the *training phase*, a series of instances are presented to the learner, with each instance labelled with a class label. After the training phase is complete, the learner is expected to have derived a competence – some hypothesis – enabling it to classify instances which it had not encountered during learning. That is, the task requires the learner to postulate an appropriate class, by the process of induction, for some set of novel unclassified cases. Machine learning is generally interested in the resulting classification accuracy: the principal concern is how well the induced hypothesis generalises to novel instances. In contrast, we will only be interested in

how the bias present in the learner impacts on the evolution of the system in the context of iterated learning.

Instance-based learning

A wide range of approaches exist for modelling the process of concept learning. I will use instance-based learning algorithms (Aha *et al.* 1991). Broadly speaking, an instance-based learning algorithm classifies novel instances by drawing an analogy with previously encountered examples which are similar to the query instance. More precisely, the class of the novel instance is predicted by retrieving the most similar previously observed instances. Instance-based learners therefore rely on the nearest-neighbour decision rule, which postulates the majority class of these neighbours. Instance-based learning is a very simple, elegant, and highly effective approach to concept learning (King *et al.* 1995).

The key operational feature of an instance-based learner is that observed instances are stored, *as is*, without any further processing, in a set termed the *concept description*. The concept description is simply a database of observed instances. Contrast this with, say, a neural network, where the concept description is reflected by a set of activation units interconnected by a web of weighted connections. Only when called to classify a novel (unclassified) instance will a instance-based learner carry out any processing. In contrast, a neural network performs a great deal of processing when updating the concept description in light of new evidence. To classify a novel query case, the standard instance-based learning algorithm simply locates the k nearest instances to the novel instance and predicts the majority class (Cover & Hart 1969).

Competence and performance

By using an instance-based learner, the competence of an agent is defined by the concept description in conjunction with the nearest neighbour decision rule. These two components specify how novel instances are classified. The competence of the learner refers to a hypothesis – a general rule – that explains the observed data. The hypothesis, on the basis of induction, will also explain more than the observed data; the explanation extends beyond the observed set of instances to also describe unseen instances by using the nearest neighbour decision rule. The performance of an agent corresponds to some set of classification decisions. The precise form of the performance (the set of classified instances), unlike the notion of competence, is independent of the mechanisms an agent employs to solve the problem. The notion

of performance of a concept learner will be identical whichever concept learning algorithm we use.

3.3.2 Iterated instance-based learning

Agents in an iterated learning model transmit their internally defined competence to other agents via external performance. Because competence is always induced from the behaviour of other agents, an agent can be thought of as a mapping from performance to performance. The role of an agent is therefore well defined. How multiple agents are organised to construct an agent-based iterated learning model, on the other hand, is open to a significant degree of variation. I will focus on one possibility: a single agent generational model. This generational model is identical to that adopted by Kirby (2002a). Here, agents are arranged in a chain. The classification performance of the first agent in the chain is the input to the second agent in the chain. This classification performance of the first agent is obtained by calling the agent to classify a number of random instances. This second agent will then derive a hypothesis. Using this hypothesis, the second agent will be prompted to make classification decisions, and this performance will form the input to the third agent in the chain. This process is repeated for a number of generations. The result is that the agents will transmit their classification competence to the next generation. Internal competence is always transmitted via external performance.

A classified set of instances defines an instance space. This space will often be divided into regions such that these sub-spaces contain instances that share a class. The space therefore represents a set of concepts. Now, if a particular instance space is fed to the first generation agent in the iterated learning model, will this instance space change down the generations? That is, will the hypotheses induced by subsequent agents accurately reflect the initial set of concepts? Another possibility is that, rather like the parlour game *Chinese whispers*, the concepts will change as a result of transmission. Before focusing on these questions, a few more details are required before the experiment is fully defined.

In the following experiment, instances are drawn from a bounded two-dimensional space: instances are pairs (x, y) such that $x, y \in \Re$ and $0 < x, y \leq 100$. Despite the instance space being bounded by the above restrictions, the set of possible instances is infinitely large because the coordinates are drawn from a real-valued number space. Distance between instances is measured using Euclidean distance. The class labels are drawn from the set $\mathcal{C} = \{0, 1, 2, 3, 4\}$. The five classes contained

in $\mathcal{C}$ are nominal: class 1 is in no way more similar to class 2 that it is to class 4. Letters could just as well be used to label the classes. These details define the concept learning problem: two dimensional instances are classified using one of five class labels.

Details of how the individual agents interact with each other are now required. The first agent, A_0, learns its concept description from an instance space composed of 100 instances. These instances are drawn at random from the space, and each instance is assigned a random class drawn from $\mathcal{C}$. A minor observation worth noting is that, because the instance space is constructed randomly, it will not necessarily represent all the class labels found in $\mathcal{C}$. By chance, some classes may not be represented at all.

Starting with A_0, each generation leads to the current agent being called to perform classification decisions on the basis of its competence. These classification decisions require the classification of novel instances drawn from the instance space. Importantly, these query instances do not contain a class label. The result of a single classification decision will be class label. Let the set $\mathcal{C}'$ be the set of classes present in the concept description at some point in the simulation. $\mathcal{C}'$ is dynamic. It can change over the course of the experiment. Importantly, the cardinality of this set can only ever decrease. Classes are never invented.

Now, after a learner is presented with the 100 instances required for training, and it has induced a hypothesis, it is called into action by being queried with 100 unclassified instances. These query instances are classified using 1-nearest neighbour classification. This decision rule, in conjunction with the observed training instances, defines a space composed of *decision surfaces*. These surfaces sit between clusters of instances that share a class, and define how query instances are classified. The result of the queries will therefore be 100 classified instances, and these instances represent the performance of the agent. This performance behaviour is then observed by the learner in the next generation. Now the cycle of arriving at a competence in light of observed performance, and then transmitting this competence by producing further performance has been defined. Figure 3.2 depicts this process.

We now have an example of an iterated learning model where classification competence is transmitted between agents by examples of performance that are derived from the competence. The model is generational. The degree to which the competence is transmitted successfully down the generations is the next question.

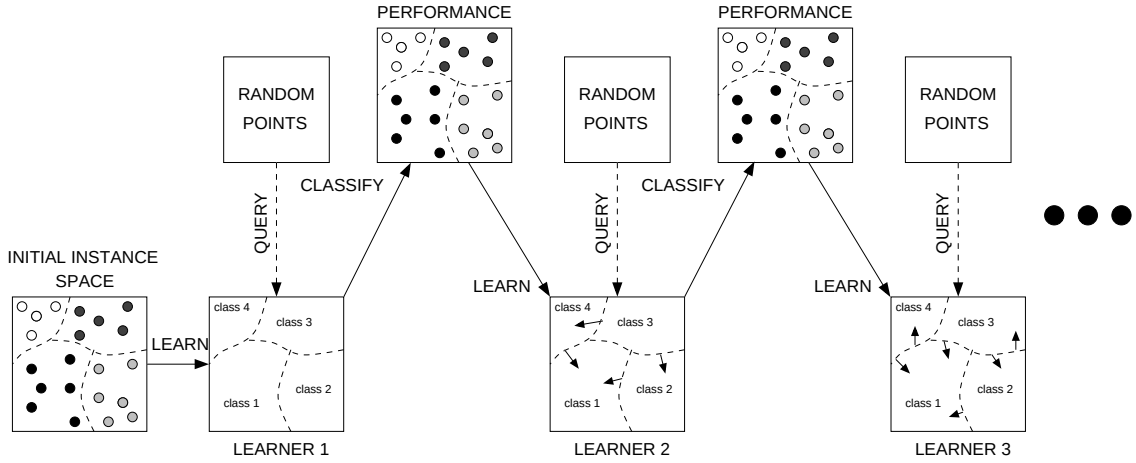


Figure 3.2: Iterated instance-based learning. An agent receives a set of query instance to be classified. The result of these classification decisions – newly classified instances – constitutes the performance of the agent. This performance is then observed by the agent in the next generation.

3.3.3 An illustration of state change

Because instances are points in a two dimensional space, and each instance has an associated class drawn from small finite set, visualising the changing state of the iterated learning model is straightforward. Figure 3.3 depicts a typical trajectory of change in the agents performance over 60 iterations. The concept description of the first classifier is composed of 100 randomly labelled instances. This state of affairs does not last long. By the third iteration, a number of distinct regions emerge that represent four classes. Some of the classes are divided into spatially distinct fragments. This initial phase represents a transition from a state of maximum entropy, where classes are equi-probable in all parts of the instance space, to one where homogeneous regions of instances that share a class emerge. Maximum entropy is a situation at odds with the assumption underpinning the nearest neighbour decision rule, where an arbitrary instance is assumed to have the class of its neighbours. For this reason, class regions become established within 3-4 generations as the bias in the concept learning agents begin to introduce non-random effects in the data.

As the simulation progresses, the homogeneous regions in the instance space appear to float aimlessly, perhaps expanding one generation and then contracting in the next. Over time, a pattern emerges, as the larger regions defining a class fragment become larger, and the smaller regions become smaller. This is what one would expect. For a region of the instance space to be represented in the performance of the agent, some of the random query points used to derive the performance data

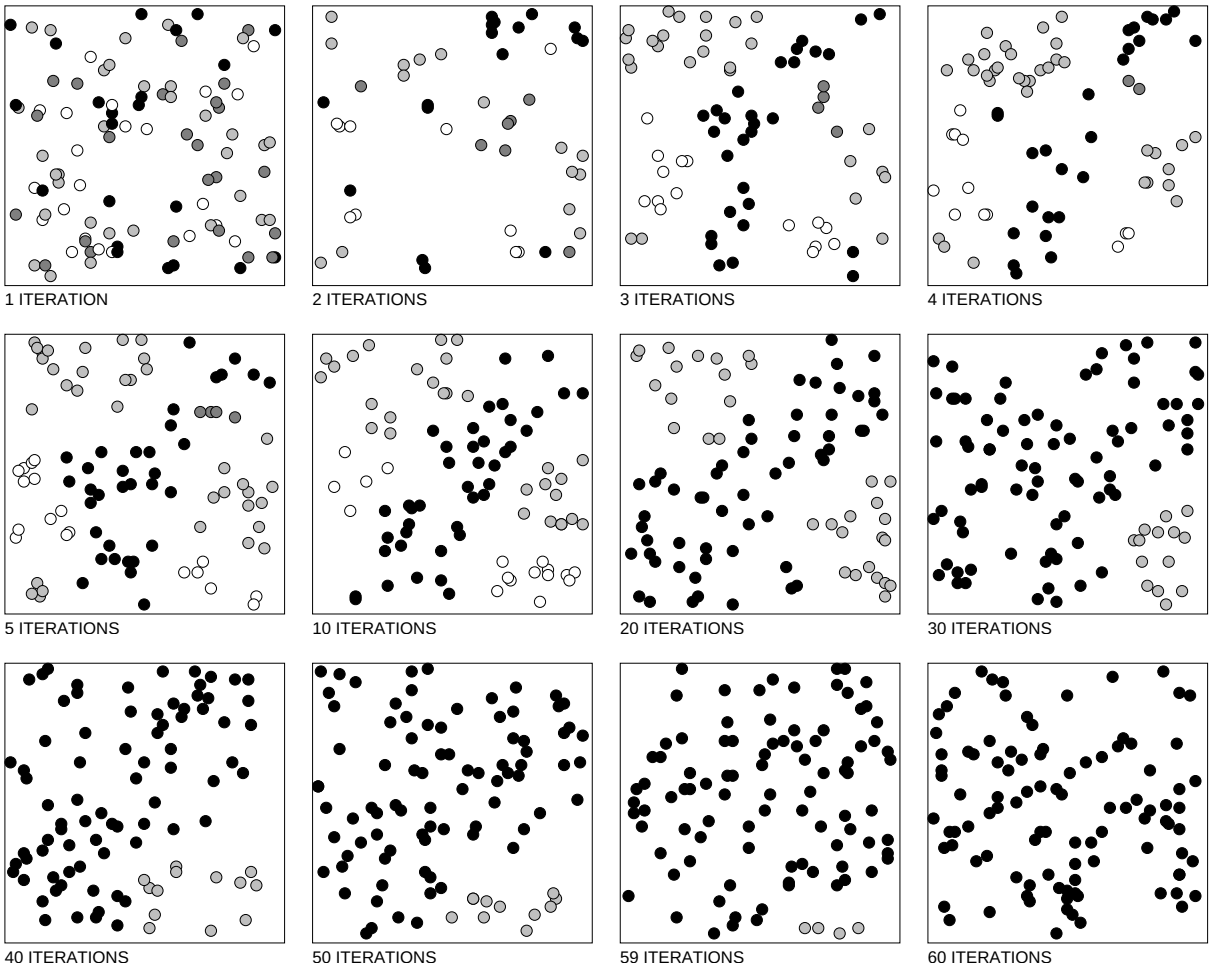


Figure 3.3: The class belonging to each point in the instance space if shown using various shades of grey. From one iteration to the next, the instance-space changes to reflect two biases. First, a bias for homogeneous class definitions, and second, a one-to-many bias from points in the instance space to one of a finite set of classes.

will have to occupy that region. The probability of a region being sampled at all, and therefore surviving by being represented to the next generation, will depend on the size of its area. This observation must follow, as sample points are constructed randomly, and therefore have a higher probability of lying in the larger regions of instance-space. As Figure 3.3 illustrates, after 20 iterations there are only two classes remaining. The larger one (black) continues to grow, and the smaller one (dark grey) shrinks and disappears. The other classes have disappeared because, for each disappearance, the class had become so small that a random sample of the instance space failed to represent it. The only steady state achievable in this model is therefore total dominance by one class, which we see occurring at the 60th generation.

3.3.4 The determinants of state change

If the whole instance space were to be sampled at each iteration, rather than just 100 random points, then whichever instance space we used to train the first concept learner would be transmitted perfectly at each iteration. Because only a subset of the instance space is sampled, certain parts of the instance space have to be approximated using the inductive properties of the learner. This is an example of a *transmission bottleneck*. Bottlenecks will be a recurring theme throughout this thesis. Bottlenecks represent limitations on transmission. They are motivated by the situation faced by language users. Our knowledge of language relates to an infinite number of utterances, yet we learn language from finite data. The number of utterances we produce over our lifetime is also finite. When considering the transmission of knowledge of language via language performance, we must therefore face the fact that transmission can never spell out knowledge of language explicitly. Performance can only ever be an impoverished subset of the set of possible instances derivable from an agent's competence. In the model, the transmission bottleneck captures this situation. Without a transmission bottleneck, performance in this model would be a point-to-point model of competence. However, because the instance space is infinitely large, it is impossible to avoid the transmission bottleneck; performance can only consist of a finite number of instances. In contrast, the competence of the learner extends to an infinite number of instances; it can classify all instances in the instance space.

The presence of a bottleneck brings the bias of the learner into play. Because certain parts of the instance space will not be explicit represented, induction is required to yield an appropriate class. This is why, as the simulation proceeds, we see the accumulated residue of past classification decisions. The iterated learning model can be thought of as revealing the bias of the agents. This particular experiment illustrates that instance-based learners have a prior expectation that instances share the class of their neighbours. This is why, as the simulation proceeds, we do not see the introduction of non-homogeneous regions of class definitions. This observation echoes findings that show instance-based learners perform badly on non-homogeneous class definitions (Brighton & Mellish 2002).

The purpose of this discussion is to understand why, by looking at a particular example, the translation between performance and competence, and then back to performance, results in state change. This change is driven by a combination of a transmission bottleneck and the bias present in the learners.

3.3.5 Explaining trajectories through state space

Understanding the behaviour of an iterated learning model will often require understanding, independent of the initial state of the model, why the system enters certain states and not others. Such an explanation can act as explanation for the origin of certain states. Ultimately, as the models become more plausible models of language evolution, these states will correspond to particular languages. Iterated learning models can therefore shed light on linguistic evolution.

Iterated learning models are complex adaptive systems. Two arbitrarily close sets of initial conditions are highly likely to result in different end states. The important point is that these end-states, although entirely different, will often occupy similar regions of the state space. So our analysis concerns an evolving ILM, or more more precisely, a series of state transitions. Rather than considering each individual state in turn our analysis will concern the trajectory, defined by a series of states, through regions of the state space. This practice is necessary as often an experiment will comprise thousands of generations. Any insightful explanation of the trajectory of state change will therefore revolve around an understanding of the topology of the state space.

To accurately discuss state transitions, I will employ a vocabulary borrowed from complex systems theory. Imagine two different systems that change over time. The first system is random; from one state to the next the probability of the system entering any one of the states is equi-probable. The state trajectory will therefore skitter around, and never remain in any one state. The second system is not random but ordered. Certain states are attractors – once the system enters such a state, it will remain there. For the first system, we cannot explain the state space trajectory in any principled way. For the second system we can: a rich descriptive vocabulary is available.

Attractors need not be single states, they may be a local set of states to which the system converges and then remains. The volume of states that reliably lead to a trajectory ending in an attractor is termed a *basin of attraction* (Kauffman 1993:176). A state space is best imagined as a landscape composed of features like basins, ridges, and peaks. State change is characterised by a movement downhill. So, as a result of this landscape, states will occupy points from which only certain trajectories are likely. We will frequently be interested in attractors. In the context of an ILM, these will be often be steady states. It is therefore worth marking out some clear terminology.

Within a state space *Liapounov stability* refers to the property of some point in the state-space: if we start near that point, then we stay near it.

Definition 4 (Liapounov stability) *A point x is Liapounov stable if there exists some $\epsilon > 0$ such that all points within the neighbourhood of x bounded by ϵ remain within a distance of ϵ from x for all $t \geq 0$.*

Therefore, remaining in the neighbourhood of x, which is bounded by ϵ, is all that is required for states within the neighbourhood to be Liapounov stable. This definition says nothing about whether the system will reach x, however. This definition is illustrated in Figure 3.4(a), where the hypersphere in state space centred on x and bounded by radius ϵ, $B(x, \epsilon)$, contains all trajectories starting within the hypersphere $B(x, \epsilon)$. Liapounov stability concerns bounding the behaviour of states surrounding a point in state space.

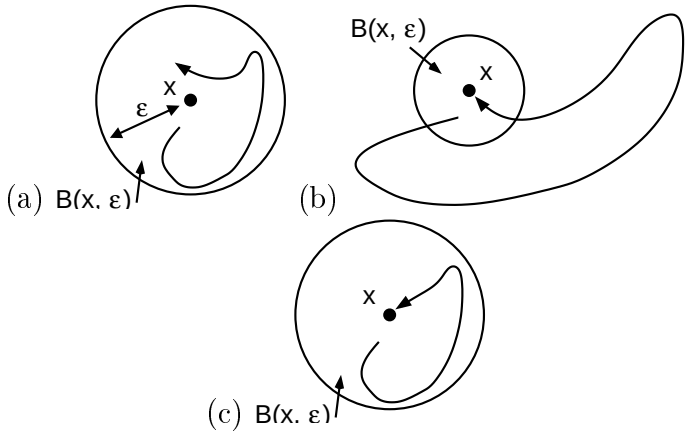


Figure 3.4: Three kinds of stability. (a) Liapounov stability, or "start near, stay near" stability. (b) quasi-asymptotic stability, where the point x is stable as $t \to \infty$, but has undefined behaviour otherwise. (c) Asymptotic stability, which is defined as both Liapounov stable and quasi-asymptotically stable.

In contrast, a quasi-asymptotically stable point is one for which points within the neighbourhood bounded by ϵ tend to x eventually as $t \to \infty$.

Definition 5 (Quasi-asymptotic stability) *A point x is quasi-asymptotically stable if for some ϵ, all points in the neighbourhood of x bounded by ϵ tend to x as $t \to \infty$.*

That is, a quasi-asymptotically stable point concerns the behaviour of the system in the limit, as $t \to \infty$. States starting within the state space surrounding x, bounded by ϵ, end up at x eventually. Importantly, this definition says nothing

about the intermediate states: they could be anywhere. Figure 3.4(b) illustrates this behaviour. A further definition naturally proceeds as follows:

Definition 6 (Asymptotic stability) *A point x is asymptotically stable if it is both Liapounov stable and quasi-asymptotically stable.*

This Definition, in contrast to that of Definition 5, makes a claim about the locality of the states visited, starting from the initial state, before it homes in on x: They must lie within the neighbourhood bounded by ϵ. Figure 3.4(c) illustrates this behaviour. Relating the definition of asymptotic stability to attractors, we can say that the largest region surrounding x for which points are asymptotically stable is precisely the basin of attraction mentioned above. Furthermore, if this region extends to the whole state space, then the point x is termed *globally* asymptotically stable (Glendinning 1994:30).

Attractors in instance-space

This model of iterated instance-based learning leads to a relatively simple state space topology. In general, if $\mathcal{C}$ is the set of possible classes, then there are $|\mathcal{C}|$ point attractors, or stable states. Each of these attractors represents dominance of a single class. Furthermore, these attractors are global asymptotic stable points. This statement can be justified by noting two properties. Firstly, imagine dividing up the space of all concept descriptions into regions, where each region contains concept descriptions with, say, n classes. Such a region will also contain concept descriptions with less than n classes. Now, these regions are Liapounov stable. Because classes can never be introduced, future states *must* remain within the region from which they started. Secondly, for a similar reason, the state space trajectory will tend to one of the point attractors over time. The point attractors are regions containing just one class. These observations justify the claim that $\mathcal{C}$ globally asymptotically stable points exist in this model.

This kind of behaviour is not an inevitable outcome of the iterated learning model. Firstly, because the classifier cannot invent new classes, once a class is no longer represented by any instances, the class cannot reappear. A pressure therefore exists for the number of classes to decrease. Secondly, due to the nature of the classification problem, there is a many-to-one bias in the mapping between instances and classes. In other words, when viewing a concept learner as a function, many instances in the domain are represented by a single class in the range. These two observations

explain why a single-class instance-space is the only steady state. First, the many-to-one bias is a bias away from diversity in the mapping from instances to classes. The classification bias leads to a pressure for fewer and fewer classes.

3.3.6 Steps towards a linguistic model

This experiment tells us little about language. Language is not a classification problem. Any attempt to draw a parallel between this model and a plausible notion of language must break down, as the mapping from instances to classes bears no relation to any linguistic property. This model does, however, demonstrate two fundamental properties of iterated learning models of linguistic evolution. Firstly, the transmission of competence via the expression of performance leads to state change across generations of agents. Secondly, the interaction between the transmission bottleneck and the learning bias of the agents determine the kind of state changes that occur. Whatever the initial state of the above model, the same kind of behaviour results. This experiment has also served to introduce iterated learning models as complex adaptive systems. I have introduced some initial concepts in explaining such systems, and these concepts will crop up in the explanations of future models.

3.4 Language as a structured mapping

We can imagine language as an infinitely large mapping from meanings to signals. Given a meaning to express, this mapping defines which signal is appropriate. Similarly, given a signal, the mapping defines which meaning the signal refers to. In this section, I will extend the model of iterated instance-based learning such that the competence of each learner represents a mapping between an infinitely large meaning space and an infinitely large signal space.

Human language, when viewed as a mapping from meanings to signals, has a property that is extremely rare in the world of natural communication systems. Human languages are *compositional*: the mapping from meanings to signals is structured in such a way that the meaning of a signal is a function of the meaning of its constituent parts (Krifka 2001). For example, the words in a sentence carry meaning themselves, and the meaning of the whole sentence is a derived from these meanings. In contrast, the communication systems used by, for example, vervet monkeys is not compositional: their signals refer to whole meanings and are regarded as holistic (Cheney & Seyfarth 1990). The mapping from meanings to signals used by vervet

monkeys is not structured in the same way that human languages are. Perhaps the only example of a non-human compositional communication system is the dance of the honey bee, where food location is communicated through a structured dance (von Frisch 1974). Importantly, the food location itself is a structured entity, composed of direction and distance. This communication system should be contrasted non-compositional communication system of bird song, where structured signals are used to refer to atomic, unstructured meanings (Hauser 1996).

It is the property of compositionality that I will explore in the following experiment. In order to do so, I will use an abstract model of language that can accommodate the properties of compositional language and holistic communication systems. In fact, this model of language is only a slight modification from the notion of an instance space discussed in the previous section. Instead of a mapping from a two-dimensional real valued number space to a set of classes, I will use a mapping from one two-dimensional number space, $\mathcal{M}$, to another two-dimensional number space, $\mathcal{S}$. These two spaces will correspond to the meaning space and signal space used by natural languages. Rather than investigating plausible models detailing the structure of these spaces, I will instead focus on the nature of mapping between the two spaces. Ultimately, this section will aim to explain how compositional mappings can emerge in iterated learning models.

A mapping that defines a compositional language will have neighbourhood preserving properties. This means that similar meanings will map to similar signals. This fact follows from the observation that, firstly, in order for meanings to be similar, they must share components. Secondly, because these meaning components are common to both meanings, parts of their corresponding signals must also share components. The signals will occupy neighbouring parts of the signal space, just as the meanings do. Compositional mappings can therefore be considered examples of *topographic mappings* (Van Hulle 2000). Topographic mappings possess the property of neighbourhood preservation; neighbouring elements in the domain map to neighbouring points in the range. This kind of mapping is ubiquitous in neurological systems. For example, many organisms possess a vision system that relies on *retinotopic* mapping: the retinal image is mapped to the visual cortex such that neighbourhood relations hold true (Van Essen *et al.* 1981).

Figure 3.5 depicts three mappings. Figure 3.5(a) is an example of holistic mapping, where points in meaning space map randomly to points in the signal space. In holistic mappings, neighbouring points in the meaning space provide no information about where a meaning maps to in the signal space. Figures 3.5(b-c) depict

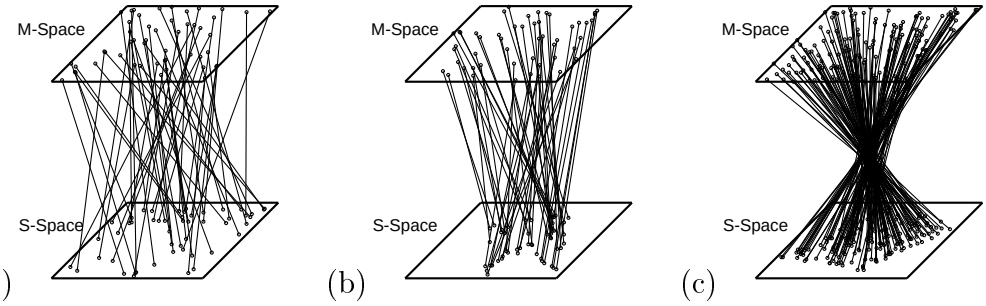


Figure 3.5: Example mappings. The mapping shown in (a) is random. The neighbours surrounding a point in the meaning space suggest nothing about the location of the corresponding signal. In (b-c), similar meanings map to similar signals.

compositional mappings, where a clear relation exists between neighbourhoods in the meaning space and neighbourhoods in the signal space. In the following experiments, meanings and signals are drawn from two-dimensional real valued number spaces, denoted by $\mathcal{M}$ and $\mathcal{S}$, respectively. More precisely, some meaning $m \in \mathcal{M}$ is the pair (x, y) where $x, y \in \Re$ such that both $0 < x \leq 100$ and $0 < y \leq 100$.

3.4.1 Linguistic competence and performance

The experiment discussed in Section 3.3 revolved around the classification of points. In this experiment, the competence of an agent relates to its ability to propose appropriate signals for given meanings. Just as before, an agent learns from input representing another agent's performance. Much of the details of the previous experiment carry over into the present discussion, but we now need to interpret them slightly differently. First, an agent's performance is obtained by prompting it to express meanings. The result is a set of meaning/signal pairs – instances of the mapping between meanings and signals. The agent in the next generation takes these instances as input, and forms its own concept description which is also a mapping between meanings and signals.

Figure 3.6 illustrates this process in more detail. The first generation agent A_0 is presented with examples of some initial mapping. For the experiments outlined here, this initial mapping is random. On the basis of this evidence, A_0 constructs a hypothesis H_0 which accounts for the performance it has observed. This hypothesis not only defines the regions of the infinitely large mapping it has observed, but also, with the aid of an associative memory, a competence that relates all meanings to an appropriate signal. There is no distinction between the notion of a hypothesis and that of competence; I will use the terms interchangeably. Now, A_0 is prompted

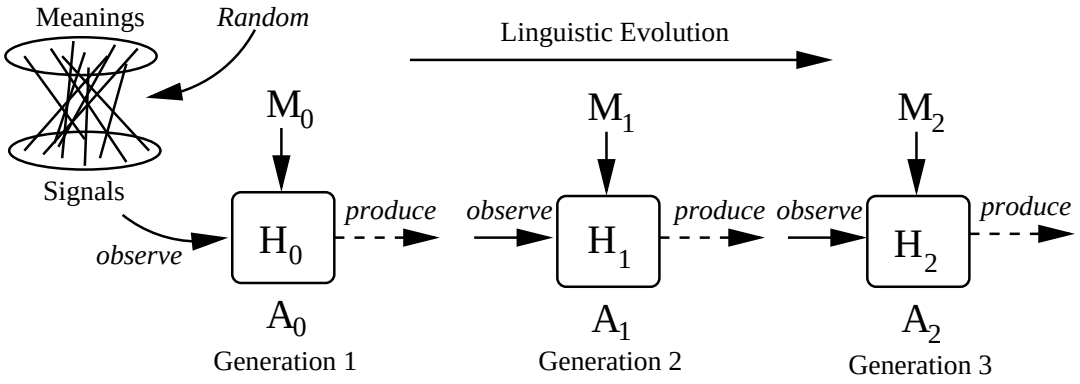


Figure 3.6: Iterated learning using a mapping between meanings and signals. From an initially random mapping, each agent is prompted with a random subset of the meaning space. The production decisions for these meanings represent linguistic performance, and form the input to the agent the next generation.

with some random subset of the meaning space, $M_0 \subset \mathcal{M}$. Appropriate signals for each element in M_0 are found, and these meaning/signal pairs constitute the performance of A_0, which will in turn form the input to the next generation agent A_1. As before, this process is repeated generation after generation.

Simplifying assumptions

Now that explanations of linguistic issues such as compositionality are being proposed, and the model is interpreted as a basic model of linguistic transmission, two important assumptions should be made explicit.

1. Agents in the model have the ability to "mind read". When an agent observes a signal, the intended meaning of that signal is also provided. This simplification avoids the problem of modelling the ascription of a meaning to a signal. An agent must associate a signal with a meaning somehow, but I regard this as a separate, non-trivial problem (Steels 1997; Steels 1998; Smith 2001; Smith 2003a). In short, I will assume that meaning transmission occurs over a noiseless channel.
2. Issues of communication are not considered. For example, communicative accuracy, the agents intentions, or any model of success or failure in language use is not considered. How much of the structure of language is due to pressures on learning? To begin to answer this question, we must strip the model of any assumptions about the functional aspects of communication. This assumption is in line with the function independence hypothesis I introduced in Chapter

2. In short, part of the argument I am developing assumes that issues relating to communication are not the principal determinants of language structure.
3. Population effects are not considered. Each agent learns from the output of only one other agent. Similarly, utterances produced by an agent are only observed by a single learner.

From a modelling perspective, these simplifications are crucial. Ultimately, we should seek the minimal set of assumptions and hypotheses with which linguistic structure can be explained.

3.4.2 An instance-based associative memory

Due to the simplicity of the instance-based learning model, learning the mapping simply requires the learner to memorise each meanings/signal pair it has observed. Moving beyond the learning process, these learners will differ substantially to those of Section 3.3 in how they produce performance. Instead of carrying out a classification task, the learner must instead postulate a signal for any given meaning such that this signal is in some sense *appropriate*. Unlike classification, this process can introduce new signals, as the range is no longer a finite set of classes, but an infinite set of signals represented as points in a two-dimensional space. In order to place this modified form of instance-based learning into the iterated learning framework, we need a model of production. Production performs a comparable function to that of classification in the previous model.

Since both the meaning and the signal spaces are infinitely large, when an agent is required to produce an appropriate signal for a given meaning m, it is highly unlikely that m will have been observed in the performance the agent is learning from. The process of production will therefore often require the agent to somehow deduce an appropriate signal for m. A solution to this problem is suggested by the nearest neighbour decision rule. Instance-based learners classify unseen instances by exploiting the solutions offered by nearby instances. Translating this approach into the problem of production, an instance-based approach proceeds by finding nearby meanings for which the appropriate signal is known. Using these examples drawn from the existing mapping, an appropriate signal should depend on the manner in which the neighbouring meanings map to the signal space.

The most obvious production scheme is the following. Given the meaning m, the three nearest neighbours of m are found, along with their corresponding signals. Now, using these three observed examples of the mapping, given by $\langle m_1, s_1 \rangle$,

$\langle m_2, s_2 \rangle$, and $\langle m_3, s_3 \rangle$, we can say that the relationship between m and its neighbours in meanings space m_1, m_2, and m_3 should somehow be reflected in the relationship between s, the signal to be produced for m, and the signals s_1, s_2, and s_3. That is, any local relationship between meanings and signals should guide our choice of s. This relationship is illustrated in Figure 3.7. The problem of production, characterised in this way, is in line with the motivation behind instance-based learning[1]. To exploit the relationship suggested by this approach, we can first represent m as linear combination of two vectors, denoted as p and q, derived from m_1, m_2, and m_3:

$$p = m_2 - m_1 \tag{3.1}$$

$$q = m_3 - m_1 \tag{3.2}$$

Representing m as linear combination of p and q we use two scalar constants A and B:

$$m = A \cdot p + B \cdot q \tag{3.3}$$

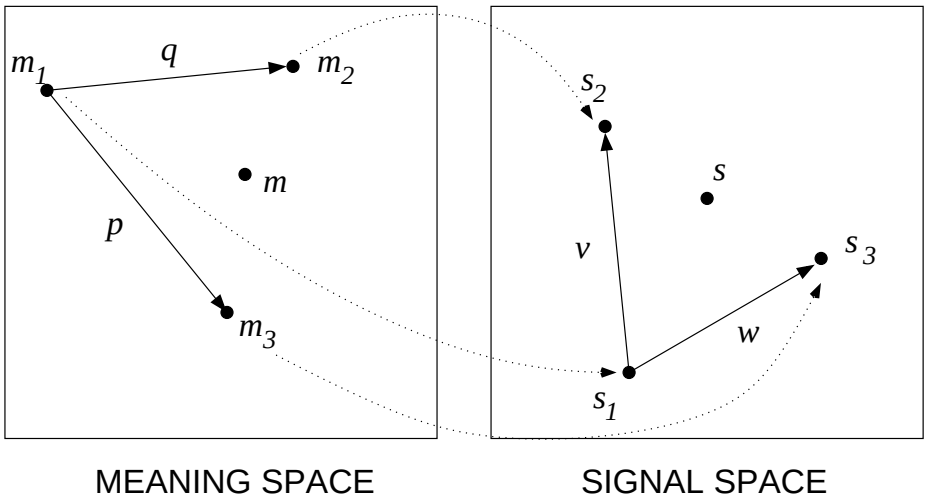


Figure 3.7: To find an appropriate signal s for the meaning m, the relationship between the three nearest neighbours of m is used to postulate the location of s in the signal space.

[1]This approach is also closely related to the problem of case-based reasoning in which similar stored solutions to a problem are adapted for solving the new problem (Kolodner 1993).

Now, A and B represent the multiplicative factors of the vectors p and q that give m when added together. Rewriting equation 3.3 we have:

$$\begin{pmatrix} m_x \\ m_y \end{pmatrix} = A \cdot \begin{pmatrix} p_x \\ p_y \end{pmatrix} + B \cdot \begin{pmatrix} q_x \\ q_y \end{pmatrix} \tag{3.4}$$

Rearranging, we get:

$$A = \frac{q_y m_x - q_x m_y}{p_x q_y - p_y q_x} \tag{3.5}$$

$$B = \frac{p_y m_x - p_x m_y}{q_x p_y - q_y p_x} \tag{3.6}$$

Next, we perform a similar operation in the signal space. Using s_1, s_2, and s_3 we construct another two vectors v and w which mirror the vectors p and q used in the meaning space.

$$v = s_2 - s_1 \tag{3.7}$$

$$w = s_3 - s_1 \tag{3.8}$$

The process of induction used to postulate the signal s proceeds by using A and B in the signal space. Because A and B capture the relationship between m and its neighbours, by the process of induction, we assume the same relationship holds between s and the corresponding signals of m's neighbours:

$$s = A \cdot v + B \cdot w \tag{3.9}$$

Any structured relationship between m and its neighbours will now be reflected in the relationship between s and its neighbours. This procedure for finding an appropriate signal given a meaning represents a bias toward structure preservation. Any existing relationship between the meaning space and the signal space is used to inform the choice of new signals. With the production mechanism in place,

the cycle of observation, induction and production is fully defined. The iterated learning model can now be analysed.

3.4.3 Analysing the evolving map

As the model cycles through the processes of production and induction, structural changes in the mapping are the primary concern. More precisely, I will attempt to shed light on the possibility of compositionality emerging in the mapping. If a compositional mapping consistently emerges, independent of the initial language structure, then the adaptive properties of the learning algorithm must form part of an explanation for compositional structure in the model. Agents in the model perform two processes: learning and production. First, the ability to recall previously encountered instances of the problem is itself a bias introduced by the model of learning I have employed. The second part of the process relates to how these nearby examples will be used, which will depend on the mechanisms underlying the production process, developed above. I will term the impact of this production process the *production bias*.

Ultimately, any explanation of the resulting mapping between meanings and signals must be related to these biases. Recall that the key measurement is the degree of compositionality. Specifically, compositionality is measured as the degree of correlation between inter-meaning distances and the distances between the corresponding signals. That is, we measure to what degree the distance between two meanings correlates with the distance between the two signals these meanings map to. A highly correlated mapping will be one where, for example, nearby meanings will map to nearby signals. The degree of correlation in a mapping containing l meaning pairs is measured by taking all pairs of meanings (m_i, m_j) (with $i \neq j$) where both i and j range from 1 to l. For each of these pairs, we find the corresponding pairs of signals (s_i, s_j) where s_k is the signal for the kth meaning. We then calculate the correlation between $\Delta m = distance(m_i, m_j)$ and $\Delta s = distance(s_i, s_j)$ for all i and j such that $i \neq j$. The correlation is computed using Pearson's product moment correlation[2] and is denoted by r. A fully distance correlated mapping will have a high r, but it is not necessarily the case that $r \approx 1.0$. The reason for this will be discussed below. Random mappings will be those where $r \approx 0.0$.

[2]The r value is computed using the following expression:

$$r = \frac{\sum_{i=1}^{l}(\Delta m_i - \overline{\Delta m})(\Delta s_i - \overline{\Delta s})}{\sqrt{[\sum_{i=1}^{l}(\Delta m_i - \overline{\Delta m})^2 \sum_{i=1}^{l}(\Delta s_i - \overline{\Delta s})^2]}} \tag{3.10}$$

3.4.4 Adaptive properties of the learning and production bias

To fully understand the consequences of the learning and production bias, I will trace the trajectory of the evolving mapping from different starting positions in the state-space. I will examine the behaviour of the system from two extreme starting positions. Firstly, to gain an understanding of the global behaviour of the system I will start the simulation from a random mapping. Secondly, an understanding of the local behaviour around regions of the state-space containing structured mappings will be achieved by starting the simulation from perfectly compositional mappings. These mappings have a correlation value of $r = 1.0$, and are constructed so that some random set of points in the meaning space map to identical points in the signal space. These two initial states correspond to the issues of *constructing* structure and *maintaining* structure, respectively (Smith 2002).

Evolving structure

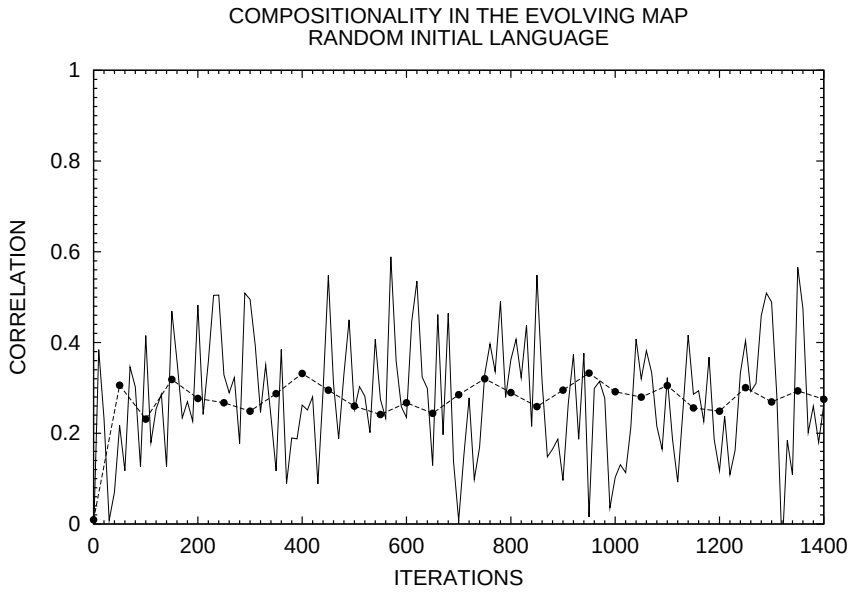
Figure 3.8(a) illustrates the extent to which the learning and production biases discussed so far result in the evolution of a compositional mapping. These results present compositionality, or distance correlation, as a function of time. Both the average behaviour over 30 independent runs and an individual run are shown in these plots.

The first observations is that, however long the system is allowed to cycle through the twin processes of production and induction, the learning and production biases alone cannot un-tangle the messy relationship between meanings and signals. The distance correlation in the mapping fluctuate about a mean value of $r = 0.3$. Notice that this mean value is higher than $r \approx 0$ reflected in random mappings: the learning and production bias guide the system away from the initial state of randomness. The initial random state of the system is best thought of as a set of multiple sub-mappings that compete for transmission. This is an important observation because the production mechanism does favour a neighbourhood preserving relationship between meanings and signals. Yet this preference does not in itself lead to the evolution of structure. Before discussing this observation further, it is worth considering whether the biases presented so far can *maintain* structure.

Maintaining structure

Figure 3.8(b) plots the distance correlation of the evolving map from an fully compositional initial mapping. This mapping is not stable, and any initial structure

(a)



(b)

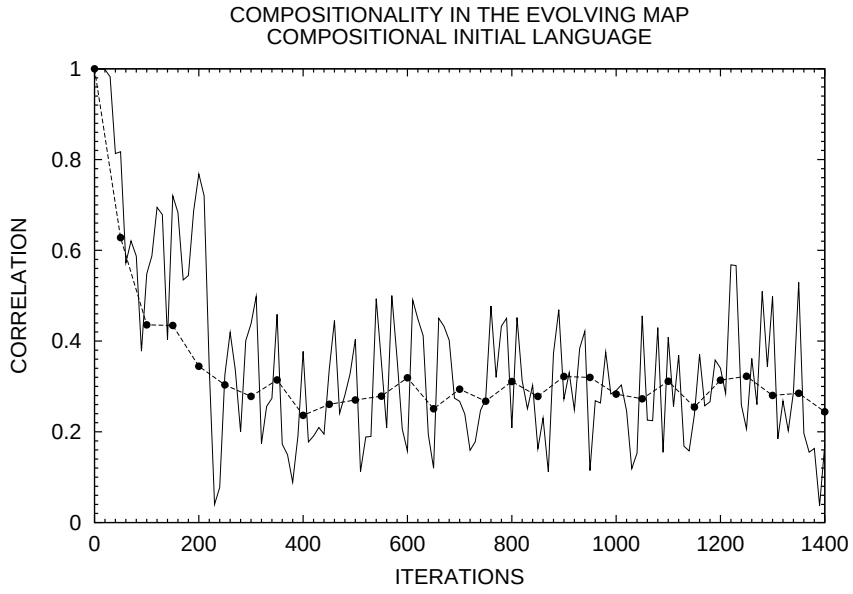


Figure 3.8: The production bias alone does not lead to highly correlated maps. The dashed line represents the average distance correlation over 20 independent runs. The full line represents an individual run. With a bottleneck of 100 instances, we see, in (a), the evolution of the system from a random initial mapping. In (b), the initial state of the system is a perfect topology-preserving mapping. Any structure degrades rapidly, and the system enters a state comparable with that shown in (a).

rapidly decays and the mapping starts to fluctuates around a mean distance correlation of $r = 0.3$, which is precisely the behaviour found when starting from an initially random state. At first sight, the degradation of structure from an initially structured mapping is a counter-intuitive result. If the production mechanism faultlessly captures its geometric motivation, then this result should not occur. An arbitrary meaning m has three nearest neighbours. In the perfectly compositional initial system, these neighbours map to exactly the same points in the signal space. Providing the relation between m and its neighbours is faultlessly parallelled in the signal space, as it should be, production for m should lead to a point in the signal space with exactly the same coordinates as m itself. This, however, does not happen.

The reason behind the disparity between the geometric theory and the model itself is that the arithmetic operations carried out by the production mechanism can introduce very slight arithmetic errors. Meanings and signals are represented by pairs of coordinates realised as floating point numbers. Rounding errors are inevitable when computing with floating point numbers. In many situations this does not pose a problem, but the production mechanism is particularly sensitive to these errors. Furthermore, when applying the production mechanism hundreds of times during each of many iterations, the errors are magnified iteration after iteration.

Apart from highlighting the problems associated with manipulating floating point numbers accurately, this observation suggests that the system is brittle in a wider sense. For example, natural systems robustly deal with noise. The instability of a perfectly compositional system suggests that if the transmission of information was subject to even slight degrees of noise, the same effect would occur: the system would pass into an unstable state. Maintaining structure under in a noise free system does not mean a great deal if the system cannot recover from perturbations. Consider the structure of the state space suggested by these results. In a noise free system, perfectly compositional mappings will be stable points. However, they will not be a Liapounov stable points, nor asymptotically stable. Instead, these stable points will be region of stability occupying just a single point, surrounded by regions of instability. In short, compositional structure in this model does not occupy a robust stable region of the state space; it is a region with no basin of attraction.

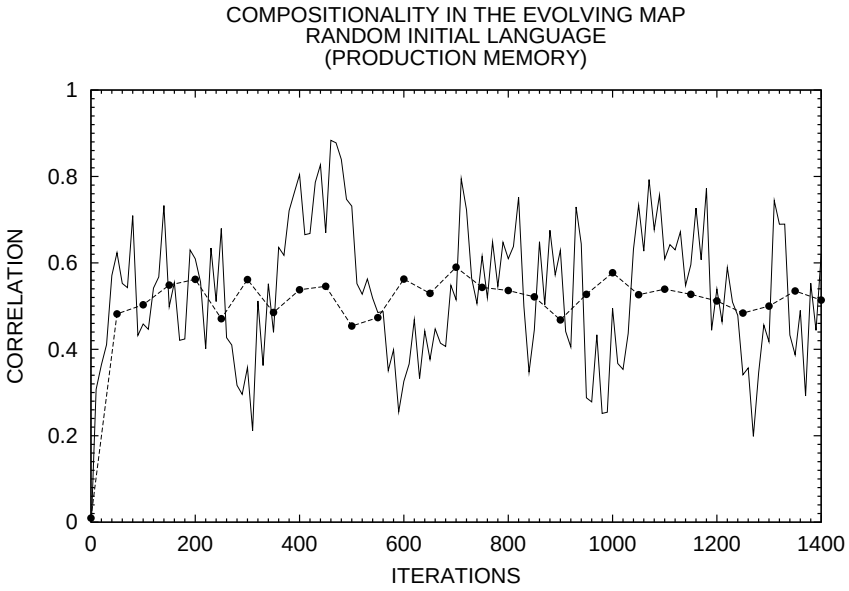
3.4.5 Production memory

If two subsequent agents agree on the mapping between meanings and signals then the system is in a stable state. Between two subsequent agents, the arithmetic errors discussed above, however small, introduce a disagreement between the two agents. To clarify this position, consider the function that defines the mapping. A perfectly compositional system is describable by a simple function. As soon as slight deviations from this mapping are introduced, then the simplest function describing the mapping increases in complexity. In short, for stability to result in the face of less than perfect transmission we must increase the probability that two subsequent learners agree on the mapping between meanings and signals. Using this observation, I will next aim to strengthen the bias toward compositionality.

One problem arising from the production bias concerns the production decisions made over an agent's lifetime. If two meanings which are local to one other are expressed, it does not necessarily follow that the two signals derived will be similar. For example, two meanings m_1 and m_2, however close they are to each other, may posses different nearest neighbours, and therefore could map to very different parts of the signal space. For example, this situation will occur when a production decision is sought for a meaning located at the boundary of two competing sub-mappings. To alleviate this problem, I will employ a *production memory*. The production memory requires that learners update their own competence in light of production decisions they make over their lifetime. The production memory will have the effect of, firstly, smoothing the differences between adjacent regions of the meaning space that map to non-adjacent regions of the signal space, and secondly, increasing consistency between learners. A production decision made by an agent, after all, is guaranteed to form part of the competence of the agent receiving the utterance. As well as making production more consistent over the lifetime of the learner, the production memory also increases the probability that two subsequent learners are in agreement over the mapping between meanings and signals.

Figure 3.9(a) illustrates the importance of this observation: it shows compositionality as a function of iterations for simulations containing agents armed with a production memory. By increasing the consistency between learners, and therefore stability, the production memory increases the degree to which structure emerges in the mapping over time. From an initially random mapping, a mean distance correlation of $r \approx 0.5$ is achieved. At times the mapping contains a significant degree of distance correlation: for approximately 50 generations the system achieves

(a)



(b)

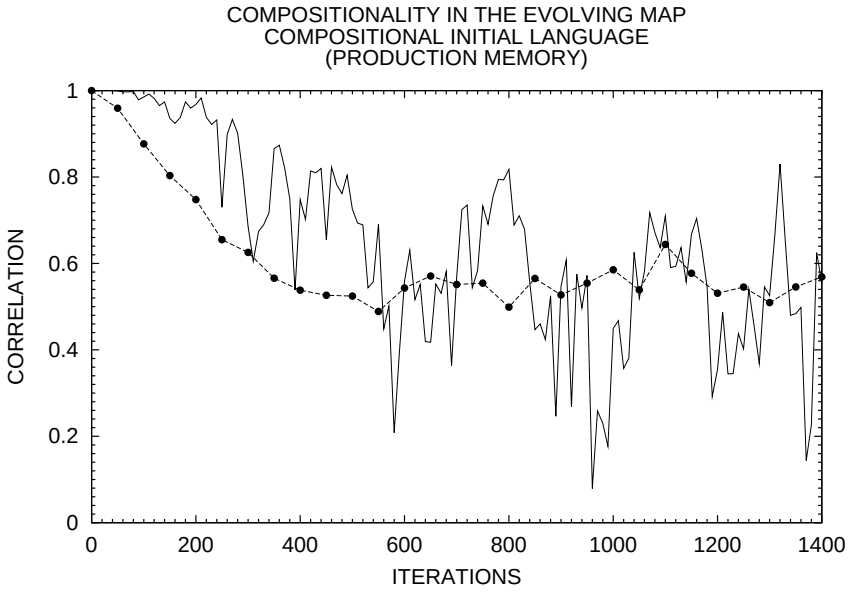


Figure 3.9: The production bias in conjunction with the production memory, where we observe an increase in the distance correlation to that obtained without a production memory. In (a), the initial state of the simulation is a random mapping. In (b), the initial compositional mapping degrades slower that observed without a production memory. The asymptotic behaviour of both systems is identical.

a distance correlation as high as $r \approx 0.8$. Starting with a perfectly compositional mapping, Figure 3.9(b) illustrates that structural decay occurs far slower than that demonstrated in Figure 3.8(b), where a production memory is not used. A production memory therefore aids in maintaining structure as well as evolving it. Nevertheless, irrespective of the initial conditions, the system is still unstable and suffers from high variation in the distance correlation.

3.4.6 Modelling the reception behaviour of others

Independent of the initial conditions, the production memory leads to an increase in the distance correlation contained in the mapping. This observations hints at the importance of deploying mechanisms to increase the degree of agreement between two subsequent learners. At this point it is useful to relate this observation to work focusing on the cultural evolution of communication systems. In contrast to the language model I have introduced, a communication system is regarded as a mapping between meaning and signals where both sets of comprise unordered atomic elements (Oliphant & Batali 1997; Oliphant 1999; Smith 2002). Such a communication system is parallelled by vervet monkeys, for example (Cheney & Seyfarth 1990), and corresponds to the notion of a vocabulary (Smith 2002). Rather than attempting to explain linguistic properties such as compositionality, computational simulations of the cultural evolution of communication systems seek to explain how coordinated mappings emerge and persist in populations of agents.

Oliphant and Batali model the process of communication with respect to a *send* function and a *receive* function (Oliphant & Batali 1997). They note that in order for maximally coordinated systems to emerge, the manner in which the mapping between meanings and signals is determined – the send function – needs to be informed by what they term the *obverter* procedure. The obverter procedure requires a learner to model its mapping from meanings to signals on the basis of the reception behaviour of the surrounding population. This reception behaviour represents the manner in which other agents map signals to meanings. For example, to find an appropriate signal for a meaning m, the obverter procedure would use a knowledge of which signal leads others to infer that the meaning m has been transmitted. Such knowledge might be gleaned from observing how others react to signals, and then somehow working out the intended meaning of these signals from this behaviour of others. Accordingly, if an agent could always read the mind of the other agents, then this procedure would be straightforward. One approach for an agent to approximate this strategy is to use its own reception behaviour as a model of the

reception behaviour of others. Here, to produce a signal for a meaning m, the agent would work out which signal, if received by itself, would maximise the probability of inferring the meaning m.

An obverter procedure therefore informs the process of production by maximising the likelihood of the proposed signal correctly being interpreted as carrying the intended meaning. Ideally, one should guide this process using a knowledge of how others perform the reverse process of translating signals to meanings.

An instance-based learning parallel of the obverter procedure

The increase in the distance correlation achieved using the production memory suggests that increasing the degree to which two adjacent agents agree on the mapping between meanings signals will lead to further stability and compositionality in the mapping. Using Oliphant and Batali's work on basic communication systems as a starting point, I will address this issue by constructing an instance-based learning equivalent of the obverter procedure.

The obverter procedure relates to how production decisions are made. Given a meaning m, we find m's 3 nearest neighbours. These meanings, m_1, m_2, and m_3, map to signals s_1, s_2, and s_3. The basic production mechanism exploits the geometric relationship between m and m_1, m_2, and m_3. The relationship is then used to postulate a signal s such that the same geometrical relationship holds between s, s_1, s_2, and s_3. Now, the obverter procedure requires that this signal s would be interpreted as referring to m, if we heard s ourselves. The obverter procedure is best thought of as a check for the legitimacy of s as an appropriate signal for m. The instance-based learning parallel of such a check is as follows.

Take the proposed signal s and find its 3-nearest neighbours in signal space. Note that these signals will not necessarily be s_1, s_2, and s_3. Denote these neighbours of s as s'_1, s'_2, and s'_3. The procedure for testing the legitimacy of s now depends on the meanings that map onto the signals s'_1, s'_2, and s'_3. If these meanings, in turn denoted by m'_1, m'_2, and m'_3 are identical to the set $\{m_1, m_2, m_3\}$ then we can confident that s would be the interpreted as meaning m. In short, the obverter procedure restricts production to the case when:

$$\{m_1, m_2, m_3\} = \{m'_1, m'_2, m'_3\} \tag{3.11}$$

Where the set $\{m_1, m_2, m_3\}$ represents the nearest neighbours of m and the set $\{m'_1, m'_2, m'_3\}$ represents the meanings that the nearest neighbours of s map to. The obverter procedure acts as a filter on production. Only those signals that are in line with the reception behaviour of the agent are used. This procedure is perhaps best understood geometrically. Figure 3.10 illustrates the procedure.

It is interesting to note that the obverter procedure I have defined above has a parallel in the wider context of nearest-neighbour pattern recognition, where several extensions to the nearest neighbour decision rule have been proposed (Dasarathy 1991; Toussaint 2002). In particular, there are two previous studies where the nearest neighbour decision rule has been extended to allow a reject option. Of these two studies, Hellman (1970) is worth noting for fact that the asymptotic behaviour of such a system reduces the error rate in classification to be at most equal to the Bayes optimal error.

The adaptive consequences of the obverter procedure

Figure 3.11(a) plots compositionality as a function of iterations for a random initial language. Both the production memory and the obverter procedure were used in these simulation runs. This result is striking. Highly compositional stable mappings emerge quickly and remain highly compositional and stable. An average distance correlation of $r \approx 0.92$ is achieved. Similarly, for simulation runs initialised with a compositional mapping, the initial mappings remain highly correlated, as depicted in Figure 3.10(b). Starting from a perfectly compositional mapping with maximum distance correlation, the initial correlation of $r = 1.0$ degrades a little to level out at the $r \approx 0.92$ level discussed above. The fact that the distance correlation falls short of $r = 1.0$ is of no significance. This slight deviance from a perfectly correlated mapping is due to the nature of the transformation from $\mathcal{M}$ to $\mathcal{S}$. For example, if the transformation is slightly stretched in one dimension, this will have the effect of reducing the distance correlation. That is, a correlation of $r = 1.0$ occurs when the mapping transforms points to exactly the same points in the meaning space. The only time this occurs is when we inject such a mapping into the model as an initial state.

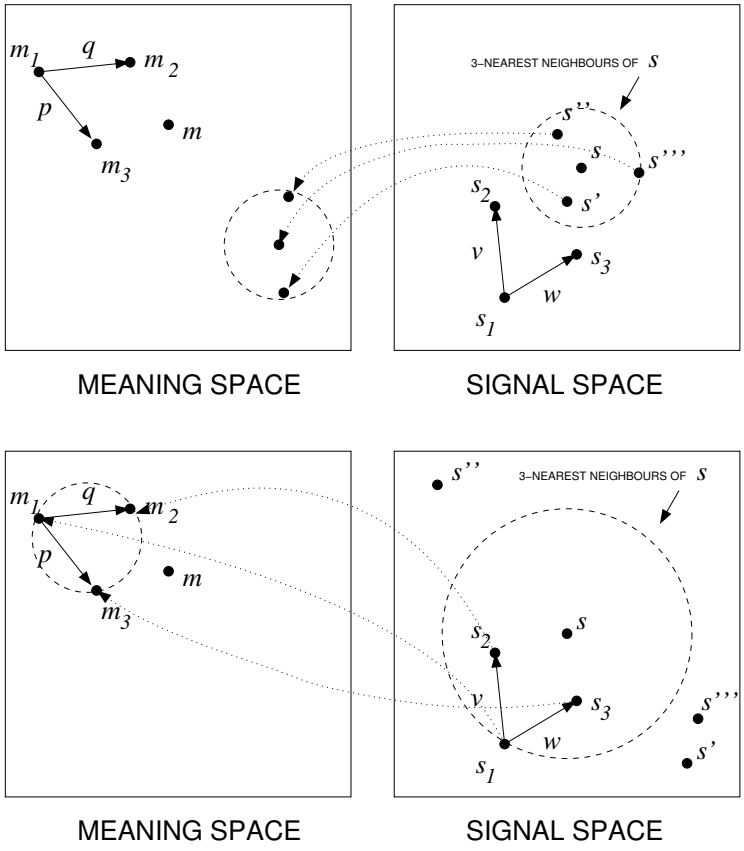


Figure 3.10: A geometrical interpretation of the obverter procedure. The topmost mapping is an example of a failed production decision because the three nearest neighbours of s do not map back to the three nearest neighbours of m. In lower mapping depicts an example of a successful production decision. Here, s's three nearest neighbours map directly back to m's three nearest neighbours.

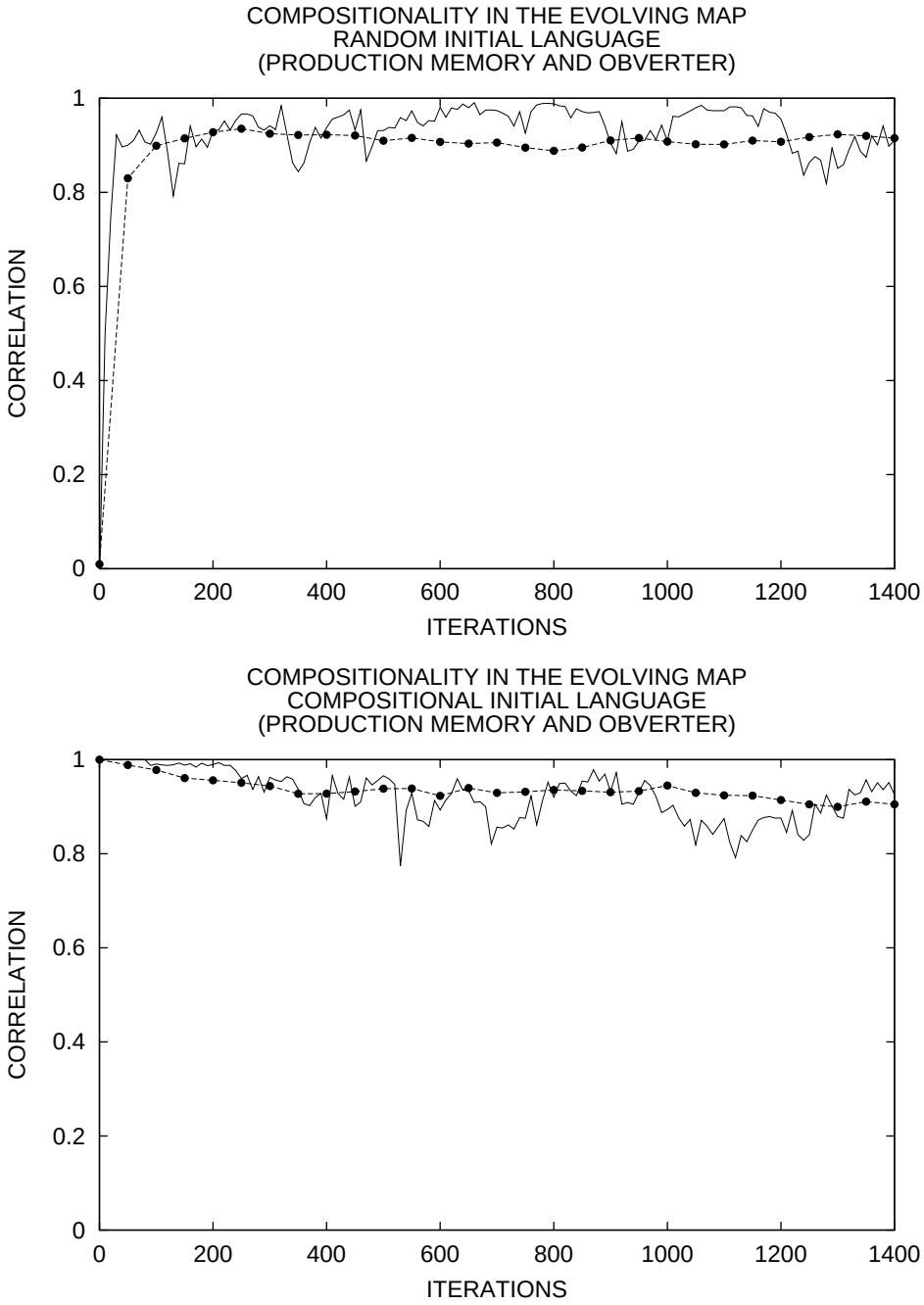


Figure 3.11: Distance correlation in the evolving map for agents armed with a production memory and the obverter procedure. In (a), from random initial mappings, a high degree of distance correlation emerges and remains. In (b), from initially compositional mappings, the same behaviour occurs. The variance in the individual simulation runs is also lower than that observed in runs without the obverter procedure.

A clearer understanding of such a transformation, as well as how the correlated map evolves over time, can be gained by studying Figure 3.13. Using a test image, we can visualise the mapping from meanings to signals by associating the test image point-to-point to the meaning space. Then, by mapping these points to the signal space, the test image is projected into the signal space. This process is clarified by Figure 3.12. The images shown in Figure 3.13 depict this projection for a test image composed of a grid and an arrow pointing from the top right corner to the bottom left corner.

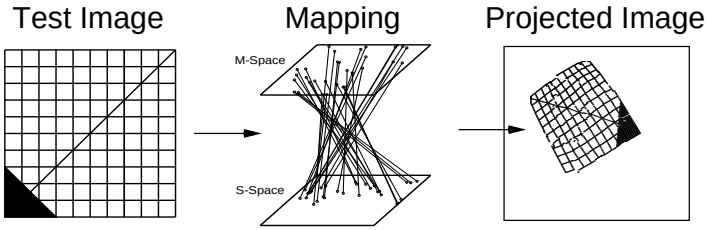


Figure 3.12: Visualising the evolving map. Points in the test image are associated with points in the meaning space. The test image is then transformed using the mapping between meanings and signals.

From an initially random state, the evolving map, depicted by the projected images, clearly illustrates that early on in the simulation multiple sub-mappings exist. After 10 iterations, the projected image contains subs-spaces of the test image projected by different transformations. Some are rotated and contracted, while others are stretched and enlarged. After 60 iterations it becomes evident that certain sub-mappings grow to such an extent that the mapping coherently projects half of the test image. The next 100 iterations lead to the merging of competing sub-mappings until they appear to merge completely after 140 iterations. By 220 iterations a near perfect representation of the test image results, albeit rotated and contracted.

3.4.7 Summary of results

Learning biases leading to the emergence of structure

The experiment I have described aims to explain how compositional structure can emerge through the interaction of learning and cultural transmission. If compositional structure in the mapping between meanings and signals is the inevitable outcome of a model, and this outcome is independent of the initial state of the mapping, then the model can inform an explanation for the origin of the compositional structure.

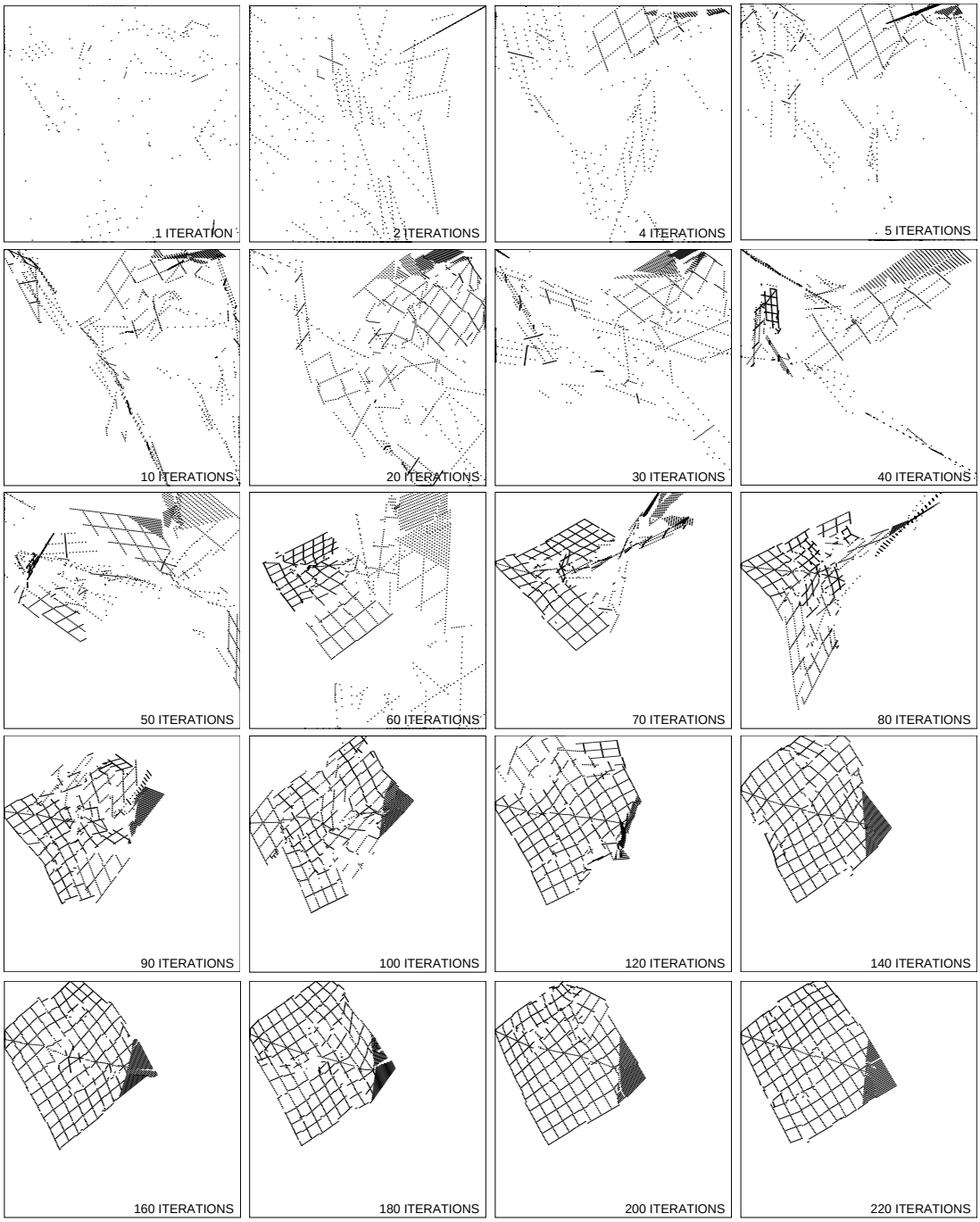


Figure 3.13: By using the mapping between meanings and signals to project the test image, we can visualise the evolving map.

By building layer upon of layer of production bias, this experiment has shown that very specific biases are required for the emergence of compositional structure. The key question relates to how the biases present in each learner have a cumulative effect on the evolving mapping. This discussion has centred around four biases, each one pulling the mapping toward certain regions of the language space. The biases themselves are best described as mechanisms, but their evolutionary impact is best explained in terms of the trajectories through the state space that they introduce. To re-cap, the mechanisms that define the biases are as follows:

Nearest-Neighbour Decision Rule. Generalisation, which enables unobserved meanings to be expressed, is achieved by recalling similar previously observed instances of production. These instances are recalled in exactly the same form as they were observed, and due to their similarity, are presumed to contain information relevant to solving the new problem.

Production Mechanism. The production mechanism takes the relevant information found by the decision rule and uses it to synthesise a new signal. The production scheme assumes that nearby production decisions should inform a novel production decision.

Production Memory. The production memory records the production behaviour of an individual over its lifetime. During an agent's lifetime, production decisions will therefore be partially informed by previous production behaviour.

Obverter Procedure. The obverter procedure introduces a reject option for the process of production. If the production process results in a signal which would be incorrectly interpreted by the agent itself, then, rather than proceeding with this production behaviour, the learner instead declines to produce.

The first two biases are a requirement for a learner to function within the iterated learning model. They define how observations are processed in order to yield the concept description, and how the concept description is interrogated to yield appropriate signals for meanings. Although these two biases together push the mapping away from maximum entropy, towards regions with a distance correlation of $r \approx 0.3$, they fail to result in the evolution of stable structured mappings, nor can they sustain such mappings when injected into the iterated learning model.

In contrast, the production memory is an additional mechanism. It increases the degree of consistency in production over the lifetime of an agent. By memorising production decisions, production for nearby meanings are more likely to be similar. This might not be the case if the nearby meanings have slightly different

nearest neighbours; their production could be guided by a different competing sub-mappings. By learning from production decisions the differences between adjacent sub-mappings is smoothed. The production memory leads to an average distance correlation of $r \approx 0.5$, for both initial states of randomness and those that are perfectly correlated.

The obverter procedure is a mechanism that excludes the possibility of homonymy in the mapping between meanings and signals. Here, the obverter procedure is translated into a bias leading to the rejection of certain production decisions. Only those production decisions consistent with the reverse mapping from signals to meanings are carried out and therefore produced. From another perspective, the obverter procedure leads to ensuring the production of signals which occupy a region of the signal space that is not projected onto by other parts of the meaning space. This is clearly a bias away from homonymy, as it decreases the probability of different meanings sharing the same signal. In relation to this bias, the model also contains a pressure to exclude synonymy. Due to, what Hurford terms, the *production bottleneck*, it can never be the case that a single meaning is mapped to two different signals (Hurford 2002:306-307). In general, the production bottleneck acts to block certain production decisions by restricting the production mechanism to only ever express one signal where in fact there could be many appropriate signals. The production mechanism used in this model, as a result of its geometrical motivation, *assumes* that a meaning can only ever map to a single signal. In mathematical terms, there is only one solution to the set of equations defining the production process.

In short, the obverter procedure represents a bias that leads to a drastic increase in the emergent distance correlation between meanings and signals. Correlations of $r \approx 0.92$ consistently emerge from both random and perfectly correlated initial states. Examining Figure 3.10 also suggests that the obverter bias leads to reduced variation in the correlation. This observation suggests that an increase in stability occurs: the emergent mapping is restricted to an area in state space consisting of similarly correlated mappings.

The importance of the obverter procedure can be related to studies of the emergence of culturally transmitted communication systems (Oliphant & Batali 1997; Oliphant 1999; Smith 2002). In this context, Smith (2003b) conducts a thorough study of the obverter procedure using associative neural networks. His conclusion is that the obverter procedure should more accurately be thought of as a bias toward a one-to-one mapping between meanings and signals. Such a bias deters the

emergence of homonyms and synonyms, and as a result, leads to coordinated and unambiguous communication systems. Exactly the same process is at work here. The bias away from synonymy is implicit in the production mechanism, and the bias away from homonymy is introduced by the obverter procedure. In ongoing work, Smith notes that a bias toward one-to-one mappings pervades all known studies on the cultural evolution of both communication systems and models of linguistic evolution (Smith 2003c). The model presented here supports this conclusion, and acts as supportive evidence for a general statement on the learner biases required for cultural emergence of coordinated mappings between meanings and signals.

Stability in the emergent mapping

We can differentiate two interpretations of stability: firstly, stability in the distance correlation of the mappings, and secondly, stability in the mapping itself. For example, two subsequent iterations in the model might exhibit exactly the same degree of distance correlation in the mapping, yet these two mappings might be entirely different; they could map meanings to totally different signals. Stability in correlation is one measurement, stability in the mapping itself is quite another. Although the possibility of this kind of disparity is unlikely, especially after close inspection of Figure 3.13, a more rigorous test for stability is worthwhile. In short, we need to be sure that stability in the distance correlation is in fact a reflection of stability in the mapping.

The definitive test for such a question must relate to showing that subsequent mappings occupy the same region of state space. The state space is the space of all functions mapping points in $\mathcal{M}$ to points in $\mathcal{S}$. Any test for locality in this space between two states is difficult to formulate. Instead, one can employ a what I will define as a *centre-point trace*. Here, the central point in the meaning space is projected onto the signal space. The position of this point is then plotted iteration after iteration. Stability in the state space will be reflected by the trace occupying a local region. Figure 3.14 depicts three pertinent traces. Figure 3.14(a) traces the position of the centre point for a simulation initialised with a random mapping, without a production memory or the obverter procedure. This trace depicts a semi-random walk: the variance in the correlation evident in Figure 3.8(a) is a reflection of instability in the mapping. Next, the same experiment carried out with a production memory results in a trace shown in Figure 3.14(b). The trace is considerably less erratic and occupies a local region. This behaviour is increasingly evident in Figure 3.14(c), which shows the trace for the same experiment augmented

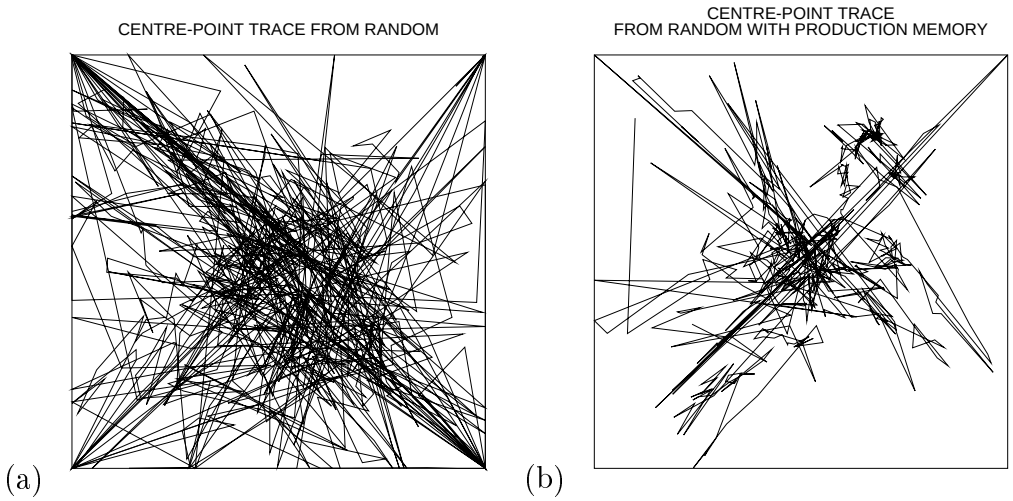
with the obverter procedure. This strongly localised trace is a reflection of the stable mapping depicted in Figure 3.13. In short, the highly correlated mappings resulting from the use of the obverter procedure are stable.

Trajectories through state space

The state space under consideration is the space of all functions mappings $\mathcal{M}$ to $\mathcal{S}$. Since both $\mathcal{M}$ and $\mathcal{S}$ are infinitely large, the state space is also infinitely large. The simulation runs discussed above show that the learning and production biases featured in the experiments determine which parts of the space are visited. Crudely speaking, we can section the space into regions that reflect the degree of distance correlation in the mapping. As we have seen, deployment of the obverter procedure consistently directs the state towards regions rarely visited, if at all, without the obverter procedure in place. Complete stability is never achieved. That is, point attractors are not a feature of these systems, in contrast to the evolving classifier discussed earlier. So the occurrence of variance in the mapping suggests that stable regions of the state space are not point attractors. They must be Liapounov stable points. The three biases mentioned previously therefore lay a probability distribution over regions of the state space marked out by degree of distance correlation. That is, the three biases explored lead to behaviour occupying three overlapping regions of the state space. These regions are arrived at independent of the initial state.

3.5 Chapter summary and discussion

I started this Chapter with the assumption that the adaptive properties of cultural transmission are a necessary component of an adequate explanation for hallmarks of language. Iterated learning is a broadly defined explanatory framework for reasoning about this assumption: I drew parallels between iterated learning and theories of cultural transmission and evolution developed by, for example, Tomasello (1999). The computational modelling of iterated learning is one route to exploring these theories. The foundational models of Batali (2002) and Kirby (2002a) stand as a starting point for a more thorough investigation of the role of learning and constraints on transmission on linguistic evolution. The qualms vocalised by Tonkes & Wiles (2002) raise real issues that future computational models must address in order to make such models more convincing vehicles for explanation.



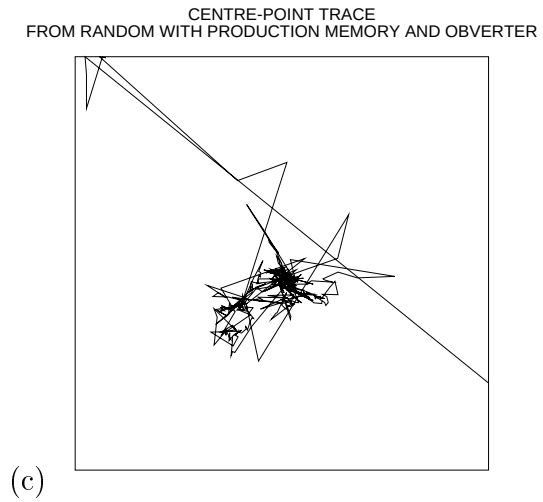


Figure 3.14: Centre-point traces for, (a), production only, (b), production and production memory, and (c), production, production memory, and the obverter procedure.

With one eye on addressing these outstanding questions, I have presented an incremental study of the emergence of compositional structure. As a starting point, the general question of transmission of internal competence via external performance has been investigated by exploring the effect of the cultural transmission of classification competence. This simulation illustrates the effect of a transmission bottleneck, and how learning biases impact on the evolved system. Quite deliberately, this model is far removed from any plausible model of linguistic evolution. Redressing this deficit, I expanded the model to account for mappings between two infinitely large structured spaces; this extension permitted an investigation into the emergence of compositionality. Although the notion of compositionality used in these experiments is a highly abstracted parallel of compositionality in language, some interesting outcomes can be identified. First, a biased learner alone is not enough to account for linguistic evolution: structured mappings do not emerge easily. Second, two categories of additional bias where required to account for the evolution of structured mappings. A production memory acts to increase the consistency of production decisions over an agent's lifetime, and ensures, to a limited degree, that subsequent agents agree of the mappings between meanings and signals. In addition, I developed an instance-based learning parallel of the obverter procedure identified by Oliphant & Batali (1997). This procedure acts as a filter on production. Only those signals which would correctly be interpreted by the agent itself are sent by the agent. The result of these additional biases is that compositional structure emerges independent of the initial conditions of the model.

These experiments have also introduced a set of explanatory mechanisms borrowed from complexity theory; state change in the models are viewed as trajectories through a state space. I have shown that structured mappings occupy stable regions of this state space under certain configurations of constraints on transmission and learning bias. These concepts will be used in remainder of the thesis.

3.5.1 Relating models to hypotheses

It is useful at this point to relate the preceeding discussion to the three hypotheses I identified in Chapter 2. The approach I have taken is in conformity with these hypotheses. Starting with the function independence hypothesis (Hypothesis 3), the experiments I have developed make no assumptions about the communicative value of language. Of course, agents in the model emit performance derived from their competence, but the purpose of this behaviour is in no way significant to the workings or interpretation of the model. One might object to this claim and argue

that the obverter procedure rests on the assumption of communicative accuracy: the potential effectiveness of the transmitted signal is taken into account. On the other hand, this procedure can also be regarded as device for maintaining a structured relation; this second interpretation is quite different to the one invoking a model of communicative function.

The innateness hypothesis (Hypothesis 1) is clearly the most problematic hypothesis to relate to these models. The agents approach the task hard-wired to deal with similarity between meanings and a capacity to relate meanings to signals. The nearest neighbour decision rule can justifiably be regarded as part of a general ability to learn. The point is this: the agents are *not* hard wired to produce compositional output. This is subtle issue, and I will defer a more thorough discussion until later.

The situatedness hypothesis, I argue, is given support by these experiments. The explanatory mechanisms required to explain why compositional structure emerges in these models comprises concepts such as the transmission bottleneck, cultural transmission, and learner biases. Only an understanding of the combination of these mechanisms can lead to an explanation of the emergence of structure. An explanation for the results of the models therefore lies outwith an explanation of the performance/competence interaction of a single agent.

Chapter 4

Towards a Model of Linguistic Evolution Based on a Simplicity Principle

4.1 Introduction

On route to developing a more plausible model of linguistic evolution, this Chapter will examine the role of simplicity in linguistic and non-linguistic cognition. Before discussing issues of simplicity, Section 4.2 proposes a model of meanings and signals. On the basis of these modelling decisions, the range of explanation of future models will be restricted. I justify this position, and focus the range of explanation to the issue of compositionality in language. Next, in Section 4.3, I will motivate an investigation into simplicity by outlining the role of simplicity in Chomsky's minimalist program, as well as the role in general issues of language acquisition. To place the discussion on a firmer footing, I will also introduce the theory of Kolmogorov complexity. Kolmogorov complexity will used as a model of simplicity throughout the thesis. Encapsulated in this theory is the minimum description length principle. This principle provides a rigorous approach to determining justifiable inductive generalisation.

Armed with a solid definition of simplicity, Section 4.4 offers a discussion of simplicity principles in human and animal cognition. I will then narrow the focus, and examine the notion of simplicity in the context of linguistic theory. Here, I will relate Chomsky's minimalist program to the theory of Kolmogorov complexity. This approach is then contrasted with theories of language acquisition based on compression and the minimum description length principle. Finally, in Section 4.5, I will lay the foundations of a model of language based on the minimum description length principle. This model will be central to the progression of the thesis.

4.2 Language as a mapping

In the last Chapter, I presented language as an abstract mapping between meanings and signals, and paid little attention to the precise nature of the domain and range of this mapping. In this Chapter I will elaborate on the language model by introducing an additional level of abstraction. The intention is to bring the model of language further in line with the notion that language is, in Chomsky's terms, "a particular relationship between sounds and meaning" (Chomsky 1972:17). For Chomsky, the faculty of language is a self-contained cognitive system interfacing with other cognitive systems external to language: one interface bridges a linguistic representation, termed Phonetic Form (PF), with "sensory-motor" systems responsible for producing and perceiving external signals; the other interface bridges another linguistic representation, termed Logical Form (LF), with those parts of the cognitive system related to "conceptual-intentional" aspects of cognition, ie., "meaning" and "thought" (Uriagereka 1998; Hauser *et al.* 2002). Figure 4.1 illustrates these relationships[1].

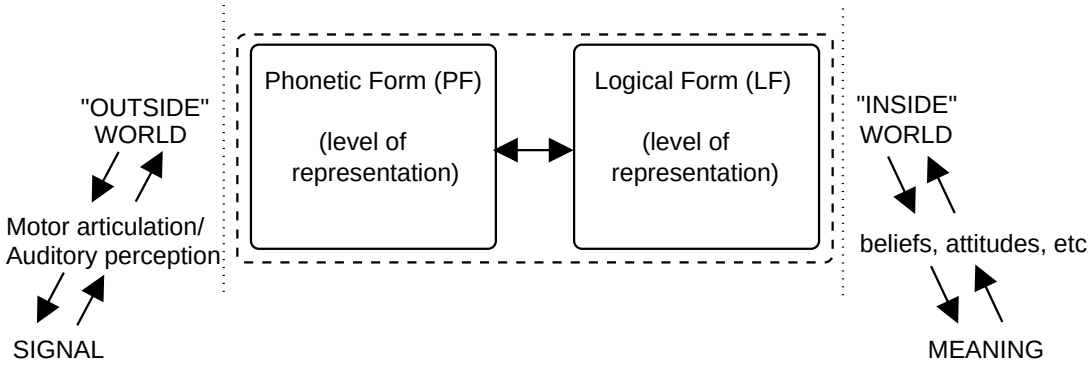


Figure 4.1: The mapping between meanings and signals in detail. PF and LF are linguistic representations, internal to the abstract computational system defining knowledge of language. They both interface with external cognitive systems. PF interfaces with cognitive systems concerning motor articulation and auditory perception, and ultimately, the production of external signals. LF interfaces with cognitive systems concerning, broadly speaking, "thought" and "meaning".

The significance of any particular linguistic representation detailed by theories of PF and LF will not be considered. When I refer to a model of language, I will be referring to the mapping between meanings and signals in an abstract sense; the model is an abstraction because only the inputs and outputs of the linguistic system are considered. Such a model will concern the structure of signals and the structure of meanings, and any structural relationship between the two. The computational

[1]Borrowed and adapted from Uriagereka (1998:93)

system responsible for the instantiation of this mapping will be treated as an entirely different issue. The model of language introduced in the last Chapter, where both meanings and signals where represented as points in a real-valued number space, is clearly unsatisfactory with respect to the relationship between meanings and signals realised by the linguistic system. Rather than mapping between two identical structures, the purpose of language is to map between two very different structures. For example, Pinker and Bloom state:

> language is a complex system of many parts, each tailored to mapping a characteristic kind of semantic or pragmatic function onto a characteristic kind of symbol sequence (Pinker & Bloom 1990:713)

In order to capture this property of language, a more accurate model of meanings and signals is required. However, this increase in accuracy will be limited to considering meanings as multi-dimensional propositional structures, and signals as sequences of symbols. Such a model of language captures the basic input-output relationship of the computational system underlying language.

4.2.1 Meaning structure and signal structure

Meanings are defined as feature vectors representing points in a *meaning space*. Throughout the remainder of this thesis, meaning spaces will be defined by two parameters, F and V. The parameter F defines the number of features each meaning will have. The parameter V defines how many values each of these features can accommodate. As a result, each feature will have the same number of values. Such a stipulation will prove useful, as many of the analyses to come will be made simpler as a result. It should be pointed out that this approach to specifying a meaning space using just two parameters is of no intrinsic importance; it serves only to simplify notation. I could just as well define a meaning space using the parameter F, along with an extra V parameters detailing the number of values each individual feature can take.

A meaning space defined by the parameters F and V can be thought of as an F-dimensional space. Each dimension will have V discrete points along its axis. For example, a meaning space $\mathcal{M}$ specified by $F = 2$ and $V = 2$ would represent the set:

$$\mathcal{M} = \{(1,1),(1,2),(2,1),(2,2)\}$$

Meanings represent structured objects of a fixed length. The values associated with each feature are drawn from a set; no notion of similarity applies to these elements, they are unordered. Similarly, the values in one feature are in no way related to values in another. Within a single feature, the only means by which values are related is through equality.

Signals are represented as a finite string of symbols drawn from some alphabet Σ. Signals can be of variable length. For example a signal space $\mathcal{S}$, defined by $\Sigma = \{\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{d}\}$, might be the set:

$$\mathcal{S} = \{\mathsf{ba},\mathsf{ccad},\mathsf{acda},\mathsf{c},\ldots\}$$

4.2.2 Structured relations between meanings and signals

The structure of meanings and signals has been defined. Of greater importance to following discussion will be the structure of the mapping *between* meanings and signals. It is the kind relationship existing between meanings and signals that makes human language so distinctive. Accordingly, it is crucial to be aware that the structure in the meanings and signals will restrict the set of mappings that are possible. As it stands, the model of language presented above allows both compositional mappings and non-compositional mappings. Importantly, it also *disallows* other aspects of the mappings between meanings and signals found in natural language. These restrictions need to be made clear and justified.

Compositionality

Compositionality is a property of the mapping between meanings and signals, rather than a property of a set of meanings, or a set of signals. A compositional mapping is one where the meaning of a signal is a function of the meaning of its parts (Krifka 2001). Such a mapping is possible given the model of language developed so far. Consider the language $L_{compositional}$:

$$L_{compositional} = \{\langle\{1,2,2\},\mathsf{adf}\rangle,\langle\{1,1,1\},\mathsf{ace}\rangle,\langle\{2,2,2\},\mathsf{bdf}\rangle,$$

$$\langle\{2,1,1\},\mathsf{bce}\rangle,\langle\{1,2,1\},\mathsf{ade}\rangle,\langle\{1,1,2\},\mathsf{acf}\rangle\}$$

This language has compositional structure because each meaning is mapped to a signal such that parts of the signal (some sub-string) correspond to parts of the meaning (a feature value). The symbol *a*, for example, represents feature value 1 for the first feature. The precise relationship between meanings and signals can vary substantially. For example, one feature value can map to two separate parts of the signal, these parts of the signal can be of variable length, and some parts of the signal can correspond to no part of the meaning. But importantly, the property of compositionality is independent of such characteristics of the mapping. Compositionality is an abstract property capturing the fact that *some* function determines how parts of the signal correspond to parts of the meaning.

As I noted in the previous Chapter, human language is compositional. Those languages with no compositional structure whatsoever I will term *holistic* languages[2]: the whole signal maps to a whole meaning, such that *no relationship* exists between parts of the signal and parts of the meaning. Here is an example of a holistic language $L_{holistic}$:

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

From holistic language to compositional language

Animal communication systems are generally regarded as holistic mappings between meanings and signals, with bees and ants standing as notable exceptions (von Frisch 1974; Reznikova & Ryabko 1986; Hauser 1996). By invoking the term *linguistic evolution* I am implying that at some point hominids, too, had some holistic protolanguage. In the context of the transition from protolanguage to modern syntactic language, two broadly defined scenarios must be considered (Hurford 2000a). First, consider the possibility that the meanings associated with signals were initially atomic in nature, with more complex semantic and syntactic structures emerging as a result of *combining* pre-existing utterances (Bickerton 1998;

[2]Strictly speaking, I should use the term *holistic communication system* since one of the defining features of language is compositionality. Nevertheless, I will continue to abuse the term *language* in this way in the interest of clarity.

Bickerton 2000; Hurford 2000b). This scenario corresponds to a *synthetic* move from protolanguage to full language. Contrast this situation with an *analytic* transition, where pre-existing complex semantic structures where associated non-compositionally with signals. These signals were then broken down or elaborated on to introduce compositional syntax (Wray 1998; Wray 2000; Kirby 2002a; Brighton 2002).

It should be noted that the language model I am developing here implies an analytic move from a holistic language to compositional language. This assumption follows from the observation that meanings always have some semantic structure: the possibility of communicating under-specified meanings or sub-components of fully specified meanings is not considered. The analytic approach I am assuming is common among models of linguistic evolution. For example, both Batali's (1998) and Kirby's (2002a) models assumes an analytic transition. In contrast, Hurford's (2000b) model is an exception in modelling a synthetic transition.

This analytic approach is closely related to an explanation for why human language contains both compositional and holistic relationships between meanings and signals. That is, in-between these two extremes I have outlined, a language can be partly compositional and partly holistic. There are many possible gradations between the two. Human language is most accurately thought of as a representing such a gradation (Wray 1998; Wray 2000). Some of our utterances exhibit compositionality. For instance, the English signal "*large olive pike*" has meaning by virtue of a compositional relation between the parts of the signal that carry meaning. In contrast, the English signal "*jump the gun*" has a holistic interpretation – the meaning is not derived from the meaning of any of the sub-signals. Of course, we can identify meaning with each of the constituents of the signal, but in English the most likely reading of this signal will not rely on such a relationship. In short, with regard to human language, a more plausible structured mapping between meanings and signals, using the language model developed above, might be L_{mixed}:

$$
\begin{aligned}
L_{mixed} = \quad & \{\langle\{1,2,2,1\},\mathsf{sghs}\rangle, \langle\{1,1,1,1\},\mathsf{ppold}\rangle, \langle\{2,2,2,1\},\mathsf{monkey}\rangle, \\
& \langle\{2,1,1,1\},\mathsf{q}\rangle, \langle\{1,2,1,1\},\mathsf{rcd}\rangle, \langle\{1,1,2,1\},\mathsf{esox}\rangle, \\
& \langle\{1,2,2,2\},\mathsf{adf}\rangle, \langle\{1,1,1,2\},\mathsf{ace}\rangle, \langle\{2,2,2,2\},\mathsf{bdf}\rangle, \\
& \langle\{2,1,1,2\},\mathsf{bce}\rangle, \langle\{1,2,1,2\},\mathsf{ade}\rangle, \langle\{1,1,2,2\},\mathsf{acf}\rangle\}
\end{aligned}
$$

This language is simply $L_{compostional} \cup L_{holistic}$, with an extra feature in each meaning differentiating between the two. Half of the utterances represent a compositional relation and half represent a holistic relation. Why does human language contain both holistic and compositional relationships? One plausible explanation for this situation, which draws heavily on the notion of an analytic transition, is that posited by Alison Wray (Wray 1998; Wray 2000). Firstly, the reason why holistic utterances are rife in natural language is due to the fact that they serve an important purpose: that of providing an efficient mechanism for producing frequent everyday utterances:

> It seems that holistic language may be picking up a shortfall between what we want to say and what we have the processing power to compute from scratch, by removing the burden of the everyday, pragmatically determined and communicably predictable, leaving the way clear for the more demanding analytic system to achieve the goals that only it can. (Wray 1998:63)

Secondly, Wray goes on to argue that this approach to forming utterances is still utilised for efficiency reasons, and represents a residue left by protolanguage:

> We can hypothesise, then, that although full human language was developing out of the older protolanguage system, it operated *alongside* it, not instead of it. (Wray 1998:57)

The model of linguistic evolution I am developing will assume such a transition: structured relationships between meanings and signals will arise due to the evolution of the signal, rather than the meanings and the signals. Agents form utterances for communicatively relevant situations which are not assumed to increase in semantic complexity over the course of linguistic evolution. Linguistic evolution occurs as a result of the signals being interpreted analytically. The biological evolution of semantic complexity is therefore an assumption of this model. It is in agreement with the view that syntax emerged partly as a result of the evolution of semantic complexity (Schoenemann 1999).

Beyond compositionality

Human language is clearly more complex than the basic compositional mapping between meanings and signals that I have outlined. Notably, language has recursive structure where utterances contain embedded and repeated sub-structures. In

contrast to the models of Kirby (2002a) and Batali (2002) discussed in the previous Chapter, I will not consider such aspects of language. Before furthering the discussion, this position needs to be justified.

By focusing on an explanation of the linguistic evolution of compositional structure, it is first necessary to strip the model of as many assumptions as possible. By proceeding to build and analyse a model that can also account for recursive structure, then any conclusion we draw concerning compositionality may be affected by assumptions made in modelling recursive structure. It is clear that any rigorous approach to modelling must proceed incrementally. Of course, it may turn out that modelling recursive structure does not impact on a study of compositionality in this way. But we must exclude this possibility by first focusing on an understanding of the linguistic evolution of compositionality. By building the picture piece-by-piece in this way, we can approach a more thorough understanding. The development of the models in the previous Chapter illustrates this process.

First I investigated a basic model of the transmission of classification competence. As a result of this experiment, it became clear that for structured states to be stable, the mapping being transmitted requires structured entities in both the domain and range. Without these basic models, which are inadequate with regard to explaining the big issues, we would lack a clear understanding of the basic issues. Second, I then extended the model to account for the evolution of a compositional relation between two abstract structured spaces. The refinements introduced above now extend the model further by taking a step toward more plausible meaning and signal spaces. To suddenly jump to a model capable of representing recursive structure would be a methodological oversight.

4.3 Simplicity: Motivation and foundations

On route to developing a more plausible and principled model of language, I have outlined a definition of a meaning space and a signal space. With a view to explaining the linguistic evolution of compositional structure, I have also discussed the explanatory limits imposed by these modelling decisions. For example, the model of meaning spaces adopted above *excludes* the study of recursive structure in language. But so far, I have not detailed how the relationship between meanings and signals is realised, or how knowledge of this relationship is acquired in light of linguistic evidence. These are fundamental problem for linguistics.

I will approach these issues from the perspective of simplicity. First, what I will refer to as *explanatory simplicity* relates to seeking the most economical explanation. For example, models developed with less parameters are preferred to those with more, and less ornate representations and process models are preferred to any others. This objective is widely appreciated and sought. In contrast, what I will refer to as *cognitive simplicity* relates to an assumption that cognitive structures fulfilling some function are consistently organised such that they are minimal in some sense. This is a vague definition which I will flesh out during the following discussion. Are these two applications of the notion of simplicity related? Not necessarily, as we can seek the most simple explanation of an inherently complex and baroque system. Explanatory simplicity is an *a priori* property of desirable explanations. The possibility of cognitive simplicity is a conjecture relating to the nature of the object being studied.

I have already put the notion of explanatory simplicity into practice: the incremental building of models is a prime example. An incremental development necessarily implies the primacy of simple explanation. The notion of cognitive simplicity needs further exploration, and its relevance to the cognitive systems underlying language needs to established. Before focusing on cognitive simplicity I will introduce the theory of Kolmogorov complexity, which provides a wide ranging and rigorous definition of simplicity. Kolmogorov complexity allows us to gain a firm grasp on concepts such as induction and information transformation from the perspective of simplicity. Using these insights, I will ask the following question: Why should we consider cognitive simplicity as a concept relevant to the study of language? At first sight, language is far from a simple system. With respect to the ostensibly basic task of relating meanings to signals, human language exhibits the hallmarks of complexity: imperfections such as redundancy and ambiguity are rife. I will argue that these complexities are actually artifacts arising from the principle of cognitive simplicity. There are two broadly defined scenarios in which simplicity can act as a driving force in determining the complexities of linguistic structure:

1. *Simplicity as the driving force for the internal transformation of representations.* The language faculty serves to mediate between LF and PF, and as such solves a problem defined by the boundary conditions imposed by cognitive demands *external* to language. Chomsky's minimalist program conjectures that the language faculty is a perfect solution to this task. This perfection relates to "considerations of economy and computational simplicity" (Lasnik 2002).

The imperfections of language are then conjectured to occur a result of the underlying drive toward internally defined perfection.

2. *Simplicity as the driving force behind induction from external linguistic stimuli.* The induction of linguistic knowledge from linguistic input is driven by a cognitive demand for the smallest consistent description of this external linguistic input. Compression of the observed data leads to generalisation, and this acquired knowledge is therefore partly determined by the form of the linguistic input. Iterated learning, in conjunction with this simplicity principle, can, as I will argue, result in the linguistic input containing residue arising from the historic use of language. This residue contains imperfections. But crucially, this residue does not have a detrimental effect on learnability, and therefore survives.

This discussion has outlined the motivation for investigating the role of simplicity in the cognitive mechanisms underlying language. Before furthering this discussion, we need a firmer grasp on the concept of simplicity. So far I have not defined precisely what I mean by simplicity. In order to do so, I will approach a fundamental and objective measure of simplicity: Kolmogorov Complexity. To place Kolmogorov complexity in context, a slight historical detour is required.

4.3.1 Simplicity and induction

Given data, induction is the process by which a hypothesis – some general law – is chosen to act as description for that data. In general, we entertain the possibility of some set of hypotheses $\mathcal{H} = \{H_1, H_2, \ldots\}$, and then given some data D, we choose an "appropriate" hypothesis $H_i \in \mathcal{H}$. In light of the observed data, some number of the hypotheses are typically *inconsistent*, the descriptions they represent will contradict the evidence represented by the data. The result of induction is therefore likely to result in the exclusion of some hypotheses in $\mathcal{H}$, leaving a set of candidate hypotheses $\mathcal{C} \subset \mathcal{H}$. The process of induction, which reduces to selecting one of these candidate hypotheses in a principled way, is an age-old and fundamental problem. The scientific process can be seen as the choice of a hypothesis in light of observations. Learning, too, can be regarded as the process of choosing a hypothesis on the basis of observations. Given that induction is such a weighty and deep problem, it is unsurprising that generations of scholars have attempted to find a general solution.

Any serious approach toward a general solution must start with the *Principle of Multiple Explanations* attributed to the Greek philosopher Epicurus, which states that one must entertain all consistent hypotheses, as rejecting a hypothesis that is consistent with what one has observed is unscientific – it is nothing more than a leap of faith. In one form or another, this principle resonates with all subsequent attempts at thinking about induction. First, it highlights the fact that there will often be a choice of hypothesis. Second, any preference in this choice must deviate from what the data alone suggests. In many respects, all subsequent attempts to solve the problem of induction relate to justifying a preference between consistent hypotheses.

Occam's Razor Principle, attributed to William of Ockham, is an appealing candidate preference: when there is a choice, always opt for the simplest hypothesis. The appeal of this principle is that overly complicated hypotheses should be rejected in favour of simpler, more elegant hypotheses. But this principle introduces a new problem: which measure of simplicity should we employ? Any measure we opt for will necessarily bias our choice of hypothesis, as no measure is theory-neutral.

Thomas Bayes offered an important insight by taking a probabilistic approach to the problem of induction. Rather than discarding inconsistent hypotheses and leaving a set of candidate hypotheses, the probabilistic approach assigns probabilities to the members of the hypothesis space, be they consistent or inconsistent. Hypotheses, therefore, are ranked according to their likelihood. These probabilities can be thought of as degrees of belief. Bayes's rule is subtle. In essence, the rule relates *prior* probability to *posterior* probability. The prior probability distribution over the hypotheses represents our initial *a priori* belief in each hypothesis, that is, the degree of belief before any evidence has been seen. In terms of probability theory, we say that $P(H)$ defines this probability distribution over hypotheses; every hypothesis has a probability, and the sum off all probabilities is unity. Now, in light of observed data and this prior distribution, Bayes's rule provides the posterior probability: the probability, or degree of belief, in a hypothesis H *given* D has been observed. More formally, Bayes's rule gives us the the posterior probability $P(H_i|D)$, the probability of H_i in light of D:

$$\mathrm{P}(H_i|D) = \frac{\mathrm{P}(D|H_i)P(H_i)}{\mathrm{P}(D)}$$

If all hypotheses are a priori equiprobable, then the most probable hypothesis given D will be the one that maximises $\mathrm{P}(D|H_i)$. This hypothesis will be the one that describes the data and *only* the data. This observation highlights the fact that prior bias is required for generalisation beyond the observed data. Bayes's rule and Occam's razor principle offer approaches to alleviating the problem posed by the principle of multiple explanations. But neither approach offers a satisfactory solution to the problem. We can object to Occam's razor principle on the grounds that there is no theory-independent criterion of simplicity: whichever notion of simplicity one employs must represent some kind of bias. Similarly, how can the prior probabilities of the candidate hypotheses, used by Bayes's rule, be determined? It is perfectly possible to assign prior probabilities to hypotheses, but how to do so in some principled and independent way is far from clear. These objections clarify an unavoidable conclusion: there must be *some* bias present in the hypothesis selection process.

4.3.2 Induction and universal computation

The principle of Occam's razor seeks the simplest consistent hypothesis. One measure of simplicity is suggested by the theory of *Kolmogorov Complexity*. Given an abstract object x, the Kolmogorov complexity, $K(x)$, is the length of the shortest computable description of x (Li & Vitányi 1997). So here, simplicity is measured relative to the shortest program capable of outputting the object in question. If the object is, for example, a random series of digits with no intrinsic regularity then the series cannot be compressed in any way: the Kolmogorov complexity will be identical to the length of series itself, or more precisely, the length of the program that outputs each digit in turn. But what about an object such as π, which is ostensibly a random sequence of digits? The Kolmogorov complexity of π is significantly lower than the series itself as a program that computes π is itself a description of π. This must be the case as a finite computational procedure can approximate the infinite series π to an arbitrary precision. The function $K(x)$ suggests a universal measure of simplicity across all computable objects. Is Kolmogorov complexity the objective measure of simplicity we seek? Given the conclusions of the above, we should ask: What kind of bias does the function $K(x)$ introduce?

An important result from the theory of Kolmogorov complexity is the *invariance theorem* (Li & Vitányi 1997:96). This theorem answers an understandable query one might have that arises from the definition of Kolmogorov complexity: The value of $K(x)$ for some x will depend entirely on the details of the manner in which

we form algorithmic descriptions. For example, different computer languages will favour certain kinds of algorithmic problem, and therefore yield different values of $K(x)$ for some fixed x. In short, our choice of description method will bias $K(x)$ in some fundamental way. This is where the depth and allure of the theory of Kolmogorov complexity strikes. The invariance theorem proves that whichever way the algorithmic description is formed, the Kolmogorov complexity is invariant within a constant factor. In other words, given any object x, $K(x)$ differs across two description methods ϕ_1 and ϕ_2 by some constant factor c. Crucially, the value c is *independent* of x. It is worth pointing out that we assume both ϕ_1 and ϕ_2 are models of universal computation (Hopcroft & Ullman 1979), and therefore, the theory of Kolmogorov complexity is restricted to apply to computable objects. Furthermore, the function $K(x)$ is not computable. Kolmogorov complexity, in its ideal and universal form presented here, is hard to deploy in any practical sense.

Despite its impracticality, Kolmogorov complexity is a useful concept when discussing the general issue of simplicity. There are several routes to relating the theory of Kolmogorov complexity to the issue of induction[3], for example, Kolmogorov complexity could serve as the measure of simplicity required to apply Occam's razor principle.

Rissanen's MDL principle

If our discussion on induction were to stop now, as a result of the gulf between theoretical elegance and practicality, we would be left in a frustrating position. We have deep and elegant theories that cannot be applied. Fortunately, we can approximate some of the intuitions of the theory of Kolmogorov complexity in such a way that they become computable. Such an approximation is realised through the *Minimum Description Length* (MDL) principle (Rissanen 1978; Rissanen 1989; Wallace & Boulton 1968). The MDL principle states that the best hypothesis for some observed data D is the one that minimises the sum of (a) the encoding length of the hypothesis, and (b), the encoding length of the data, when represented in terms of this hypothesis. These encoding lengths are measured in bits. More formally, for some hypothesis space $\mathcal{H} = \{H_1, H_2, \ldots\}$ we have an optimal scheme for encoding hypotheses denoted by C_1. Such a scheme must uniquely code each hypothesis. Then, the encoding length of a hypothesis H is defined as $L_{C_1}(H)$. Next, the data must be encoded with respect to this hypothesis. Depending on

[3]The principal means of relating the theory of Kolmogorov complexity to induction is via Solomonoff prediction, but I will omit a discussion of this branch algorithmic complexity for reasons of brevity. See, for example, Li & Vitányi (1997:324)

the structure of the hypothesis, different codes will specify different data items. The coding scheme denoted by C_2 defines how data is represented in terms of the hypothesis, that is, $L_{C_2}(D|H)$ is the length in bits of encoding the data using H. Note that coding schemes C_1 and C_2 are general coding schemes – they must apply across all hypotheses in $\mathcal{H}$ and all bodies of data that members of $\mathcal{H}$ can encode. We can now define the MDL principle:

Definition 7 (Minimum Description Length Principle) *Given the hypothesis space $\mathcal{H} = \{H_1, H_2, \ldots\}$, and optimal coding schemes C_1 and C_2, the best hypothesis, $H_{MDL} \in \mathcal{H}$, is defined as:*

$$H_{MDL} = \min_{H \in \mathcal{H}} \{L_{C_1}(H) + L_{C_2}(D|H)\} \tag{4.1}$$

This principle is appealing on several grounds due to its generality, simplicity, and roots in the theory of Kolmogorov complexity. Importantly, it has also been applied to countless engineering problems (Quinlan & Rivest 1989; Cameron-Jones 1992; Gao *et al.* 2000; Schmidhuber 1997). The details of how the principle is applied to a specific problem will be discussed below. But in general, applying the MDL principle requires one to construct a hypothesis space, encoding schemes, and some search strategy to locate H_{MDL} in the hypothesis space given some data. It is worth spending time in understanding how and why the principle works.

The MDL principle relies on the use of a two-part code to represent the hypothesis chosen and the effect that this hypothesis has in representing the observed data. A description composed of a two-part code is recurring theme in the theory of Kolmogorov complexity. It has a natural reading. In basic terms, given some data D, two contrasting ways of encoding D exist:

1. We could encode D as is. That is, D is considered a description of itself.
2. We could encode D as the conjunction of (a), a *general description*, G, that also describes observations not found in D, and (b) a *code* C that precisely identifies each observation found in D in terms of G.

Both these approaches to encoding describe the observed data perfectly, and nothing else. Now, if D is completely random then the first approach yields the smallest possible encoding length of D. This result follows because there can no general description of random data. But if D contains regularities, then the shortest possible encoding of D may be achieved by using the second description scheme. Here, any

regularity in the data is encoded in compressed form in G, a general description. As a result of this compression, the cost of encoding G may be significantly smaller than the cost of encoding D. But on its own, G is incapable of describing the data and only the data. To combat this deficit, the code C is used to identify each and every item in D in terms of G. Importantly, the *combined* cost of encoding both G and C may be *smaller* than the cost of encoding D. As a result, we have a compressed description of D. So, providing that enough regularity is present in D, then this second encoding method yields a shorter description of D. MDL uses this intuition to select a hypothesis. The hypothesis chosen by MDL corresponds to the general description G, and therefore may describe more data than that which was observed. This is how MDL can result in generalisation through induction. This is the intuition behind MDL.

The MDL hypothesis selection process has been defined in terms of a set of hypotheses $\mathcal{H}$ and optimal coding schemes L_{C_1} and L_{C_2}. The fact that the MDL of some data D is an approximation to the Kolmogorov complexity of D can now be understood by noting that the bias introduced by our choice of $\mathcal{H}$ and the coding schemes may deviate from the bias present in the function $K(x)$. To tie up this discussion, it is worth pointing out the concept of *Ideal MDL* (Vitányi & Li 2000). Ideal MDL is a reformulation MDL that stipulates that:

$$H_{MDL} = \min_{H \in \mathcal{H}} \{K(H) + K(D|H)\} \tag{4.2}$$

That is, when considering the encoding lengths of the hypothesis and the data given the hypothesis, then the encoding lengths are precisely the Kolmogorov complexity of these constituents. In other words, with the standard definition of MDL, the coding schemes L_{C_1} and L_{C_2} are not required to yield identical lengths to those reflecting the Kolmogorov complexity of the objects being encoded. This discrepancy is a result of the practical casting of MDL: computing $K(H) + K(D|H)$ is an undecidable problem. Deviating from the theoretical clarity and universality of ideal MDL is therefore a fact of life when applying MDL to practical problems. This discrepancy will introduce some kind of bias, an issue that I will return to later.

4.4 Simplicity: Cognition and language

The theory of Kolmogorov complexity is an abstract one concerning simplicity, compression, and induction from data. But is this theory relevant to the notions of

simplicity and induction we find useful to invoke when explaining aspects of cognition? It seems a basic fact that much of cognition involves compression. Perception, for example, requires complex and continuous stimuli to be filtered and understood. We don't act on the basis of the whole body of information arriving at our sensory surfaces. Some process of compression or filtration must be occurring. Another example is the problem of language acquisition. As soon as we entertain the idea that the acquisition process involves induction and generalisation, then the question of choosing between multiple consistent hypotheses must be answered. So, can we justifiably carry the notion of Kolmogorov complexity, and the related insights into the problem of induction, into a study of cognition?

One problem is that the concepts I have introduced have clear theoretical foundations, yet suffer from practical limitations such as the non-computable nature of the function K where $K(x)$ is the Kolmogorov complexity of the object x. Furthermore, even if the required calculations of K were both computable and tractable, its utility as a tool for the explanation of cognitive systems is by no means self-evident. In this section I will build the case for using the concepts developed above in a study of cognition. Firstly, through restricting its applicability, I will argue that a model of simplicity based on the theory of Kolmogorov complexity need not inherit all the theoretical restrictions. Secondly, I will discuss evidence that suggests that organisms behave in line with the concepts discussed above.

4.4.1 Simplicity and cognition

Chater and Vitányi (2003b) and Chater (1999) argue that simplicity is a relevant and widely applicable principle in the cognitive sciences. Stressing the importance of Kolmogorov complexity as a measure of simplicity, they present a convincing case for simplicity and information complexity being a unifying principle by citing examples ranging from low-level vision to high-level cognitive functions such as similarity and categorisation. Far from being a post hoc justification for the relevance of simplicity, Chater and Vitányi cite examples of research that have been driven by issues of simplicity and make novel predictions as a result (Feldman 2000; Hahn *et al.* 2003).

To illustrate the relation between simplicity and cognition, I will discuss two studies. The first study reports theoretical work done in extending Shepard's universal law of generalisation. This work suggests the theory of Kolmogorov complexity as a valid approach to investigating high level cognitive function. The second study analyses the communication system of ants, and demonstrates that these superficially basic

organisms are in fact capable of compression. So here, the theory of Kolmogorov complexity is used to shed light on models of both human and animal non-linguistic cognition.

Simplicity and compression applied to high-level cognition

In order to arrive at a general measure of psychological distance, Chater and Vitányi (2003a) apply the theory of Kolmogorov complexity. They aim to generalise Shepard's Universal Law of Generalisation, which states that the perception of similarity between objects is a negative exponential function of the distance between the items in psychological space (Shepard 1987). That is, according to some objective measure of similarity between physically measurable stimuli such as, for example, phonemes, the corresponding distance in psychological space is negatively correlated with the degree of perceived similarity.

Leaving aside the details of how this Law is tested empirically, a fundamental question relates to how distance is measured in psychological, or "mental" space. The notion of Euclidean space is frequently used as a model of psychological space, yet this practice restricts the applicability of Shepard's Law. Chater and Vitányi attempt to combat this problem by proposing a universal measure of distance between two objects x and y. The measure they propose is the length of shortest program that transforms x to y. This program could, for example, execute a representational distortion of x such that y is the result.

The key point here is that the notion of "shortest program" and therefore Kolmogorov complexity, is presumed to be psychologically real in some sense. That is, the theory of Kolmogorov complexity is a valid means of explanation for an act of cognition. In functional terms, similarity is reduced to the problem of finding the shortest (computational) procedure that relates one entity to another. Aware of the problems that this proposal introduces, specifically the issue of arriving at such a shortest program, Chater and Vitányi note:

> If we are interested in gaining insight into some complex system, it is common to attempt to formulate a radical (and knowingly unrealistic) idealization, that one hopes capture the minimal assumptions needed to make theoretical progress. (Chater & Vitányi 2003a)

The point is this: due to the non-computable nature of the function $K(x)$, a skeptic might, quite rightly, argue that humans cannot possibly arrive at measures derived

from Kolmogorov complexity. At a practical level, the skeptic is missing the point. Undoubtedly, humans cannot solve the *general* problem of computing $K(x)$, but it is perfectly possible that, given constraints on x, we can approximate $K(x)$. Adopting this stance, Chater and Vitányi go on to show that Shepard's Law can be derived when psychological distance is measured in this way. In doing so, the aim is to expand the applicability of Shepard's law. The important point here is that the theory of Kolmogorov complexity can usefully be translated into an explanatory framework one might use in explaining cognition. Applying Kolmogorov complexity in its purest and general form is not possible due to theoretical limitations. This does not exclude the use of Kolmogorov complexity, in a restricted form, informing a computationally tractable model of some cognitive processes.

Simplicity and compression in basic animals

In analysing the communication systems of ants, Ryabko and Reznikova (1996) demonstrate that organisms far more basic than humans can compress information (see also Reznikova & Ryabko (1996)). The fact that Kolmogorov complexity offers explanatory force in explaining cognition in animals must suggest that humans are capable of similar cognitive processes.

In these experiments, ants are placed in an environment configured so that the communication of food location between ants can be studied. Communication between ants is achieved using contact – the experiment is organised in such a way that the laying of pheromone trails can be ruled out as a means of communication. The environment consists of a tree-structured maze, with each junction in the maze presenting two possible paths, left (L) and right (R). A scout ant enters the maze at the root of the tree. Food is left at the end of one of the paths through the maze. The experiment then centres on the scout ant, aware of the food location after searching the maze, informing the other ants of the food location. That is, the scout ant finds the food itself by travelling through the maze, taking L/R decisions at each junction it encounters.

On returning, the experiment is orchestrated so that the scout ant transmits information detailing the food location to a stationary team of foragers through contact. The duration of this contact is timed. Assuming that the communicative act consists of a series of messages drawn from the set $\{L, R\}$, Ryabko and Reznikova showed how the compressibility of the path is negatively correlated with the time required for the message to be transmitted. That is, apparently random paths such as $LRLLRL$ (for a 32 leaf maze) took longer to transmit than compressible paths

such as *LLLLLR*: transmission duration correlates with message complexity. The ants must therefore be compressing the information required for informing the foragers of food location. Note also that this is an example of an innate compositional communication system.

4.4.2 Simplicity and the language faculty

The work discussed in the previous section suggests that simplicity is an appropriate concept to invoke when explaining non-linguistic cognition. I will now focus on language. Before focusing on the role of compression and generalisation during language acquisition, I will discuss how the concept of simplicity can shed light on theories of the language faculty.

The minimalist program

When considering Chomsky's minimalist program it is first useful to distinguish *methodological minimalism* and the *substantive thesis* (Chomsky 1995; Chomsky 2002; Lasnik 2002). Methodological minimalism refers to the practice of seeking explanations in such a way that the *range* of phenomena explained is of secondary importance to the *depth* of understanding one obtains of some core set of salient phenomena[4]. This practice has a long history, and is not specific to linguistics:

> Well, all of this is part of what you might call the "Galilean style": the dedication to finding understanding, not just coverage. (Chomsky 2002:103)

The substantive thesis, in contrast, concerns the computational processes of the faculty of language itself. The minimalist program sets out to investigate the possibility that the language faculty is an example of a perfect design. Of course, such a statement requires qualification: what is the linguistic system perfect *for*? It would seem that language is not a perfect communication system, as natural language exhibits ambiguity. Instead, the proposed perfection relates to the fact "that the human language faculty might be a computationally perfect solution to the problem of relating sound and meaning" (Lasnik 2002:434).

[4]Epstein & Hornstein (1999) also cite *methodological economy* as a characteristic of the minimalist program. This means, for example, applying principles such as Occam's razor. My interpretation of methodological minimalism is the one I discuss in the main text, and the one articulated by Chomsky (2002). Methodological economy, so defined, is hard to argue against. In contrast, what Chomsky terms methodological minimalism constitutes a more controversial claim.

The language faculty represents a solution to a problem that is specified by the boundary conditions imposed by interfaces with non-linguistic cognitive systems. On the one hand LF interfaces with conceptual-intentional aspects of non-linguistic cognition. On the other, PF interfaces with sensory motor systems capable of constructing and perceiving signals. These constraints are external to the faculty of language. So, if the language faculty is "well designed for interaction with the systems that are internal to the mind" (Chomsky 2002:107), then why does language exhibit imperfections and redundancy? What we might call *the minimalist assumption* leads to the conjecture that these *external* imperfections are artifacts caused by considerations of simplicity internal to the language faculty. For example, Chomsky states:

> This is the research direction: try to show that the apparent imperfections in fact have some computational function, some optimal computational function. (Chomsky 2002:118)

Given that the imperfections of language are a result of the interaction between internal boundary conditions and a pressure for cognitive simplicity, then given alternative boundary conditions, language, as well as the language faculty, might have been "perfect", and lack the imperfections we see now. This hypothetical observation, in conjunction with the preceding discussion, illustrates clearly the principle of detachment discussed in Chapter 2: the determinants of language structure and form can be understood entirely from the perspective of *internal* cognitive pressures and considerations.

The precise nature of the proposed perfection of the language faculty is frustratingly unclear. Advocates of the minimalist program are far from providing a concrete definition of perfection, but terms such as "computational simplicity" with respect to "general properties of organic systems" (Lasnik 2002:432) go some way to clarifying the issue, in contrast to general statements about "economy principles" (Marantz 1995:351), and the fact that the language faculty represents grammars that are "organised frugally to maximise resources" (Epstein & Hornstein 1999:xi), or that these grammars are "optimally economical" (Uriagereka 1998:523).

Although there is no a priori reason to presume Kolmogorov complexity is relevant to the minimalist program, or language in general, Lasnik's comments certainly suggest Kolmogorov complexity might be a plausible notion of simplicity with respect to minimalist theorising. Putting into use the theory of Kolmogorov complexity

would suggest that the language faculty might be the smallest piece of biological machinery that meets the requirements of the interface conditions mentioned above. In particular, and in a similar vein to Chater and Vitányi's (2003a) proposal of transformational distance, this situation suggests that the language faculty could be the shortest program capable of transforming between PF and LF. Such a proposal may go some way to explain the imperfections of language by appealing to the two-part code discussed earlier. Here the shortest description of the transformation is composed of, first, an elegant, simple, and regular transformation relating certain aspects of LF and PF, in conjunction with, second, corrections to this transformation that introduce the observed imperfections of language.

In short, the adoption of simplicity considerations by the minimalist program is far from an idle desire to seek more economical explanations. It makes a claim about the driving force behind imperfections in language: Can the underlying considerations of cognitive simplicity lead to these observed imperfections?

> If you are interested in the minimalist questions, what you'll ask is exactly that: why are they there? I think there is at least a plausible suggestion: they are there as perhaps an optimal method of implementing something else that must be there (Chomsky 2002:113)

Because Chomsky is at pains to stress the irrelevance of inductive generalisations during language acquisition, simplicity must relate to solving an internally defined problem – one specified by interface conditions will other cognitive sub-systems. As soon as we entertain the idea of inductive generalisation playing a role in the acquisition of language, then the impact of simplicity considerations may be very different. Before turning to this issue, I will clarify a possible source of confusion.

The simplicity measure for grammars

It worth pointing out that the "simplicity measure" used in Chomsky's early work on transformation grammar is an entirely different issue to that of simplicity in the minimalist program. Given primary linguistic data D, there may be several consistent hypotheses – grammars – permitted by a linguistic theory. Part of an explanatory theory of acquisition must therefore contain a procedure for ordering mutually consistent hypotheses:

> Two grammars for the same language might be equally complete and correct, and we wish our general theory to provide a criterion of choice

> between them. Of two theories that cover the same facts, it is usual to choose that one which is simpler (or more economical). (Bach 1964:178)

A number of criteria with which to measure grammatical simplicity exist. For example, among several possibilities, Bach considers the number of tokens used in the description of the grammar (Bach 1964:178). But importantly, the precise nature of the simplicity measure is an empirical question. Both Bach and Chomsky are at pains to point out that we cannot assume that some *a priori* measure of simplicity exists. The simplicity measure for a particular theory must be found by further linguistic enquiry, rather than by some independent and absolute measure such as, for example, Kolmogorov complexity. To avoid such confusion, the term *evaluation procedure* was adopted (Chomsky 1965; Chomsky & Halle 1968). The evaluation procedure is required to achieve explanatory adequacy for some theory of language – such a theory must explain why particular grammars are chosen over others for particular primary linguistic data. The evaluation procedure is *specific* to a particular grammatical theory; it is not intended to be employed as a criterion with which to judge competing theories of language. This early use of simplicity in Chomsky's writings is therefore of limited relevance to the current discussion. In direct contrast to this position, I will now discuss grammar acquisition as a process directed by a simplicity bias.

4.4.3 Simplicity and language acquisition

Compression and generalisation

I have discussed work that relates simplicity principles to cognition. So far, none of these studies have shed light on the issue of generalisation through induction, an issue suggested by the MDL principle. In contrast, Wolff (1982) focuses on precisely this problem but draws no parallel between his work and the theory of Kolmogorov complexity, or MDL[5]. Noting that the result of language acquisition is an intricate knowledge of a complex linguistic system, Wolff argues that simplicity considerations, or to use Wolff's terminology, "cognitive economy", should be invoked to explain this complexity:

> By exploring the properties of simple mechanisms which have complex implications and by relating these properties to known features of language development we may, in the interests of theoretical parsimony,

[5]For related work on using simplicity and MDL in grammar induction see Langley & Stromsten (2000) and Grünwald (1996).

> hope to find simplicity behind the apparent complexity of language acquisition. (Wolff 1982:58)

For Wolff, simplicity means compression: syntactic forms are derived from data as a result of generalisations arising from a preference for compact representations. But beyond simply the interests of theoretical parsimony, Wolff believes that simplicity has a cognitive reality, and constitutes a broadly applicable principle. On issues of simplicity, Wolff notes:

> They have an obvious relevance to information storage, but their impact can be at least as great on information transmission — from one person to another or from one part of the nervous system to another. So great are the potential advantages of data compression principles in increasing the volume of data that may be handled by a given system, or in reducing costs for a given body of data, that it is hard to imagine how any biological information processing system could evolve without incorporating at least some of the principles in one form or another. (Wolff 1982:61)

Wolff seeks to build a computational model of the language acquisition process. Given some text generated from a grammar G, the system induces some grammar G'. The fundamental question Wolff investigates is the degree to which generalisations made by G' mirror those made by G. Noting that young children often over-generalise to forms like *goed* and *mouses*, the issue of the *extent* of generalisation is the principal concern. Wolff identifies the degree of grammar compression to be the determiner of the extent of generalisation. Accordingly, acquisition effects such as overgeneralisation are best investigated from the perspective of compression.

Two measures are employed to track the effect of compression. Given some data D, S_g defines the size, in bits, of the grammar induced from D. The degree of compression of the data achieved using the grammar is denoted as CC. This measurement is crucial; it is calculated on the basis of, (a), the the number of bits required to encode the data in terms of the induced grammar, v, and (b), the number of bits required to encode the data in unprocessed form, V. The measurements S_g and CC are but a short step from corresponding directly to the two factors used by the minimum description length principle.

With respect to Equation 4.3, Wolff's S_g is identical to $L_{C_1}(H)$, the length in bits of encoding the hypothesis. Furthermore, Wolff's CC is a function of $L_{C_2}(D|H)$.

Because CC is represents a compression rate, rather than a number of bits, Wolff's scheme departs from the MDL principle as the sum of these two quantities lacks a clear interpretation. Nevertheless, Wolff remarks on the utility of considering these two expressions:

> A trade-off exists such that an improvement in CC may be achieved at the expense of a larger grammar and *vice versa*. In these circumstances it would seem necessary to assign weights to CC and S_g representing their relative importance biologically, so that an optimum balance may be found between the two. *A priori*, however, we do not know what these weights should be. (Wolff 1982:65)

The elegance of the MDL principle is a result of its simplicity: there is no need to specify weights, as the trade-off is achieved by minimising the sum of the two key quantities. In short, Wolff uses the same measurements of the MDL principle, but stops short of a thorough reconstruction of MDL by failing to consider the *total* cost of encoding the data in terms of the hypothesis and the encoding length of the hypothesis itself.

Recall that Wolff is interested in the processes causing overgeneralisation. What Wolff terms the *range* of the grammar is the variety of terminal strings it can generate. The key insight that Wolff makes is that by restricting the degree of compression of the grammar by considering the trade-off between CC and S_g, the range, or degree of generalisation, can be regulated. Such regulation, determined by the degree of grammar compression, can be used to explain overgeneralisations made during acquisition.

During language acquisition from a text, the system builds a grammar incrementally. Starting with the most frequently represented components, words, the system induces increasingly more complex (but less frequent) structure represented in the text. Overgeneralisations occur as a result of the initial phase of the induction of syntactic patterns which, when first discovered, result in the grammar overgeneralising. As more complex patterns are induced, the range of the grammar is regulated due to the increased cost of representing the data in terms of the hypothesis: small grammars no longer lead to maximum compression. Wolff claims that this intermediate phase, where overgeneralisation occurs, is observed in infants over the course of their linguistic development.

In relation to this issue, Clark (2001) notes that the frequency of structures occurring in an input text is inversely proportional to their complexity. As a result, this observation must place a bound the degree of complexity of "witnesses" to syntactic parameters. That is, certain structures in the input text must provide evidence that suggest an appropriate setting of a parameter value – these structures act as witnesses. If a target grammar is to be reliably learned in finite time from a finite text, then there must be a bound on the complexity of the structures acting as witnesses. Similarly, this bound must inform any theory of variation across languages:

> If parameter values must be expressed on extremely compact structure, then the set of parameters we can incorporate into our theory of UG will be tightly constrained, thus placing a limit on the kind of linguistic variation we can observe in principle. (Clark 2001:136)

Clark goes on to propose that the notion of the complexity of a parameter can be regarded as the Kolmogorov complexity of the simplest structure that expresses that parameter – its witness. Simple parameter values will therefore be more frequent in the input text, and be set early on in the language acquisition process. Similarly, complex parameters will be set later in the acquisition process. This ordering of acquired complexity is mirrored in Wolff's theory: smaller (less complex) grammars, with a wide range, are acquired before more complex grammars with a narrower range.

In the discussion that follows, I will draw on these insights, which correspond directly to those suggested by the MDL principle, to justify a model of language induction developed below. It should be noted, however, that the work of Wolff and Clark deal only with inducing grammars from sets of signals, or strings. Their work therefore has little *direct* relationship with the model of language I have been developing, which requires us to also consider the role of meanings.

4.4.4 Language, simplicity, and evolution

Some of the fundamental problems in linguistics can be informed by simplicity principles. Now is a useful point at which ask the following: Are the simplicity principles discussed so far a unifying theme in the study of language, or do they refer to unrelated, mutually exclusive, questions? In particular, is the notion of simplicity in the minimalist program related to the notion of simplicity with respect to language acquisition? This is a significant question. Due to the reticence of Chomsky in

accepting the role of inductive generalisations in language acquisition, these interpretations of the role of simplicity appear quite disparate. The minimalist program adopts the notion of simplicity to make a claim about the relationship between pressures determining the structure of the faculty of language, and imperfections in language. Simplicity guides *internally* directed pressures on cognitive organisation that lead to *external* complexities. By internal, I mean internal to the language faculty, and by external, I mean external to the language faculty.

In contrast, if language acquisition is driven by the pressures of simplicity, specifically the compression of representations, then simplicity guides cognitive organisation on the basis of *externally* defined structures found in the environment, that is, the linguistic input. The artifacts of this process, as we have seen, correspond to degrees of generalisation which are not necessarily a reflection of the smallest grammar consistent with the input: generalisations are a result of grammar exhibiting far from optimal design, due to the underlying pressure to represent the *observed* data minimally. The minimalist position can claim to account for the observed imperfections of language. Can this alternative application of simplicity principles do the same? In Chapters 6 and 7, I will argue that it can, when placed in the context of iterated learning. One conclusion arising from this theory is that the two applications of the notion of simplicity can in fact be regarded as the same process, but only if generalisation is accepted as part of the process of language acquisition. Both perspectives assume that the cognitive structures relevant to language are optimal. One perspective proposes that the task is determined by an expression of the genes. The other proposes that this task is (at least partly) specified by the linguistic input.

Before discussing the details of the model I will use to investigate this possibility, it is worth considering some recent work tackling many of the issues I have raised. In the context of iterated learning, compression, and language evolution, Teal & Taylor (2000) investigate change in regular languages as a result of repeated cultural transmission. These regular languages are represented by finite state automata (FSA) (Hopcroft & Ullman 1979). At each iteration in the model, the "smoothness" of the language is tracked by measuring the average string-edit distance between words contained in the language. Using the minimum description length principle, the FSAs are compressed at each iteration. Teal and Taylor note that the degree of generalisation achieved at each iteration increases as a result of compression: more strings can be generated by the induced FSAs as the experiment progresses. Contrary to their expectation, they also found that the average string edit distance

did *not* change as a result of linguistic evolution. Linguistic evolution in no sense leads to smoother string spaces. Teal and Taylor's main finding was that languages change more rapidly when they are incompressible. Therefore, as linguistic evolution proceeds, the compressibility of the languages increases. This observations will tally with results obtained later in the thesis.

Although Teal and Taylor's work focuses on the key issues I will be investigating, their work has little impact on the discussion that follows. Rather than treating language as a set of signals, I will be treating language as mapping between meanings and signals. As a result, very different structural properties will be of interest. For example, compositionality is property of a mapping, rather than a string space. In this respect, the conjecture I have made regarding imperfections arising as a result of simplicity bias in language learning cannot be investigated in a model such as that presented by Teal and Taylor. In order to progress, I will have to develop a new model.

4.5 Towards a model of linguistic evolution

Within an iterated learning model, agents act as a conduit for language. An agent observes a subset of the language of the previous generation. The process of learning from this subset, and then producing utterances for the next generation, is a complex one. At one level an agent simply maps one language onto another. At a more detailed level, the function which defines this mapping is composed of a learning mechanism and a production mechanism. Learning, in this context, is the process of arriving at a hypothesis which explains the observed language. Production is the process that, given a hypothesis, completes the mapping from meanings to signals. The production mechanism defines how the hypothesis is interrogated to yield signals.

In Section 4.2, I introduced the rudimentary aspects of a model of language by detailing a model of meanings and a model of signals. But how should the relationship between meanings and signals be realised? And how is knowledge of this relationship acquired in light of linguistic evidence? Informed by the discussion of simplicity in Sections 4.3 and 4.4, I will now lay the foundations of a model of linguistic evolution driven by a simplicity principle. The motivation for developing this model relates to the question of justifiable generalisation. How biased do linguistic agents have to be in order to effect a cumulative evolution towards structure? Such questions are hard to answer because the degree of inductive bias

of a learner is hard to quantify. I will use the minimum description length principle as the selection criterion when choosing a hypothesis in the light of linguistic data. In doing so, the range of *possible* generalisations derivable from the data will not always be represented by the hypothesis. Instead, generalisation is only performed when the induced regularity in hypothesis contributes to the smallest description of the observed data, and nothing else. By using the MDL principle, the issue of overgeneralisation can be addressed. If models of linguistic evolution are to carry any explanatory weight, then issues of justifiable induction must be tackled.

To construct such a model based on MDL, two widely recognised problems need to be overcome (Chater & Vitányi 2003a):

1. The *representation problem* requires designing a representation capable of accounting for possible hypotheses in some principled way. As MDL selection is in part determined by the structure of the hypothesis space, then this choice is important and needs to be justified.
2. The *search problem* requires a procedure for finding the hypothesis with the minimum description length. The hypothesis space, in conjunction with a coding scheme, defines which hypothesis leads to the minimum description length of some data D. But *finding* this hypothesis is an entirely different issue, requiring a search strategy. If the search space is infinitely large, then heuristics may have to be employed.

Overcoming these two problems will mean that we have an effective induction algorithm. In order to place an agent in an iterated learning model, several other issues need to be resolved: How are signals derived for meanings which have never been observed in conjunction with a signal? What happens when the hypothesis cannot express a novel meaning?

The remainder of this section introduces an approach to hypothesis selection using the MDL principle, and formulates an initial attempt to answer these further questions concerning the use of the model in an iterated learning framework. Many of the details of this model, however, will be developed and analysed as the thesis unfolds.

4.5.1 Hypothesis selection by minimum description length

To re-cap, in order to deploy the MDL principle the following components need to be defined:

A hypothesis space $\mathcal{H}$. Given some data D, the most appropriate hypothesis needs to be drawn from a set of possible hypotheses.

Coding schemes C_1 and C_2. The coding schemes tell us how many bits are required to uniquely encode a particular hypothesis (L_{C_1}), and how many bits are required to encode each observed data item in terms of the hypothesis (L_{C_2}).

Hypothesis selection rule. The selection rule underlying the MDL principle was discussed earlier. Recall that it is defined as follows:

$$H_{MDL} = \min_{H \in \mathcal{H}} \{L_{C_1}(H) + L_{C_2}(D|H)\} \tag{4.3}$$

The precise form of the data D is totally defined by the meaning and signal spaces introduced in Section 4.2. Therefore, the data will always be a finite set of meaning/signal pairs $D \in \mathcal{M} \times \mathcal{S}$. I will now tackle the remaining issues in turn.

The hypothesis space

I will introduce a novel model for mapping strings of symbols to meanings, which I term a Finite State Unification Transducer (FSUT) (Brighton & Kirby 2001; Brighton 2002). An FSUT is a variation on the basic notion of a finite state transducer. A finite state transducer maps one regular language to another by attaching output symbols to each state transition within the transducer. An FSUT extends this mechanism by mapping meanings to signals, where meanings and signals take the form discussed at the beginning of this Chapter. A finite state unification transducer is therefore an augmented *Mealy machine* (Hopcroft & Ullman 1979:42). This model can also be considered an extension to the scheme used by Teal & Taylor (2000); the extensions accommodate variable length signals and, more importantly, meanings. Given some observed data, the hypothesis space consists of all FSUTs that are consistent with the observed data. Observed examples of meaning/signal associations are never discarded. Both compositional and non-compositional languages can be represented using the FSUT model.

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

A, which receives a set of meaning/signal pairs during language acquisition. For example, an observed language might be the set:

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

This language is compositional. It was constructed using the following lookup table, which relates feature values to the corresponding substrings in the signal:

	value 1	value 2
feature 1	a	b
feature 2	c	d
feature 3	e	f

So for example, the sub-signal corresponding to feature value 2 for the first feature is "b". Figure 4.2(a) depicts a FSUT which models L_{comp}. We term this transducer the *Prefix Tree Transducer* – the observed language and only the observed language is represented by the prefix tree transducer.

4.5.2 Compression

The power of the FSUT model only becomes apparent when we consider possible generalisations made by merging states and edges:

1. *State Merge.* Two states q_1 and q_2 can be merged to form a new state if the transducer remains consistent. All edges that mention q_1 or q_2 now mention the new state.
2. *Edge merge.* Two edges e_1 and e_2 can be merged if they share the same source and target states and accept the same symbol. The result of merging the two edges is a new edge with a new meaning label. Meanings are merged by finding the intersection of the two component meanings. Those features which do not have values in common take the value "?" – a wild-card that matches all values. As fragments of the meanings may be lost, a check for transducer consistency is also required. Without this consistency check, some observed meaning/signal pairs will not be accounted for by the resulting transducer.

The check for consistency ensures that compression of the transducer is not "lossy": compression can only ever lead the transducer to explain more than the observed

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

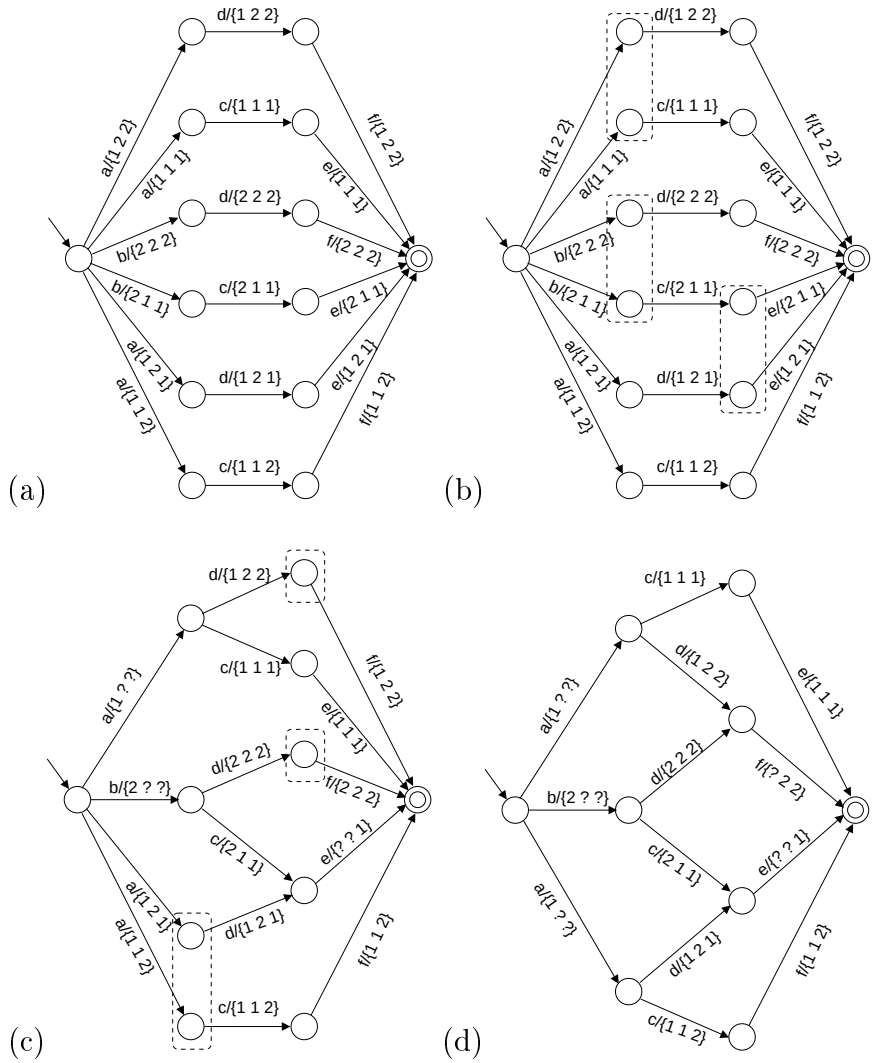


Figure 4.2: Given the compositional language L, the Prefix Tree Transducer shown in (a) is constructed. By performing edge and state merge operations, outlined in (b) and (c), the transducer can be compressed. The transducer shown in (d) is compressed, but does not lead to any generalisations.

data, it can never explain *less* than the observed data. Information is never thrown away or filtered. Figure 4.2(b) and (c) illustrate some possible state and edge merge operations. The transducer resulting from these merge operations is show in in Figure 4.2(d). Figure 4.3 depicts the fully compressed transducer, which is found by performing additional state and edge merge operations. Note that further compression operations are possible, but they lead the transducer to express meanings in such a way that they are inconsistent with the observed language. For example, all the states could be merged into a single state that could act as both the start state and the accepting state. But by performing these merge operations, the ordering of the symbols is lost, and as a result, a signal can generated for a meaning that differs from the signal observed with that meaning. Transducer behaviour must always be consistent with the observed data.

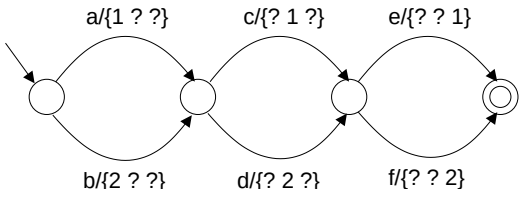
Prefix tree transducers are compressed by applying the merge operators described above. For the time being, the details of how these operators are applied is not important. I will return to the issue in Chapter 6. In short, the search problem one must confront when deploying MDL is now cast as series of merge operators that lead from a prefix tree transducer to a compressed transducer.

4.5.3 Production and Invention

To express a meaning, a search through the transducer is performed such that an appropriate series of edge transitions are found. These edges represent a set of meaning fragments – some of the meanings may contain wildcard feature values, and will be under-specified. If the unification of the set of meanings yields the meaning to be expressed, then the meaning can be expressed, and the resulting signal is formed by concatenating the symbols attached to each edge. The ordering of the symbols in the signal reflects the ordering of the edge traversals when passing through the transducer. The fully compressed transducer can potentially express meanings which are not present in L_{comp}. The language L^{+}_{comp}, shown in Figure 4.3, contains all the meaning/signal pairs which can be expressed by the fully compressed transducer. By compressing the Prefix Tree Transducer, the structure in the compositional language has been made explicit, and as result, generalisation can occur.

Generalisation can lead to a transducer expressing meanings it has not observed in conjunction with a signal. Thus, within the context of iterated learning, any structured relation between meanings and signals will have a chance of surviving to

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



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

Figure 4.3: Given the compositional language L a series of state and edge merge operations, beginning with those shown in Figure 4.2, result in the compressed transducer shown above. As a result of compression, the transducer can express more meanings than those contained in L. All members of language L^{+} can be expressed.

the next generation. When structure is lacking in the observed data generalisation will not be possible, and the result will be that some meanings cannot be expressed. Furthermore, even if generalisation is possible, the set of meanings that can be expressed may still be a subset of the meanings the transducer is called to produce utterances for.

For an agent to be fully functional within the iterated learning model, some invention procedure is required. Invention is the process by which an agent produces an utterance containing a meaning which cannot be produced using the usual method. The process of invention is necessarily unprincipled, and more often than not will require random signals or sub-signals to be constructed. For the time being this issue is not important, but it will be of fundamental interest in the discussion of Chapter 6.

4.5.4 Encoding Lengths

In order to apply the MDL principle we need an appropriate coding scheme for: (a) encoding the hypotheses, and (b) encoding the data using the given hypothesis.

These schemes correspond to C_1 and C_2 introduced in Equation 4.3. The requirement for the coding scheme C_1 is that some machine can take the encoding of the hypothesis and decode it in such a way that a unique transducer results. Similarly, the coding of the data with respect to the transducer must describe the data uniquely. To encode a transducer $T = (Q, \Sigma, F, V, \delta, q_0, q_F)$ containing n states and e edges we must calculate the space required, in bits, of encoding a state ($S_{state} = log_2(n)$), a symbol ($S_{symbol} = log_2(|\Sigma|)$), and a feature value ($S_{fvalue} = log_2(V)$). The number of bits required to encode a meaning relies not only on S_{fvalue}, but also the cost of encoding the wildcard specifier. The number of bits used to encode an arbitrary meaning $m = \{f_1, \ldots, f_F\}$ is given by:

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

where f_i denotes the value of the ith feature, and

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

That is, $Z(f_i)$ represents the number of bits required to encode either a feature value or the wildcard specifier. The initial bit is used to differentiate between these two possibilities. Denoting the meaning associated with the ith edge by m_i, the encoding length of the transducer is then:

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

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

$$S_{prefix} = S_{state} + 1 + S_{symbol} + 1 + S_{fvalue} + 1 + S_F + 1$$

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

$$L_{C_1}(h) = S_{prefix} + S_T \tag{4.4}$$

Recall that $L_{C_1}(h)$ defines the length of the encoding of the hypothesis h using the coding scheme C_1. This quantity is termed the Grammar Encoding Length (GEL) (Teal & Taylor 2000). Similarly, the length of the encoding of the data, in terms of the hypothesis h, $L_{C_2}(D|h)$, is termed the Data Encoding Length (DEL). The DEL is far simpler to calculate than the GEL. For some string s composed of symbols $w_1 w_2 \ldots w_{|s|}$ we need to detail the transition we choose after accepting each symbol with respect to the given transducer. The list of choices made describes a unique path through the transducer. Additional information is required when the transducer enters an accepting state as the transducer could either accept the string or continue parsing characters, as the accepting state might contain a loop transition. Given some data D composed of p meaning/signal pairs, $L_{C_2}(D|h)$ is calculated by:

$$L_{C_2}(D|h) = \sum_{i=1}^{p} \sum_{j=1}^{|s_i|} \{\log_2 z_{ij} + F(s_{ij})\} \tag{4.5}$$

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

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

4.5.5 Bottlenecks and the problem of learning from incomplete data

The transmission bottleneck is a crucial parameter in many iterated learning models. In Chapter 3, both models relied on a transmission bottleneck. Until now, the precise nature of the bottleneck was of limited importance; the resulting behaviour of the models presented is contingent on the presence of a bottleneck, but invariant over the severity of the bottleneck. Both *semantic* and *production* bottlenecks will be used in the iterated learning model I am developing here (Hurford 2002).

Semantic bottleneck. The semantic bottleneck refers to the degree of exposure of the meaning space represented in the data an agent learns from. Often, the bottleneck severity is defined by the number of utterances, R, a learner is exposed to. Because the utterances are formed from a random sampling of the meaning space, the number of individual meanings observed will be less than R. Another approach to specifying the severity of the bottleneck is *meaning space coverage* (Brighton

2002). Coverage refers to the proportion of the meaning space an agent can expect to observe. If utterances are constructed by randomly sampling a meaning space containing n meanings, then coverage, c, is related to R and n as follows:

$$c = 1 - (1 - \frac{1}{n})^R \tag{4.6}$$

So when sampling from n meanings R times with replacement, we will on average see a proportion of the meaning space c ($0 \leq c \leq 1$). Rearranging, we can also say the following:

$$R = \frac{log_2(1-c)}{log_2(1-\frac{1}{n})} \tag{4.7}$$

Both these equations assume that the meaning space is finite, which is the case given the language model I have defined. In Chapters 5 and 6, I will investigate the significance of bottleneck size. The bottleneck size will be discussed in terms of the value c or the value R, depending on the context.

Production bottleneck. A production bottleneck is also employed in the model. Recall that a production bottleneck refers to the situation where a single signal is expressed when several candidate signals may exist. Hurford claims that such a bottleneck is required in any plausible model of production:

> Note that it would in fact be unrealistic **not** to implement a production bottleneck. Communication in real societies involves singular speech events, in which a speaker finds a single way of expressing a particular meaning. There is no natural communicative situation in which a speaker rehearses **all** her forms for a given meaning. (Hurford 2002:306)

The production bottleneck in this model is implemented by the production mechanism selecting the first successful path through the transducer for expressing the given meaning. Importantly, this modelling decision excludes the possibility of synonymy.

4.5.6 A summary of modelling decisions made so far

It is worth consolidating the key modelling decisions made so far:

Meanings and signals. At the beginning of this Chapter I defined the notions of meaning space and a signal space. Meanings are feature vectors defined by two parameters F and V. Signals are variable length strings of symbols drawn from an alphabet Σ.

Representing the mapping. Finite state unification transducers represent the mapping between meanings and signals. They define how meanings are *transformed* into signals.

Hypothesis selection. Given a series of observed meaning/signal pairs, the prefix tree transducer represents this data exactly as it was presented. The set of possible hypotheses is the set of all FSUTs consistent with the observed meaning/signal pairs. Hypothesis selection is performed by compressing the prefix tree transducer. The most appropriate transducer is the one that results in the smallest description length of the data. As a result, compression is guided and limited by considerations of simplicity: we seek the smallest description of the observed data, and only the observed data. This description tells us which hypothesis to use. This hypothesis may explain more than the observed data.

Mapping novel meanings to signals. In order for the model to function within the iterated learning framework, the process of relating previously unseen meanings to appropriate signals must be well defined. Two mechanisms have been introduced. First, when generalisation is possible, novel meanings can be assigned an appropriate signal through a principled use of the induced hypothesis. When the hypothesis cannot inform the production of the novel meaning, then some kind of random invention must be performed. Invention, at this stage of the analysis, need not be fully defined.

Further details will be discussed later, but as I will show in the next Chapter, some fundamental insights into linguistic evolution can be gleaned from the model as it stands.

4.6 Chapter summary

This Chapter has resulted in the development of a model. The motivation behind this model stems from a discussion of the role of simplicity, Kolmogorov complexity, and induction in cognition. In particular, I have focused on the role of the minimum description length principle in language learning. Before investigating this model further, it is worth reflecting on how I have arrived at certain modelling decisions.

First of all, I have restricted the range of explanation by formulating a basic model of meanings and signals. Meanings are represented as feature vectors, and therefore exclude the possibility of recursive structure. Signals are strings of symbols. The model is therefore well placed to investigate compositionality. I have also adopted the view that the emergence of compositionality must arise through an analytic transition, where structured meanings are first associated non-compositionally with signals. Linguistic evolution will result in the signals becoming structured in relation to the meanings.

I have discussed the role of simplicity in both linguistic and non-linguistic cognition. Engaging with the theory of Kolmogorov complexity, I then considered the role of simplicity in language. Both Chomsky's minimalist program, and theories of language acquisition based on inductive generalisation, relate directly to the theory of Kolmogorov complexity. Motivated by these issues, I have formulated a model of language acquisition based on the minimum description length principle. This model realises the mapping between meanings and signals. In light of observed data, the model also suggests the most appropriate hypothesis according to the minimum description principle. In this way, the process of generalisation is guided by a rigorous application of a simplicity principle: the observed data is represented in compressed form, and this representation suggests an appropriate hypothesis.

The model I have developed will prove central to the progression of the thesis. When required, I will build on the existing model. However, as it stands, several fundamental results can be established.

Chapter 5

Stability Conditions through Static Analysis

5.1 Introduction

A model capable of representing the mapping between meanings and signals is in place. In addition, I have introduced a selection criterion detailing which mapping is appropriate in the light of some arbitrary body of data. Now several interesting questions arise. Within the context of iterated learning, which set of model parameters consistently lead to the evolution of structured mappings? And what do these structured mappings look like? Rather than proceed by building a detailed iterated learning model, this Chapter will attempt a short-cut. Instead of modelling the dynamics of change explicitly, the regions of state-space that interest us will be analysed first. At this stage of the discussion, the arrival of stable and structured mappings is only an assumption. The purpose of this analysis is therefore, first, to see whether this assumption can be justified, and second, to derive the conditions for stability. Recall that some fleeting occurrence of a structured language is of no real interest; a random walk through the state space will guarantee such a state of affairs. Rather, the issue of interest concerns consistent linguistic evolution towards stable linguistic structure.

The following analysis depends on two abstractions. The first abstraction introduces the possibility of simplifying the learning process to a degree that permits us to draw bounds on the generalisation process. I will introduce the notion of *optimal generalisation* which captures the most powerful generalisation bias possible given certain compositional language structure. That is, under the assumption of compositional input, the highest degree of generalisation to unseen utterances consistent with the input. The only restriction on the range of compositional languages

covered by this abstraction is the stipulation that synonymy is not present in the language. Given the presence of the production bottleneck discussed in the last Chapter, this simplification is consistent with the model being developed.

The notion of optimal generalisation rests on the MDL principle in conjunction with certain assumptions about the probability distribution in the data. The second abstraction leads us to consider only two language types. The FSUT model can account for holistic language, compositional language, and all gradations between these two extremes. But by only considering these extremes, we can gain a general insight into the issue of stability. In doing so, general results concerning the stability payoff gained by structured language can be found. By introducing these two abstractions, I will show how a substantial part of the parameter space of the ILM can be mapped and understood. In short, the conditions determining stable states can be understood without invoking a full-blown agent based model. Furthermore, these shortcuts will allow us to understand regions of the parameter space that are computationally intractable to explore through simulation modelling.

5.2 The parameter space, linguistic structure, and stability

Only under certain conditions will structured languages be adaptive and therefore consistently observed. The lack of a transmission bottleneck, for example, will weaken the pressures causing language adaptation. The absence of any structure in the meaning space will also rule out any kind of structured mapping between meanings and signals. We can intuit such conditions quite easily; but other conditions are more subtle and cannot be arrived at without thoroughly mapping the parameter space. For a rigorous understanding of the process of iterated learning, knowledge of the behaviour of the system under a wide range of conditions is required. In doing this, general statements can be made, rather than specific observations.

The parameter space, sketched in Figure 5.1, represents every possible configuration of parameters for a particular model. This space can be divided into regions that reflect the degree of structure present in the languages that are observed to occur under those conditions. I have depicted three such regions, the first characterising random mappings. Part of this region must contain those parameter combinations in which no transmission bottleneck is present. It must also contain those parameter conditions that define meaning spaces with only one feature, as these conditions represent structureless meanings. But what kind of parameter combinations characterise the set of conditions that lead to structured languages being observed? This

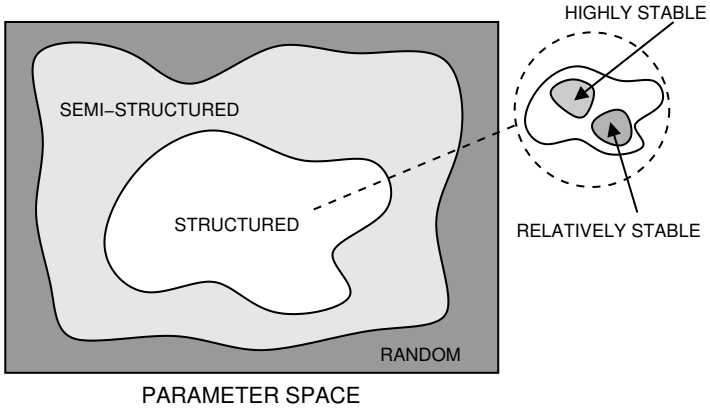


Figure 5.1: Key regions of the parameter space. The space of all parameter combinations can be divided into three distinct regions. First, those regions that consistently lead to structured mappings. Second, those regions that lead to random mappings. And third, an intermediate level, where mixed systems emerge. Within each region, stability can vary.

is the question I will answer in this Chapter. Of the three regions identified, the region comprising structured languages is the most salient region in the context of this discussion. Ultimately, we aim for a firm understanding of the conditions for compositional structure.

I have broken up the space of parameter combinations, and noted the kind of the languages that are *likely* to be observed in the context of iterated learning. What exactly does this mean? The statement suggests that after multiple runs of the iterated learning model, certain kinds of languages are observed more often than others, and it is these frequently observed languages that characterise that part of the parameter space. By frequency of observation, I mean the behaviour of the system after each iteration. Within the regions defined above, another set of sub-divisions is possible. Parts of the region characterised by structured languages will represent the case when structured languages are *always* observed. Similarly, another part may just represent the parameter combinations which lead to compositional structure being observed with greater than chance likelihood. A more accurate and insightful way of labelling regions of the parameter space is to consider the degree of *stability* in the mappings. To see a certain language structure often implies that the structure is more stable than the ones we don't see.

Now we can rephrase the question. The parameter conditions that steer the system to compositional regions of the language space are best thought of as the parameter conditions that maximise the stability of compositional structure. It is interesting to note that the region of the parameter space for which we expect to observe

random mappings occur can also be subdivided in this way. For example, imagine the case when no transmission bottleneck is in place. If we injected a random mapping into the mind of one agent, then this mapping would be stable through the duration of the experiment. Because every subsequent agent would observe the whole language, no deviation from the original mapping would occur. Other parts of the region where we expect to see random mappings may be very unstable: a new random mapping is observed after each iteration. This point is this: different degrees of stability can occur irrespective of the degree of structure in the mapping. We are interested in the conditions for stable structured mappings, because these conditions may shed light on how and why linguistic structure emerges.

There are two routes to understanding the parameter space of the iterated learning model. We can run simulation models with various parameter combinations, or we can carry out an analytic exploration. For the remainder of this chapter, I will choose the analytic path. An analytic examination of the parameter space, when possible, offers several advantages over simulation modelling. Certain parameter combinations will prove computationally intractable to explore by simulation. In addition, the analytic path naturally yields a solid explanation. With a set of simulation results, a post hoc explanation is required that is often hard to formulate and can lack the rigour of analytic explanation. The first step in the analytic exploration of the parameter space will require abstracting the process of learning. I will now describe such an abstraction which I term *optimal generalisation*. Crucially, this abstraction can be related directly to twin concepts of hypothesis selection via MDL, and the space of all finite state unification transducers.

5.3 Optimal generalisation

The notion of optimal generalisation is a simple one. A learner capable of optimal generalisation can generalise correctly to all utterances in a language, and only utterances in the language, given the minimum amount of evidence. But what does the *minimum* amount of evidence mean? We can say that given less than the minimum amount of evidence, then generalisation to all utterances becomes impossible for an arbitrary compositional language, whatever bias the learner brings to the learning process. This definition is a first approximation to optimal generalisation; a more precise and general one will follow. This definition is also an empirical definition contingent on us knowing about all possible learning biases. In this section, I will define more precisely what I mean by the minimal degree of evidence. Furthermore, there is one part of puzzle missing: Is optimal generalisation applicable in

any practical sense? Is there some learning algorithm that represents the notion of optimal generalisation? I will show that optimal generalisation is precisely a degree of generalisation suggested by MDL hypothesis selection under certain conditions, and fortunately these conditions are precisely the ones we are interested in.

5.3.1 Definition

Recall that from the last Chapter that a compositional language is partially defined by the lookup table used to construct it. A lookup table relates each feature value with a single sub-signal, and therefore defines how the signal is built. A simplifying assumption is implicitly introduced when using a lookup table to define a compositional language. The possibility of synonymy is excluded, as the table can only define a single signal for a given meaning. As I noted above, the model I am developing discounts the presence of synonmy through the application of a production bottleneck, so this simplifying assumption is consistent with the model. Now, given the expectation of a compositional language, what degree of exposure to the language is required before the lookup table can be filled? The earliest point at which the lookup table could be constructed is when all feature values have been observed. Disregarding the details of the procedure for reconstructing the lookup table, this is the minimum degree of exposure before reconstruction is possible at all. Optimal generalisation is the ability to express all the meanings for which all the feature values have been observed. It is worth pointing out that optimal generalisation achieves the best degree of generalisation given the *available* data. This data will not necessarily lead to the derivation of the *whole* lookup-table; the data may be sparse, in which case optimal generalisation will achieve the highest degree of generalisation given the data available.

In a trivial sense, the strongest possible generalisation bias is one which results in some fixed compositional language L_C being induced, *whatever* the input. But any such scheme would be inconsistent with all observed languages other than L_C. I will consider only hypotheses that are consistent with the observed data. Compositional languages, for our purposes, are those where a feature value appearing in a meaning is associated with a single sub-signal. A lookup table relating every feature value to a sub-signal totally defines the compositional language. This is a slight simplification, as the lookup table does not define the order in which sub-signals are assembled. In the following discussion, I will assume that the ordering of the feature values in the meaning is reflected in the construction of the signal. To express a meaning, the sub-signal corresponding to each feature value in the meaning is located in

the lookup table. The signal is formed by concatenating these sub-signals. Now, if an insufficient body of evidence to build the whole lookup table is observed, expressivity will be sub-optimal. Some meanings that need to be expressed will be meanings which contain unobserved feature values: the entry in the lookup table will be missing. In short, optimal generalisation is the ability to express all meanings which are built from *observed* feature values.

For the purposes of a more formal definition, consider some compositional language, L_C, as a mapping between a meaning space $\mathcal{M}$ and a signal space $\mathcal{S}$. Within the meaning space $\mathcal{M}$, I will define $v_{x,y}$ as the xth value of the yth feature. In other words, each feature has an associated set of possible values; the expression $v_{x,y}$ serves to identify one of these values. The corresponding sub-signal, $s_{x,y}$, for this feature value is denoted by $\omega(v_{x,y})$. The function of a lookup table is to define the association between each feature value and each sub-signal; it therefore represents the function ω. In general, a learner observes a series of R meaning/signal pairs: $\{p_1, p_2, \ldots p_R\}$. Each pair, p_i, is made up of a meaning m_i and a signal s_i, that is, $p_i = \langle m_i, s_i \rangle$. We can now formally define optimal generalisation:

Definition 8 (Optimal Generalisation) *An arbitrary meaning $m_k \in \mathcal{M}$, where $m_k = (v_{a_1,1}, v_{a_2,2}, \ldots, v_{a_F,F})$, can be expressed providing each feature value $v_{a_j,j}$ has been observed at least once. More precisely, $\omega(v_{a_j,j})$ is defined providing that $v_{a_j,j}$ is observed in one of $\{p_1, p_2, \ldots p_R\}$.*

The defining characteristic of optimal generalisation is the assumption that the signal corresponding to each observed feature value can always be deduced. Optimal generalisation therefore serves as an upper bound on the degree of inductive bias over arbitrary compositional language. No algorithm exists that can construct the function ω with fewer examples. Optimal generalisation defines a theoretical possibility; a lower bound on the amount of exposure required before generalisation is at all possible. The possibility of realising optimal generalisation in any practical sense is an entirely different issue.

5.4 Preliminary abstractions

Recall the general question: When is linguistic structure most likely to emerge and remain stable? We can approach a preliminary understanding by examining the relationship between compositional languages and the resulting hypothesis structure induced by MDL. Then, by contrasting this hypothesis structure with that induced

for holistic language, a tentative claim can be made about the relative stability of compositional language over languages in *general*. Figure 5.2 illustrates this approach: an understanding of the whole parameter space is approached from an understanding of the parameters effecting the two extreme language types. The angle of attack is therefore to draw a coarse-grained map of the parameter space. If needed, this map can then be understood further through simulation.

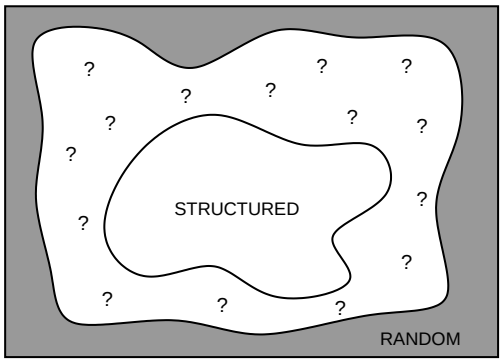


Figure 5.2: A simplified view of the parameter space. By comparing the stability of compositional language with holistic language, a tentative claim about the stability of compositional language with respect to *all* languages can be made.

The analysis has the following structure. Firstly, the relationship between language structure and induced hypotheses is outlined. MDL hypothesis selection completely defines this mapping between language space and hypothesis space. Secondly, a model of generalisation can be developed by examining the structure of the induced hypotheses. As result of these two observations, we can deduce a clear relationship between language type and resultant degree of expressivity. The major hurdle in making headway into such an understanding is an argument supporting the relationship between optimal generalisation and MDL induction for compositional languages. This argument rests on two assumptions: (1) meaning are observed with equal probability, and (2), a situation resembling the poverty of the stimulus is in place.

5.4.1 Relating language structure to induced hypothesis

I will consider two classes of language structure. These two classes occupy the extremes of the scale between structure and randomness. At one extreme is compositional language structure, and at the other is holistic language. The compositional languages occupy some subset $\mathcal{L}_{comp}$ of the language space $\mathcal{L}$. Similarly, holistic languages occupy a subspace denoted by $\mathcal{L}_{holistic}$. Hypothesis selection is the process that maps examples of language, utterances, to hypotheses that described these

examples. Faced with members of $\mathcal{L}_{comp}$, hypothesis selection via MDL will induce hypotheses of a certain structure. Of the set of all hypotheses, that is, all FSUTs denoted by $\mathcal{H}$, these induced hypotheses occupy a subspace denoted by $\mathcal{H}_{comp}$. Similarly, faced with holistic languages drawn from $\mathcal{L}_{holistic}$, MDL hypothesis selection will induce hypotheses occupying some subset of $\mathcal{H}$ denoted by $\mathcal{H}_{holistic}$. Figure 5.3 illustrates this relationship between the language space and the hypothesis space. Ultimately, the aim of this section is to deduce the extent to which members of $\mathcal{H}_{comp}$ and $\mathcal{H}_{holistic}$ generalise to novel utterances. That is, to what degree do the induced hypotheses explain more than that which was observed.

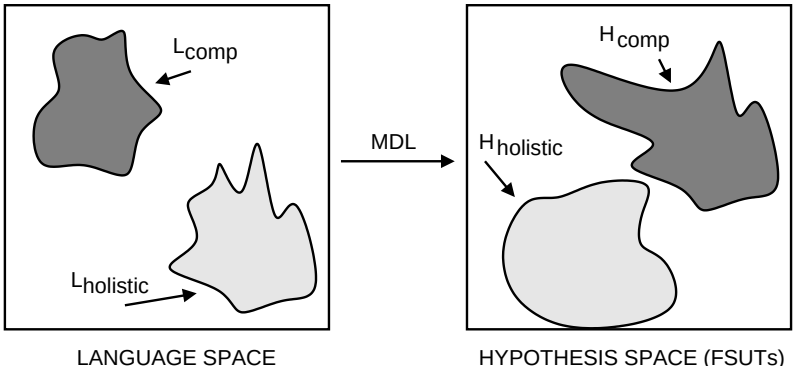


Figure 5.3: Hypothesis selection maps languages to hypotheses. Here, we only consider two language classes: $\mathcal{L}_{comp}$ and $\mathcal{L}_{holistic}$. The aim is to understand the structure of the corresponding induced hypotheses, $\mathcal{H}_{comp}$ and $\mathcal{H}_{holistic}$, respectively.

We can arrive at such an understanding by analysing the structure of the transducers that make up the sets $\mathcal{H}_{comp}$ and $\mathcal{H}_{holistic}$. In order to arrive at this understanding, we need to consider the smallest encoding lengths of transducers in $\mathcal{H}$ that are consistent with the bodies of data drawn from $\mathcal{L}_{comp}$ and $\mathcal{L}_{holistic}$.

Transducer structure and expressivity for compositional languages

Compositional languages represent a structured mapping between meanings and signals. This structure in the mapping permits compression. So, not surprisingly, the structure present in compositional languages results in structured transducers being induced. As I discussed in Chapter 4, the transducer structure depicted in Figure 5.4(a) represents, given certain assumptions about the distribution of utterances, the hypothesis with the smallest encoding length given the compositional language L_{comp}:

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

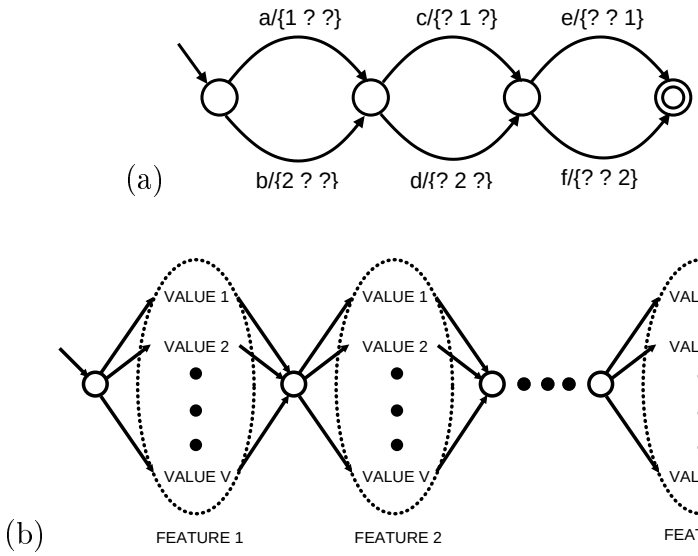


Figure 5.4: Maximally compressed transducers. (a) The smallest consistent transducer given the compositional language L_{comp}. (b) The general structure of a maximally compressed transducer. Each feature is coded as separate region of the transducer. A traversal through the transducer passes through each feature.

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

Smaller hypotheses exist, but they will fail to maintain the structure of the observed signals – they will lead to production decisions becoming inconsistent with the observed data. Recall from the last Chapter that one of the conditions of compression is that consistency with the observed data must be maintained. For this reason, further compression will never occur. The *general* transducer structure for a compositional languages, show in Figure 5.4(b), leads to optimal generalisation. This fact becomes clear when we note that this general structure corresponds directly to the notion of a lookup table. Recall that a lookup table is used to construct a compositional language; it defines, for each feature value, the sub-signal to use when constructing the signal. In one sense the lookup table acts as a hypothesis used to generate the language. Accordingly, this same hypothesis can be used to describe the language. The set of compressed transducers is therefore identical to the space

of lookup tables. In short, compressed transducers are the FSUT equivalent of a lookup table.

An important result, used throughout this chapter, is that the degree of generalisation achieved by compressed FSUTs is precisely that of optimal generalisation. Due to the equivalence between compressed FSUTs and lookup tables, those meanings that have signals defined by the lookup table are also expressible by the compressed FSUT. As we have seen, these meaning are those for which all feature values have been observed. The expressivity of a compressed FSUT is therefore a function of the number of feature values observed.

Transducer structure and expressivity for holistic languages

Holistic languages contain no structure in the mapping between meanings and signals. For this reason, they cannot be compressed[1]. Figure 5.5(a) depicts a hypothesis induced for the holistic language L_{comp}. This hypothesis is the prefix tree transducer, and represents the data verbatim. Figure 5.5(b) illustrates the general structure of a prefix tree transducer. A non-branching path through the transducer encodes each utterance. The result of this structure is that generalisation cannot occur: the transducer can only produce signals for meanings it has observed in conjunction with a signal – it acts as a basic memory.

In contrast to compressed transducers, where optimal generalisation is possible, prefix tree transducers can express only those meanings that have been observed. The expressivity of a holistic transducer is therefore linearly related to the number of distinct observations. The expressivity of compressed transducers, as function of the number of utterances, is super-linear.

5.4.2 Justification for the proposed MDL hypothesis selection

So far I have simplified the range of languages we are interested in. This simplification has a knock-on effect: the range hypotheses selected, given these languages, results in a simplification of our treatment of the hypothesis space. Importantly, the simplified treatment of the hypothesis space rests on two assumptions about the distribution of meanings and the amount of evidence available before hypothesis selection is carried out. Until now, I have given no justification for simplifying

[1]This is not strictly true. Compression can occur, but the result of compression will not lead to generalisation. At least, the probability of generalisation is small. For example, a randomly constructed holistic language could contain compositional structure by chance.

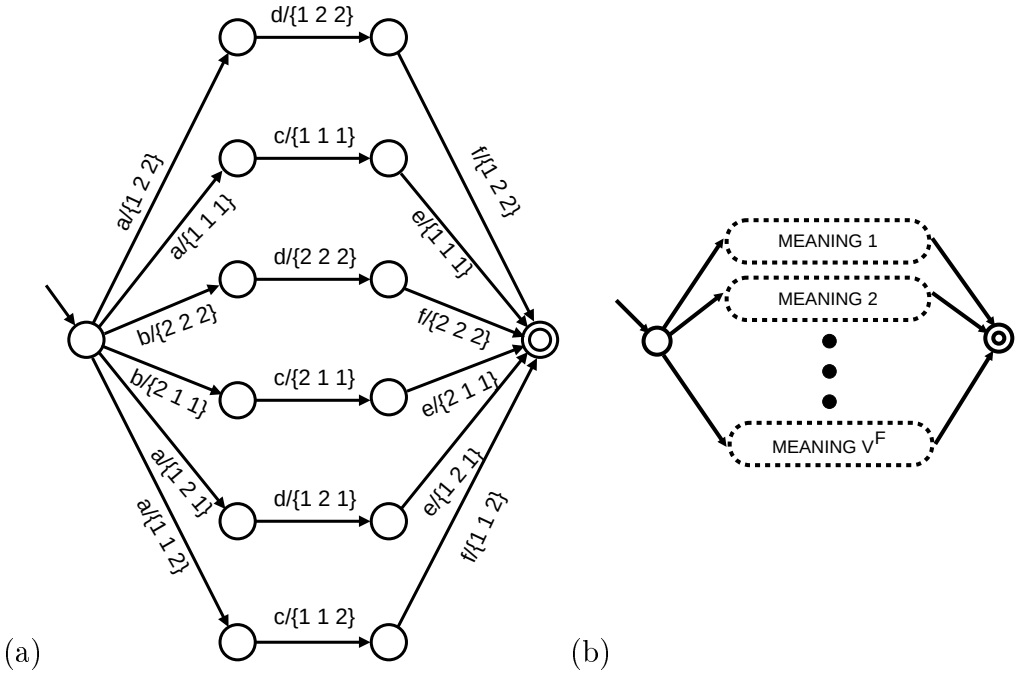


Figure 5.5: Prefix tree transducers. (a) The prefix tree transducer for the language L. Each distinct observation is represent by separate path through the transducer: the data is represented exactly as it was seen. (b) The general structure of a prefix tree transducer.

the hypothesis space in this way. This section provides such a justification. The impatient reader can skip to Section 5.5, and avoid the details of this justification without losing track of the central argument.

Recall that the process of MDL hypothesis selection seeks the smallest combined cost of, (1), encoding the hypothesis, and (2), encoding the data in terms of this hypothesis. The calculation therefore divides naturally into the grammar encoding length (GEL) and the data encoding length (DEL). The choice of transducer provides us with the GEL, and the structure of this transducer dictates the cost of encoding the data, and therefore the DEL. At this point, it is worth noting the precise manner in which the DEL is calculated. Each *individual* utterance needs to be represented in the encoding of the data. If an utterance is observed n times, then the calculation of the DEL must take into account the cost of specifying the path through the transducer, for that utterance, n times. This is why the relative frequency of utterance observation impacts on MDL selection.

GEL and DEL for compositional language

Given a compositional mapping between meanings and signals, the smallest possible GEL obtainable is that suggested by the general compressed transducer. Any further reduction in the GEL, through compression, will result in the transducer becoming inconsistent with the data. This conclusion is intuitive. A compressed transducer structure is such that each feature is encoded, and within the coding of each feature, each feature value is explicitly coded precisely once. This structure corresponds to that of a lookup table. If one doubts that compressed transducers are not of the minimum size, one must also doubt that there is a more efficient method of specifying a compositional language than that achieved using a lookup table. Given that the minimum GEL is achieved by a compressed transducer, the effect of the DEL will be important in selecting the MDL transducer. If we were to neglect this part of the calculation, our hypothesis selection procedure would reduce to Occam's razor – pick the smallest consistent hypothesis – and our argument would be complete.

I will now consider the data encoding length. In contrast to the GEL for compressed transducer, the DEL of a compressed transducer is not the smallest possible. Assuming every feature value for each feature has been observed then the cost of specifying a path through the transducer is $F \cdot log_2(V) = log_2(F^V)$. This fact follows by noting that each section of the transducer representing a feature contains V possible paths. Specifying this path requires $log_2(V)$ bits of information. As there are F such choices to made, the cost is $F \cdot log_2(V)$. Now, this cost of encoding an individual utterance is larger than the minimum because a compressed transducer also accounts for unseen utterances: induction explains more than what was observed, and these extra utterances require the number of possible routes through the transducer to increase. This fact is reflected in the increased cost of encoding the data. We will see, however, that this increase in DEL is in fact quite small with respect to the GEL.

GEL and DEL for holistic language

The GEL of a prefix tree transducer is, in contrast to that of the compressed transducer, the largest possible transducer considered. Each *distinct* observation – a meaning/signal pair – is represented by a unique path through the transducer. Any relationship with other observed meaning/signal pairs, however similar, is not represented or exploited: no compression is carried out. Multiple observations of the

same distinct pairing, however, *are* compressed: despite multiple observations of the meaning/signal occurring in the input data, the pair is represented by a single path in the transducer[2]. The GEL can only decrease as a result of compression. Because our notional holistic language is assumed to represent a random mapping, compression cannot occur. However, the DEL resulting from a prefix tree transducer is the smallest possible. Given that p distinct meaning/signal pairs have been observed, there will be p possible paths through the transducer. The cost of specifying a single route through the transducer is therefore $log_2(p)$ bits. Such a code is optimal; it leads to the smallest average path specification length, as each meaning is assumed to be equiprobable (Shannon & Weaver 1949:6-24). If the probability distribution over meanings is non-uniform, such that some meanings are highly probable, then the most efficient code, or Shannon-Fano code, would be a variable length code (Li & Vitányi 1997). Fortunately, given our assumption of a uniform probability distribution over meanings, the most efficient code is one where paths through the transducer are represented by bit-strings of the same length.

Towards a general result

To recap, an argument for relating language structure to transducer structure is required. This argument falls into two parts. First, it is clear that a randomly constructed holistic language can only ever lead to an uncompressed (prefix) transducer being induced. But can we be so sure about the fact that a compressed transducer, say T_c, will always selected by MDL given compositional input? That is, given a compositional language, under what circumstances would some other, hitherto unconsidered transducer structure, say $\hat{T}$, be preferred by the MDL selection process? For MDL to prefer $\hat{T}$ over T_c, we can say that two conditions must be met:

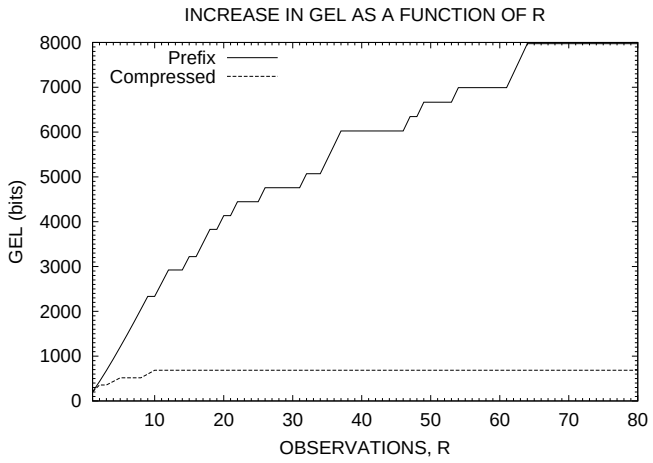
1. The alternative transducer, $\hat{T}$, must lead to an increase in the GEL of T_c by, say, b bits.
2. This increase in GEL, providing a decrease in degree generalisation occurs, must result in the DEL of $\hat{T}$ being smaller than that of T_c by at least b bits.

That is, due to our observation that compressed transducers have the minimum GEL, the alternative transducer $\hat{T}$ must, by definition, require more bits to encode. This increase in grammar encoding length must therefore be matched by a larger

[2]As the pairing is stored only once, repeated observations do not affect the GEL. These multiple observations *do* affect the DEL as ultimately the description of the data must describe the data perfectly. That is, the total DEL of distinct meaning/signal pair will be proportional to the number of times it was observed.

decrease in the cost encoding the data in terms of $\hat{T}$. Under our two assumptions, this situation cannot arise.

(a)



(b)

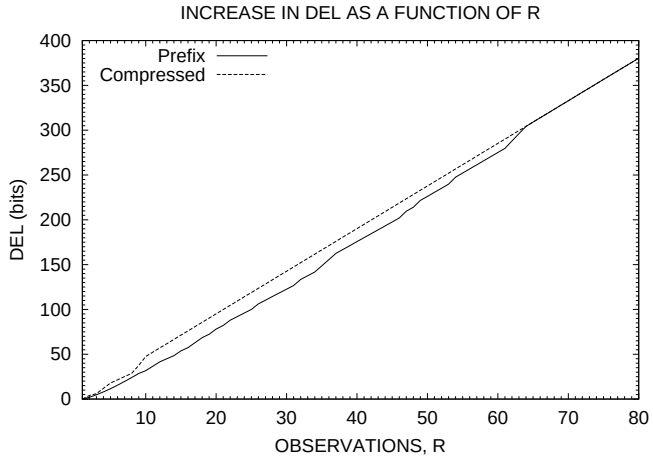


Figure 5.6: Encoding lengths, in bits, as a function of observations. (a) As R increases, so does the grammar encoding length of both prefix and compressed transducers. A large difference in encoding exists between the two transducer types. (b) The increase in data encoding length as a function of observations. These results were extracted from a simulation, described in more detail in the next Chapter.

First, note Figure 5.6(a-b). These graphs clarify the relationship between DEL and GEL for compressed versus prefix tree transducers. Here, the GEL and DEL are shown for compressed and prefix tree transducers as a function of the number of observations. The most striking fact is that the magnitude of the GEL is much larger than that of the DEL. In fact, there is little difference between the DEL of the compressed transducer and the DEL of the prefix transducer. This observation supports the view that compressed transducers offer the smallest encoding length: they have the minimum GEL, and have a DEL that is only just larger than the minimum DEL.

Continuing the discussion of the hypothetical transducer $\hat{T}$, note that for $\hat{T}$ to preferred over T_c, a small increase in the size of the transducer, b bits, must be accompanied by a larger decrease in the cost of encoding the data. Figure 5.6(a) hints at the fact that the magnitude of changes in the GEL will dwarf any changes in magnitude in the DEL. Given two assumptions, there are a number of grounds on which to argue that such a circumstance will always occur. These two assumptions are the following:

1. All meanings are equiprobable.
2. Not all meanings are observed.

The most immediate proof of the non-existence of the hypothetical $\hat{T}$ follows by noting that any increase in the grammar encoding length – which is essential if we are to investigate any other transducer than the compressed transducer – can be made arbitrarily large by increasing the size of the signals. The longer the signal, the larger the number of states required to encode the signal, and hence the larger the grammar encoding length. This increase in size of the signals, in contrast, does not effect the data encoding length at all. The data encoding length is reflection of the number of choices made when traversing the transducer. The signal length has no bearing on this branching factor. Therefore, *any* increase in the grammar encoding length can, assuming sufficiently large signals, be larger than any decrease in data encoding length. Such a set of circumstances are perfectly feasible. It would simply require, for a given meaning space, that signals have some minimum length. In practice, the restrictive conditions suggested by this proof are not required: large signals are not required for our simplified exploration of the hypothesis space. Later on, a more thorough treatment of the hypothesis space will support this observation. But for now, the argument given above serves the purpose of justifying our simplifications. How general is this result? It worth noting that this argument follows as result of the nature of the hypothesis space. If I had chosen some other model of language, then the result may not hold true.

5.4.3 Breaching the assumptions

It is important to consider just how dependent on our assumptions the previous result is. In this section I will consider probability distributions over meanings for which the above results do not hold. Recall that the difference between the DEL of a prefix transducer, DEL_{prefix}, and that of a compressed transducer, DEL_{comp}, is usually much smaller than the difference between the corresponding grammar

encoding lengths GEL_{prefix} and GEL_{comp}. As I have shown, this disparity is due to transducer compression resulting in the removal of many transducer states, but the degree to which this loss of states increases the cost of encoding the data is usually small.

Rather than compare the encoding length of T_c with $\hat{T}$, as we did above, I now compare T_c with T_h, a prefix tree transducer. Given a compositional language, for a prefix tree transducer to be preferred over a compressed transducer, the number of occurrences of each meaning/signal pair, $K(i)$, for $1 \leq i \leq p$, must be large. In other words, a large amount of data can lead to the data encoding length swamping any changes in the grammar encoding length. Assuming a uniform distribution over meanings, this situation will not occur unless each meaning signal/pair is observed many times. For this to happen, meaning space coverage, c, must approach 1, i.e., we expect the whole language to be observed. Notice that for this situation to occur, an infinite number of observations would be required. Obviously, this situation is not modelled here, as I assume there is a transmission bottleneck, and therefore a limited coverage of the language is expected.

It is interesting to reflect on why MDL would pick a prefix tree transducer in the face of such a large body of evidence. In terms of MDL, the justification for this relationship rests on the belief that the more evidence we have for a set of observations, the less likely novel observations are to occur. In the limit, MDL will disallow any induction: a huge amount of data containing little variation is precisely the grounds on which to expect no further variation.

Figure 5.7 illustrates how, for a compositional language, hypothesis selection depends on the probability distribution over meanings. First, given a transmission bottleneck ($c < 1.0$) and a uniform distribution over meanings, compressed machines are always chosen. If we define EL_{prefix} as $DEL_{prefix} + GEL_{prefix}$ and EL_{comp} as $DEL_{comp} + GEL_{comp}$, then $EL_{prefix} - EL_{comp} > 0$ holds irrespective of coverage. Contrast this with situation where we fix $K(i) = 100$ for $1 \leq i \leq p$, also shown in Figure 5.7. For low coverage values, the inequality $EL_{prefix} - EL_{comp} > 0$ no longer holds — prefix tree transducers are preferred for low coverage values. In practice, given a uniform distribution over meanings, this situation cannot occur unless $c \to 1.0$.

However, if we consider a distribution such as that resulting from Zipf's law, then this situation can occur when $c < 1.0$. A Zipfian distribution, in this context, would mean that the frequency of meanings decay as a power function of their

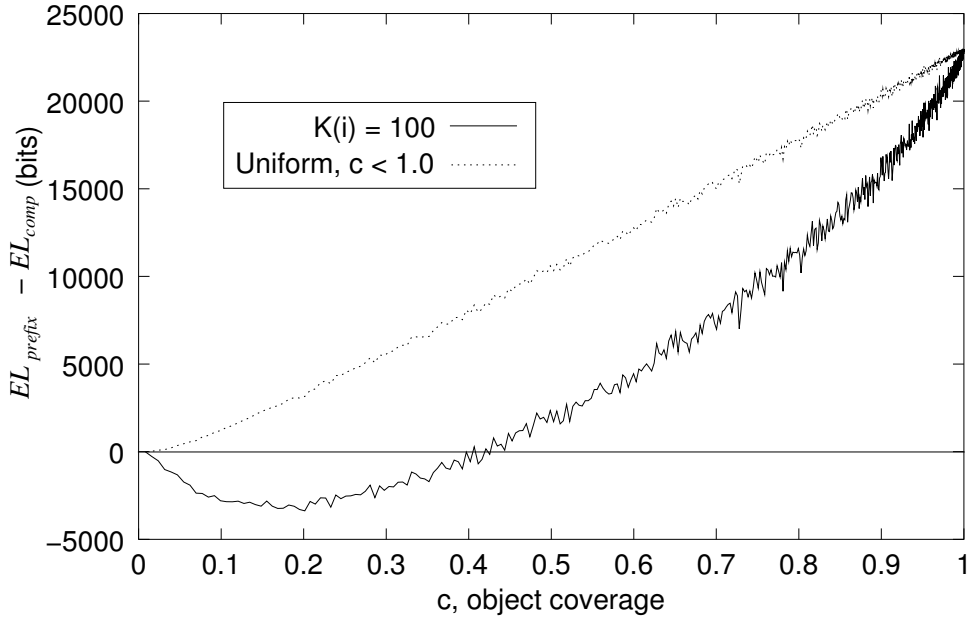


Figure 5.7: When $EL_{prefix} - EL_{comp} < 0$, prefix transducers are chosen by MDL for compositional language. With a uniform distribution and a transmission bottleneck, this does not occur. Only when, for example, we fix $K(i) = 100$ will prefix tree transducers be chosen over compressed machines.

rank (Zipf 1936; Kirby 2001). In this case, as c increases $K(i)$, for some values of i, will become huge. In this situation, the rule that compressed transducers are always selected for compositional language no longer holds. By introducing strong statistical effects into the data, hypothesis selection by MDL begins to diverge from that of Occam's Razor. The occurrence of communicatively relevant situations are unlikely to conform to a uniform distribution (Bullock & Todd 1999). The assumption of a uniform distribution made in this analysis therefore restricts the structure of languages we consider. The introduction of a non-uniform distribution would be an important extension to the model.

5.5 Conditions for language stability

The analytic route to understanding the ILM taken in this chapter will now begin to provide fundamental insights into the iterated learning model. By calculating the expected behaviour of agents over a wide, and previously intractable range of parameter combinations, general statements can be made about the emergence of structured languages in the iterated learning model. So far, I have detailed the

manner in which a simplified view of the language space, in conjunction with assumptions about the distribution of meanings in the meaning space, allows a simplified treatment of MDL hypothesis selection. In turn, by simplifying the hypothesis selection process, those parts of the hypothesis space in question have been identified and the resulting expressivity quantified. In this section, I use these results to develop mathematical and monte carlo models. These models allow us to get a grasp on the expected behaviour of the agents in the ILM over a wide range of parameter combinations. At this point, the key parameters of the iterated learning model used in the following analysis are worth noting:

- The number of features in the meaning space, F.
- The number of values per feature, V.
- The number of utterances on which the hypothesis selection process is based. This parameter represents the severity of the transmission bottleneck. It can either be measured in terms of the number of utterance observations, R, or the proportion of the meaning space we expect to be observed, termed *coverage*, and denoted by c.

Many of these results rely on drawing a parallel between the degree of generalisation achievable given a certain language structure and the resultant *stability* of that language structure. Recall that stability is, ultimately, the measurement we are interested in. Language stability tells us about the frequency of observation of particular language types within the context of iterated learning.

5.5.1 Relating transducer structure to expressivity

The relationship between learning data and the induced hypothesis structure has been tackled at a general level; given data that is assumed to contain a certain structure, the structure of the induced hypothesis, and the resulting degree of expressivity, has been established. In order to model this process explicitly, an algorithm is required such that data defined by the parameters F and V is observed subject to the restrictions imposed by the parameter c. For the two extreme language structures, we then require a measure of expressivity, E_{comp} for compositional languages and $E_{holistic}$ for holistic languages. That is, the expressivity resulting from the generalisation process will depend on the structure of the input language. Next, I will formalise this process, and present a mathematical model relating the parameters F, V and c, to the measures E_{comp} and $E_{holistic}$. In short, the argument so far needs to be formalised and investigated.

Calculating E_{comp}

The learning task begins with a learner observing a series of R meaning/signal pairs $(p_1, p_2, p_3, \ldots, p_R)$. Each of these meaning/signal pairs are drawn from the set $\{p_i : p_i \in \mathcal{M} \times \mathcal{S}\}$. The observed set of meaning/signal pairs is denoted as $\mathcal{A}$. Notice that $|\mathcal{A}| \leq R$ is highly likely to hold true, as observations of the language are random with replacement, and therefore some pairs will be observed more than once. For this reason, even though the parameter R defines the number of observations, we will frequently measure the amount of evidence on which the learning process is based in terms of meaning space coverage c. The c value occupies the range $0 \leq c \leq 1$, and represents the proportion of the meaning space observed. As meaning/signal pairs are observed at random and with replacement, the parameter R tells us little about the diversity, or coverage, of the observed data. Now, our objective, given this problem, is to find the number of meanings in $\mathcal{M}$ that can be expressed: that is, the number of meanings for which a signal can be constructed. The *minimum* number of expressible meanings will be $c \cdot |\mathcal{M}|$ as this is the number of meanings observed in conjunction with a signal. As a result of generalisation, the learner will be able to express more than this minimum number. The resulting expressivity will depend on generalisation, and therefore on the structure of the language observed.

First I will describe how the set of expressible meanings can be calculated, given a compositional language. Recall that the degree of expressivity, given a compositional language, is that defined by optimal generalisation. The degree of expressivity is therefore calculated by constructing a lookup table detailing how each feature value is to be expressed. This table is simply a $F \times V$ matrix denoted as O. As each meaning/signal pair a_i is observed, F elements of the O matrix can be filled in. Assuming a maximally favourable series of observations, the matrix can be completely filled with as little as V observations. However, this situation is unlikely. In determining the degree of expressivity, rather than being concerned with building signals, we will be concerned more with simply determining if they can be built. The actual signals are of no consequence. Hence, some entry $O_{i,j}$ in the matrix will either true or false:

$$O_{i,j} = \begin{cases} \texttt{true} & : \text{ if the jth value for the ith feature is observed} \\ \texttt{false} & : \text{ otherwise} \end{cases} \tag{5.1}$$

When the entry $O_{i,j}$ is true, the sub-signal for the jth value of the ith feature is known. When $O_{i,j}$ is false, we can say that no meaning/signal pair has been observed that had the jth feature value present in the ith feature. On receiving some observation $a_i = \langle m, s \rangle$ the matrix is updated as follows. The feature values contained in the observed meaning are logged as present in the O matrix. More formally, if $m = (v_1, v_2, \ldots, v_F)$, then the O matrix is updated as follows.

$$O_{i,v_i} = \texttt{true} \text{ for i} = 1 \text{ to F} \tag{5.2}$$

To determine if some arbitrary meaning $m = (v_1, v_2, \ldots, v_F)$ can be expressed, the O matrix is interrogated; the O matrix defines the set of expressible meanings, denoted as M_{comp}. More formally, members of M_{comp} have the following property:

$$M_{comp} = \{m \in \mathcal{M} : \bigwedge_{i=1}^{F} O_{i,v_i}\} \tag{5.3}$$

Notice that the criterion $\bigwedge_{i=1}^{F} O_{i,v_i}$ will always be true for a meaning that has been observed, and therefore logged using the update rule. But induction is made possible because this expressivity criterion also permits other meanings, which have not been observed, to be expressed providing the appropriate sub-signals are known. The cardinality of M_{comp}, $|M_{comp}|$, defines the expressivity achieved by the learner. That is,

$$E_{comp} = |M_{comp}| \tag{5.4}$$

Calculating $E_{holistic}$

Next, the set $M_{holistic}$ needs to be constructed. This task is substantially simpler. Rather than constructing an O matrix, the observed meanings are just stored along with their associated signals. Only meanings which have been observed, and therefore stored, will be expressible.

$$M_{holistic} = \{m \in \mathcal{M} : \langle m, s \rangle \in \mathcal{A}\} \tag{5.5}$$

This result implies:

$$E_{holistic} = |M_{holistic}| = c \cdot \mathcal{M} \tag{5.6}$$

To recap, given a series of R observations of meaning/signal pairs, two sets of meanings, M_{comp} and $M_{holistic}$, are constructed. M_{comp} is the set of meanings that can be expressed assuming the language is compositional; it is constructed using the O matrix in conjunction with the decision procedure defined in Equation 5.3. $M_{holistic}$ is the set of meanings that can be expressed assuming that the meaning/signal pairs are drawn from a holistic language. Recall that the aim of this model is to relate the parameters F, V, and R (or c) to the resulting degree of expressivity E_{comp} and $E_{holistic}$.

Expected behaviour: A probabilistic model

The meaning/signal pairs that make up the input to the learning process are drawn at random from the space of possible meanings. In the case of a compositional language, the signals are constructed for the random meanings using a lookup table. For a holistic language, the signals are constructed at random. As we have seen, however, the signals are of no consequence in this model. The task is split into two: processing compositional language and processing holistic language. By employing a separate model of learning for each case, the signal structure is assumed in the model, and therefore does not need to be explicitly modelled. Because the input to the learning process is drawn at random, the resulting expressivity values will differ across specific instances of iterated learning. To gain an understanding of the expected behaviour of the models, I will present a probabilistic model from which expected values can be calculated.

Holistic language Let $L_{holistic}$ be a holistic language where every meaning is drawn from a meaning space with F features and V values per feature. In total, there will be V^F possible meanings, denoted by M. Given a holistic system, the number of features and values are not of any consequence; instead the value M, derived from F and V, proves relevant. Given that there are M distinct meanings, and R observations with replacement of these meanings, then the probability that some meaning m will be a member of the observed set of meanings O_m is the following:

$$\Pr(m \in O_m) = 1 - (1 - \frac{1}{M})^R$$

The number of observed meanings, and therefore, the expected expressivity achieved, $\overline{E}_{holistic}$, is this probability multiplied by the total number of meanings M:

$$\overline{E}_{holistic} = \Pr(m \in O_m) \cdot M \tag{5.7}$$

Notice that, because we are sampling with replacement, then the inequality $\overline{E}_{holistic} \leq R$ always holds.

Compositional language Now, for a compositional language L_{comp}, a similar analysis follows but now F and V are relevant. Take an arbitrary feature, and note the V possible values. Let O_v be the set of these values actually observed after R meaning/signal pair observations. After each meaning observation, for this single feature value, there is a $\frac{1}{V}$ chance of it being observed. For an arbitrary feature value, the probability that this value is observed after R meaning observations is therefore:

$$\Pr(v_i \in O_v) = 1 - (1 - \frac{1}{V})^R$$

Using this observation, the number of feature values we expect to observe will be:

$$\overline{V}_{obs} = \Pr(v_i \in O_v) \cdot V$$

The number of meanings that can be expressed, assuming the model of learning discussed above, is precisely the number of combinations of picking a value for each feature in the meaning. We know, from the above equation, that for each feature we have observed V_{obs} different values. The total number of distinct meanings we can express will therefore be:

$$\overline{E}_{comp} = (\overline{V}_{obs})^F \tag{5.8}$$

Another way of thinking about this equation is as follows. Using the O matrix, how many different meanings can we construct? Because this mathematical model deals in expected values, the expected number of feature values observed for each feature is identical. Contrast this observation with a particular O matrix; the number

of feature values for each feature will not necessarily be the same. The preceding analysis provides us with the expected values of the general model developed above. The basic model relates F, V and R to E_{comp} and $E_{holistic}$, and the probabilistic analysis relates F, V, and R to $\overline{E}_{comp}$ and $\overline{E}_{holistic}$. In the discussion that follows, references to expressivity values will refer to the expected values $\overline{E}_{comp}$ and $\overline{E}_{holistic}$, rather than values of individual runs.

Parameters affecting expressivity

The model developed above reveals some fundamental insights into how expressivity increases as a function of the number of observations. That is, our choice of F and V are strong determinants of the degree of expressivity given R observations. First, Figure 5.8 illustrates the rate at which expressivity increases as a function of the number of observations, both in the case of compositional language and holistic language. In this example, a meaning space defined by $F = 3$ and $V = 3$ is used. As one would expect, through learning, the expressivity achieved with holistic language, $E_{holistic}$, as input increases asymptotically to 27. That is, there are 27 possible meanings, and the number of observations before we expect to see all of them is infinitely large. After 150 observations, as Figure 5.8 suggests, the expressivity approaches 27. To guarantee an expressivity of 27 meanings, however, would require an infinite amount of evidence. For a compositional language, the relationship between expressivity and the number of observations differs substantially. The maximum expressivity of 27 meanings is approached after 20 observations; the rate at which expressivity increases is significantly faster in the case of compositional input.

The rate of increase in expressivity is dependent on the language type (compositional or holistic). The structure in the compressed language allows generalisation. Importantly, the rate of increase in expressivity can differ substantially *between* compositional languages too. To illustrate this point, I will consider two meaning spaces, both of which contain 256 meanings; $\mathcal{M}_1$ is constructed using 8 features with each feature taking 2 values ($F = 8, V = 2$), and $\mathcal{M}_2$ is built using 2 features with 16 values per feature ($F = 2, V = 16$).

The rate of growth of expressivity as function of observations for $\mathcal{M}_1$ and $\mathcal{M}_2$ is significantly different. Figure 5.9(a-b) illustrates this difference. Using 8 features leads to maximum expressivity occurring before 20 observations. In contrast, somewhere in the region of 140 observations are required when only 2 features are used. Given the problem of representing 256 meanings, we have a choice between using

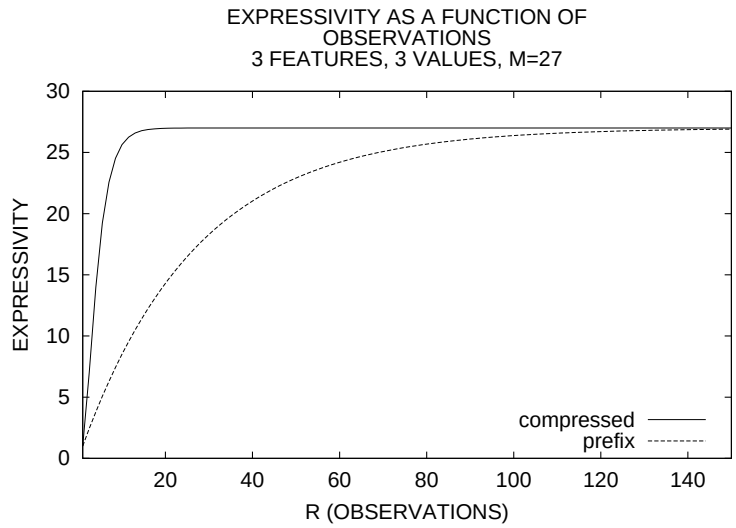


Figure 5.8: Expressivity growth as a function of observations for compressed and holistic languages. Maximum expressivity is achieved with fewer observations given a compositional language.

fewer features and more feature values, or more features and fewer feature values. There is a trade-off. The meaning spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ represent the two extremes in describing 256 meanings. Of course, a single feature with 256 values could be used, but this situation does not permit compositional structure: each meaning is discriminated purely on the basis of a single value. These results show that preferring more features and fewer feature values leads to a faster growth in expressivity. This phenomenon can be explained by noting that with each observation of a meaning/signal pair, F feature values are observed. As a result, preferring the position in the trade-off reflecting the case that more features are preferred to more feature values will result in the O matrix filling at a faster rate. In short, the feature values are observed at a faster rate when a rich feature structure is used.

These expected values were computed using the mathematical model described above. As observations are drawn at random, deviation from these expected values is inevitable. Figure 5.10 depicts two scatter plots detailing the results of 20 individual runs of the simulation described at the beginning of this section (Equations 5.1 to 5.8). Before maximum expressivity is consistently reached, a window of variation occurs. For $\mathcal{M}_1$, the window of variation is narrow, it occurs in the first 20 observations. The degree of variation within this period is great, because a single missing value will have a large impact on the resulting degree of expressivity. There are only 16 feature values (two for each of the 8 features). Contrast this observation with that of $\mathcal{M}_2$, which has 32 feature values. The degree of variation

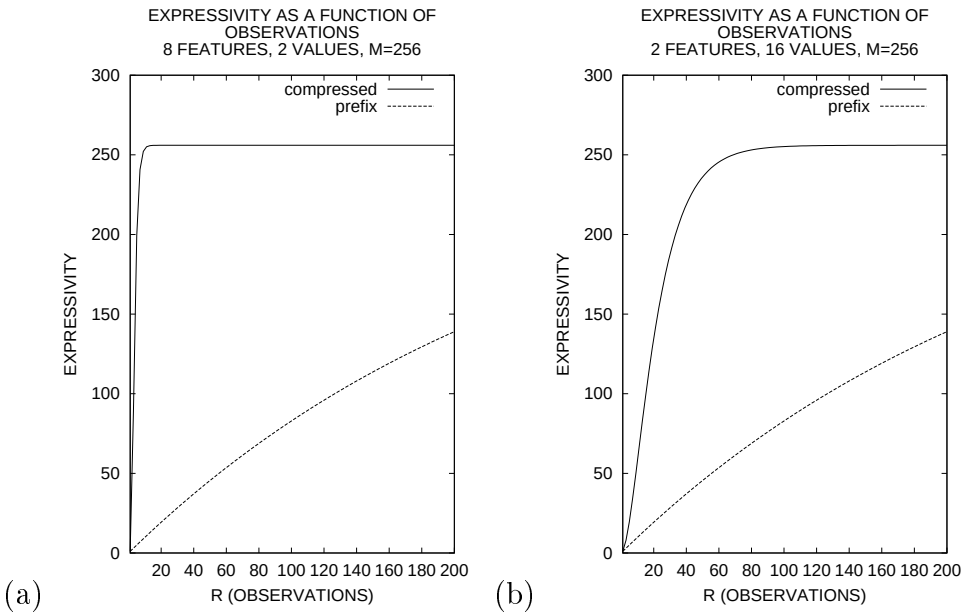


Figure 5.9: Expressivity growth as a function of observations for compositional and holistic languages. In representing 256 meanings, there are several choices in meaning space structure. In (a), the meaning space contains the maximum number of features, and as a result, the expressivity growth is faster than that observed for a meaning space using fewer features, shown in (b).

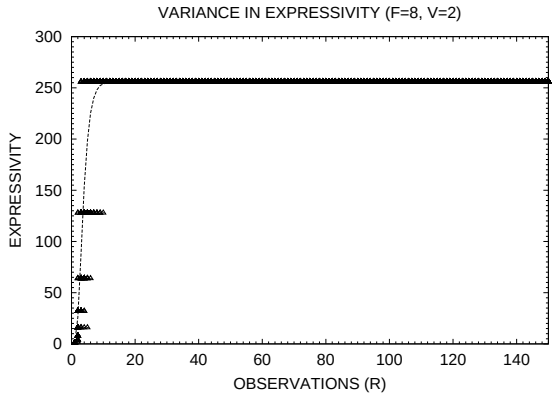
is less extreme, but occurs over a longer period. After 140 observations maximum expressivity is consistently achieved.

5.5.2 Relating expressivity to stability

Expressivity growth in itself can tell us little about the behaviour of an iterated learning model. But clearly expressivity does play a pivotal role in linguistic evolution in the ILM. If meanings can always be expressed in some principled way, then randomness cannot enter the system. To draw an analogy with simulated annealing, the temperature of the evolutionary system can never increase when expressivity is high. Expressivity is therefore related to stability. If meanings cannot be expressed, then invention of some form is required. With a necessarily random component, invention must make the system less stable due to the introduction of random, unstructured regions of the mapping between meanings and signals.

Between two agents, the stability of the mapping between meanings and signals is the degree to which these agents agree on the mapping between meanings and

(a)



(b)

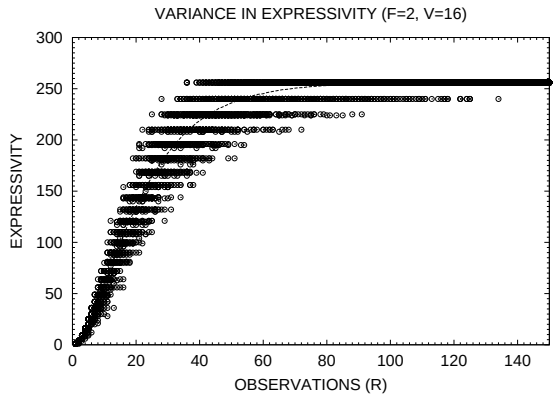


Figure 5.10: Variation in expressivity achieved, for 20 independent runs, as function of observations. For the same meaning spaces shown in Figure 5.9, we see variation in expressivity occurring over periods of contrasting length. In (a), where the maximum number of features are used to differentiate 256 meanings, high variation occurs over a small period. In (b), where the minimum number of features are used, less variation occurs, but for a longer period.

signals. Stability, in this static analysis, must therefore be proportional to expressivity. In the following analysis I make this assumption in defining the *pairwise stability* of compositional language, $S_{compressed}$, and holistic language, $S_{holistic}$:

$$S_{compressed} \propto E_{compressed} \tag{5.9}$$

$$S_{holistic} \propto E_{holistic} \tag{5.10}$$

Pairwise stability differs from the property of stability used elsewhere. Two subsequent states may lead to the same, or similar, hypothesis being induced. But this may be a temporary state of affairs precipitated by a fortunate series of random observations. Any rigorous treatment of stability will depend on the dynamics of the

system. Here, I will carry out a static analysis. Given this restrictive perspective, the concept of pairwise stability is justifiable. It should be noted, however, that properties of pairwise stability will prove to translate perfectly well into a dynamic analysis.

To recap, I have equated the degree to which a language confers expressivity when learned, to the degree to which the language is stable over generations. As each agent is physically identical, and agents differ only by virtue of the observations they experience, assuming that the degree of expressivity we predict is achievable consistently and irrespective of the distribution of utterances, then stability must indeed be highly correlated with expressivity.

Now the key question can be asked. Under what conditions is compositional language more stable than holistic language? This question relates to the probability of some language L passing through an agent unscathed to the next generation. In the presence of a transmission bottleneck, a holistic language is highly unlikely to survive intact. In general, there are two cases when a holistic language can survive:

1. When two subsequent agents are called to express the same set of meanings. This situation is extremely unlikely as it requires two random samples of the meaning space to be identical.
2. By chance, random invention maintains the mapping between meanings and signals. Similarly, this situation is possible, but highly improbable.

So both these cases are highly unlikely. Rather, at each iteration a holistic language L_h will change. Only $c \cdot M$ of a possible M meanings in the mapping between meanings and signals will persist at each generation. The further introduction of meaning/signal pairs, via invention, will lead to a new language L'_h being generated by the next generation, where, on average, $c \cdot M$ of the pairs in L_h will also exist in L'_h. In contrast, depending on c, and the degree of feature structure in the meaning space, a compositional language can survive from one generation to the next. That is, through generalisation, the mapping between meanings and signals will persist.

I will term the degree to which compositional language is more stable than holistic language as the *relative stability* of compositional language over holistic language, denoted by S. S is defined as:

$$S = \frac{S_{compressed}}{S_{compressed} + S_{holistic}} \tag{5.11}$$

This expression tells us how much larger $S_{compressed}$ is than $S_{holistic}$. In turn, relative stability measures the degree to which compositional language, as a result of generalisation, results in higher expressivity than holistic language. Note that, assuming a holistic language can never lead to a higher degree of expressivity than that of a compositional language, the inequality $0.5 \leq S \leq 1.0$ holds.

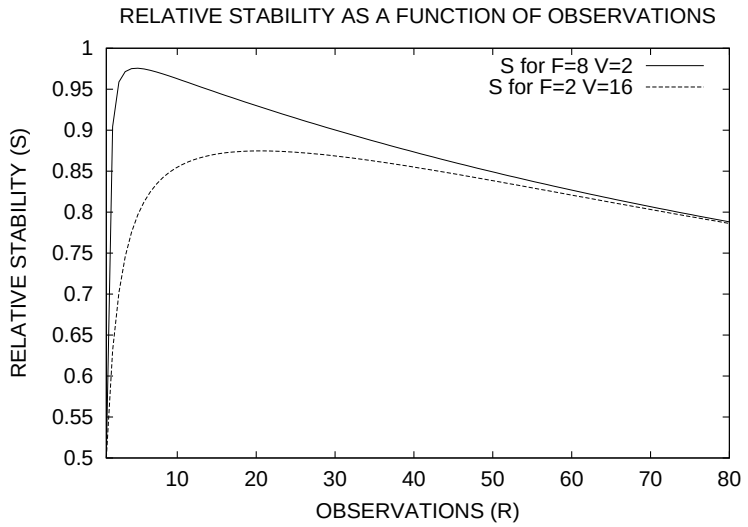


Figure 5.11: Relative stability of compositional language over holistic language for the two meaning spaces $\mathcal{M}_1$ and $\mathcal{M}_2$.

Figure 5.11 plots S (relative stability) as a function of R (number of observations) for the meaning spaces $\mathcal{M}_1$ and $\mathcal{M}_2$, each capable of coding 256 meanings. First we see that S is largest for small R, which translates into low meaning space coverage (c). This is when compositionality begins to allow generalisation from few observations. These few observations result in low expressivity for holistic language. As a result, compositional language is far more likely to be stable at low meaning space coverage, where the bottleneck is tight. As more observations become available to the learner, holistic language asymptotically approaches the degree of expressivity conferred through generalisation. That is, as $R \to \infty$, then $S \to 0.5$, and compositional language structure fails to offer any advantage. In the limit, the whole language is observed, and instability cannot occur.

Relative stability allows us understand the point at which the stability advantage gained through generalisation is maximal. If we consider language stability as an issue of survival, with languages competing to be transmitted, then low c and high meaning space structure provides the the most favourable conditions for compositional language to survive: they lead to high expressivity at precisely the time that

holistic languages lead to low expressivity. This relationship is explored further in the next section.

5.5.3 Mapping the parameter space

At this point it is worth stepping back from the above experiments to reconsider the assumptions, and question how much has been learned. The preceding experiments assume that every meaning in the meaning space represents a communicatively relevant situation. An agent has an internal meaning space which is used to label events occurring in an external environment. In a very general sense, this labelling can be related to the issue of categorical perception. The environment prompts the agent with communicatively relevant situations which are represented internally by the agent as structured representations. These representations are then externalised as signals. In one sense, the number of features and the number of values per feature define the problem by defining the structures that need to be signalled. An unfortunate side-effect of this assumption is that our choice of F and V determine the problem to be solved. Yet, at the same time, we seek to vary F and V to determine their relationship with other parameters. The two meaning spaces $\mathcal{M}_1$ and $\mathcal{M}_2$ defined earlier were chosen precisely because they contain the same number meanings, and therefore permit direct comparison; they solve the same problem, that of representing 256 events. Finding combinations of F and V such that V^F is constant is restrictive.

For this reason, an extension to the model will be presented in which a subtle change to the notion of the environment allows a valid comparison between any combination of F and V. So far, two issues have been rolled into one. I will now separate them.

Objects

At a general level, agents must solve the problem of communicating a signal for some set of N objects $\{\omega_1, \omega_2, \ldots, \omega_N\}$, and these objects are labelled with a meaning drawn from a meaning space defined by F and V. This situation is depicted in Figure 5.12. Previously, I have only considered the case when $V^F = N$. In the following analysis, I will breach this restriction. In doing so, the relationship between meaning space structure, the number observations on which learning is based, and the relative stability of the language types, will be further understood. By teasing apart these two issues, the environment in which the agent operates

becomes clearer. Agents are presented with an environment containing N objects. These objects represent communicatively relevant situations: they define the space of possible linguistic events. Each of these objects is labelled with meaning drawn from a meaning space defined by F and V. Depending on the values of F and V, the meaning space will contain M possible meanings ($M = V^F$). Notice that if $M > N$ then some of the meanings will not be used – they will never be used to label the objects in the environment. Now consider the other extreme, where $M < N$. Here, some of the objects will have the same meaning, the upshot being that a competence in expressing one object will enable the agent to express another. The fact that fewer objects can be discriminated is not important.

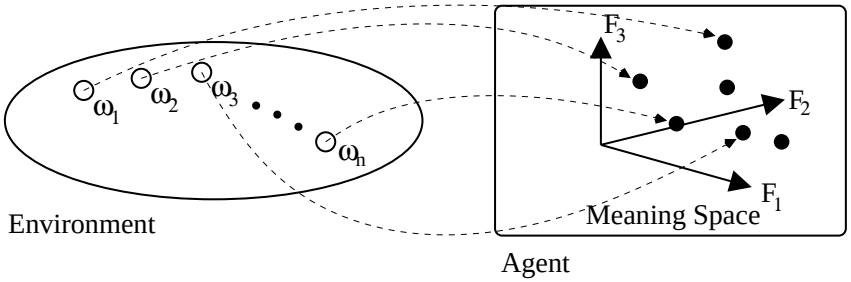


Figure 5.12: Agents are required to conceptualise a set of n objects in the environment. These objects map to an internal meaning space. This mapping is pre-determined – the agents play no role in constructing it. Furthermore, this mapping does change over time, and each agent in the model is subject to the same mapping.

Each agent conceptualises the objects in the environment as meanings. The meaning space should therefore be regarded as a way of constructing labels for these objects. It is wrong infer that this relationship models the act of perception, although it is a useful to imagine the relationship between agent and environment as follows: The meanings are internal to the agent, and are used to represent the objects existing in the environment. Neither the environment nor the meanings used to label objects in the environment change over time, and each agent is subject to the same mapping between objects and meanings. Meanings are monotonically consistent labels for objects. Signals are non-monotonic and evolving labels for meanings. By introducing an environment, F and V can be varied without restriction. Armed with this ability, along with the understanding of the parameter space already developed, we are now in a position to map the parameter space more thoroughly.

Monte Carlo simulation

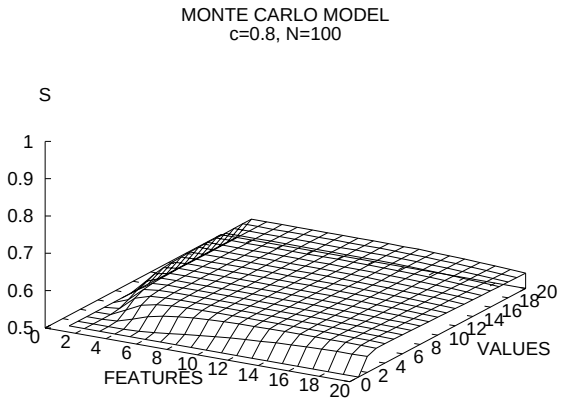
In an environment, the n objects are first assigned a meaning drawn randomly from the meaning space. The mapping between objects and meanings remains true for the duration of the experiment. The transmission bottleneck is then applied with respect to the objects, rather than directly to points in the meaning space. That is, an agent arrives at its knowledge of language after R observations where each observation is an observation of a single object. As before, observations are drawn at random – objects are equiprobable. On the basis of these observations, the method described earlier (Equations 5.1 to 5.8) is employed to calculate the expressivity achieved for the two language types. The results presented below reflect the average over 1000 independent runs. The introduction of the environment therefore only changes the manner in which meanings are observed. How the learner uses these meanings is precisely as described above.

In the following analysis, the advantage of introducing an environment becomes apparent when we consider that varying meaning spaces can be used to solve the same problem. The four surfaces shown in Figure 5.13 and Figure 5.14 cover a significant part of the parameter space. For a given bottleneck size, each surface shows the relative stability (S) of compositionality for different meaning spaces given 100 objects. In Figure 5.13(a) a relatively wide bottleneck represented by a coverage of $c = 0.8$ is imposed. Here, 80% of the objects are expected to be observed. This translates into 160 observations. The S values are low for all meaning spaces. In Figures 5.14(a-b) the surfaces illustrate how, for certain meaning spaces, the relative stability increases significantly as the bottleneck is tightened. Notice that relative stability is at a maximum value $S \approx 0.91$ when $c = 0.1$ (11 observations), but only for certain combinations of F and V.

The parameters F and V are critical in determining S. As we saw earlier, every observation made by a learner leads to the acquisition of F feature values, as every object has F components in its meaning label. The larger F is, the faster the rate at which feature values are observed. Furthermore, the larger F is, the smaller V needs to be in order to represent a given number of objects. For this reason, features with binary values will always lead to higher R than any other combination of F and V.

Achieving a high S value also requires a tight bottleneck. With fewer observations, a compositional language can be generalised from, and therefore outstrip the expressivity achieved given a holistic language. In short, for a holistic language, fewer

(a)



(b)

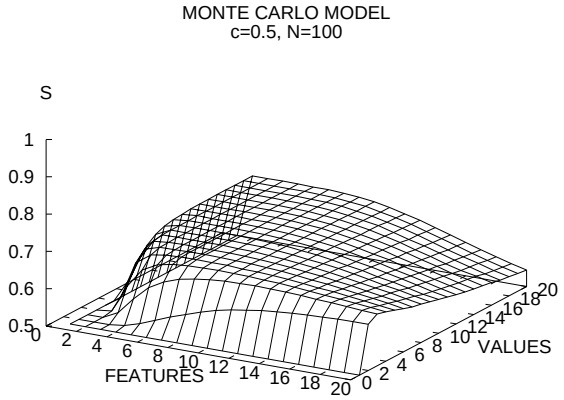
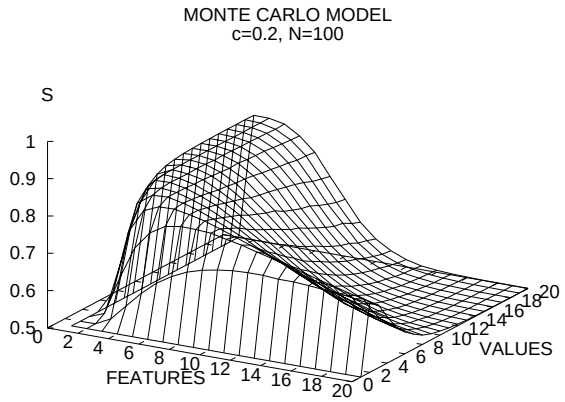


Figure 5.13: Relative stability of compositional language over holistic language. For 400 meaning spaces, (a) shows S for a bottleneck coverage of 0.8, and (b) shows S for a bottleneck coverage of 0.5. Low relative stability is a feature of wide bottlenecks.

observations means that fewer objects can be expressed. This is not necessarily the case with a compositional language: few observations can lead to high expressivity through generalisation. It is in precisely this region of parameter space that compositional language is maximally stable in comparison to holistic language.

With few features and few values, S is low. In the limit, we can label each object with meanings containing one feature. In this situation, meanings can only be differentiated on the basis of the feature values: the structure in the meaning labels reduces to an unstructured set. Compositionality is therefore not possible and we attain $S \approx 0.5$. However, as the feature complexity increases, so does S, due to generalisation becoming possible. This increase does not continue indefinitely. Both high F and high V lead to a decrease in S. The reason for this is simple. Such

(a)



(b)

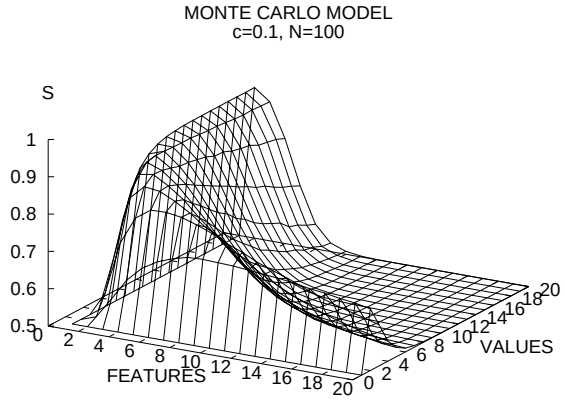


Figure 5.14: Relative stability of compositional language over holistic language. For 400 meaning spaces, (a) shows S for a bottleneck coverage of 0.8, and (b) shows S for a bottleneck coverage of 0.5. High relative stability occurs as the bottleneck is tightened.

parameter combinations create vast meaning spaces. If only 100 objects need to be labelled, and the meaning labels are drawn from a space containing 20^{20} meanings, then the probability that observed objects will be labelled such that they share feature values becomes very small. Quite simply, compositionality cannot function if the co-occurrence of meaning components is negligible.

Expressivity

I have identified the regions of parameter space that lead to high relative stability. Using this information, we would like to conjecture that it is precisely these regions in which compositional language is most likely to emerge and remain stable. The

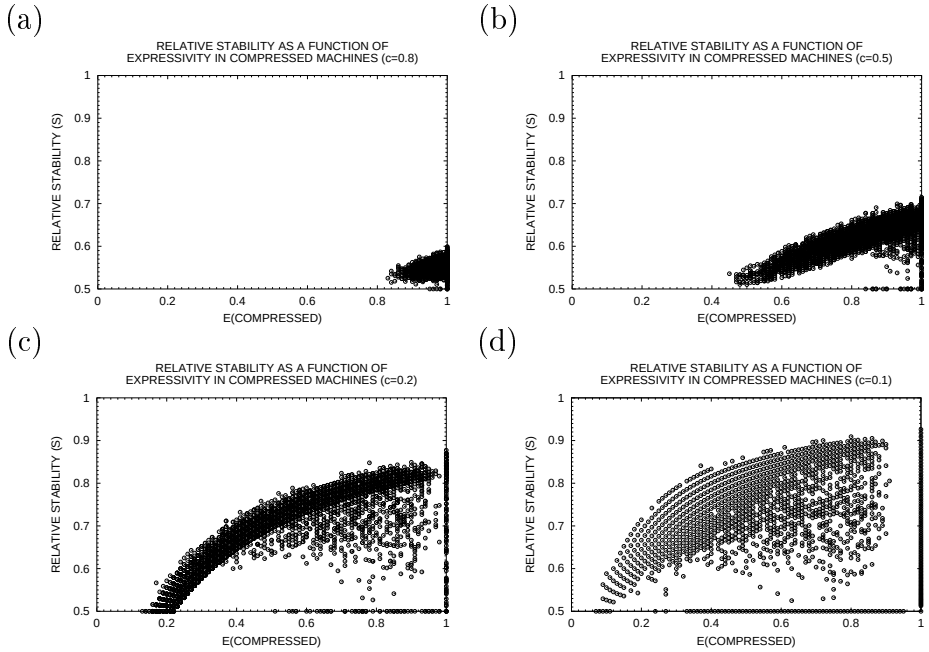


Figure 5.15: For 1000 independent runs of the model, these plots depict the correlation between expressivity of the compressed transducer and the relative stability of the compositional language over holistic language. For sparse language exposure – a tight bottleneck – high expressivity correlates with high relative stability.

pairwise stability argument, ideally, should carry over to stability in the dynamic sense.

Before this argument can be developed, we need to reconsider these results in light of the following possibility. High relative stability does not necessarily result in stability in a meaningful way if the absolute expressivity is low. A compositional language may lead to maximally improved expressivity over a holistic language, but unless that expressivity is high, then it is hard to argue for stability.

For four different bottleneck sizes, Figure 5.15(a-d) depicts the relationship between the expressivity achieved from compositional input and relative stability. To strengthen our interpretation of the results presented in Figure 5.13 and Figure 5.14, we would like high expressivity for compositional language to be related with high relative stability. That is, when compositional language confers increased expressivity in relation to holistic language, the degree of this increased expressivity should be at a maximum. For compositional language, if high S occurs with less than maximum expressivity, then maximum stability for compositional language will not occur when relative stability is at a maximum.

The relationships depicted in Figure 5.15(a-d) represent each of the 400 meaning spaces analysed in Figures 5.13 and 5.14. Notice that that the expressivity derived from compositional language is scaled to occupy the range $0 \leq E_{comp} \leq 1$. Two patterns need to be explained. First, data points sitting on the x-axis ($S = 0.5$) correspond to the case when the number of features is 1. Here, maximum expressivity can be achieved, but the language cannot have any structure. As a result, relative stability can never be greater than 0.5. Second, the data points sitting on the y-axis ($E_{comp} = 1.0$) correspond to maximum expressivity when $V = 2$.

In each plot, the highest S values corresponds to situation when $E_{comp} = 1.0$. This observation indicates that maximum expressivity does indeed relate to maximum relative stability, for certain meaning spaces. The remainder of the scatter plots illustrate a general trend. For a tight bottleneck, which corresponds to values of $c = 0.1$ and $c = 0.2$, there is a clear relation between high expressivity and high relative stability. For a wide bottleneck, which corresponds to values of $c = 0.5$ and $c = 0.8$, this relationship is less clear. High relative stability, S, cannot occur for these bottleneck values. In this situation, stability is a feature of a wider range of language structures.

5.5.4 Summary of results

Two models have been developed. As a function of the number of utterance observations, the probabilistic mathematical model highlighted some preliminary insights into the rate of increase in expressivity between holistic language and compositional language. Depending on the choice of F and V, the mathematical model also illustrates that expressivity growth differs between compositional languages. Then, by invoking the concept of objects, and therefore refining the notion of the environment, I used the monte carlo model to conduct a more thorough exploration of the parameter space. Relative stability is a measure introduced to reflect the increase in stability conferred by compositional language with respect to holistic language. Several important results have been identified, and these results are best grouped as follows:

Expressivity growth In these experiments, learning always occurs with respect to an impoverished body of data: a bottleneck is always in place. The proportion of the whole language that can be expressed, either through generalisation or as a result of observation, will depend on the sparsity of the observed data. Due to the structure in the mapping from meanings and signals, compositional language structure is

required for generalisation to occur. In contrast, holistic language contains no structure, and therefore generalisation is not possible. As the body of data available to the learning process increases, so does the expressivity of the learner. As one might expect, the rate of expressivity growth will depend on the language structure, with compositional language leading to a higher degree of expressivity than holistic language given the same amount of evidence. Importantly, different degrees of expressivity growth occur *between* compositional languages too.

Meaning space complexity For the purposes of this discussion, individual compositional languages are defined by a meaning space. How the signals are configured is not relevant: all we need to know is that the language has compositional structure. For two compositional languages, the rate of expressivity growth will depend on the structure of their meaning spaces, defined by the parameters F and V. Different combinations of F and V will lead to expressivity growing, as a function of the number of observations, at different rates. If the environment contains N objects, then several choices of F and V will be available. By preferring the number of features over the number of values per feature, generalisation leads to expressivity increasing at a faster rate; every observation confers information about F features. Expressivity increase is most rapid when binary feature values are used; in such a situation F is maximised.

Structural extremes When representing N objects, too much structure in the meaning space (high F and V) will lead to the generalisation ability of the learner to suffer. In this situation, due to a huge meaning space, the likelihood of observing co-occurring feature values will be low. In such a situation, generalisation becomes impossible: the set of observations fail to contain enough similarity. The other extreme, one where F and V are too low, will result in too little structure being present in the meaning space for generalisation to confer significant advantage. For example, when $F = 1$, compositionality is not possible as objects are differentiated solely in terms of feature values. There is only one dimension of variation. When considering meaning spaces, structural extremes tend to result in low relative stability.

Stimulus poverty The most striking result concerns the conditions under which compositional structure offers the maximum degree of expressivity relative to holistic language. With a low degree of language exposure, a situation corresponding to a tight transmission bottleneck, high expressivity can be achieved from compositional input. In contrast, for holistic language, limited language exposure leads to

low expressivity. So rather than sparsity in the data resulting in instability, a tight bottleneck is precisely the condition under which compositional language attains maximum relative stability.

5.6 Chapter summary and discussion

Language is a mapping between meanings and signals. The FSUT model allows us to represent a series of meaning/signals pairs derived from some language. These FSUTs stand as hypotheses that describe the linguistic input. The process of hypothesis selection, through compression, makes generalisation beyond the observed data possible. This is the learning process: a competence is derived on the basis of evidence, and this competence can capture the ability to handle situations not necessarily explicitly observed. The learning process therefore maps bodies of linguistic evidence to hypotheses. In this Chapter I have concentrated on just two classes of language, compositional and holistic language, and related these two language types to a notion of stability within the ILM.

This analysis pits structured mappings against random mappings within the ILM and aims to understand the regions of the parameter space that determine the emergence and persistence of structure. In one sense this analysis is a short-cut. If structured language is to emerge and remain stable, then an understanding of the conditions for stability will inform any wider understanding. This discussion is therefore a static analysis of a dynamic process, or in the vocabulary of Chapter 2, represents a detached understanding of a situated process.

The relationship between observed utterances and induced hypothesis is defined by the MDL hypothesis selection process, which operates over the space of FSUTs. The degree of generalisation achievable by certain regions of the hypothesis space can be modelled in terms of three parameters: F, V, and R. In the case of holistic language, generalisation cannot occur, and the set of expressible meanings derived from the induced hypothesis is identical to the set of observed meanings. In the case of compositional language, I introduce the notion of optimal generalisation and show that the FSUTs selected by MDL, subject to certain assumptions, yield a degree of generalisation identical to optimal generalisation.

The upshot of this short-cut is a model relating F, V and R to the degree of expressivity. Although the space of FSUTs in conjunction with the MDL principle, too, offers a model compression and expressivity, such simulation models are complex

and make mapping some parts of the parameter space computationally intractable. Using this short-cut, in conjunction with the derived models, I have shown how much of the parameter space can be mapped, and the relation between model parameters and expressivity can be understood in comparatively simple terms.

The utility of the models developed here is dependent on the relation between expressivity and stability. Stability occurs when subsequent agents in the ILM induce the same, or similar, hypotheses. For this to happen, enough of the mapping must be externalised such that generalisation to the remainder of the mapping is possible. Of course, if the whole mapping is externalised, then stability is guaranteed and occurs independent of language structure. But here I have concentrated on stability in the presence of a transmission bottleneck. High expressivity through generalisation is required for stability to occur in this situation. The higher the expressivity, therefore, the greater the degree of stability. In this respect, the notion of stability I am referring to is Liapounov stability: from a state of high expressivity, subsequent states are also likely to have high expressivity, and therefore stability.

By noting this relationship between expressivity and stability, the conditions for stability must be related to the conditions for expressivity. In the context of the dynamics imposed by the ILM, languages compete for survival. The transmission bottleneck radically alters the conditions for survival. As I have noted, without a bottleneck, all languages are equally likely to survive. As the severity of transmission bottleneck increases, so does the pressure for structured, generalisable languages. The measure of relative stability, developed above, captures this fact. Relative stability indicates how much more stable compositional structure is than holistic language. It therefore tells us the degree to which compositional language is likely to survive transmission in comparison to holistic language. The fundamental insight gained in this chapter is the *range* of conditions for high relative stability.

The determinants of high relative expressivity are meaning space structure and severity of the transmission bottleneck. Sufficient structure is required for generalisation to occur, and too much structure results in the co-occurrence of feature values being improbable; the result is that generalisation cannot begin to function. A high degree of severity of the transmission bottleneck leads to randomness becoming less and less stable. Strikingly, it is precisely under these conditions that compositional language starts to become generalisable, and the combination of these two effects makes compositional structure maximally advantageous for low bottleneck values. These results are striking because they are analogous to the situation on which the argument from the poverty of stimulus is based. This situation is discussed

further in Chapter in 7. So before relating these results more firmly to the context of linguistic evolution, I will use them to inform an investigation into the dynamic properties of the FSUT hypothesis space and MDL hypothesis selection.

Chapter 6

Dynamic Analysis

6.1 Introduction

Several assumptions underly the model of linguistic evolution developed in the previous two Chapters. Perhaps the most fundamental assumption is that linguistic evolution is possible at all, given that hypothesis selection is governed by the minimum description length principle. Under certain conditions, the minimum description length principle diverges from Occam's razor, and fails to suggest the hypothesis that leads to the greatest degree of generalisation. In this Chapter I will present a full simulation model of linguistic evolution based on the minimum description length principle. In order to apply the mechanisms developed so far, several assumptions must be realised in practice. This Chapter, therefore, presents a practical casting of the the theory developed in the previous two Chapters.

Section 6.2 details the remaining aspects of the model, such as hypothesis selection, signal production, and signal invention. In Section 6.3 I discuss an unsuccessful model of linguistic evolution that highlights the importance of bias in the invention algorithm. After developing an invention algorithm based on a simplicity principle, Section 6.4 discusses a model that results in successful linguistic evolution, where compositional language consistently evolves. This model relies on pre-existing signal diversity. To strengthen the model, Section 6.5 develops the model further, such that linguistic evolution toward compositionality is possible from an initial blank hypothesis. This model relies on a mixture of random and non-random invention. The precise structure of the evolved systems is discussed in Section 6.6, where I note that linguistic evolution consistently results in sub-optimal transducers containing redundant transitions. To relate these results to the intuitions of the previous two

Chapters, Section 6.7 examines the role of bottleneck size, and as such confirms the results of Chapter 5. Finally, I will discuss the results of this Chapter with respect to models of linguistic evolution in general. In short, the intuitions developed so far will prove accurate. But as well confirming existing results, the simulation models that follow bring to the fore some fundamental issues relating to the evolution of compositionality.

6.2 Iterated Learning and MDL in practice

I have used the FSUT model to represent the relationship between meanings and signals. Then, from the space of possible FSUTs, the minimum description length principle has been used to determine which FSUT is chosen to act as a hypothesis given some observed data. On the basis of these modelling decisions, I have presented some foundational results that rely on a number of assumptions. First, given some data D, I have assumed that the transducer with the minimum description length can be found. Second, I argued that under certain assumptions about the underlying probability distribution of the data, that the expressivity of the induced transducer can be straightforwardly determined. Third, I justified why expressivity is in some sense proportional to stability in the iterated learning model. Fourth, I made the simplifying assumption that the conditions under which compositional structure offers increased stability with respect to holistic structure can act as *general* conditions that shed light on the emergence and stability of compositional language.

During the course of the last Chapter, I presented arguments justifying each of these abstractions. In order for the intuitions of the last Chapter to be considered solid, faith in the *conjunction* of the assumptions is required. In this Chapter I will present a full simulation model that strengthens the conclusions of the last Chapter, as well as uncovering several new questions. The purpose of this Section is to fill in the remaining details of the existing model such that it can be tested through simulation. The outstanding issues concern the details of how one language is mapped onto another by an agent. Figure 6.1 illustrates these issues, and makes clear the components in need of thorough explanation. In particular, how exactly is the FSUT with the minimum description length found? How exactly is generalisation achieved? What happens when an agent needs to invent a signal? By answering these questions, the key questions regarding the possibility of, and the conditions for, the linguistic evolution of compositionality can be addressed.

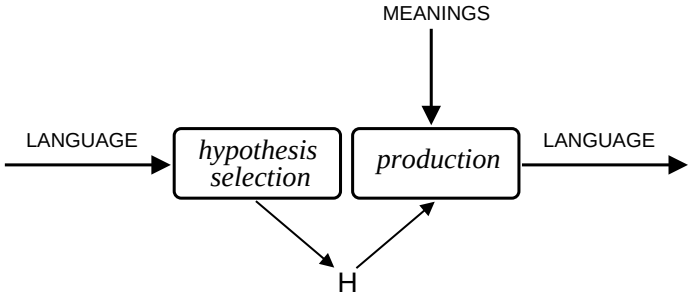


Figure 6.1: The key components of an agent. In order to place the existing model into a full simulation model, the processes of hypothesis selection and signal production need to be fully defined.

6.2.1 Hypothesis selection

Given a hypothesis space $\mathcal{H}$, the MDL principle can be thought of as a function that maps the data onto some hypothesis, H_{MDL}, found in $\mathcal{H}$. Until now, examples of this mapping have been derived analytically. That is, given assumptions about the structure of the data, along with a knowledge of the probability distribution over this data, the FSUT chosen (which corresponds to H_{MDL}) can be derived through analysis. In general, we would like a mechanism for finding H_{MDL} automatically without placing restrictions on the kind of languages we can consider, nor on the underlying probability distribution over utterances. As I have noted previously, this is practical problem. I have not needed to tackle this issue until now. By solving this problem, a considerable part of the information processing task of an agent will be fully realised (see Figure 6.2).

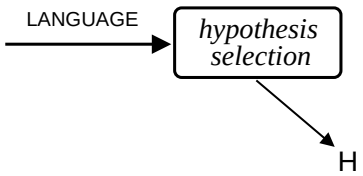


Figure 6.2: Hypothesis selection. Given linguistic data, MDL tells us which hypothesis is appropriate , H_{MDL}, but it does not tell us how to find this hypothesis.

Given some data D and the hypothesis space $\mathcal{H}$, the process of automatically locating H_{MDL} reduces to a search problem. The brute force approach would proceed by checking each member of $\mathcal{H}$ in turn and noting, with respect to the hypothesis being checked, the description length of D. When I refer to the description length of D, I will be referring to the sum of *both* the data encoding length and grammar encoding length. The hypothesis chosen would then be the one that results in the

smallest description length of D. Unfortunately, checking every hypothesis in $\mathcal{H}$ is intractable, as $\mathcal{H}$ is too large. A more refined search strategy is required.

The solution I will employ proceeds by first constructing the prefix transducer in light of the data, D. By applying compression operators, this transducer is incrementally compressed until one of two situations occur:

1. The transducer cannot be compressed further.
2. The description length begins to increase.

A transducer is maximally compressed when any further removal or modification of the edges results in the transducer becoming inconsistent with the observed data; the data can no longer be represented using the hypothesis. When further compression leads to the description length of D increasing, then this event is an indication that additional compression will result in the data encoding length becoming too high with respect to the grammar encoding length: the sweet-spot in the trade-off between data encoding length and grammar encoding length has been overshot. This is the general picture: compression of the prefix tree transducer progresses until some stopping criterion is met.

This general picture, however, masks several problems. A traversal of the search space, which is achieved through the application of compression operators, does not necessarily present a smooth path to H_{MDL}; the description length of the data does not monotonically decrease as a result of applying the compression operators. The search space contains many local minima. For example, imagine merging two states which in turn introduce the possibility of applying edge merge operators. These edge merging operations may not be possible without merging the states first. This two-step procedure – a state merge followed by several edge merges – may result in a substantial decrease in the description length of the data. This decrease in encoding length reflects the introduction of structure into the transducer. The problem is that the initial state merge operation will often lead to an increase in the data encoding length, and this increase dwarfs any decrease in the grammar encoding length. That is, in order to get closer to H_{MDL}, a temporary increase in the description length is required. This is an example of a local minima occurring between two compression operators. Frustratingly, this phenomena also occurs across relatively long series of compression operator applications. To combat this situation, I will employ two mechanisms: (1) a less brittle search strategy, and (2), compound compression operators.

Searching the hypothesis space using a beam search

Rather like a genetic algorithm, a beam search maintains a population, or set, of hypotheses subject to a turnover (Mitchell 1997:250). I will apply the concept of a beam search by maintaining a set of hypotheses that change over time. As the search progresses, members of the set are replaced with new hypotheses derived from the application of the compression operators. More formally, the procedure BEAM-SEARCH is a function that maps the prefix tree transducer H_{prefix}, the size of the beam n, and the set of compression operators $\mathcal{O}$, to the hypothesis $H_{MDL} \in \mathcal{H}$:

$$\text{BEAM-SEARCH}(H_{prefix}, n, \mathcal{O}) = H_{MDL}$$

The procedure BEAM-SEARCH begins the search for H_{MDL} by first constructing the beam, $\mathcal{B}$, which consists of n hypotheses $\{H_1, H_2, \ldots, H_n\}$. Initially, the beam contains n copies of the prefix tree transducer. As the search proceeds, three items of information are maintained: (1) the mapping DEAD-ENDS defines those members of the beam that cannot be compressed any further; (2) the variable MIN-TRANSDUCER details which member of the beam represents the hypothesis found so far that yields the smallest description length of the data; (3) the variable MAX-TRANSDUCER details which member of the beam leads to the largest description length of the data. At each step in the search, a random member of the beam, H_{rand}, is chosen and then each merge operator is applied at random until one of them is successful. If none of the compression operators can be applied, then H_{rand} is flagged as a dead-end. Otherwise, after a successful application of a compression operator, the resulting transducer H'_{rand} replaces the transducer in the beam represented by MAX-TRANSDUCER, so long as the description length of the data achieved using H'_{rand} is less than that achieved using MAX-TRANSDUCER. By doing this, H_{rand} may still be present in the beam, and therefore other compression operators can be applied if H_{rand} is chosen again.

The beam search is complete when all beam members are marked as dead-ends, or when all compression operators lead to an increase in the description length. The hypothesis represented by MIN-TRANSDUCER is then chosen, and offered as a candidate solution, $\hat{H}_{MDL}$. Because the process is stochastic – operators are applied at random to random members of the beam– there is no guarantee that the hypothesis returned is in fact H_{MDL}:

$$\hat{H}_{MDL} \approx H_{MDL} = \min_{H \in \mathcal{H}} \{L_{C_1}(H) + L_{C_2}(D|H)\} \tag{6.1}$$

In Chapter 5 it was assumed that H_{MDL} could always be found. Here, then, is an example of an assumption made in Chapter 5 that will not necessarily be realised in practice.

Compression operators

There are two fundamental compression operators: the *edge merge* operator and the *state merge* operator.

Edge merging Two edges can be merged providing that they have the same source and target states, they accept the same symbol, and the intersection of their meanings is non-empty. That is, the meanings on each edge to be merged must have some feature values in common. Figure 6.3(a) depicts the result of the edge merge operator. The result of such an edge merge is one less edge, and therefore the grammar encoding length is guaranteed to decrease. Also, by merging edges a decision that before needed to be encoded for each item traversing these edges no longer needs to be encoded: that branch in the transducer no longer exists. For this reason, edge merging always leads to a decrease in data encoding length. An edge merge operation can, however, result in the transducer no longer being consistent with the data. Such a situation can occur when the subsumption of the two meanings introduces a wildcard specifier, and the feature values lost by introducing this wildcard are not specified elsewhere in the transducer. For this reason, an edge merge is only considered applicable when the resulting transducer can continue to parse all observed utterances.

State merging When two states s_1 and s_2 are merged and replaced by a new state s, all edges that either begin or end in either s_1 or s_2 now begin or end in s. Figure 6.3(b) depicts this operator in action. As as result, the grammar encoding length will decrease as one less state needs to be encoded. The data encoding length, in contrast, will usually increase because the new state may have more outward edges than therefore. In this case, more bits are required to describe paths through the transducer traversing s.

Compression is achieved by applying a succession of compression operators. The order in which these operators are applied is subject to restriction. Given a prefix

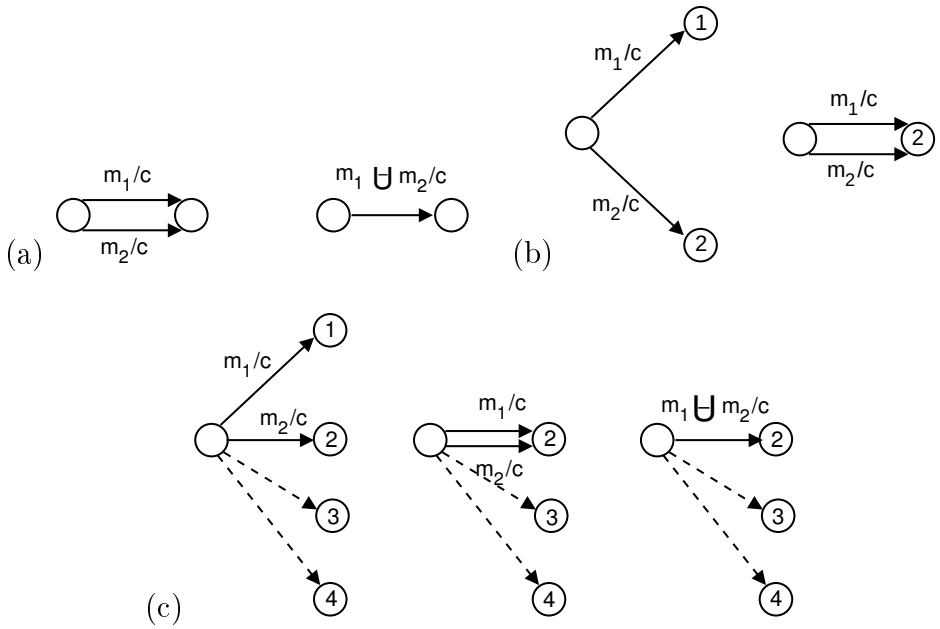


Figure 6.3: Compression operators. In (a), the result of an edge merge is shown. Similarly, in (b), the state merge operator is shown. In (c), an example of compound merge operator shows how multiple merging decisions are carried out in one operator application. Compound merge operators increase the efficiency of the search by deferring the calculation of encoding lengths until the several operators have been applied.

tree transducer, state merge operations must be applied before edge merging becomes an option. This pattern of operator application is a recurring theme during the compression process. So far, the basic edge merge operator and state merge operator have been described. By extending the collection of operators, the compression process can be made more reliable and efficient An extension to the set of compression operators becomes apparent when we note that the operators shown in Figure 6.3(a-b), from a focus state, only operate in one direction: forward. By adding operators that apply backwards as well, the number of possible focus states to which a compression operator can apply is increased. The second extension to the set of operators arises from the observation that more than two edges can be merged as a result of state merge. For this reason, *compound operators* are required. A compound operator performs a local search through the hypothesis space, and picks the set of edge merging decisions that maximise the degree of compression. A compound merge operator is just a series of state and edge merge operators, only the order in which they are performed is optimised such that the encoding lengths are not computed after each operator application. As with state and edge merge operators, two directions of compound operators are possible: inward and outward.

Compound merge Given a focus state f, and a set of outward edges, the compound merge operator chooses the best, in terms of MDL, set of edge merge operations. For example, five edges accepting the symbol ω could be merged, and the result of this merge will lead to a generalised meaning. At first sight, this may be an appealing compression operation. On the other hand, the encoding length may be minimised further by only merging three of these edges. The compound merge operator attempts all permutations of edge merge operations, and the resulting state merge operations that become possible as a result. Figure 6.3(c) illustrates a basic example of this process. The sequence of operator applications that minimise the description length are chosen.

To recap, the basic edge and state merge operators are all that is required to compress a transducer. The problem is, they introduce local minima into the search space. To alleviate this problem, compound merge operators are used to jump the local minima by carrying a local subspace search, rather than taking the chance of a fortunate series of explorations occurring during the beam search. In this way, a trade-off between local intelligent search and wider stochastic search is achieved.

6.2.2 Production

Production is the process that, given a hypothesis H and meaning m, yields an appropriate signal, s, for m. At this point, MDL no longer plays a role. That is, the process of interrogating the hypothesis, and the notion of an *appropriate* interrogation, is outside the remit of the MDL principle. The hypothesis chosen by MDL may *suggest* a set of appropriate generalisations that can be made on the basis of the hypothesis, but it is important to note that some bias will necessarily be introduced into the production process.

One of three mechanisms can be used to express a meaning. First, if a meaning m has been observed in conjunction with a signal then m will always be expressible; production is always consistent with observation. Observed meaning signal pairs are therefore, in the sense that they can always be expressed, *memorised*. Second, providing there is compositional structure in the observed language, then it may be possible to *derive* a signal for a meaning on the basis of the induced transducer structure. That is, some meanings can be expressed as a result of induction. Third, in the absence of compositional structure, the signal for an unobserved meaning cannot be derived. In this case, the agent can either *invent* a signal, or decline to produce at all. When called to express a meaning, an agent can therefore invoke

one of three mechanisms: memorisation, derivation, or invention. Even though the processes of memorisation and derivation are made possible through different circumstances, they are, in practice, implemented using a single mechanism. Both require a search through the transducer.

Production as depth-first search

Given a hypothesis H and a meaning to express, m, a path through the transducer is sought such that the unification of the meanings encountered along the path uniquely specify the meaning m. Accordingly, each meaning fragment should be consistent with m, and when unified, fully specify m. The consistency criterion requires that the meaning fragments do not contain feature values that differ from the value of that feature found in m. To fully specify m, the result of unification should not leave any feature values unspecified. While the meanings encountered along the path determine the validity of the path, the symbols encountered impose no such restriction. These symbols are simply concatenated. For example, to express the meaning $\{2\ 1\ 2\}$, using the hypothesis represented by the transducer shown in figure 6.4, the following series of operations are invoked:

$$\{2\ ?\ ?\}/b \quad \uplus \quad \{?\ 1\ ?\}/c \quad \uplus \quad \{?\ ?\ 2\}/f \quad = \quad \{2\ 1\ 2\}/bcf$$

where $\uplus$ is an operator denoting meaning unification and symbol concatenation. Note that all other paths through the transducer shown in Figure 6.4 will result in inconsistency with $\{2\ 1\ 2\}$. This need not be the case; there is no restriction on FSUTs such that each meaning has a unique path though the transducer. That is, synonymy is possible, but ruled out as a result of the production bottleneck discussed earlier.

Finding an appropriate path through the transducer, that is, one satisfying the criterion discussed above, is achieved by performing a depth first search. Beginning with the start state, each consistent path is followed until the meaning is fully specified and the accepting state is reached. The search is depth-first as a single path is explored until it is proven to lead to either success or failure. It is worth noting when the language is random, and the induced hypothesis resembles the prefix tree transducer, then no choices will exist along the paths through the transducer used for production. For compositional languages, where a fully compressed transducer is induced, then, as we seen in Figure 6.4, a choice is available at each step through

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

(b)

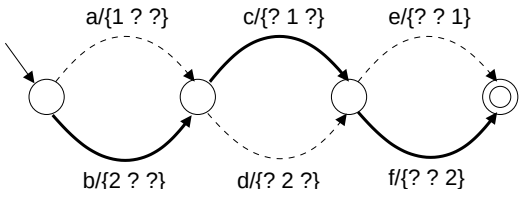


Figure 6.4: Given the language L, shown in (a), the transducer shown in (b) is induced. The meaning {2 1 2}, even though it was not observed, can be expressed as a result of the induced structure.

the path. It is only when choices exist, and sufficient structure is present in the language, that production through depth first search will lead to generalisation.

6.2.3 Invention

The issue of invention will prove a very important mechanism in the iterated learning models that I will explore. Until now, the issue of invention has been pushed aside. The simplicity of the models developed in Chapter 3 was partly a result of the twin processes of production and invention being rolled into a single mechanism. The mapping between the two spaces was well defined, and as a result, basic vector geometry solved both problems. In Chapter 5, the issue of invention was not invoked at all.

Before concentrating on the details of invention, the role of invention in an iterated learning model is worthy of some discussion. In general, unstructured languages lead to a high rate of invention. For the notion of iterated learning to have any degree of explanatory purchase, the rate of invention must decrease over time. This decrease is indicative of linguistic evolution toward structure. In this sense, invention drives linguistic evolution, just as mutation drives biological evolution. The intuition is that invention will, by chance, introduce structure where it was previously absent. The learning process latches onto these chance events, and provides the traction required to effect cumulative evolution. This is a general argument highlighting the processes driving state change in an iterated learning model.

Given the above argument, we might propose the following basic invention mechanism: when a meaning m cannot be expressed via any principled means, then

propose a random signal. Unfortunately, a successful invention mechanism, as I will show, is unlikely to be as straightforward as this. Such an invention scheme is extremely unlikely to lead to the chance introduction of structure. Even if the symbols used to construct the signal are picked selectively so that, for example, only symbols used in expressing similar meanings are chosen, then the chance introduction of structure will still be very low. In the following discussion I will discuss alternative approaches to invention. After investigating several candidate invention schemes, it will become apparent that the general motivation behind invention in the iterated learning model, discussed above, needs to be revised. A single chance invention, although in theory sufficient to introduce structure, is unlikely to be sufficient in practice. A more accurate scenario is one where a *series* of chance events are required for invented structure to be induced by a learner. Given the assumption of a totally random invention scheme, a *series* of chance events is less probable than a single chance event. As a result, in order for invention to function in any practical sense, and consistently direct evolution toward structure, an invention scheme cannot be entirely random. Biased invention is required.

Partially random invention

How can invention be biased? What I will term *partially random invention* attempts to exploit any existing structure in the hypothesis when suggesting an appropriate signal. For invention to be required, the usual means of expressing some meaning, say, m, must fail to apply. Now, given the meaning m, the most similar meaning that can be expressed, m_n, is found. The similarity between two meanings is defined as the inverse of the Hamming distance between them. The production of the signal for m_n is represented by a series of edge traversals, each one detailing, a symbol at a time, how the signal is built using the transducer. Partially random invention exploits this information by assuming that the production of m should follow a similar path. The series of edge transitions used to produce a signal for m will be identical to those used to produce a signal m_n, but for those transitions which carry a meaning inconsistent with m. For those symbols in m_n's signal built using these inconsistent transitions, a random symbol is used in constructing the signal for m. The result is that m's signal will be the same m_n's signal where the transitions are consistent with m, and random otherwise.

Partially random invention is motivated by the need to retain as much structure in the language as possible. If the meaning m was expressed using a totally random signal, then any structural relationship existing between similar meanings and their

signals is unlikely to be reflected. Introducing randomness into the system should be a last resort as it is likely to destroy any existing structural relationship between meanings and signals. Such a policy will surely increase the likelihood of the system occupying trajectories toward structure: structured parts of the language will be used whenever possible, and are therefore likely to survive to the next generation.

In the models I will develop, invention is regulated. Rather than swamping the evolving language with random signals, invention is only performed with a probability of 0.5. This value represents the *invention rate* (Kirby 2002a). All the simulations described below have an invention rate of 0.5. When invention is not performed, the agent will not produce any signal for the given meaning. The number of utterances produced by an agent will therefore vary.

6.2.4 A full simulation model

I have now finalised the specification of a full iterated learning model. The agents are capable of hypothesis selection, production, and invention. A practical investigation of the issues raised in Chapter 5 is now possible. Before exploring the consequences of this simulation model, it is important to point out some practical limitations.

Unlike the static analysis of Chapter 5, certain parts of the parameter space will prove to be computationally intractable to explore. Although the simulation model will prove to offer considerable insight, it should be noted that many of the experiments I will present take in the order of weeks to complete. The combination of calculating encoding lengths, testing for transducer consistency, and searching the hypothesis space require a great deal of time. This is why the analysis of Chapter 5 is so important; issues such as average case behaviour across multiple runs, and the exploration of high dimensional meaning spaces is not possible, and are in some sense not the purpose of the simulation model. Instead, the simulation model should be seen as, first, validating the conclusions of Chapter 5, and second, shedding light on the nature of the evolved systems.

6.3 Simulating linguistic evolution

Now all the model components required for a full-blown simulation run are in place. An agent induces a hypothesis by applying MDL hypothesis selection over the space of FSUTs; production is achieved by searching the transducer for a successful production path; partially random invention is invoked when production is not

possible. In order to apply the model, a number of parameters need to be chosen. The following parameter settings will be used:

Parameter	Value
MaxStringLength (l_{max})	15
BottleneckSize (R)	32
NumberOfFeatures (F)	3
NumberOfValues (V)	4
AlphabetSize ($\|\Sigma\|$)	20
MaxIterations	200
BeamSize	30

It is worth dwelling on the significance of these parameters. The majority of the parameters define the language the first agent will learn from. Specifically, the first agent in the simulation is passed R utterances drawn from a random language L_{rand}. This language is constructed by associating R random meanings with signals of a random length between 1 and l_{max}. Each symbol in the signal is drawn randomly from the alphabet Σ, which contains 20 distinct symbols. By initialising the simulation in this way, an agent is presented with linguistic input it may never actually generate itself. For example, because the language L_{rand} is totally random, synonymy in the mapping between meanings and signals is likely: the same meaning may be given different random signals. The meaning space contains 64 meanings. The bottleneck is set at 32 (random) observations of the meaning space. As a result, the expected proportion of the meaning space to be observed, the coverage, will be 0.396, which corresponds to 25.344 distinct meanings.

6.3.1 Linguistic evolution

Initialised with the random language L_{rand}, the simulation is run for 200 iterations. Figure 6.5(a) aids an understanding of the evolving system: the minimum description length of the evolving language is plotted against the expressivity of the resulting transducer. Expressivity is defined as the number of meanings in the meaning space that can be expressed through production, but not invention. Figure 6.5(a) represents a vector plot detailing the transitions made in this subspace.

Two observations need to be made here. First, the initial state of the system (point A) changes significantly after the first iteration. Initially, because utterances are drawn from the language L_{rand}, the first agent is guaranteed to observe 32

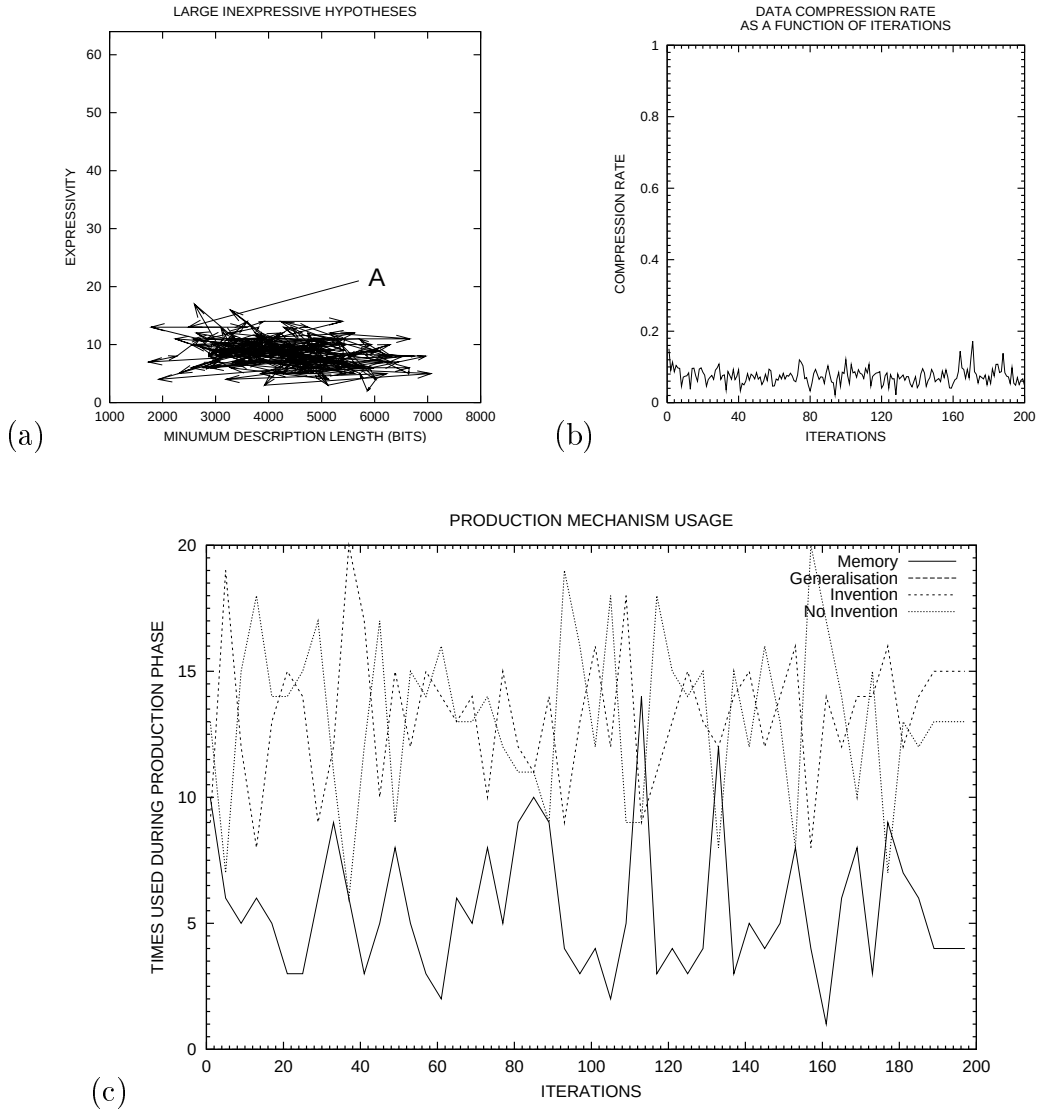


Figure 6.5: Linguistic evolution resulting from partially random invention. In (a), the fact that the system is limited to large relatively inexpressive hypotheses is illustrated. The size of the induced hypotheses is comparable to the prefix tree transducer, a fact illustrated in (b), where the compression rate is shown to be consistently very low. In plot (c) the production mechanisms driving state change are shown. As the system evolves, three production mechanisms drive state change. Memorisation, invention, and no invention.

meaning/signal pairs. In contrast, the hypothesis induced from this data, due to the data being random, will be unlikely to associate each of the 32 randomly sampled meanings with a signal. In short, the language generated by the first agent in the simulation will be abbreviated due to the occurrence of non-invention decisions. This is why the initial jump from point A occurs. Point A represents the case when the transducer associates 32 random meanings with signals. Due to the option of not inventing at all, this state can only occur at the first generation.

After the first iteration, the system begins a semi-random walk through the space of possible languages. This fact illustrated by Figure 6.5(b), where the compression rate achieved as result of the induced hypotheses is plotted as a function of iterations. An average compression rate of approximately 0.06 occurs across all 200 iterations. Figure 6.5(c) illustrates the production mechanisms underlying this behaviour. At each iteration, this graph plots the number of times each production method is used during the expression of the 32 observed meanings. Across all 200 iterations, meanings are expressed using a combination of memory, invention, and no invention. For reasons that will shortly become evident, compositional languages are stable when injected into the model as it stands. The key point here is that these languages cannot be constructed from an initial state of randomness: the invention bias is not strong enough. Of course, if the simulation were to be run for an infinite number of iterations, then a compositional language would crop up and persist. Such a situation is of no real interest to this discussion: we are interested in the consistent emergence of compositional structure in finite time. To further our understanding of linguistic evolution, we must now focus on the issue of the invention bias. This experiment has highlighted the fact that learning is just one component of an iterated learning model.

Recall that in developing the iterated instance-based learning experiment detailed in Chapter 3, a similar problem was faced. The solution was the obverter procedure, which went some way to ensuring that production bias reflected the learning bias. To progress the experiment, I will consider a similar approach to invention.

6.3.2 Rethinking invention

The best invention decisions are those that cause additional production paths to be induced in the transducer of another agent. Ideally, these paths, which would otherwise be absent, will present hitherto unavailable routes leading to generalisation. By focusing on this scenario, we can say that such invention decisions must alter

the existing mapping between meanings and signals in such a way that *compression* is more likely. This must be the case, as generalisation is only possible as a result of a significant degree of compression. It is this observation which will motivate the following invention procedure.

The Compressor Scheme

The fundamental question is as follows: Given a meaning m that cannot be expressed, which signal should be invented for m such that the probability of compression is maximised? This issue should be contrasted with the objective of invention *maintaining* structure, which was investigated above. In this respect, invention should really be performing the opposite, it should be *introducing* opportunities for compression.

First of all, in order to progress, we need to make the production competence of the transducer explicit: all utterances expressible through memory and derivation are found. That is, for each meaning in the meaning space, the production decision of the transducer is sought; the vocabulary is defined as this set of E expressible meanings. Of course, some of the meanings, including m, will not be expressible; for invention to be required at all implies expressivity is sub-optimal. The vocabulary is therefore a set of meaning/signal pairs:

$$\mathcal{V} = \{\langle m_1, s_1\rangle, \langle m_2, s_2\rangle, \ldots, \langle m_E, s_E\rangle\}$$

Expressible meanings are enumerated such that m_i and s_i denote the meaning and signal of the ith member of $\mathcal{V}$. The part of the mapping, defined by the transducer, that relates m_i to s_i is determined by a path through the transducer, which is defined as a series of edge traversals. If $|s_i| = l$ then mapping m_i to s_i requires a path containing l edges, where an edge takes the general form f/c such that f is a (possibly under-specified) meaning consistent with the meaning being expressed, and c is a symbol to be concatenated with the previously parsed symbols. In general, using the f/c notation, we can represent how the meaning m_i is mapped to the signal s_i, at each stage in the path through the transducer, in the following way:

$$m_{\{i,1\}}/s_{\{i,1\}} \rightsquigarrow m_{\{i,2\}}/s_{\{i,2\}} \rightsquigarrow \cdots \rightsquigarrow m_{\{i,l\}}/s_{\{i,l\}}$$

Where $m_{i,j}$ is the meaning fragment contained in the edge traversed after generating the jth symbol of the signal s_i. The invention scheme I am about to define is most easily understood by viewing the transducer as a series of transducer paths for all the expressible meanings.

The reason for representing the transducer competence in this way can be understood by noting that segments of transducer paths represented by the vocabulary will be selected to form the path for the signal to be invented. More precisely, fragments of paths found in the vocabulary can be chosen even though they are not linked by a path in the transducer itself. That is, certain configurations of path fragments cannot be arrived at by performing, say, a depth first search through the transducer. Figure 6.6(a) illustrates the general layout of the paths through the transducer using this scheme.

To understand how this representation of the transducer competence is used to invent a signal for an inexpressible meaning, consider the following question: In expressing m, which initial edge, located in the vocabulary, should be chosen? First, the policy of choosing an *existing* edge needs to be justified, and second, some criterion for edge selection, from the possible candidates, needs to be developed.

The justification for using existing edges is motivated by a simplicity principle. From the perspective of compression, introducing random symbols into the language is likely to be detrimental. To maximise compression, the assumption that the transducer already contains a sufficiently diverse set of symbols makes more sense. This assumption can be related to a simplicity principle: we wish to avoid multiplying the complexity of the hypothesis beyond that which is required.

The second problem that needs to be addressed concerns edge selection. Given a set of candidate edges, which one should be chosen? This choice is partly determined by the depth, within the transducer, of the edges to be considered. If the result of invention is to be an increased tendency for compression, then the point in the transducer where compression is to occur needs to aligned in relation to edges used to form the invented signal. By aligned, I mean, for example, that the first symbol in the invented signal must be chosen *relative* to existing edges that encode the first symbol in other signals. The edge representing the first symbol in the invented signal must therefore be chosen from the same, or similar, depth in the transducer.

The edge selection process therefore revolves around narrowing the set of candidate edges to those that appear at the same depth as the symbol to be generated. Given

the meaning m to express, the first symbol used in expressing m is therefore chosen from those edges at depth 1. This candidate set is of edges at depth 1, D_1, is represented by:

$$D_1 = \{m_{\{1,1\}}/s_{\{1,1\}}, m_{\{2,1\}}/s_{\{2,1\}}, \ldots, m_{\{E,1\}}/s_{\{E,1\}}\}$$

Figure 6.6(b) shows how the set D_1 is extracted from the edge layout discussed above. Choosing an edge from the candidate set requires a ranking over these candidate edges. The lowest ranked edges will be those that have no feature values in common with the meaning to express, m, as these edges cannot introduce a generalisation involving m – the intersection between the two meanings will be a fully unspecified meaning composed only of wildcards. An unspecified meaning can only occur as a result of compression, but it can never lead to generalisation. The highest ranked edges will be those that maximise the probability of *introducing* the possibility of compression. These edges will be those that are fully specified – they will contain no wildcards. An edge with a meaning fragment containing wildcards can only exist as a result of compression. The key idea is to avoid these edges, as it is more fruitful to attempt to introduce compression into points in the transducer where compression has not occurred previously, rather than reinforce the result of existing compression. Compressed regions of the transducer will already be more likely to survive transmission.

To choose a single edge from the set of candidates we need to consider how edges can be ranked. Here, the idea is to prefer those edges with the smallest intersection with m, the meaning to expressed. The ranking function is defined as:

$$\text{rank}(m, m_{\{i,j\}}) = n : \text{where } n \text{ feature values of } m_{\{i,j\}} \text{ match with } m$$

Importantly, feature values match only when specified; wildcards never count as a matching feature value. This ranking prefers a minimal intersection between meanings. Those meaning fragments scoring zero are filtered out as an empty intersection between meanings cannot lead to generalisation, for the reason described above. Among edges with the same rank, the winner is chosen at random. To recap, given a depth d in the transducer, candidate edges D_d are chosen, and ranked according the ranking function. The symbol of the highest ranked edge is then

(a)

Meaning	Transducer path	Signal
m_1	$\{m_{\{1,1\}}/s_{\{1,1\}} \rightsquigarrow m_{\{1,2\}}/s_{\{1,2\}} \rightsquigarrow \cdots \rightsquigarrow m_{\{1,\lvert s_1\rvert\}}/s_{\{1,\lvert s_1\rvert\}}\}$	s_1
m_2	$\{m_{\{2,1\}}/s_{\{2,1\}} \rightsquigarrow m_{\{2,2\}}/s_{\{2,2\}} \rightsquigarrow \cdots \rightsquigarrow m_{\{2,\lvert s_2\rvert\}}/s_{\{2,\lvert s_2\rvert\}}\}$	s_2
$\vdots$	$\vdots$	$\vdots$
m_e	$\{m_{\{e,1\}}/s_{\{e,1\}} \rightsquigarrow m_{\{e,2\}}/s_{\{e,2\}} \rightsquigarrow \cdots \rightsquigarrow m_{\{e,\lvert s_e\rvert\}}/s_{\{e,\lvert s_e\rvert\}}\}$	s_e

(b)

m_1	$\{\overbrace{m_{\{1,1\}}/s_{\{1,1\}}}^{D_1} \rightsquigarrow$	$\overbrace{m_{\{1,2\}}/s_{\{1,2\}}}^{D_2} \rightsquigarrow \cdots \rightsquigarrow$	$\overbrace{m_{\{1,\lvert s_1\rvert\}}/s_{\{1,\lvert s_1\rvert\}}}^{D_{\lvert s_1\rvert}}\}$	s_1
m_2	$\{m_{\{2,1\}}/s_{\{2,1\}} \rightsquigarrow$	$m_{\{2,2\}}/s_{\{2,2\}} \rightsquigarrow \cdots \rightsquigarrow$	$m_{\{2,\lvert s_2\rvert\}}/s_{\{2,\lvert s_2\rvert\}}\}$	s_2
$\vdots$		$\vdots$		$\vdots$
m_e	$\{\underbrace{m_{\{e,1\}}/s_{\{e,1\}}} \rightsquigarrow$	$\underbrace{m_{\{e,2\}}/s_{\{e,2\}}} \rightsquigarrow \cdots \rightsquigarrow$	$\underbrace{m_{\{e,\lvert s_e\rvert\}}/s_{\{e,\lvert s_e\rvert\}}}\}$	s_e

(c)

$$
\begin{aligned}
D_1 &= \{m_{\{1,1\}}/s_{\{1,1\}}, m_{\{2,1\}}/s_{\{2,1\}}, \ldots, m_{\{E,1\}}/s_{\{E,1\}}\} \\
s_1 &= s_{\{\min_{m_k \in D_1}\{\text{rank}(m,m_k)\neq 0\},1\}}
\end{aligned}
$$

Figure 6.6: Steps toward inventing a signal that is likely to lead to compression. For each expressible meaning, the mapping generated by the transducer is represented as a series of paths. The set of all such paths is termed the vocabulary, and shown in (a). The invented signal is built a symbol at a time. Each such symbol is drawn from candidates occupying a certain depth in the transducer, shown in (b). For a given depth, the symbol chosen depends on the result of applying a ranking function, shown in (c).

chosen as the dth symbol of the invented signal. Figure 6.6(c) illustrates how this symbol is derived from the edge layout.

Finally, two possibilities need to be considered. First, signals in the vocabulary may be of varying length, so the size of the sets $D_1, D_2, \ldots$ may be variable. Second, what happens if none of the meanings attached to the edges in D_d are consistent with m? The first observation is not important, as the length of an invented signal will be determined by the length of the series of the edges at varying depths that inform the production of m. The result is that similar meanings will tend to have signals of a similar length. The second problem, too, can be alleviated simply by not proposing a symbol at depths that contain no relevant information about m.

Invention as self-organisation

The invention procedure I have described is based on a simplicity principle. When called to produce a signal for meaning that cannot be expressed, instead of introducing randomness into the system, existing segments of the transducer are utilised. The assumption is that the alphabet of symbols present in the initial random language is sufficient to construct signals for a structured language: introducing new symbols, according to this assumption, would be multiplying the complexity of the hypothesis beyond that which is required. Rather than looking beyond the hypothesis, we look within it. The prior bias guiding invention therefore resides, as we would wish, in the hypothesis. In this respect, linguistic evolution is best thought of as the self-organisation of a pre-existing set of basic elements (symbols). To conclude this section, it is worth consolidating the processes underlying signal production. Four processes define how a meaning is mapped to a signal:

1. *Memorisation.* If the meaning to be expressed has been observed in conjunction with a signal, then this signal is used to express the meaning.
2. *Derivation.* The transducer may contain a path that, through the process of generalisation, enables an unobserved meaning to be expressed.
3. *Invention.* Using the compressor scheme described above, a signal is invented that will be likely to introduce further transducer compression.
4. *No invention.* With a certain probability, the invention scheme is bypassed, and no signal is proposed. The result is that production is not performed.

The effect on linguistic evolution of the compressor scheme has not yet been examined. It is this issue I will now turn to.

6.4 Linguistic evolution as self-organisation

To move forward, the effect of introducing a bias toward compression when inventing signals needs to be understood. Given that the bias toward compression introduced above differs substantially to the partially random invention investigated earlier, one would expect linguistic evolution to follow very different trajectories. In the following experiments, precisely the same parameters setting are used to those defined in the previous Section. Furthermore, Figure 6.7 illustrates the same measurements as those shown in Figure 6.5. Strikingly, Figure 6.7 reveals that very different states are achieved as a result of the newly developed invention bias.

Figure 6.7(a) illustrates an entirely different trajectory, one where a series of transitions lead to small, stable, and expressive hypotheses. Starting at an expected expressivity of approximately 22 meanings (point A), the system makes two distinct jumps to a stable state where we find small hypotheses capable of expressing all 64 meanings. The compressor scheme consistently directs linguistic evolution toward compositional systems.

The first major transition through the state space takes the system to point B, where expressivity increases slightly, but the minimum description length of the language decreases by a factor of 3. From requiring approximately 6000 bits to encode the evolving language, linguistic evolution results in transducers being induced with an MDL of approximately 2000 bits. The lack of increase in expressivity is a reflection of the transducers organising themselves in such a way that significant compression results, but an increase in expressivity is not achieved. The second transition, from point B to point C, is very different in nature. Here, for a small decrease in the MDL of the emergent language, a significant increase in expressivity occurs. This is an important transition, as it results in the system entering a stable region of the state space. Although a few deviations away from this stable region occur early on (to point D), the system settles into a steady state characterised by high expressivity. Figure 6.7(b) reflects these transitions. The compression rate rises in two stages, mirroring the points B and C.

Figure 6.7(c) illustrates the production mechanisms underlying these transitions. The first 20-30 iterations correspond to the transition from point A to point B. During this period meanings are expressed on the basis of production from memory and production through invention. The transition is driven by the introduction of generalisation during production. For the remainder of the simulation, production from memory and production through generalisation keep the system in a stable state; only occasionally is invention used, and this is a result of a highly skewed sample of the meaning space leading to the induction of transducer that is not fully generalised.

Invention using the compressor scheme causes a fundamental change in dynamics. First, the induced transducers become increasingly compressible directly as result of invention – as one might expect from the motivation underlying the compressor scheme. Once transducers become compressible, the emergence of paths through the transducer that share states with other paths begin to appear. From such a situation, only a small step is required to make inductive generalisations. Before

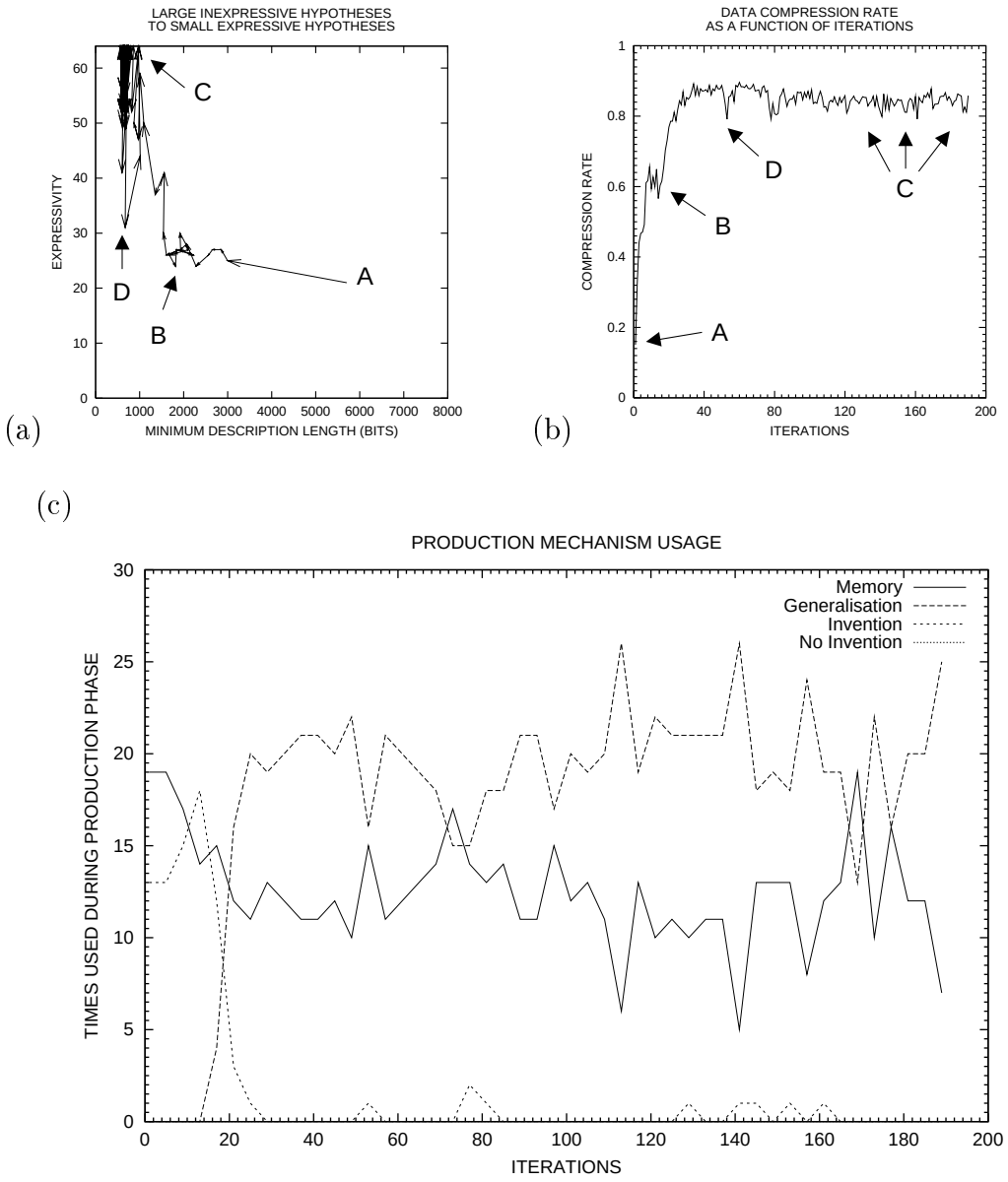


Figure 6.7: Linguistic evolution arising from the application of the compressor scheme. Plot (a) illustrates how the minimum description length of the emergent language is related to expressivity. Plot (b) illustrates how the compression rate increases over time. Plot (c) details which production mechanisms are used as the language evolves.

examining in detail the emergent transducers, it is worth sketching an interpretation of the state space depicted in Figure 6.7(a).

6.4.1 Possible trajectories

It is clear that when analysing the model of linguistic evolution developed so far, only certain regions of the state space are visited. In this section, I will divide up the state space into distinct regions, discuss the kind of languages and transducers that characterise each region, and explain why certain regions are visited and not others.

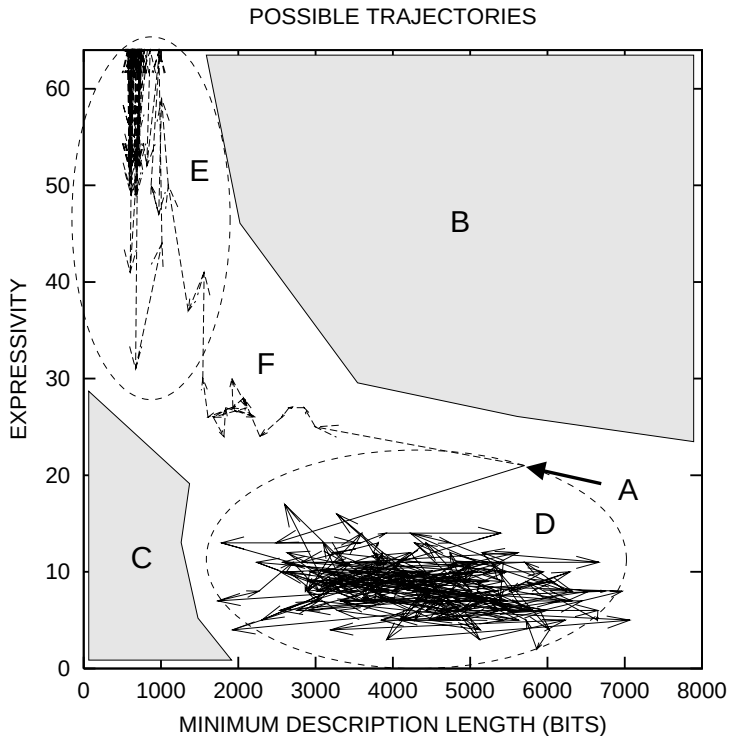


Figure 6.8: Regions of the state space. The language L_{rand} places the system at point A. Depending on the invention scheme used, the system either moves toward region D, or begins a series of transitions leading to region E. The system never enters region C or region B.

Figure 6.8 depicts several distinct regions of the state space. Both the trajectories achieved using partially random invention and invention through the compressor scheme are included in this map. Both these trajectories share a starting point, point A. Region B is not visited by either trajectory. Why is this? Region B represents the presence of large expressive transducers. Large expressive transducers are perfectly possible. However, when a transmission bottleneck is in place,

as is the case here, to achieve maximum expressivity requires a compressed transducer structure of kind discussed in the previous Chapter. It is possible that these transducers could be elaborated on and expanded without loss of expressivity, but these modifications will not be found in the hypothesis chosen by the minimum description length principle. In short, the presence of a bottleneck restricts the set of transducers capable of maximum expressivity to be those transducers that can be compressed to a significant degree.

Like region B, both trajectories fail to enter region C. Region C represents the occurrence of very small hypotheses with low expressivity. States in this region are not possible, but for a different reason to that of region B. Small expressive hypotheses are possible, so why not small inexpressive hypotheses? For example, we could imagine constructing a small inexpressive transducer by doctoring a compressed transducer such that its ability to generalise is crippled in some way. Such a modification is not possible when we consider the requirement of data consistency: the small transducer we imagine must be consistent with the observed data. There must be a lower bound on the size of transducers consistent with the observed data, and region C represents the case where this lower bound is breached. In short, region C represents states corresponding to inconsistent compressed transducers. Trajectories will not enter this region because the data consistency criterion must be met by all induced hypotheses.

The regions of key interest – those that are visited – are regions D and E. Region D represents unstructured and therefore incompressible languages. The initial language L_{rand} puts the system into the fringes of region D, with a thorough exploration of region D requiring partially random invention. Invention using the compressor scheme quickly moves the system away from region D, as each invention decision is likely to introduce structure to the language, and therefore make compression more likely. Once compression occurs, a path toward region E is inevitable. The area between region D and region E represents a smooth transition between a decrease in minimum description length and an increase in expressivity.

6.4.2 Analysing the major transitions in linguistic evolution

With a clearer understanding of the range of transitions that occur, it worth looking in detail at the various stages of linguistic evolution. The starting position for both the trajectories discussed so far is point A shown in Figure 6.8.

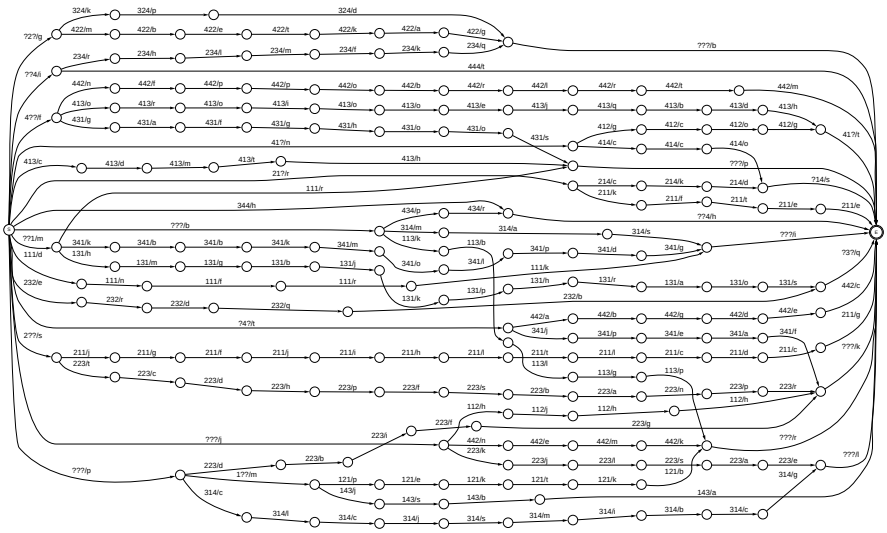


Figure 6.9: The transducer corresponding to point A. This transducer represents the hypothesis induced from the initial, randomly constructed language.

Figure 6.9 depicts the transducer representing point A. This transducer is the one induced from the initial language, L_{rand}. Some compression is evident, but not nearly enough for generalisation to function. It is quite normal for a slight degree of compression to occur when faced with a random language, as all meanings will have some symbols in common. This compression may appear significant, but it should be stressed that constructing a language that leads to absolutely no compression is in fact quite hard. In contrast, for compression to lead to generalisation requires very specific and highly improbable structural relationships between signals and meanings.

From point A, I will trace the effects of linguistic evolution toward region E by examining the series of induced transducers. Figure 6.10 depicts the transducers at various stages of evolution toward region E. Figure 6.10(a) represents the transducer corresponding to point F. In this transducer, we see the beginnings of significant structure emerging. The first symbol in each signal appears to discriminate between feature values in the second feature. This structural relationship acts as a seed for further discrimination, which will ultimately result in generalisation.

Between point F and region E, the self-organisation of the emerging language becomes increasingly more evident. In 6.10(b) additional discrimination within the transducer occurs, and certain meanings can be expressed even though they were not observed – significant generalisation is occurring. Figures 6.10(c-d) illustrate the occurrence of further discrimination and generalisation, as the state of the system climbs up and skitters around region E. Figure 6.10(e) represent the final stable

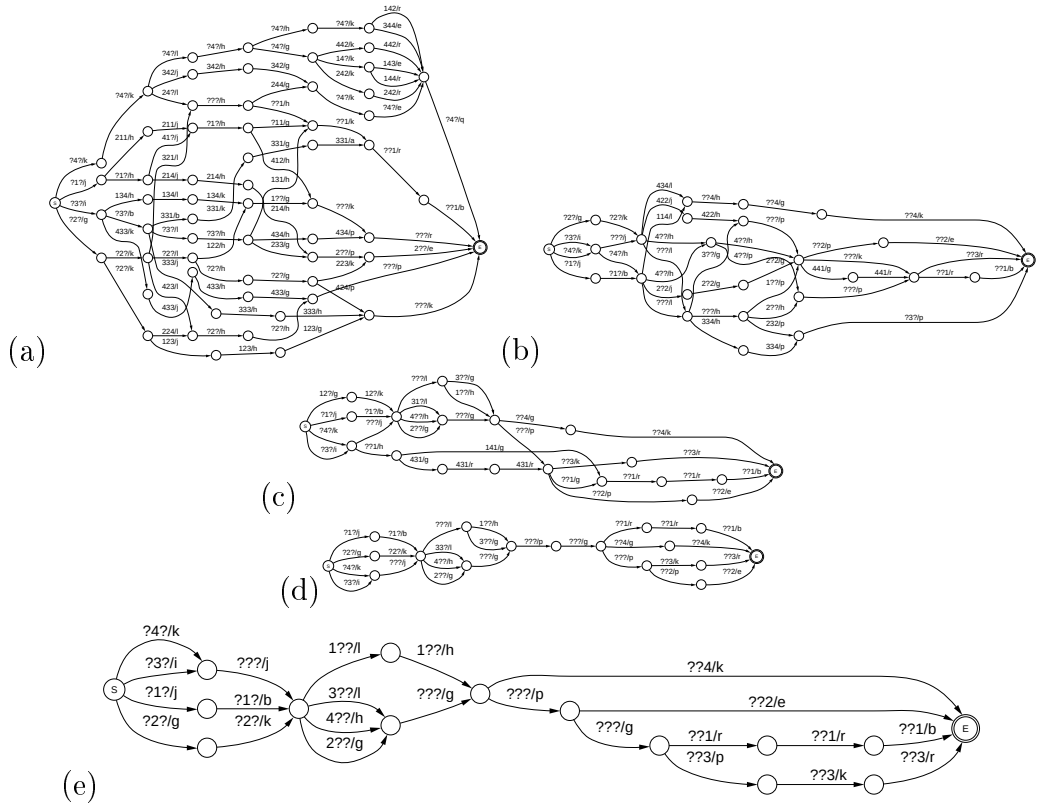


Figure 6.10: Steps toward stability. The transducer in (a) corresponds to the point F in the state space. The states occurring along the transition from point F to region E are represented the transducers in (b-d). The final stable state is represented by transducer in (e).

state, characterised by a fully compressed and maximally expressive transducer. Although this transducer exhibits a large amount of redundancy, this redundancy does not effect its ability to be induced generation after generation. Initially, as the state approaches region E, some variation occurs across iterations before the steady state is arrived at. This suggests the stable regions of the state space are Liapounov stable: if the system were to start in this region, it would stay in that region.

This Section has shown how the compressor scheme, used during invention, results in the self-organisation of the initial language. The assumption underlying this approach is that enough signal diversity exists in the initial language. On the basis of this assumption, a simplicity principle is justified: never diversify the language beyond that which is required. To strengthen the results so far, I will now briefly investigate an alternative approach to initialising the system. Here, the compressor scheme will be tested by starting small: I will fix the initial state of the system such that the required signal diversity is *not* present to begin with.

6.5 Starting small

The preceding experiments start with the first agent in the simulation observing utterances from a randomly constructed language. In doing so, the first agent differs from all other agents in that its input originates from a non-biased source. All subsequent agents in the simulation receive their input from the output of other learners. I have shown how the resulting evolutionary trajectory soon begins to fall into a path partially determined by an array of adaptive pressures present in the model, such as bottleneck size, hypothesis selection criterion, and the like. For this reason, a random initial state serves to illuminate how certain parts of the state space are masked by constraints imposed by the model. The diversity of symbols contained in the randomly constructed initial language of the first iteration is sufficient for the linguistic evolution to proceed through self-organisation.

A model in which the linguistic evolution of structure is invariant over initial conditions is stronger than one that relies on specific properties being present in the initial conditions. So far, the model has presupposed signal diversity. An alternative to this approach is motivated by the need to start a simulation from a blank slate. This situation is modelled by adding a random invention mechanism to the existing invention mechanism. In this way, when an agent is called to express a meaning, and no principled production decision is possible, invention will proceed as defined but for an occasional *randomly* invented signal. Agents themselves introduce signal diversity, and the language is constructed from scratch. In contrast to the existing experiments, every learner in the model receives output from another learner. The first iteration will now start with a learner represented by an empty transducer – a blank hypothesis. When called to express meanings, the random invention mechanism will emit random signals.

6.5.1 Introducing signal diversity

Introducing random invention poses a problem. An empty transducer containing no transitions will obviously have no option but to invent random signals. But once the language grows from these invention decisions, a choice in which invention scheme to apply is introduced. The solution I will adopt proceeds by invoking random invention with a probability 0.1. That is, when invention is required, random invention is occasionally performed. The signal invented is totally random, and will be of length between 1 and l_{max}. The symbols used to construct the signals are

drawn from a pre-defined alphabet containing the first 20 lowercase letters of the alphabet. In this respect, *symbol* diversity is built into the model.

6.5.2 Three phases

The introduction of random signals during linguistic evolution may, in the first instance, seem foolhardy. Any evolved structure will, one might presume, be destroyed by the introduction of randomness. But contrary to this intuition, the introduction of random signals has no such corruptive influence. The insertion of signals with a (random) structure, at odds with the existing signals, is only temporary due to higher frequency application of the compressor scheme. Those random signals introduced into the language soon decay as a result of the invention decisions of the compressor scheme. The result of random invention serves precisely the purpose we intended: to introduce signal diversity.

In short, the compressor scheme drowns out the corruptive influence of random invention. The result is that emergent languages initially grow to a certain size, before a period of self-organisation occurs, and linguistic evolution proceeds as described in the previous Section. The initial growth of the language is due to the first agent in the simulation uttering the random signals. As the simulation progresses, and the compressor scheme comes into play, the occasional random inventions have little effect. Consistent linguistic evolution results.

Figure 6.11 illustrates the key transitions from the initial iteration, represented by a blank hypothesis, through large inexpressive hypothesis, and finally at small expressive hypotheses. These state transitions mirror those found in the previous experiments, but for the initial phase, where the evolving language increases in size due to random invention. In contrast to the discussion of the state space in Section 6.4.1, the region representing small inexpressive hypotheses is fundamental. Rather than representing an unobtainable region of the state space, this region is now the initial state made possible by starting with a blank hypothesis.

Figure 6.12 represents a different experiment where a meaning space of $F = 3$ and $V = 3$ is used. Here, the description length of the emergent language is shown as a function of iterations. The three phases leading to compositionality are clearly evident. At each phase, an example transducer is shown to illustrate the state of the system.

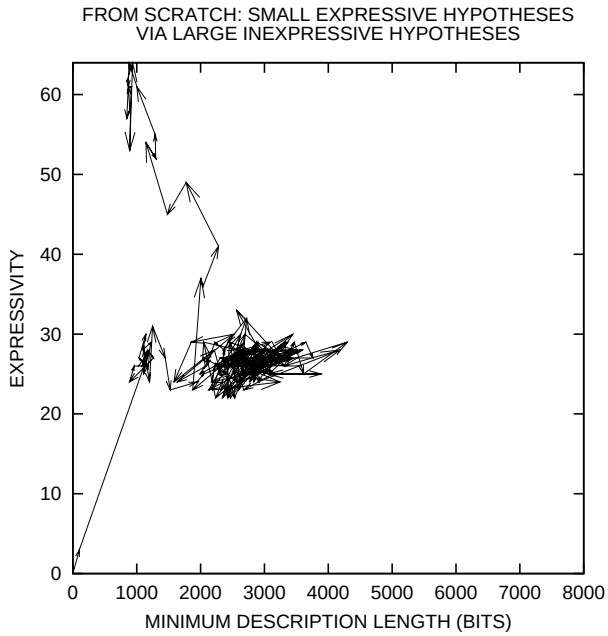


Figure 6.11: Starting small. From a blank hypothesis, random invention leads to the construction of signals for each observed meaning. Once signal diversity is present, the compressor scheme introduces structure into the signals. Compositional languages emerge as a result.

These experiments have shown that the robust linguistic evolution of compositional structure occurs under two contrasting circumstances:

1. *Self-organisation of a predefined signal space.* Given a randomly determined diversity of signals, can linguistic evolution untangle this structureless signal space?
2. *Construction and self-organisation of signal space.* Starting from an empty mapping between meanings and signals, can linguistic evolution construct a compositional language through an interaction of the contrasting pressures to introduce randomness, and introduce compressible signals?

As I have shown, both these question are answered affirmatively. Furthermore, these experiments demonstrate that compositional structure can emerge under the constraints applied by the minimum description length principle. These results are important, and validate the conclusions of Chapter 5, an issue I will return to shortly. At this point in the discussion, it is worthwhile examining the precise nature of the evolved languages.

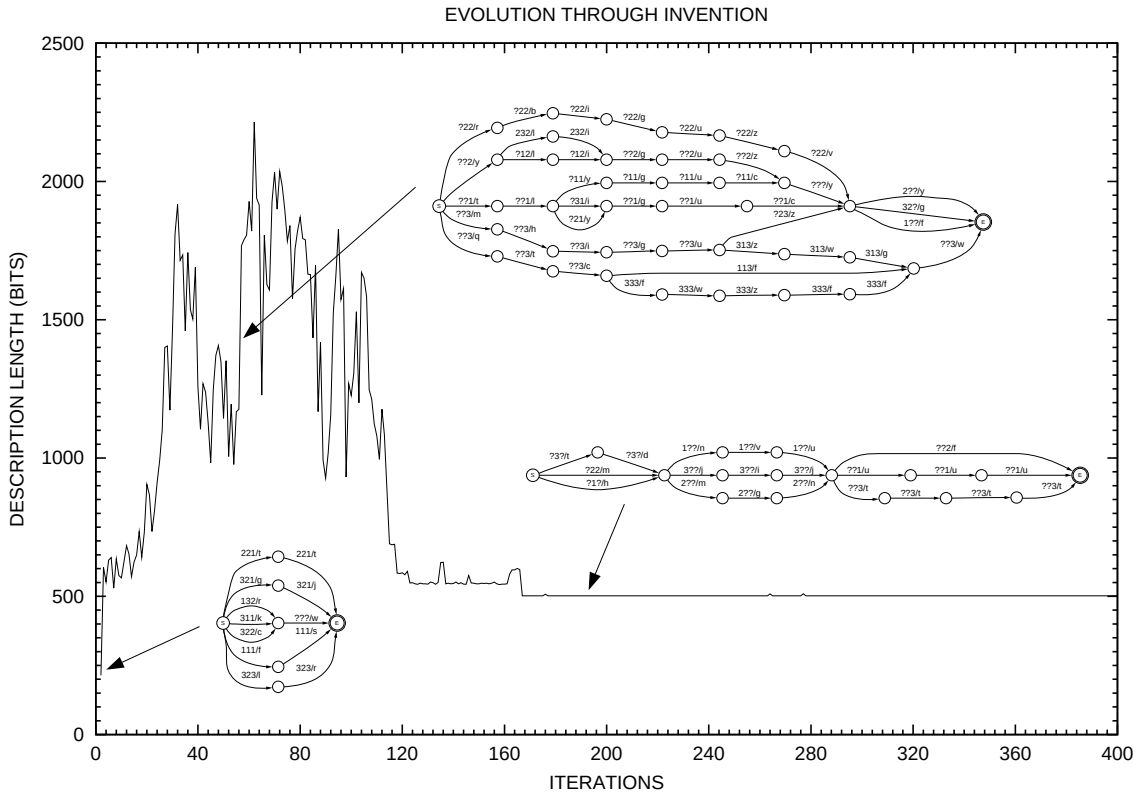


Figure 6.12: Three phases in linguistic evolution. This plot shows how minimum description length rises as random signals are introduced. Self-organisation then results, and compression leads to structured language emerging. Three transducers are shown that illustrate these transitions.

6.6 Analysing evolved languages

Stable languages must pass through each agent unscathed; the agents act as a conduit, compressing the language they encounter and reconstructing it through production. The hypothesis induced must faithfully represent the whole language, even though only a subset of the language has been observed. Compression occurs as a result of applying the minimum description length principle: a bias exists for selecting the smallest hypothesis while at the same time taking the cost of encoding the observed data into account. We might then conclude that the languages that evolve are *optimal*, and represent the most efficient way of forming signals for the meanings. The results so far suggest a quite different scenario: the evolved languages are far from optimal. These results are in contrast to the assumptions of Chapter 5, which took an analytic and idealised view of linguistic evolution. Although valuable results can be gleaned from such static analysis, fundamental complexities arise as a result of full blown simulation modelling. In this section I

will examine the evolved systems, and discuss the complexities that arise from the simplicity principle.

6.6.1 Optimality and redundancy in evolved systems

As the static analysis suggests, the optimally compressed transducer, T_{min}, given the parameters F, V, and c is one where each signal is F characters long. If the signals were any shorter, then they would be incapable of being related to all F features individually. If the signals were any longer, then some features would be represented by more than one symbol, and would therefore be sub-optimal. Furthermore, to minimise the size of the alphabet, only V characters are required. Given such an alphabet, each feature value can be discriminated within the signal. There is no reason to stipulate different sub-signals are required for different features: homonymy between features plays no role in the ability of MDL to induce a compressed transducer. The structure of the notional T_{min} is therefore well defined. It is the smallest transducer possible given the parameters F and V.

None of the evolved systems resulting from the simulation model are in harmony with the idealised T_{min}. All evolved systems exhibit redundancy. Figure 6.13 depicts some representative examples of transducers that define stable evolved languages. These transducers represent a diverse set of structural characteristics. Focusing on a single feature, the following structural characteristics are observed:

Optimal coding. A single symbol is used to refer to a single feature value for every feature value within the feature. This, as noted above, is the defining characteristic of the optimal transducer T_{min}. For example, the first feature of the transducer depicted in Figure 6.13(a).

Sub-optimal coding. As many as four symbols are used to represent each feature value. Feature value coding is therefore sub-optimal. Cases where both sub-optimal and optimal feature value encodings also exist. That is, variable length feature coding for a single feature. Figure 6.13(c) shows a transducer containing a sub-optimal coding of the second feature, and Figure 6.13(d) illustrates a coding for the second feature containing both sub-optimal and optimal coding.

Intra-feature redundancy A subset of the feature values are prefixed or suffixed by signal fragments carrying no meaning. These signal fragments appear to offer no help in discriminating between feature values. In this respect, they are truly redundant. Evidence of a prefix signal fragment, for the third feature, is

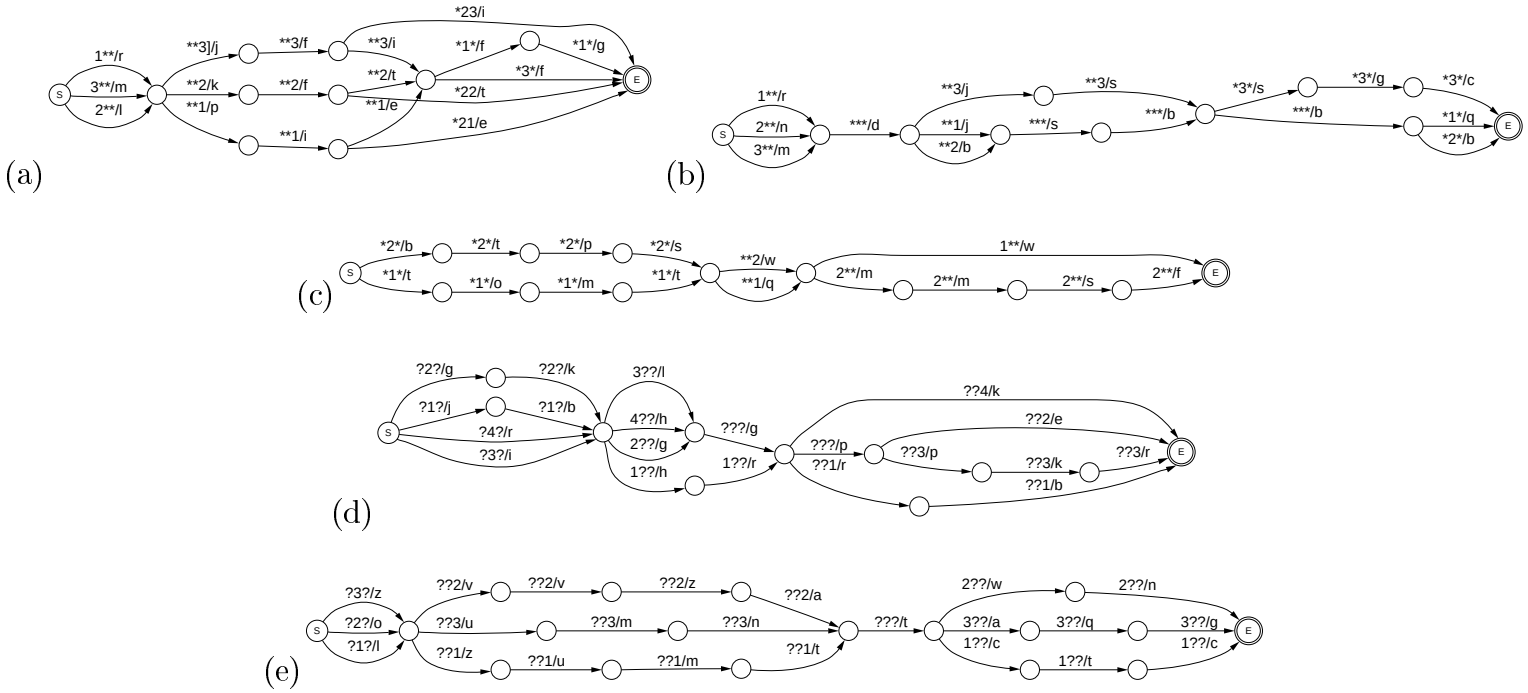


Figure 6.13: Example transducers representing stable compositional languages.

shown in Figure 6.13(d). A signal fragment representing a suffix is also shown in Figure 6.13(d), in the first feature value encoding.

As well as redundancy within a feature specification, redundancy also occurs *between* regions of the transducers representing features:

Inter-feature redundancy. Where signal fragments carrying no meaning lie between two sections of the transducer, each one describing a feature. Examples of inter-feature redundancy can be seen in Figure 6.13(b) and (e).

After examining these transducers, it should also be evident that the ordering of features within the transducer is arbitrary. Until now, for the sake of brevity, this possibility has not been considered; the representation of feature values within the transducer has been assumed to be mirror the feature ordering in the meanings. Alternative orderings have not been discussed, but the model accommodates them. In an abstract sense, this variation reflects word order variability in natural language.

Transducers as grammars

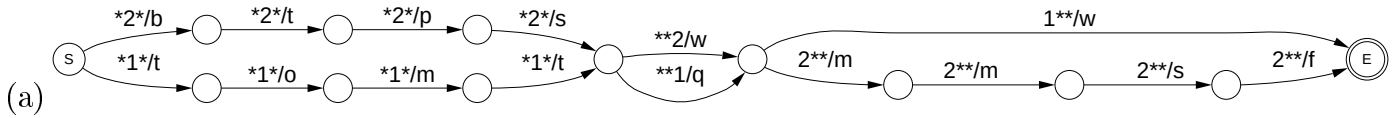
Transducers represent grammars. Perhaps a clearer insight is gained when considering such models of the evolved languages. I will consider two transducers, depicted in Figure 6.14(a-b), and their grammars G_1 and G_2, respectively. The grammar G_1 represents the mapping between meanings and signals for a meaning space defined by $F = 3$ and $V = 2$:

G_1: S/[x,y,z] → A/y B/z C/x

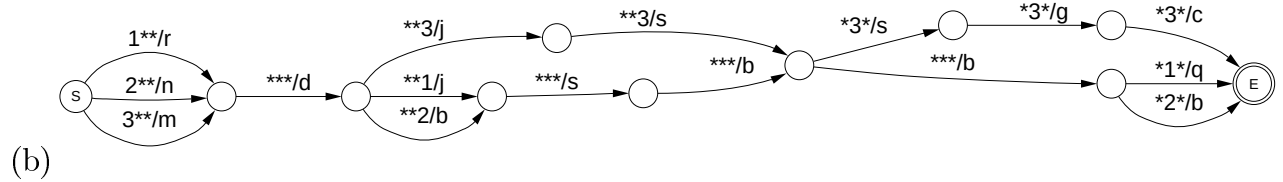
A/1 → tomt	B/2 → q
A/2 → btps	C/1 → w
B/1 → w	C/2 → mmsf

The third feature in the meaning, z, is coded optimally using a single symbol for each feature value. The second feature, y, exhibits sub-optimal coding as four symbols are used to represent each feature value. The first feature, x, is also coded sub-optimally, with variable length codes being used. Apart from these observations, however, the transducer conforms to the general structure of the compressed transducers introduced in Chapter 5.

The grammar G_2 represent a language for a meaning space defined by $F = 3, V = 3$:



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

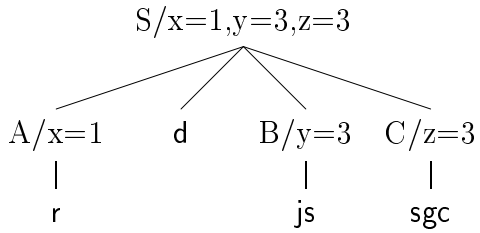


G_2: S/x,y,z → A/x d B/y C/z
A/1 → r
A/2 → n
A/3 → m
B/y → D/y sb
B/3 → js
D/1 → j
D/2 → b
C/3 → sgc
C/z → b E/z
E/1 → q
E/2 → b

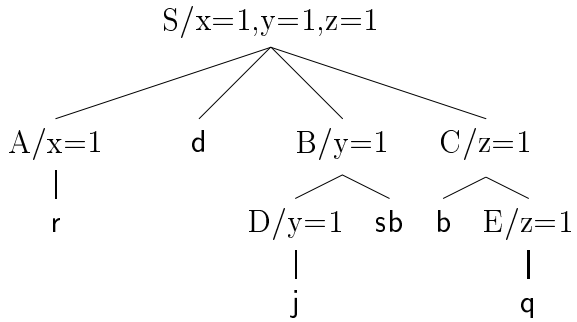
Figure 6.14: Two evolved languages. (a) Shows a transducer, and the corresponding grammar, containing redundant transitions, variable length signals, and several syntactic categories. (b) shows a language with variable length substrings.

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

Here, as before, both optimal and sub-optimal codings are used to represent feature values. Both intra- and inter-feature redundancy is also present. Consider the parse tree for the meaning $[1, 3, 3]$:



Inter-feature redundancy is illustrated by the presence of the d symbol in-between the categories A and B. All signals will therefore contain this redundant signal fragment; it contributes absolutely nothing to the mapping between meanings and signals. In contrast, only some meanings will exhibit intra-feature redundancy. Consider the parse tree for the meaning $[1, 1, 1]$



This parse tree illustrates the presence of intra-feature redundancy, where both the second and third feature are coded using extra symbols.

6.6.2 Extending the language taxonomy

The evolved languages, those languages we observe to be stable, are adaptive within the context of iterated learning. In Chapter 2, when discussing the iterated learning

process, rather than a specific model, I defined the set of observed communication system, $C_{adaptive}$, to be those we observe as a result of adaptive pressures introduced by iterated learning. I noted that these systems are a subset of the set of possible communication systems $C_{possible}$, defined by the biological machinery supporting language.

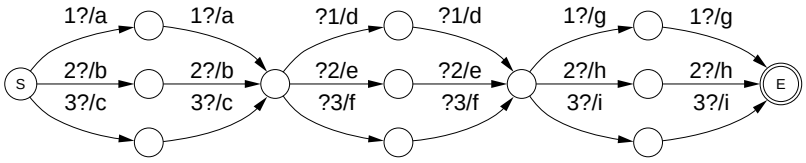
In the context of this model, the set of transducers we observe, say, $T_{adaptive}$, corresponds to the set $C_{adaptive}$: they are adaptive in the model of iterated learning I have developed. At this point, it is worth considering the set $T_{possible}$, the set of possible transducers. To do so, imagine all possible compositional language structures. Certain structural designs have already been seen, and reside in $T_{adaptive}$: these consist of the basic optimal transducer T_{min}, defined above, in conjunction with certain augmentations such as variable length signals, and redundancy occurring between features and within features. I will now consider the set $T_{possible} - T_{adaptive}$, which requires us to ask: Which compositional language designs are not adaptive within the model?

Consider the transducer T_{double} depicted in Figure 6.15(a). This transducer represents a compositional mapping between a two dimensional meaning space and signal space, but it differs from all previously discussed transducer structures. The first feature is represented twice in the signal, which means that two regions of the transducer must agree when expressing each meaning. The language represented by this transducer is shown in Figure 6.15(b).

In order for this transducer to emerge, it must be stable. This means that given the language L_{double}, T_{double} must be induced. But this is not the case, rather the transducer T'_{double} is induced, which is shown in Figure 6.15(c). Unlike T_{double}, T'_{double} has failed to generalise from the input, and therefore will not be stable. Why is this? T'_{double} represents a local minimum in the space of FSUTs consistent with L_{double}. The fact that T'_{double} is not induced is therefore a failure on the part of the search strategy. According to the MDL principle, T_{double} *should* be induced, but the bias present in the search strategy represents a constraint on the set of hypotheses that can be learned. In particular, in order for T'_{double} to be further compressed, a long series of compression operators need to be applied with the intermediate transducers all leading to a greater MDL than that of T'_{double}. Only at the end of this series of operator applications will T_{double} be arrived at.

A further exploration of the set $T_{possible} - T_{adaptive}$ yields transducers of the form shown in Figure 6.16(a). The structure of this hitherto unconsidered transducer,

(a) T_{double}:



(b) $L_{double} = \{\langle\{1,1\}, \mathsf{aaddgg}\rangle, \langle\{1,2\}, \mathsf{aaeegg}\rangle, \langle\{1,3\}, \mathsf{aaffgg}\rangle$
$\langle\{2,1\}, \mathsf{bbddhh}\rangle, \langle\{2,2\}, \mathsf{bbeehh}\rangle, \langle\{2,3\}, \mathsf{bbffhh}\rangle,$
$\langle\{3,1\}, \mathsf{ccddii}\rangle, \langle\{3,2\}, \mathsf{cceeii}\rangle, \langle\{3,3\}, \mathsf{ccffii}\rangle\}$

(c) T'_{double}:

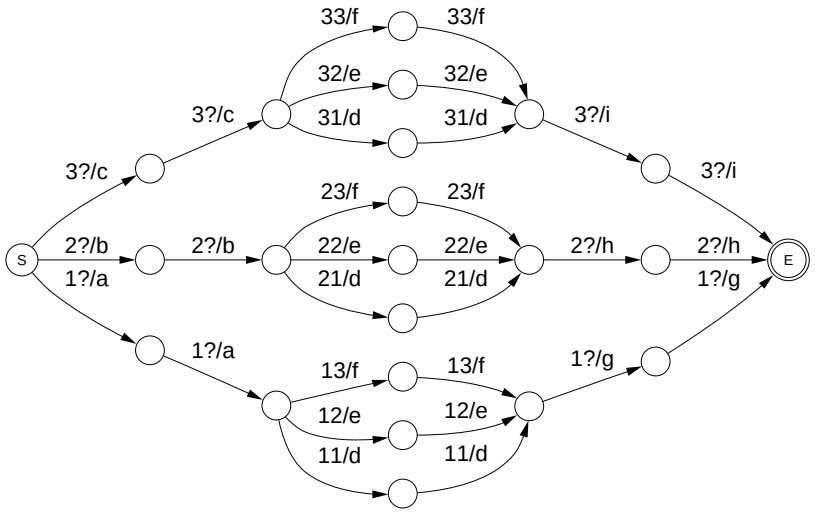
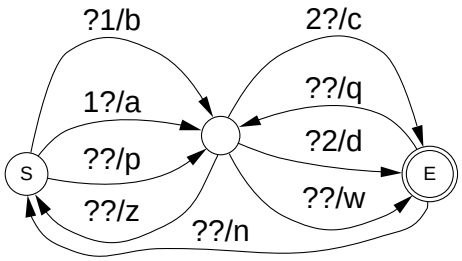


Figure 6.15: Non-adaptive transducers. In (a), T_{double} is shown, which is a member of $T_{possible} - T_{adaptive}$. In (b) the language generated by T_{double} is depicted, and denoted by L_{double}. Rather than L_{double} leading to the induction of T_{double}, it instead leads to the induction of T'_{double} shown in (c). This is why T_{double} is not adaptive.

$T_{unordered}$, is curious: the stipulation that distinct parts of the transducer code distinct features is breached. Relying on transitions with unspecified meanings, feature values for each feature are located in different parts of the transducer. Several transitions with unspecified meanings need to introduced in order for the transducer to express all meanings in the meaning space. As a result of these extra transitions, the transducer can express an infinite number of signals for a given meaning. Figure 6.16(b) shows the beginning of the infinitely large language generated by $T_{unordered}$, denoted as $L_{unordered}$.

(a) $T_{unordered}$:



(b) $L_{unordered} = \{\langle\{1,1\}, \mathsf{azbw}\rangle, \langle\{1,2\}, \mathsf{ad}\rangle, \langle\{1,3\}, \mathsf{pcnbw}\rangle\langle\{2,1\}, \mathsf{pcqd}\rangle, \ldots\}$

(c) $L'_{unordered} = \{\langle\{1,1\}, \mathsf{azbw}\rangle, \langle\{1,2\}, \mathsf{ad}\rangle, \langle\{1,3\}, \mathsf{pcnbw}\rangle\langle\{2,1\}, \mathsf{pcqd}\rangle\}$

(d) $T'_{unordered}$:

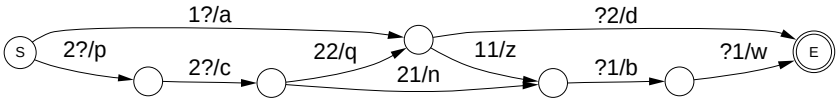


Figure 6.16: A non-adaptive transducer. The language generated by the transducer $T_{unordered}$ is not stable under the conditions of a transmission bottleneck.

$T_{unordered}$ is unstable for a two reasons. First, the transducer cannot be induced from impoverished data, several examples of signals for each meaning need to seen before the empty transitions are induced. Second, this situation requires that language on which the transducer is induced exhibits a high degree of synonymy. This is not possible due to the production bottleneck. For example, Figure 6.16(d) is induced from an impoverished sample of the language $L_{unordered}$, $L'_{unordered}$.

Determinants of $T_{possible} - T_{adaptive}$

In general context of human linguistic evolution, a number of factors will determine the set $C_{possible} - C_{adaptive}$. In these simulations, two categories of constraint influence the membership of $T_{possible} - T_{adaptive}$. First, constraints on learning. Deficiencies in the search of the hypothesis space exclude transducers such as T_{double}. Second, constraints on transmission such as the presence of semantic and production bottlenecks. $T_{unordered}$ is an example of a transducer excluded by transmission constraints.

6.7 Stability conditions and linguistic evolution

This Chapter has progressed the model of linguistic evolution developed in Chapters 4 and 5. Rather than an analytic argument based on abstractions, intuitions about linguistic evolution have been explored through computational modelling. The key advance is to show, through a practical casting of the model developed so far, that linguistic evolution is possible given the MDL model of induction. Before closing this Chapter with a discussion, I will investigate the degree to which the computational simulation model agrees with the analysis of Chapter 5. Given the set of assumptions underlying the the stability conditions developed in Chapter 5, one pertinent question is the following: Do these stability considerations hold true in the full simulation model?

First of all, consider the parameters used in Section 6.4, where I first demonstrated the linguistic evolution of compositionality. These experiments were based on a meaning space defined by $F = 3$ and $V = 4$. The transmission bottleneck (R) was set at 32 observations, which translates to an expected meaning space coverage of 0.396. Using the results developed so far I will now show how one of the fundamental results of Chapter 5 holds true. To do so, I will run an experiment with a slightly less complex meaning space. Using a meaning space defined by $F = 3$ and $V = 3$, I will conduct an otherwise identical simulation. Critical to the coming argument, both experiments have the *same* expected meaning space coverage. The key observation will be that, under these revised conditions, linguistic evolution fails to converge on a stable compositional language. This result will illustrate the importance of meaning space structure. The greater the degree of complexity in the meaning space, the higher the rate of feature value coverage, given the same degree of meaning space coverage. What does this mean? It means that for a given meaning space coverage, we should expect linguistic evolution toward compositionality to be more likely in

the presence of a rich meaning space structure. This result is fundamental, and was motivated by the discussion of optimal generalisation.

So, to recap, the next experiment illustrates the importance of meaning space complexity by performing an experiment with the same expected meaning space coverage to experiment in Section 6.4, but with a less complex meaning space. These two results will represent points either side of the minimal conditions required for robust linguistic evolution.

6.7.1 Narrow bottlenecks, low coverage

With a meaning space defined $F = 3$ and $V = 3$, 27 meanings are possible. To achieve an expected meaning space coverage of 0.4, which is slightly higher than that expected in the experiments detailed in Section 6.4, a bottleneck of 14 random observations is required.

In this experiment, the simulation will start from a blank slate. Both the compressor scheme and random invention will therefore be used at each generation. Over 1000 iterations, Figure 6.17(a) traces the degree of expressivity achieved by the transducer at each iteration. The expected minimum expressivity value of $(0.4 \cdot 27) \approx 11$ (representing the expression $c \cdot V^F$), which is the degree of expressivity one would expect in the absence of generalisation, represents the average expressivity achieved for the first 950 iterations. Compositional language has not evolved. However, on the 954th iteration, maximum expressivity is achieved: a jump from an expressivity level of approximately 11 to an expressivity level of 27 occurs. The linguistic evolution of compositionality has taken place. At this point, Figure 6.17(b) illustrates how the MDL of the emerging language has bottomed-out: a fully compressed transducer has been induced. Soon after this occurrence, the compositional system degrades.

To understand why this has happened, consider the vertical dashed lines in both plots occupying Figure 6.17. These points in the simulation represent events where the feature value coverage sinks below 1. Such an event represents the fact that the set of meanings contained in the observed utterances have failed to represent every feature value. The occurrence of this event indicates that the bottleneck is set too low for consistent linguistic evolution to function. During 1000 iterations, the feature value coverage dips below 1 five times. Point E represents such an event, and this event has the effect of retracting the transition to compositionality occurring just before it. Notice that this situation does not occur (or at least it is

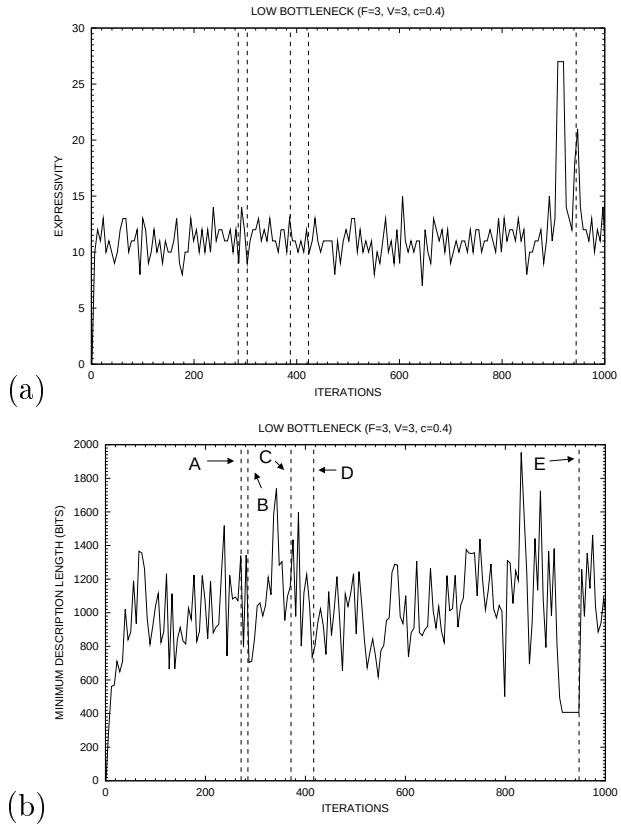


Figure 6.17: Linguistic evolution on the cusp of instability. At low bottleneck values, plot (a) shows that linguistic evolution may occur, but retaining the evolved system may be unlikely. The degradation of evolved systems occurs as a result of the occasional skewing of the observed coverage of feature values. In plot (b), the dashed lines mark out the points where feature value coverage dips below 1. The result, in the case of the event labelled E, is that any evolved compositionality degrades.

very unlikely), in the experiments of Section 6.4, where the same meaning space coverage is expected, but a meaning space of higher complexity exists.

While this observation serves to back up the intuitions of Chapter 5, namely the importance of meaning space structure, it also illustrates a shortcoming. In Chapter 5, conclusions were made on the basis of *expected* values. Over long simulations, the expected behaviour of the system will deviate occasionally. Linguistic evolution at low bottleneck values can therefore place the system on the cusp of instability. Of course, for meaning spaces more complex than those investigated here, the problem will be less severe. The key point is that expected behaviour can be misleading with respect to robust and consistent cumulative evolution.

6.7.2 Wide bottlenecks, high coverage

I have shown how a narrow bottleneck, represented by low expected meaning space coverage, highlights the importance of meaning space complexity. Setting the bottleneck too tight can reverse any progress toward the linguistic evolution of compositional structure. Another fundamental condition for the emergence of compositionality is the very presence of a bottleneck: without a bottleneck, production can function solely on the basis of recalling observed meaning/signal associations, and as result, no state change or evolution will occur.

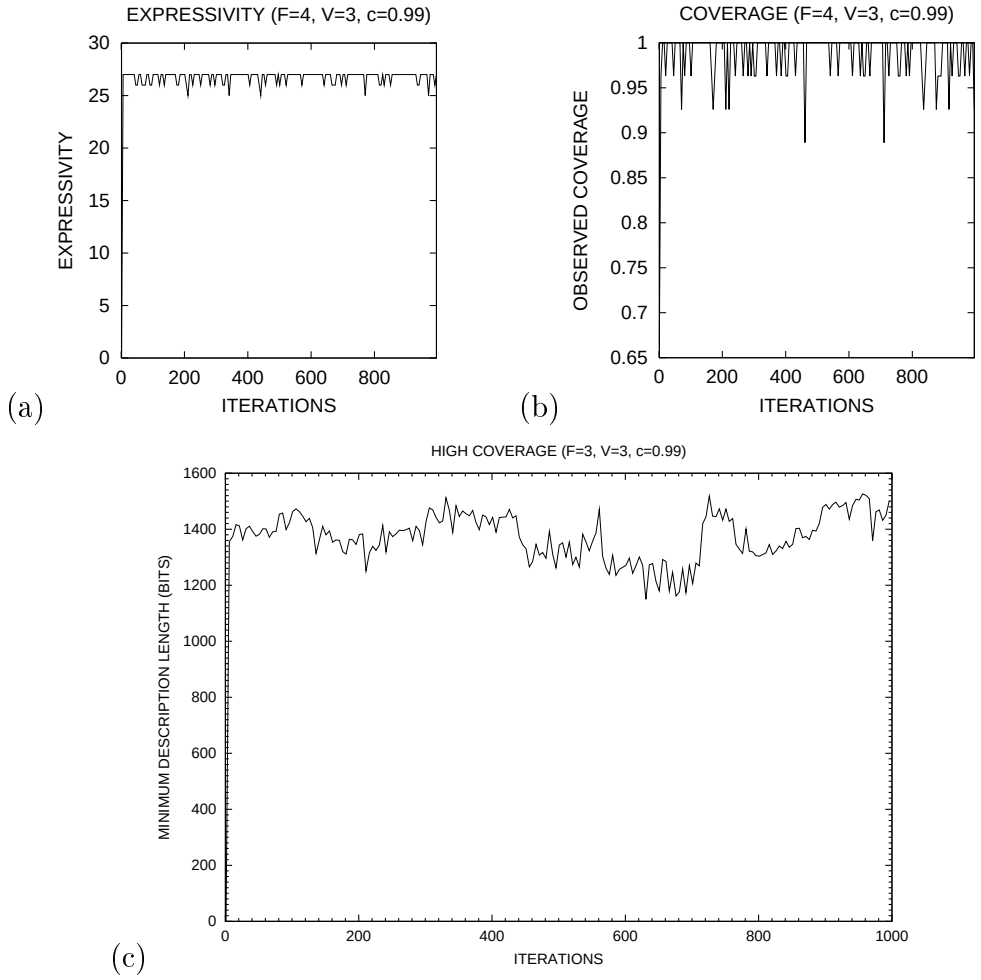


Figure 6.18: Linguistic evolution under conditions of high meaning space coverage. With an expected coverage of 0.99, representing a wide transmission bottleneck, invention will occur infrequently. This fact is highlighted in (a), where high expressivity suggests a low rate of invention. In (b), actual meaning space coverage is shown to be variable, and as a result, invention is called into play. Plot (c) shows a lack of cumulative linguistic evolution, where the minimum description length of the emerging language fluctuates.

The conclusions of Chapter 5 suggest that for large coverage values, compositionality offers a decreased stability advantage over holistic language. How does this observation relate to the models of linguistic evolution in this Chapter? The intuition is confirmed, and Figure 6.18 highlights this fact. With an expected meaning space coverage of 0.99, little cumulative evolution occurs over the course of 1000 iterations. Examining Figure 6.18(a) illustrates that expressivity is consistently reaching the maximum of 27. Figure 6.18(b) illustrates why: meaning space coverage is, as we would expect, consistently high. But as meaning space coverage is occasionally below 1.0, invention will occur. As a result, we would expect that linguistic evolution will take place, only at a much slower rate than previously observed. An interesting observation is revealed by Figure 6.18(c), where over the course of 1000 iterations, the minimum description length of the evolving language does not decrease monotonically, but fluctuates. If the simulation were run indefinitely, however, then we would expect compositionality to emerge eventually. In one sense this is a trivial observation, as compositionality is the inevitable result of evolution over a sufficiently long period, so long as a bottleneck is in place.

These simulation results suggest that a wide bottleneck acts to lengthen the inevitable trajectory toward compositionality. That is, for the rapid linguistic evolution of compositional structure to occur, then the invention pressure imposed by the bottleneck must be high – coverage must be low. Invention introduces a bias toward compression, and therefore compositionality. As long as invention is guaranteed to occur, then a compositional language can be the only stable state. This statement, however, makes a strong assumption. Given a compositional language as input, I have assumed that a compressed transducer will be chosen by MDL, and as a result, generalisation will occur and stability will result. Drawing on the analysis of Chapter 5, there are circumstances in which this will not happen. In the limit, given enough data, MDL will refrain from recommending compressed transducers as the data encoding length will dominate the grammar encoding length. The power of the model developed in this Chapter has therefore only been touched on. With a non-uniform, statistically skewed source of utterances, the results of this Chapter may be different. Although this model has not been explored fully in this respect, it has, however, helped to establish some fundamental results that strengthen the impact of existing models of linguistic evolution. At this point, any further progress toward understanding the process of linguistic evolution requires some discussion.

6.8 Discussion

6.8.1 Occam's razor and MDL

Previous models focusing on the linguistic evolution of compositional structure stress the importance of induction from data. In his introduction to a recent edited volume of computational models, Briscoe notes that "contributors all develop learning algorithms which [...] incorporate Ockham's Razor in some form; that is, a broad preference for the *smallest* grammar and/or lexicon" (Briscoe 2002:9). This observation is consistent with computational models of linguistic evolution in general. The model developed here is important because it relies on a learning algorithm that does not necessarily conform to Occam's razor. By invoking the minimum description length principle, the hypothesis chosen will not *a priori* be the smallest. The fact that, during the previous discussion, the hypothesis chosen *has* been the smallest hypothesis consistent with the data does not detract from the model. Instead, the results presented so far serve to bolster the validity of previous models. With the exception of the simulation by Kirby (2001), previous models of linguistic evolution have assumed a uniform distribution over communicatively relevant meanings. The vast majority also assume the presence of a transmission bottleneck. In light of these observations, the experiment I have developed justifies the widespread deployment of Occam's razor. The discussion in Chapter 5 suggests that the policy of Occam's razor is cast into doubt when meanings are observed subject to a non-uniform distribution. But is this situation worthy of consideration? The work of Alison Wray, mentioned in Chapter 4, suggests that it is. If holistic utterances are also the most frequent utterances, then the relationship between the frequency of meanings and the structure of the resulting utterances is likely to linked. Frequency effects may be a strong determiner of the pressures leading human language to contain both holistic and compositional utterances.

6.8.2 Degrees of bias

When considering models of linguistic evolution, a common concern is that agents are heavily biased toward compositional structure: the emergence of compositionality is therefore unsurprising and not particularly illuminating. When analysing the model of Kirby (2002a), as I noted in Chapter 3, Tonkes and Wiles observe "[i]t seems to us that the chosen induction algorithm is highly biased towards language-like, compositional structures" (Tonkes & Wiles 2002). Of course, it is a basic fact that learning requires bias. Any criticism, as Tonkes and Wiles readily accept, must

focus on the *degree* of bias. But how can this debate enlighten our understanding of linguistic evolution? The implication of such criticisms is that we should be willing to accept certain learning algorithms and reject others as implausible as a result of being too biased. On what basis acceptability should be judged is not clear. Nevertheless it is worth spending some time discussing the extent of the bias present in the model I have developed. Broadly speaking, three issues are relevant here:

1. *Bias in the hypothesis space*: By using the minimum description length principle, the choice of hypothesis space becomes critical. In the limit, it is possible to smuggle in any kind of bias we wish through engineering the hypothesis space. This should not be considered a weakness in the model, but rather a point of clarification. To argue the model I have developed is "heavily biased" requires making a claim about the implausibility of the hypothesis space.
2. *Bias in hypothesis selection*: Without bias, a learning algorithm can only induce the hypothesis that describes the observed data and only the observed data. Under certain circumstances, however regular the observed data, MDL will recommend such a hypothesis. For this reason, MDL can be considered less biased than a hypothesis selection principle such as Occam's razor. Occam's razor, given highly structured and compressible data, will always propose the smallest hypothesis. The result of this policy, when given structured data, is generalisation *whatever* the circumstances. In this respect, the model I have presented will, all things being equal, be less biased than an algorithm that adopts Occam's razor principle.
3. *Consistency bias*: Should the result of learning always be consistent with the observed data? What we might term *lossy* learning occurs when a generalisation is made that explicitly contradicts the observed data. The model I have developed is always consistent, unlike, for example, the models developed by Batali (2002) and Smith (2003c). Whether or not lossy learning is a more plausible model of language learning in humans is another issue. The point here is that lossy learning must, all other things being equal, be more biased than consistent learning.

These points are not meant to amount to a claim for lack of bias in the model I have developed. Rather, they relate to presenting a convincing case that the issue of bias has been transfered, primarily, to a consideration of the hypothesis space. One of the aims of the model I have developed is to make any bias explicit, and therefore readily understandable.

The issue of lossy learning, in contrast to consistent learning, is a problematic one. Humans undoubtedly indulge in deviating from what the the linguistic evidence they observe suggests. The model of consistent learning I have developed, on the other hand, is not meant as a plausible model of human linguistic behaviour. Rather, the model acts to explain the adaptive properties of a well-understood and rigorously justified model of learning, namely MDL hypothesis selection. One result of this approach is that residue is likely to persist, rather than be filtered by a biased and lossy learner. This issue is worthy of further discussion.

6.8.3 Imperfection and frozen accidents

Directly as a result of the introduction of randomness, either through the process of random invention, or the initial random state of the system, the stable compositional systems that emerge from these models contain the residue of the randomness. Sub-optimal coding of feature values, and redundant transitions seem to occur *without fail*. How significant is this observation? It is not clear how these "imperfections" relate to the imperfections Chomsky seeks to explain through the minimalist program. In relation to this observation, it should be noted that I am not claiming that *all* the perceived imperfections of language are the result of residue. For example, Chomsky's minimalist explanation may tell part of the story. Disregarding this important open question, these experiments do highlight the fact that simplicity in induction does not imply simplicity in the evolved states. The model of linguistic evolution I have presented does, however, illustrate that justifiable and consistent learning will not wash out remnents of randomness.

This observation is related to the concept of a *frozen accident*, often invoked in complexity theory (Gell-Mann 1992; Kauffman 1993). The fact is that these experiments consistently lead to evolved systems containing imperfections that can properly be regarded as frozen accidents. To be absolutely clear on this point, imperfections in language may be partly explained as frozen accidents, but it is unlikely they are solely explainable in these terms. It is useful at this point to distinguish between imperfections in language in the sense used by Chomsky, which generally suggests some superfluous structural role (that, nevertheless, fulfils some function), and the imperfections we observe in the simulation model. The imperfections evident in the evolved compositional language of the model I will term imperfections in the *narrow sense*.

Interestingly, compositional structure is suggested by these models to be a prime example of a frozen accident. Or, more accurately, the result of a series of frozen accidents; each one building on a cumulative base of structure that precipitates generalisation. Accordingly, two forms of frozen accident are present in the evolved systems I have discussed. First, frozen accidents that contribute to compressibility, and second, frozen accidents that do not. Compositional structure is an example of the first kind, imperfections (in the narrow sense) are an example of the second.

6.8.4 Future developments

Before closing this Chapter, it is worth contemplating possible extensions to the model. Four areas of future development immediately suggest themselves:

1. *Non-uniform distributions*: As I have discussed previously, the probability distribution over the set of meanings is likely to have a profound effect on the hypothesis selection process. It is hard to argue against the fact that users of language communicate some meanings more often than others. Will the frequent utterances be those lacking compositional structure? Is Alison Wray correct in her assertion that holistic language results in "removing the burden of the everyday, pragmatically determined and communicably predictable, leaving the way clear for the more demanding analytic system to achieve the goals that only it can?" (Wray 1998:63). The model developed is poised to investigate these questions.
2. *The effect of alphabet size*: The size of the alphabet from which symbols are drawn has not been the target of investigation. It is quite reasonable to imagine that the diversity available from the choice of alphabet could have an influential impact on the evolved systems.
3. *Invention bias*: As the progression of this Chapter suggests, the nature of the evolved systems is heavily determined by the invention scheme adopted. I have investigated the impact of one successful invention scheme. Undoubtedly there are others, and further investigation needs to be carried out.
4. *Avoiding local minima*: Section 6.6 highlighted the fact that deficiencies in the hypothesis selection procedure can be rule out certain kinds of language structure. I have assumed that, for example, features associated with disjoint sub-signals would be adaptive if the search procedure mirrored MDL hypothesis selection precisely. These assumptions need to be tested to assess the adaptive properties of varying language structures.

6.9 Chapter Summary

This Chapter has seen the application of the model developed in Chapters 4 and 5. Rather than a static analysis, full dynamic simulations have been carried out. First of all, the insights suggested by the static analysis have been verified. On top of this strengthening the results of Chapter 5, the computational simulations have resulted in evolved systems that are sub-optimal, and contain varying forms of redundancy. In order to construct a full simulation model, several of the assumptions of Chapter 5 have been addressed and tested.

First, MDL hypothesis selection, rather being applied precisely as theory dictates, has been carried out automatically and subject to approximation. This required a search through the hypothesis space. Using compression operators in conjunction with a beam search, MDL hypothesis selection has been approximated. The search is not perfect, and this fact is borne out by the fact that certain compositional structures cannot be induced.

Second, invention, an issue skirted around until now, has been discussed and investigated. The partially random invention scheme I considered initially failed to lead to the evolution of compositional structure. To address this problem, I have developed an invention scheme that maximises the probability of introducing signals that facilitate compression. Given an initial state of randomness, this invention scheme leads to the linguistic evolution of compositional structure. One deficit in the invention scheme is its reliance on an initial state containing sufficient signal diversity. To combat this problem, I deployed both random invention and the compressor scheme together. The problems found previously with the random invention scheme were alleviated when used in conjunction with the compressor scheme, and the result was the consistent evolution of compositional structure from an initially blank hypothesis.

Third, I have analysed the evolved systems and found that the simplicity principles underlying both induction and invention led to compositional systems which contain sub-optimal feature codings and redundant transitions. I have explained these anomalies as frozen accidents, and noted that this is one possibility in explaining imperfections in language.

Fourth, I briefly related the simulation model to issues of a bottleneck size. I demonstrated that for consistent evolution, the assumptions of Chapter 5 had to be weakened slightly, as small deviations from expected values can block linguistic

evolution. Large bottlenecks lead to slowing down of the evolution process. The absence of a transmission bottleneck excludes the possibility of linguistic evolution.

To conclude, the impact of this model in a wider context is to show that Occam's razor is entirely justifiable and in line with hypothesis selection by minimum description length when certain conditions are met. These conditions are those that are widely modelled, such as the presence of bottleneck and a uniform distribution over communicatively relevant meanings. The future development of the model, however, can shed light on the linguistic evolution in non-uniform environments. Of the several avenues of further investigation, this is possibly the most important, and one for which the model is well poised.

Chapter 7

Conclusions

7.1 Introduction

It is important to relate the insights suggested by the models I have developed to the process of iterated in learning in a wider sense. First, I will interpret the key results of this thesis and frame them in a plausible evolutionary context. Second, in light of the results I have outlined, I will revisit the three hypotheses introduced in Chapter 2. This concluding discussion will aim to justify the claim that simplicity can be considered a fundamental driving force in linguistic evolution.

7.2 Summary of results

The models I have developed suggest several conditions that must met in order for linguistic evolution to occur. If these results are to carry any explanatory force, then they must be related to existing explanations for the emergence of linguistic structure. The aim of this Section is to strengthen the results developed so far by considering these results in light of related work. In doing so, the three hypotheses outlined in Chapter 2 can be approached in a more concrete fashion. Ultimately, the aim is to relate the models to the issue of linguistic universals and linguistic nativism.

7.2.1 Stability conditions

Semantic complexity

It would seem a basic fact that in order for an organism to process a compositional language, it must possess some minimal degree of semantic structure. This is both a requirement of the definition of compositionality, and an observation that the models I have developed imply. Furthermore, beyond the requirement of some basic level of semantic structure, the results of Chapter 5 suggest that certain degrees of semantic complexity make the linguistic evolution of compositional structure more likely.

When considering the linguistic evolution of compositionality, within certain bounds, more semantic structure is better than less semantic structure. A general justification for this statement runs as follows. Because generalisation functions on the basis of observed regularities in the data, then the more likely that these regularities have been fully observed, the more likely generalisation is to occur. With increased semantic structure, more regularity will be observed for some fixed set of observations. On this basis, we can argue that generalisation thrives on regularity, so more regularity leads to increased degrees of generalisation. In the experiments I have conducted, this argument has been presented as one where, given a certain degree of *coverage* of the meaning space, generalisation will be greater in the case of increased semantic complexity.

One consequence of this observation is that some level of semantic complexity must pre-exist the conditions under which cumulative linguistic evolution occurs. This is not to say that semantic complexity evolved and then linguistic evolution began, but rather in order for cumulative linguistic evolution to occur, some *base* of semantic complexity must be present. Whether or not this base has evolved in any way subsequently or co-evolved alongside linguistic evolution is another issue. In short, the results require the pre-existence of a semantic base. This conclusion is closely related to the work of Schoenemann, where he states: "I would argue that most cognitive categories are independent of, and exist prior to, language itself." (Schoenemann 1999:319). It is this semantic complexity that the parameters F and V attempt to model. The abstract environment that the agents "perceive" is divided up according to the number of number channels – or features – and the acuity of these channels is determined by the number of feature values per feature. It is clear this is the kind semantic complexity that Schoenemann imagines:

> Different organisms will divide up the world differently, in accordance with their unique evolved neural systems, but they all divide up the world in *some* way. These "divisions" are what I mean by "cognitive categories," or "semantic units." Increasing semantic complexity therefore refers to an increase in the number of divisions of reality which a particular organism is aware of and can respond to in some meaningful way. (Schoenemann 1999:318)

Furthermore, Schoenemann's argument leads to the conclusion that "syntax is more properly understood as an emergent characteristic of the explosion of semantic complexity that occurred during hominid evolution." (Schoenemann 1999:309). This argument requires us to consider why humans have language and, for example, other primates do not. Schoenemann believes that humans are alone in their possession of a sufficiently high degree of semantic complexity:

> Humans are by far the most highly encephalized of all primates, which would suggest that we also have the most complex and varied reconstruction of external reality. This is simply another way of saying that hominids experienced a massive increase in semantic complexity beginning at least 2.5 million years ago with the first unequivocal increases in hominid encephalization. (Schoenemann 1999:321)

The reason for this increase may not be related to language, for example, an argument along the lines proposed by Dunbar proceeds by noting that hominids began to form large groups sizes that in turn lead to a requirement for comprehending complex social relationships (Dunbar 1996).

The predictions made the models I have developed relate directly to Schoenemann's argument. One side-effect of this increase in semantic complexity is the increased likelihood that the set of possible meanings an organism can accommodate, or treat as communicatively relevant, will increase at an alarming rate. This observation is directly related to the next condition identified by the models for the linguistic evolution of compositional structure.

Stimulus poverty

What we might term *stimulus poverty* is controlled in the models by the severity of the transmission bottleneck. The meaning space defines a range of distinct meanings that an agent could be called to produce a signal for. In all the experiments

I have conducted, this range exceeds the number of meaning/signal pairs on which the learning process is based. The results of Chapter 5 indicate that compositional language is more likely to be observed than holistic language under precisely these conditions. In Chapter 6, this result was confirmed. Moreover, because the likelihood of compositionality is greater than that of holistic language, compositional languages will, in the limit, emerge and remain stable. In short, the increased stability payoff of compositional language translates into the eventual emergence and retention of compositional systems.

In terms of an evolutionary scenario, these conditions, where the number of *possible* meanings is far greater than the number of meanings that can feasibly be observed, occur as a result of an increase in semantic complexity. Relatively small increases in semantic complexity, such as an extra feature, or an extra value per feature, will lead to a large increase in the number of possible meanings. This increase is necessarily a super-linear function of the number features or values added. Schoenemann's discussion of the increase in semantic complexity therefore implies the introduction of a situation resembling the poverty of the stimulus, especially if we consider his claim that hominids experienced a *massive* increase in semantic complexity.

The models discussed in Chapters 3, 5 and 6 all suggest that stimulus poverty is a driving force for linguistic evolution. Stimulus poverty places an agent in a situation in which it is likely that certain meanings are required to be expressed, and these meanings are likely to not have been observed in conjunction with a signal. So far from representing a problem, the poverty of the stimulus could be a fundamental determinant of the linguistic evolution of compositional structure, that is, as Zuidema (2003) points out, "the poverty of the stimulus solves the poverty of stimulus".

But is the situation represented by the presence of a transmission bottleneck really the phenomenon widely termed as *the poverty of the stimulus*? In other words, is this adoption of this term in the context of linguistic evolution a provocative abuse of the term first introduced by Chomsky?

> There are many arguments for the innateness hypothesis. But the most significant one in Chomsky's writings, and the one that has most affected the field, is the argument from the poverty of the stimulus. (Wexler 2001)

Two broad interpretations of the term exist. First, it refers to the situation in which a learner is faced with the problem of acquiring some grammar G, in light of some body of data D. This first interpretation we can label the *strong* interpretation, as it characterises the situation as one where no learner can recover G solely on the basis of D (Wexler 1991; Wexler 2001). Under this strong interpretation of the term *poverty of the stimulus*, the term has been misused in this thesis, as learners patently can recover hypotheses in light of the sparse data available to them. Of course, as I have noted earlier, whether this strong interpretation is at all likely with respect to human language learning is a currently unresolved empirical question.

Second, the *weak* interpretation of the term discussed by, for example, Marcus (2001), is simply the problem of induction discussed in Chapter 4, applied to the problem of acquiring language. In this weak sense, it is quite justifiable to talk of stimulus poverty in the context of these models. So, by adopting the weak use of the term *the poverty of the stimulus*, I am referring to a situation imposed by (1) an increase in semantic complexity, and (2), a restriction on the size of the body of data available to a learner during language acquisition. In order to induce a hypothesis that allows generalisation, which is a requirement to "solve" the poverty of the stimulus, some inductive bias is required.

7.2.2 Inductive bias, compression, and linguistic evolution

A number of modelling decisions I have made suggest that the bias toward compositional language is less likely than previous models of linguistic evolution. The consistency criterion I have imposed is important. This restriction limits the changes that can be made to an evolving language by the agents in the model. Furthermore, by employing the minimum description length principle, two issues have been addressed.

First, inductive generalisations have in one sense been restricted to those that are "justifiable". The MDL hypothesis selection criterion can fail to lead to generalisation, even if strong regularities exist in the data. Before carrying out these experiments, the occurrence of linguistic evolution under these restrictions was an open question. I have demonstrated that linguistic evolution does occur, although this is partly a result of (1) a uniform distribution over meanings, and (2), the fact that learning is based on sparse data. Setting aside the issue of the probability distribution over utterances, the fact remains that, within a particular simulation run, statistical effects can occur as a result of the random sampling of meanings. We

have seen this situation in practice: low bottlenecks highlight the fact that unrepresentative samples of the meaning space do occur. In relation to this observation, we can conclude that occasionally the data *will* contain deviation from the uniform distribution. The experiments I have conducted show that MDL is not sensitive to these issues, and cumulative linguistic evolution occurs regardless. In short, it is not entirely correct to argue that a uniform distribution over meanings justifies the deployment of Occam's razor principles, as a uniform distribution cannot be guaranteed.

Second, I have appealed to the process of compression as cognitively plausible strategy. Arguably, such a strategy is is ubiquitous is neurological systems. The argument presented in Chapter 4 justifies this claim.

7.3 Three hypotheses revisited

7.3.1 Innateness hypothesis

The innateness hypothesis asserts that humans have a biologically determined set of predispositions that impact on our ability to learn and produce language. Broadly speaking, those aspects of the model that are not culturally transmitted are candidate innate mechanisms. For example, each agent has a meaning space supporting the structured encodings of communicatively relevant situations, a hypothesis space containing the set of possible representations that relate meanings to signals, and the ability to compress representations.

A natural question to ask is the following: Which mechanisms contained in an agent are innate, and are these innate mechanisms language-specific? These are problematic questions. In one respect, they are outside the range of this thesis. The principal concern is the degree to which hallmarks of language can be *explained* as adaptive in the context of linguistic evolution. Nevertheless, one can make a strong case for arguing that the restrictions imposed by the hypothesis space are innate, as is the ability to form signals on the basis of this representation, for example. Although compression, as I have discussed, is likely to be rife in non-linguistic cognition, it does not follow that the *ability* to compress is domain-general. For this reason, the compression of a specific linguistic representation may be an innately specified competence. In short, these questions are fundamental, requiring further investigation. This is why I have been cautious in my criticism of Chomsky's minimalist program.

Although, clearly, this thesis is at odds with Chomsky's view that language acquisition is not a process of inductive generalisation, it is in some respects in agreement with the view that the computational system responsible for transforming between meaning and form is optimal in some sense. Accordingly, I consider there to be *some* overlap in the notion of simplicity driving linguistic evolution and the notion of simplicity invoked in minimalist theorising. In short, the innateness hypothesis I have defined can only be refined through a substantial amount of further research: we simply do not know whether the required representations and processes underlying compression are language-specific or not. But on the other hand, the results of this thesis could be interpreted as a small step toward to justifying the claim that linguistic processing is guided by domain-general cognitive processes.

7.3.2 Situatedness hypothesis

The Situatedness hypothesis claims that *some* language universals cannot be explained on the basis of an understanding of the biological basis for language. I have argued that compositionality – an absolute universal – is an example that serves to support this hypothesis. Now, this assertion can be criticised by arguing that this is precisely what this thesis achieves: in essence I have argued that signal structure reflects the structure of the meaning space. If the meaning space is biologically determined, then the observed features of language will simply reflect variation in how these meanings can be mapped to signals. A detached understanding of the biological basis for language therefore can tell us about absolute universals. This conclusion is tempting, but it fails to appreciate that we need an explanation of *how* signals come to be structured in this way. This is the motivation behind the situatedness hypothesis.

Beyond the conclusions of this thesis, the situatedness hypothesis makes a claim about the nature of explanations of universals in general. Support for this hypothesis requires elaborating the models developed here to include, for example, recursive structure, a consideration of processing constraints, and changes to the language learning environment such as non-uniform distributions over the set of communicatively relevant situations. Another possible source of constraints may come from consideration of language function.

7.3.3 Function independence hypothesis

The function independence hypothesis claims that *some* language universals can be explained independent of the communicative function of language. Agents in the models I have developed aquire knowledge of language on the basis of the linguistic performance of other agents. One could argue that the very act of constructing performance is a communicative act. However, any measure of the utility or effectiveness of this performance is not modelled. In this respect, communicative function is not modelled. It is important to note that I am not claiming that communicative function plays *no* role in determining the structure of language, but to establish this fact is outside the remit of this thesis.

7.4 Key contribution

Finally, the key contribution made by this thesis relates to strengthening the view that language adapts to be learnable, and this pressure restricts language to exhibit certain structural properties and not others. It is these structural properties, possibly among others, that we consistently observe, and as a result, treat as universal properties of language.

I have strengthened this view by invoking a simplicity principle – the minium description length principle – that approximates the notion of Kolmogorov complexity. This contribution is significant as it frames the generalisations made learners as justifiable. The simplicity principle also relates to the general ability of organisms to compress representation. On this basis, I have argued that simplicity should be regarded as a driving force in linguistic evolution.

Bibliography

Aha, D. W., D. Kibler, & M. K. Albert. 1991. Instance-based learning algorithms. *Machine Learning* 6.37–66.

Andersen, Henning. 1973. Abductive and deductive change. *Language* 40.765–793.

Bach, Emmon. 1964. *An introduction to transformational grammars*. Holt, Rinehart and Winston.

Batali, John. 1998. Computational simulations of the emergence of grammar. In *Approaches to the Evolution of Language: social and cognitive bases*, ed. by James R. Hurford, Michael Studdert-Kennedy, & Chris Knight, 405–426. Cambridge: Cambridge University Press.

——. 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, ed. by E. Briscoe, 111–172. Cambridge: Cambridge University Press.

Bickerton, Derek. 1998. Catastrophic evolution: the case for a single step from protolanguage to full human language. In *Approaches to the Evolution of Language: Social and Cognitive Bases*, ed. by James R. Hurford, Michael Studdert-Kennedy, & Chris Knight, 341–358. Cambridge: Cambridge University Press.

——. 2000. How protolanguage became language. In *The Evolutionary Emergence of Language*, ed. by Chris Knight, Michael Studdert-Kennedy, & James R. Hurford, 264–284. Cambridge: Cambridge University Press.

Boyd, Robert, & Peter J. Richerson. 1985. *Culture and the Evolutionary Process*. University of Chicago Press.

Brighton, Henry. 2002. Compositional syntax from cultural transmission. *Artificial Life* 8.25–54.

——, & Simon Kirby. 2001. The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In *Advances in Artificial*

Life (Proceedings of the 6th European Conference on Artificial Life), ed. by J. Kelemen & P. Sosik. Heidelberg: Springer-Verlag.

——, ——, & Kenny Smith. 2003a. Cultural selection for learnability: Three hypotheses concerning the characteristic structure of language. Technical report, LEC, University of Edinburgh. To appear in a volume arising from the Evolution of Language Conference.

——, ——, & Kenny Smith. 2003b. Situated cognition and the role of multi-agent models in explaining language structure. In *Adaptive Agents and Multi-Agent Systems*, ed. by D. Kudenko, E. Alonso, & D. Kazakov. Springer.

——, & Chris Mellish. 2002. Advances in instance selection for instance-based learning algorithms. *Data Mining and Knowledge Discovery* 6.153–172.

Briscoe, E. 2000. Grammatical acquisition: Inductive bias and coevolution of language and the language acquisition device. *Language* 76.245–296.

—— (ed.) 2002. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Brooks, R. A. 1999. *Cambrian Intelligence*. Cambridge, MA: MIT Press.

Bullock, S., & P. M. Todd. 1999. Made to measure: Ecological rationality in structured environments. *Minds and Machines* 9.497–541.

Cameron-Jones, R. M. 1992. Minimum description length instance-based learning. In *Proceedings of the Fifth Australian Joint Conference on Artificial Intelligence*, 368–373, Singapore. World Scientific.

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

——, & P. M. B. Vitányi. 2003a. The generalized universal law of generalization. *Journal of Mathematical Psychology* . in press.

——, & P. M. B. Vitányi. 2003b. Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences* 7.19–22.

Cheney, Dorothy, & Robert Seyfarth. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. Chicago, IL: University of Chicago Press.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

——. 1967. Recent contributions to the theory of innate ideas. *Synthese* 17.2–11.

——. 1972. *Language and Mind*. Harcourt Brace Jovanovich, enlarged edition.

——. 1975. *Reflections on Language*. New York: Pantheon.

——. 1980. *Rules and Representations*. London: Basil Blackwell.

——. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

——. 2002. *On Nature and Language*. Cambridge: Cambridge University Press.

——, & Morris Halle. 1968. *The sound pattern of English*. London: Harper and Row.

Christiansen, Morten, 1994. *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure*. University of Edinburgh PhD dissertation.

——, & J. Devlin. 1997. Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In *Proceedings of the 19th Annual Cognitive Science Society Conference*, ed. by M.G. Shafto & P. Langley, 113–118. Lawrence Erlbaum Associates.

Clancy, W. J. 1997. *Situated Cognition*. Cambridge: Cambridge University Press.

Clark, Robin. 2001. Information theory, complexity, and linguistic descriptions. In *Language Acquisition and Learnability*, ed. by Stefano Bertolo, 126–171. Cambridge: Cambridge University Press.

——, & Ian Roberts. 1993. A computational model of language learnability and language change. *Linguistic Inquiry* 24.299–345.

Cover, T. M., & P. E. Hart. 1969. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* IT-13.21–27.

Cowie, Fiona. 1999. *What's within? Nativism Reconsidered*. Oxford: Oxford University Press.

Croft, W. 1994. *Language Typology*. Cambridge: Cambridge University Press.

Dasarathy, Belur. 1991. *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. Los Alimos, CA: IEEE Computer Society Press.

Dawkins, Richard. 1982. *The Extended Phenotype*. Oxford: Oxford University Press.

Deacon, Terrence W. 1997. *The Symbolic Species*. W. W. Norton and Company.

Dreyfus, Hubert L. 1972. *What computers still can't do*. Cambridge, MA: MIT Press, 2nd edition.

——, & Stuart E. Dreyfus. 1990. Making a mind versus modelling the brain: Artificial intelligence back at a branch point. In *The Philososphy of Artificial Intelligence*, ed. by Margaret A. Boden, 309–333. Oxford: Oxford University Press.

Dryer, Matthew. 1992. The Greenbergian word order correlations. *Language* 68.81–138.

Dunbar, Robin. 1996. *Grooming, Gossip and the Evolution of Language*. London: Faber and Faber.

Elman, J. L., E. A. Bates, M. H. Johnson, A. Karmiloff-Smith, D. Parisi, & K. Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Elman, J.L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48.71–99.

Epstein, Samuel David, & Norbert Hornstein. 1999. Introduction. In *Working Minimalism*, ed. by Samuel David Epstein & Norbert Hornstein, ix–xvii. Cambridge, MA: The MIT Press.

Feldman, J. 2000. Minimization of boolean complexity in human concept learning. *Nature* 407.630–633.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annual Eugenics* 7.179–188.

Gao, Qiong, Ming Li, & Paul Vitányi. 2000. Applying MDL to learn best model granularity. *Aritificial Intelligence* 121.1–29.

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

Gell-Mann, Murray. 1992. Complexity and complex adaptive systems. In *The Evolution of Human Languages*, ed. by J. A. Hawkins & M. Gell-Mann, volume X of *SFI Studies in the Sciences of Complexity*, 3–18. Addison-Wesley.

Gibson, Edward, & Ken Wexler. 1994. Triggers. *Linguistic Inquiry* 25.407–454.

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

Grünwald, P. 1996. A minimum description length approach to grammar inference. In *Symbolic, Connectionist and Statistical Approaches to Learning for Natural Language Processing*, ed. by S. Wermter, E. Riloff, & G. Scheler, number 1040 in Lecture Notes in Artificial Intelligence, 203–216. Berlin, Germany: Springer-Verlag.

Hahn, Ulrike, Nick Chater, & Lucy B. Richardson. 2003. Similarity as transformation. *Cognition* 87.1–32.

Hare, M., & J. L. Elman. 1995. Learning and morphological change. *Cognition* 56.61–98.

Haugeland, J. 1997. What is mind design? In *Mind Design 2: Philosophy, Psychology, Artificial Intelligence*, ed. by J Haugeland, 1–28. Cambridge, MA: MIT Press, 2nd edition.

Hauser, Marc D. 1996. *The Evolution of Communication*. Cambridge, MA: MIT Press.

——, Noam Chomsky, & W. Tecumseh Fitch. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science* 298.1569–1579.

Hellman, Martin E. 1970. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics* 6.179–185.

Hopcroft, John E., & Jeffrey D. Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley.

Hurford, James R. 1989. Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* 77.187–222.

—— 1990. Nativist and functional explanations in language acquisition. In *Logical issues in language acquisition*, ed. by I. M. Roca, 85–136. Foris Publications.

—— 2000a. Introduction: The emergence of syntax. In *The Evolutionary Emergence of Language*, ed. by Chris Knight, Michael Studdert-Kennedy, & James R. Hurford, 219–230. Cambridge: Cambridge University Press.

—— 2000b. 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. Hurford, 324–352. Cambridge: Cambridge University Press.

—— 2002. Expression/induction models of language evolution: dimensions and issues. In *Linguistic Evolution Through Language Acquisition: Formal and Computational Models*, ed. by E. Briscoe, 301–344. Cambridge: Cambridge University Press.

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

Jones, S, R Martin, & D Pilbeam (eds.) 1992. *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press.

Kauffman, Stuart A. 1993. *The origins of order: Self organization and selection in evolution*. Oxford: Oxford University Press.

King, R. D., C Feng, & A Sutherland. 1995. Statlog: Comparison of classification algorithms on large real-word problems. *Applied Artificial Intelligence* 9.289–333.

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

——. 2000. Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, ed. by C. Knight, M. Studdert-Kennedy, & J. R. Hurford, 303–323. Cambridge: Cambridge University Press.

——. 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.

——. 2002a. Learning, bottlenecks and the evolution of recursive syntax. In *Linguistic Evolution through Language Acquisition: Formal and Computational Models*, ed. by E. Briscoe, 173–203. Cambridge: Cambridge University Press.

——. 2002b. Natural language from artificial life. *Artificial Life* 8.185–215.

Kolodner, Janet L. 1993. *Case-Based Reasoning*. San Mateo, CA: Morgan Kaufmann.

Krifka, M. 2001. Compositionality. In *The MIT Encyclopaedia of the Cognitive Sciences*, ed. by R. A. Wilson & F. Keil. Cambridge, MA: MIT Press.

Langley, P., & S. Stromsten. 2000. Learning context-free grammars with a simplicity bias. In *Proceedings of the Eleventh European Conference on Machine Learning*, 220–228. Springer-Verlag.

Lasnik, Howard. 2002. The minimalist program in syntax. *Trends in Cognitive Sciences* 6.432–437.

Li, M., & P. M. B. Vitányi. 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.

Marantz, Alec. 1995. The minimalist program. In *Government and Binding Theory and the Minimalist Program*, ed. by Gert Webelhuth, 349–382. Oxford: Blackwell Publishers.

Marcus, Gary. 2001. Poverty of the stimulus arguments. In *The MIT Encyclopaedia of the Cognitive Sciences*, ed. by R. A. Wilson & F. Keil. Cambridge, MA: MIT Press.

Markman, Arthur B., & Eric Dietrich. 2000. Extending the classical view of representation. *Trends in Cognitive Sciences* 4.470–475.

Marr, D. 1977. Artificial intelligence: A personal view. *Artificial Intelligence* 9.37–48.

——. 1982. *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman.

Matthews, Peter H. 1997. *The Concise Oxford Dictionary of Linguistics*. Oxford: Oxford University Press.

Michalski, R. S. 1992. Concept learning. In *Encyclopedia of Artificial Intelligence*, ed. by Stuart C. Shapiro, 249–259. Wiley-Interscience, 2nd edition.

Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill.

Mundinger, P. 1980. Animal cultures and a general theory of cultural evolution. *Ethology and Sociobiology* 1.183–223.

Newmeyer, Frederick J. 1998. *Language Form and Language Function*. Cambridge, MA: MIT Press.

—— 2002. Uniformitarian assumptions and language evolution research. In *The Transition to Language*, ed. by Alison Wray. Oxford: Oxford University Press.

Nichols, Johanna. 1984. Functional theories of grammar. *Annual Review of Anthropology* 13.97–117.

Niyogi, Partha, & Robert Berwick. 1997. Evolutionary consequences of language learning. *Linguistics and Philosophy* 20.697–719.

Nowak, M. A., & N. L. Komarova. 2001. Towards an evolutionary theory of language. *Trends in Cognitive Sciences* 5.288–295.

O'Grady, William, Michael Dobrovolsky, & Francis Katamba. 1997. *Contemporary Linguistics*. Longman, 3rd edition edition.

Oliphant, Michael. 1999. The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior* 7.371–384.

——, & John Batali. 1997. Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter* 11.

Pfeifer, R., & C. Scheier. 1999. *Understanding Intelligence*. Cambridge, MA: MIT Press.

Pinker, S., & P. Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13.707–784.

Pullum, G. K., & B. C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19.9–50.

Quartz, S. R., & T. J. Sejnowski. 1997. The neural basis of cognitive development: a constructivist manifesto. *Behavioral and Brain Sciences* 20.537–596.

Quinlan, J. R., & R. Rivest. 1989. Inferring decision trees using the minimum description length principle. *Information and Computation* 80.227–248.

Reznikova, Z., & B. Ryabko. 1986. Analysis of the language of ants by information-theoretical methods. *Problems of Information Transmission* 22.245–249.

——, & ——. 1996. Transmission of information regarding the quantitative characteristics of an object in ants. *Neuroscience and Behavioural Physiology* 26.397–405.

Rissanen, J. 1978. Modeling by shortest data description. *Automatica* 14.465–471.

—— 1989. *Stochastic complexity and statistical inquiry*. World Scientific.

Ryabko, B., & Z. Reznikova. 1996. Using Shannon entropy and Kolmogorov complexity to study the communicative system and cognitive capacities in ants. *Complexity* 2.37–42.

Saffran, Jenny R., Richard N. Aslin, & Elissa L. Newport. 1996. Statistical learning by 8-month-old infants. *Science* 274.1926–1928.

Schmidhuber, J. 1997. Discovering neural nets with low Kolmogorov complexity and high generalization capability. *Neural Networks* 10.857–873.

Schoenemann, P. Thomas. 1999. Syntax as an emergent characteristic of the evolution of semantic complexity. *Minds and Machines* 9.309–346.

Shannon, Claude E., & Warren Weaver. 1949. *The mathematical theory of communication*. University of Illinois Press.

Shepard, R. N. 1987. Towards a universal law of generalization for psychological science. *Science* 237.1317–1323.

Smith, Andrew D. M. 2001. Establishing communication systems without explicit meaning transmission. In *Advances in Artificial Life: Proceedings of the 6th European Conference on Artificial Life*, ed. by Jozef Kelemen & Petr Sosík, number 2159 in Lecture Notes in Artificial Intelligence, 381–390. Springer-Verlag.

—— 2003a. Intelligent meaning creation in a clumpy world helps communication. *Artificial Life* 9.

Smith, K. 2002. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.

—— 2003b. Natural selection and cultural selection in the evolution of communication. *Adaptive Behavior* 10.25–44.

——, 2003c. The transmission of language: models of biological and cultural evolution. PhD Thesis, University of Edinburgh.

——, S. Kirby, & H. Brighton. forthcoming. Iterated learning: a framework for the emergence of language. In *Self-organization and Evolution of Social Behaviour*, ed. by C. Hemelrijk. Cambridge: Cambridge University Press.

Steels, Luc. 1997. Constructing and sharing perceptual distinctions. In *Proceedings of the European conference on machine learning*, ed. by M. van Someren & G. Widmer, Berlin. Springer-Verlag.

——. 1998. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103.133–156.

Teal, Tracy K., & Charles E. Taylor. 2000. Effects of compression on language evolution. *Artificial Life* 6.129–143.

Tomasello, Michael. 1999. *The cultural origins of human cognition*. Harvard: Harvard University Press.

Tonkes, Bradley, & Janet Wiles. 2002. Methodological issues in simulating the emergence of language. In *The Transition to Language*, ed. by Alison Wray, 226–251. Oxford: Oxford University Press.

Toussaint, Godfried, 2002. Proximity graphs for instance-based learning. Article in preparation.

Uriagereka, Juan. 1998. *Rhyme and Reason: An Introduction to Minimalist Syntax*. Cambridge, MA: MIT Press.

Van Essen, D. C., J. H. R. Maunsell, & J. L. Bixby. 1981. The middle temproal visual area in the macaque: Myeloarchitecture, connections, functional properties and topographic organisation. *Journal of Compuational Neuroscience* 199.293–326.

Van Hulle, Marc M. 2000. *Faithful Representations and Topographic Maps: From Distortion- to Information-Based Self-Organization*. Wiley-Interscience.

Vitányi, P. M. B., & M. Li. 2000. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory* 46.446–464.

von Frisch, Karl. 1974. Decoding the language of the bee. *Science* 185.663–668.

Wagner, Laura. 2001. Defending nativism in language acqusition. *Trends in Cognitive Sciences* 6.283–284.

Wallace, C. S., & D. M. Boulton. 1968. An information measure for classification. *Computing Journal* 11.185–195.

Wexler, K. 1991. On the argument from the poverty of the stimulus. In *The Chomskyan Turn*, ed. by A. Kasher, 253–270. Cambridge: Blackwell.

Wexler, Kenneth. 2001. Innateness of language. In *The MIT Encyclopaedia of the Cognitive Sciences*, ed. by R. A. Wilson & F. Keil. Cambridge, MA: MIT Press.

Wilkins, W. K., & J. Wakefield. 1995. Brain evolution and neurolinguistic preconditions. *Behavioral and Brain Sciences* 18.161–226.

Winograd, T, & F Flores. 1986. *Understanding Computers and Cognition*. Adison-Wesley.

Wolff, J. Gerard. 1982. Language acquisition, data compression, and generalization. *Language and Communication* 2.57–89.

Wray, Alison. 1998. Protolanguage as a holistic system for social interaction. *Language and Communication* 18.47–67.

——. 2000. Holistic utterances in protolanguage. In *The Evolutionary Emergence of Language: social function and the origins of linguistic form*, ed. by Chris Knight, Michael Studdert-Kennedy, & James R. Hurford, 285–302. Cambridge University Press.

Zipf, G. K. 1936. *The Psycho-Biology of Language*. London: Routledge.

Zuidema, Willem. 2003. How the poverty of the stimulus solves the poverty of the stimulus. In *Advances in Neural Information Processing Systems 15*, ed. by Suzanna Becker, Sebastian Thrun, & Klaus Obermayer, Cambridge, MA. MIT Press. in press.