The Transmission of Language: models of biological and cultural evolution

Kenneth Smith
M. A. (Hons), M. Sc.

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

to

Theoretical and Applied Linguistics,
University of Edinburgh

March 2003

by

Kenneth Smith
M. A. (Hons), M. Sc.

Abstract

Theories of language evolution typically attribute its unique structure to pressures acting on the genetic transmission of a language faculty and on the cultural transmission of language itself. In strongly biological accounts, natural selection acting on the genetic transmission of the language faculty is seen as the key determinant of linguistic structure, with culture relegated to a relatively minor role. Strongly cultural accounts place greater emphasis on the role of learning in shaping language, with little or no biological adaptation.

Formal modelling of the transmission of language, using mathematical or computational techniques, allows rigorous study of the impact of these two modes of transmission on the structure of language. In this thesis, computational models are used to investigate the evolution of symbolic vocabulary and compositional structure. To what extent can these aspects of language be explained in terms of purely biological or cultural evolution? Should we expect to see a fruitful interaction between these two adaptive processes in a dual transmission model?

As a first step towards addressing these questions, models which focus on the cultural transmission of language are developed. These models suggest that the conventionalised symbolic vocabulary and compositional structure of language can emerge through the adaptation of language itself in response to pressure to be learnable. This pressure arises during cultural transmission as a result of 1) the inductive bias of learners and 2) the poverty of the stimulus available to learners. Language-like systems emerge only when learners acquire their linguistic competence on the basis of sparse input and do so using learning procedures which are biased in favour of one-to-one mappings between meanings and signals. Children acquire language under precisely such circumstances.

As the second stage of inquiry, dual transmission models are developed to ascertain whether this cultural evolution of language interacts with the biological evolution of the language faculty. In these models an individual's learning bias is assumed to be genetically determined. Surprisingly, natural selection during the genetic transmission of this innate endowment does not reliably result in the development of learning biases which lead, through cultural processes, to language-like communication – there is no synergistic interaction between biological and cultural evolution. The evolution of language may therefore best be explained in terms of cultural evolution on a domain-general or exapted innate substrate.

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.

Kenneth Smith

Acknowledgements

Acknowledgements sections are usually over-long and toe-curlingly awful. I've managed to keep this one short, but have singularly failed to remove the cringe factor. Sorry.

Thanks firstly to my wife Becky, who has supported me with good grace and charm through my interminable stint as a student. You are the best.

For financial support, my parents deserve a mention, as do the Carnegie Trust for the Universities of Scotland. The department of Theoretical and Applied Linguistics also stumped up several sums of cash so I could go off to conferences and summer schools. I hope everyone feels they've got their money's worth.

On an academic level, thanks first of all to Simon Kirby and Jim Hurford for being excellent supervisors. Thanks are also due to Henry Brighton, who has worked closely with me on some of the models outlined in this thesis, in particular the model in Chapter 5. Andrew Smith also deserves a mention, for the helpful feedback he's given me over the years, for providing an enjoyable working environment in the Junior LEC office, and for not objecting too strongly to his nickname.

Thanks to Cedric McMartin and the rest of the computing support team. I've been thrashing every processor in the department for the past three years, and I couldn't have done so without your help.

Finally, no thanks to Scotland (for being useless at football, but not quite useless enough to stop me caring), UEFA (for running the Champions' League every year) and FIFA (I could have done without World Cup 2002, to be honest).

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Right now, waves of electromagnetic radiation in the 10kHz to 100GHz range are blasting out into space at the speed of light, carrying news of a revolutionary innovation. This radiation is emitting from Earth and has already reached over 1000 nearby stars. Encoded in the frequency and amplitude of these waves of radiation is evidence of an epoch-making new mode of communication which has developed relatively recently on this planet. This innovation has changed the face of our world and signalled our presence to the wider universe. The innovation is language, and the electromagnetic signal is generated by the more recent and comparatively inconsequential inventions called radio and television.

Only a tiny minority of the 6 billion-strong human population of Earth has more than a basic level of understanding of radio technology, and fewer still have a good grasp of the physics of electromagnetic radiation. Yet virtually all of humanity has an unconscious, profound and subtle knowledge of a language. Knowledge of language is universal. Humans acquire their knowledge of language apparently effortlessly, in all manner of circumstances, with only the most severely deprived individuals remaining language-less into adult life.

Sophisticated communication systems abound on Earth. Most species have a system of signals which enable them to communicate with other members of their species, or with members of other species. Bees communicate the distance and direction of pollen sources to other bees using an elaborate dance (von Frisch 1974). Numerous species of bird use song for marking out territory and for attracting mates (Hauser 1996). Bottlenose dolphins have personalised “signature whistles”, and can rapidly learn to produce the signature whistles of other dolphins (Tyack 1986; Janik & Slater 1997). It has been suggested that dolphins may use these whistles for addressing each other, either aggressively or affiliatively (Janik 2000).

Moving to species more closely related to humans, most primate species use systems of facial expressions and vocalisations to perform social functions such as issuing threats (see Hauser (1996) for review). Vervet monkeys have a system of alarm calls whereby different calls are given for different predator species, and vervets respond to these calls in a manner appropriate for best evading the associated predator (Cheney & Seyfarth 1990). Gibbons produce structured calling bouts, with calls of males in particular exhibiting structured combination of individual notes into "phrases", with phrases being repeated during a calling bout (Raemaekers *et al.* 1984). In addition to facial expression, chimpanzees use systems of gestural communication (Tomasello 1990; Tomasello 1996; Tomasello *et al.* 1997) for performing various social functions.

These communication systems appear to echo some aspects of language. What, then, makes language special? How did human language come to exhibit these unique features? The aim of this thesis is to answer this "how" question, with the aid of a set of computational models. This investigation forms the bulk of the thesis, from Chapter 2 onwards. However, before addressing the question of how language came to be as it is, it is necessary to tackle the first question above — what feature or combination of features makes language unique? This review, in Section 1.1, serves to highlight the features of language which then have to be explained later in a theory of language evolution. Section 1.2 outlines two competing theories of how humans come to acquire a language. These theories, in combination with the potential sources of relevant evidence outlined briefly in Section 1.3, form the basis of a number of broad theories of language evolution. These theories are discussed in Section 1.4. Finally, in Section 1.5, I highlight the need for theories of language evolution to be based on formal models. This forms the motivation for the remainder of this thesis, which is primarily concerned with insights gained from formal modelling approaches.

1.1 The uniqueness of language

Two steps are then necessary to clarify the difference between language and other naturally-occurring communication systems[1]. Firstly, we must identify the features of language

[1] For the purposes of this chapter it is sufficient to rely on the rather naive notion of communication as a process by which individuals transmit information to other individuals. There is a great deal of debate about the proper definition of communication. Many definitions (e.g. those of Krebs & Dawkins (1984), Millikan (1984) and Hauser (1996)) appeal to the notion of functionality and design by natural selection. However, such notions are somewhat controversial with respect to language, as we shall see in Section

which require explanation. This is done in Section 1.1.1, based on a cross-section of introductory linguistics textbooks. Secondly, we must identify which of these features are present in other communication systems, in particular the communication systems of non-human animals. This analysis is carried out in Section 1.1.2. Based on these analyses, taxonomies of communication systems with respect to distinctive features can be developed. A taxonomy developed in Oliphant (2002) is reviewed in Section 1.1.3. This taxonomy is refined in Section 1.1.4. An explanation of the uniqueness of human language amounts to an explanation of the features which are present in language but not in non-human communication.

1.1.1 Design features of language

Most introductory linguistics textbooks include a list of *design features* of language (a term introduced by Hockett, e.g. Hockett (1960a), Hockett (1960b)) which the authors regard as important properties of language. Collating these proposed design features from a selection of five introductory texts (Burling 1992; Fromkin & Rodman 1988; Hudson 2000; O'Grady *et al.* 1996; Trask 1995), removing those which are proposed in a single text, yields the following list of seven design features of language:

OPEN-ENDEDNESS: The set of sentences that could be produced or understood by a user of a language is infinite. (Burling, Fromkin & Rodman, Hudson, O'Grady *et al.*, Trask)

CULTURAL TRANSMISSION: Language is learned (in some sense to some degree — see Section 1.2) by language users from other language users, as opposed to being genetically transmitted. (Burling, Fromkin & Rodman, Hudson, O'Grady *et al.*, Trask[2])

ARBITRARINESS: Typically, the form of a signal is arbitrarily related to its meaning. (Fromkin & Rodman, Hudson, O'Grady *et al.*, Trask)

DUALITY OF PATTERNING: Small numbers of meaningless elements (phonemes in spoken languages) are combined to form large numbers of meaningful elements (words). (Burling, Hudson, O'Grady *et al.*, Trask)

DISPLACEMENT: Language can be used to communicate about things in places, times or even possible worlds removed from the actual communicative act. (Burling, Hudson, O'Grady *et al.*, Trask)

1.4 — it has been argued that language is not designed for communication. I will provide a definition of communication which is sufficient for my purposes in Chapter 2, and consider some alternative definitions in Chapter 4.

[2] Trask in fact doesn't include cultural transmission in his introductory chapter, where he introduces the other design features of language. However, he does introduce it later in the book.

STIMULUS FREEDOM: Language users can potentially produce any signal they want at any time, and are not bound to producing a signal only when the appropriate stimulus is present. (Fromkin & Rodman, O'Grady *et al.*, Trask)

DISCRETE: The units of a language are distinguished from each other categorically, as opposed to grading into one another. (Burling, O'Grady *et al.*)

The design feature of open-endedness can be further decomposed into two features:

COMPOSITIONALITY: The meaning of an expression (excluding idioms and irregulars) is a function of the meaning of its parts and the way in which they are combined (Cann 1993; Krifka 2001).

RECURSIVENESS: An expression of a particular type can be a subpart of a larger expression of that type (see e.g. Burling (1992), Haegeman (1994), Hudson (2000)).

Recursiveness allows the creation of an infinite number of utterances. Compositionality makes the interpretation of previously-unencountered utterances possible — in a recursive compositional system, if you know the meaning of the basic elements and the effects associated with combining elements, you can deduce the meaning of any utterance in the system.

1.1.2 Design features of animal communication

In the process of identifying these design features of language, Burling (1992), Fromkin & Rodman (1988), Hudson (2000), O'Grady *et al.* (1996) and Trask (1995) explicitly contrast them with apparent design features of non-human communication systems. Some design features are identified as occurring in the communication systems of non-human animals, whereas other design features are identified as being unique to language.

OPEN-ENDEDNESS: All agree that this is unique to language.

CULTURAL TRANSMISSION: All agree that, in addition to language, song is culturally transmitted in some species of bird. In addition, Hudson suggests that the vervet alarm call system is culturally transmitted (apparently wrongly, as we will see below). With the exception of bird song, all other non-human communication systems are taken to be genetically transmitted. This is supported by a large cross-species review, which concludes that "although call structure [in non-human primates, our closest extant relatives] changes ontogenetically, no study has provided convincing evidence that acoustic experience is causally related to such changes" (Hauser 1996:315).

ARBITRARINESS: There is some disagreement on arbitrariness. Trask suggests that it is common, Hudson and O'Grady *et al.* suggest that it is only observed, outside of humans, in vervets. In fact, alarm calling systems seem to be fairly common, both in primates (for example, Diana monkeys (Zuberbuhler *et al.* 1997), Campbell's monkeys (Zuberbuhler 2001) and ringtailed lemurs (Pereira & Macedonia 1991)) and birds (various species of passerine birds (Marler 1957) and domestic chickens (Evans *et al.* 1993)).

DUALITY OF PATTERNING: All agree that this is unique to language.

DISPLACEMENT: All agree that bee dance is the only non-human communication system that allows displacement, and even then only with respect to the domain of foraging.

STIMULUS FREEDOM: Fromkin & Rodman, O'Grady *et al.* and Trask claim that all non-human communication systems are stimulus-bound, although Trask does admit to some anecdotal evidence of stimulus-free communication in animals.

DISCRETE: Burling claims that all non-human communication consists of graded signals. However, O'Grady *et al.* give the obvious counter-examples of bee dance (most species use one dance for food sources close to the hive and a completely different dance for distant food sources) and primate call systems (such as the vervet alarm call system).

How might non-human communication systems be classified according to our two additional features of recursiveness and compositionality? Bird song and gibbon calls consist of the repetition of notes and structured groupings of notes and therefore might be viewed as recursive, although only in a weak, conjunctive sense. However, these structured calls are typically interpreted as being used for territory maintenance, pair bonding or sexual advertisement (Raemaekers *et al.* 1984; Hauser 1996) and are therefore probably not compositional — no subpart of the signal stands for a subpart of the meaning, where the meaning could be interpreted as "this is my/our territory", "we are a couple" or "I would make a good mate". Furthermore, there is evidence that the gibbon calls may be largely genetically determined, although this has been contested (see Janik & Slater (1997) for a brief review). This characterisation of bird and gibbon song as structured, perhaps even weakly recursive, yet non-compositional may have to be revised if a better characterisation of what these animals are communicating about reveals that, in fact, certain aspects of their signals reflect certain components of the message they are trying to convey.

In contrast, bee dance could not be interpreted as recursive, but it seems somewhat compositional. The meaning of the dance (the direction and distance to the food source) is dependent on the angle of the straight portion of the dance (which gives direction relative

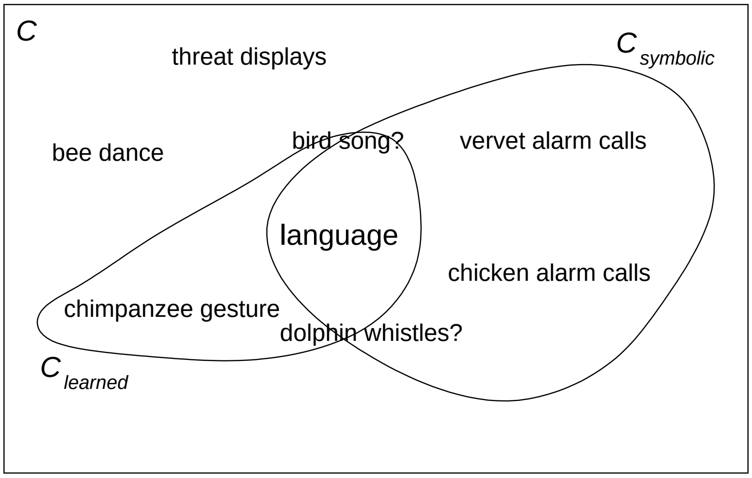


Figure 1.1: Oliphant's taxonomy of communication systems. The outer square, C, represents the space of all extant communication systems. Within this, there are some communication systems which are symbolic (the area labelled as $C_{symbolic}$), and some which are learned ($C_{learned}$). According to Oliphant, only (human) language is a member of $C_{symbolic}$ and $C_{learned}$.

to the sun) and the length of the straight portion of the dance (which gives distance). It is not clear that the meaning of the dance is dependent on the way these two parts are put together, as they are essentially part of the same action. There is also no evidence that the bee dance is culturally transmitted.

1.1.3 Oliphant's taxonomy of communication systems

Oliphant (2002) provides a taxonomy of communication systems based on two characteristics, symbolicism (related to arbitrariness) and learnedness (related to cultural transmission). Oliphant defines a symbol as "a sign that refers to the object that it denotes in a way that is arbitrary with respect to the process of conventionalization that established it" (Oliphant 2002:313). Non-symbolic systems can either be characterised as iconic (reference is by resemblance) or indexical (reference is by causal relation). Learned systems are acquired experientially (this typically involves cultural transmission, but may involve other processes, as discussed below with reference to chimpanzee gestures), whereas non-learned systems are specified genetically (and therefore genetically transmitted). Oliphant's taxonomy is illustrated in Figure 1.1. Oliphant's main conclusion is that language is the only communication system which is both learned and symbolic.

There are several features to note in Oliphant's taxonomy. Firstly, chimpanzee gestural communication is classified as learned but non-symbolic. Such gestural communication systems are formed by what Tomasello (1996) calls *ontogenetic ritualization*, whereby actions which are initially part of a process become ritualised shortcuts which stand for the whole process. Tomasello (1990) gives the example of an infant chimpanzee which moves its mother's arm in order to reach a nipple to feed. Over time, the action of the chimpanzee touching its mother's arm becomes a ritualised shortcut for this process, signalling to the mother that the infant needs to feed. The meaning and the signal are associated indexically, by being part of a causal chain, and are not arbitrarily associated. Furthermore, Tomasello *et al.* (1997) provide evidence that such ritualised signals are not culturally transmitted in chimpanzee populations once they become established.

Similar arguments based on indexicality apply to the evolution of innate non-symbolic systems such as primate facial expressions, threat displays and the bee dance, where the signal has presumably evolved from what was initially a preparatory stage for the full action of attacking, taking off or whatever.

Oliphant claims that alarm call systems, such as that of the vervet monkey, are innate and symbolic. As mentioned above, Hudson (2000) claims that the vervet system is culturally transmitted and therefore learned, so a clarification here is worthwhile. Seyfarth & Cheney (1986) and Hauser (unpublished data, cited in Hauser 1996, p306) report that the acoustic morphology of calls in vervets is essentially independent of age, "suggesting that experience plays a relatively insignificant role in shaping acoustic morphology" (Hauser 1996:306). However, vervets do become more specific with the use of their alarm calls as they mature. Infant vervets are as likely to give the eagle alarm call for a martial eagle as for non-predatory birds. Infant vervets have even been seen to give the eagle call on seeing falling leaves (Cheney & Seyfarth 1990). However, adult vervets give the eagle call in the appropriate circumstance (seeing a martial eagle) much more frequently than in inappropriate circumstances. Immature vervets must therefore learn when to apply their innately-given alarm calls. It is interesting to note that this categorisation process also appears to be partially innately coded — the eagle alarm call in immature vervets is only given to objects in the air, usually birds. Oliphant therefore argues that the mapping between objects in the air and the eagle call is innately specified and not learned, while learning somewhat refines the 'meaning' portion of the mapping.

Oliphant's classification of such alarm call systems as symbolic is interesting. Bee dance is classified as non-symbolic because it is hypothesised to have evolved from phylogenetic ritualization of preparatory movements for takeoff — the signal is causally related to the meaning. Oliphant presumably envisages an evolutionary scenario where the vervet call signals evolved from a situation where individuals, on sighting a predator, gave some involuntary vocalisation then took appropriate evasive action. Over time these vocalisations became phylogenetically ritualised into the call system. The only difference between this scenario and that proposed for the evolution of bee dance is that the hypothesised involuntary proto-alarm call is not causally related to the evasive action, and therefore arbitrary with respect to that action, which becomes the meaning of evolved call. However, we can imagine a scenario under which the vocalisation in the proto-alarm call system was part of the causal chain involved in performing the action of evading the predator — for example, anatomical constraints might require vervets to clear the air from their lungs before they can stand up. This would make the alarm call non-symbolic in Oliphant's classification. His classification appears to depend on knowing the evolutionary history of a signalling system, which we cannot do with any certainty.

There are two notable omissions from Oliphant's taxonomy — bird song and dolphin signature whistles. Certain types of bird song are at least partially learned, and may be interpretable as arbitrarily related to their meaning and therefore symbolic. Dolphins learn the signature whistle of other dolphins. If a dolphin's signature whistle reflects aspects of its physical characteristics then we might classify such learned whistles as iconic, given that there is a resemblance between a dolphin's whistle and its physical characteristics. However, if a dolphin's whistle is arbitrary then there is only an arbitrary relationship between the learned whistle and its meaning (the dolphin that produces that whistle) and therefore dolphin whistles would fall into the learned, symbolic portion of Oliphant's analysis alongside human language and, possibly, bird song. I would not like to commit to such an analysis, but merely raise it as an example of the problems which potentially arise when we attempt to distinguish language from non-language on categorial assessments in a limited number of dimensions.

1.1.4 Refining the taxonomy

Figure 1.2 gives a taxonomy of communication based on Oliphant's taxonomy, integrating the terms of three of the "design features" outlined above — cultural transmission, arbitrariness (for which I adopt Oliphant's definition of a symbol) and compositionality. Cultural transmission is preferred to Oliphant's "learnedness" classification. Compositionality is included without recursiveness.

According to this taxonomy, language is the only naturally-occurring communication system which is culturally transmitted *and* symbolic *and* compositional. The following chapters of this thesis will attempt to explain the evolution of communication systems which exhibit these properties.

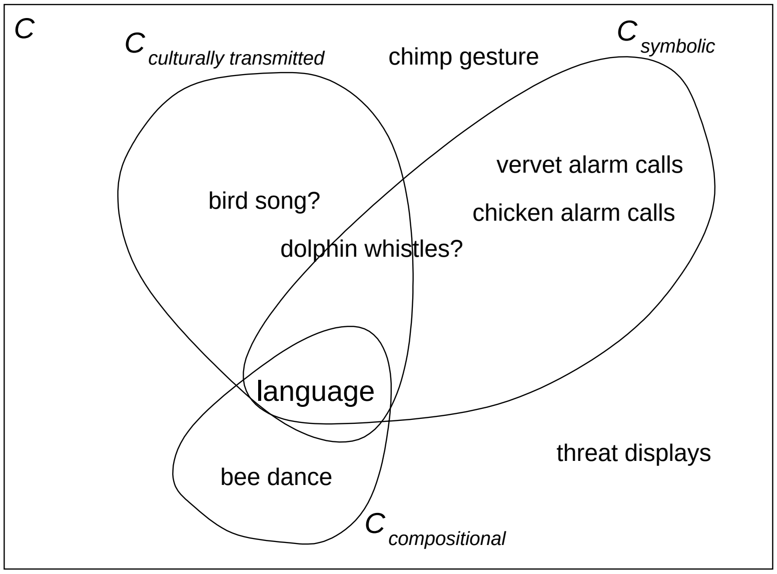


Figure 1.2: A new taxonomy of communication systems. Language is unique in occurring at the intersection of $C_{culturally\ transmitted}$, $C_{symbolic}$ and $C_{compositional}$ — only language is culturally transmitted, symbolic and compositional.

1.2 Theories of Language

We have established that language is unique. More specifically, language is unique in being culturally transmitted, symbolic and compositional. The concern of much of cognitive science, and linguistics in particular, has been to characterise what it means to have knowledge of a language with these features — how is language represented in the mind and how is this representation of language acquired?

Theories of language can be crudely divided into two paradigms — the Nativist paradigm, discussed in Section 1.2.1, and the Empiricist paradigm, discussed in Section 1.2.2, which encompasses various anti-Nativist positions

1.2.1 The Nativist paradigm

The dominant paradigm in linguistics, formed and directed by Noam Chomsky (e.g. Chomsky (1965), Chomsky (1972), Chomsky (1986), Chomsky (1995)), views language from the standpoint of individual psychology. Under this view, the domain of inquiry is the state of the mind of a person who knows a language – Internalised Language, or I-Language. Externalized Language, or E-Language, is then the epiphenomenal set of actual or possible behaviours produced by application of this internalised knowledge to a set of contingent situations in the external world.

In conjunction with this focus on I-language, Chomsky and others have advanced the notion of Universal Grammar (UG). “UG may be regarded as a characterisation of the genetically determined language faculty . . . an innate component of the human mind that yields a particular [I-]language through interaction with presented experience . . . UG is a theory of the ‘initial state’ of the language faculty, prior to any linguistic experience” (Chomsky 1986:3-4). Under a more refined analysis, UG may be decomposed into two parts — a “specification of permitted types of rules and permissible interactions among them” (Chomsky 1986:52) or “a system of principles associated with certain parameters of variation” (Chomsky 1986:221) (the hypothesis space for possible I-languages) and a Language Acquisition Device (Chomsky 1965) which specifies how presented experience determines the selection of a particular I-language from the range of possibilities admitted by the specification of permitted hypotheses. In addition to being innately specified, UG is viewed as being language-specific — “a system of grammatical relations and a system of morphological agreement . . . are totally useless to any other [non-linguistic] cognitive capacity; they are exquisitely specialized” (Jackendoff 2002:264).

This theory of UG was advanced to account for the seemingly insoluble problem of language acquisition. Fundamental to theories of UG is the observation that children must learn language from “meager and unspecific evidence” (Chomsky 1986:149). Pullum & Scholz (2002) provide a fairly rigorous survey of the ways in which the evidence presented to the child has been claimed to be meager and unspecific. These can be summarised as:

1. Children are not specifically or directly rewarded for their advances in language learning.
2. Children’s data-exposure histories are finite, but they acquire an ability to produce or understand an infinite number of sentences.
3. Children’s data-exposure histories are highly diverse, yet language acquisition is universal.
4. Children’s data-exposure histories are incomplete in that there are many sentences they never hear, yet can produce and understand.
5. Children’s data-exposure histories are solely positive — they are never given details of what is ungrammatical.
6. Children’s data exposure histories include numerous errors, such as slips of the tongue and false starts.

Such arguments are known as arguments from the *poverty of the stimulus*. Given that the evidence presented to language learners is so impoverished, the only possible explanation for universal language acquisition is that some of the knowledge of language must be prespecified in UG — UG constrains and guides the language acquisition process in such a way that, even given the paucity of evidence available to children, language is reliably acquired. As Chomsky puts it, language acquisition is interpreted as the "growth of cognitive structures [linguistic competence] along an internally directed course under the triggering and partially shaping effect of the environment" (Chomsky 1980:34). To put it in Jackendoff's terms, in order to explain the fact of language acquisition "we need a way for children to be . . . hampered by preconceptions — in fact hampered by the very same preconceptions as every other child. It is just that children's preconceptions happen to give them the right solutions" (Jackendoff 2002:84).

1.2.2 The Empiricist paradigm

The Nativist position on language acquisition can be summarised as the application of domain-specific, innate acquisition procedures to an impoverished set of evidence which leads to selection of a hypothesis from a domain-specific, innately prespecified space of hypotheses. The alternative approach, which I will call the Empiricist approach, can be characterised as follows: application of domain-general acquisition procedures to a (possibly rich) set of evidence leads to the selection of a hypothesis from a domain-general space of hypotheses.

This position in its most extreme form is expounded by Geoffrey Sampson (Sampson 1997), who argues that "biological constraints on language are limited to matters which are 'trivial' because they follow from properties of our speech and sense organs which are known to be genetically fixed" (Sampson 1997:25). For Sampson there is no innate LAD and no prespecified space of possible hypotheses, nor even an innate propensity to want to learn a language. Less extreme forms of the anti-nativist view are expressed by Popper, who suggests that "[t]he capacity to learn a language — and even a strong need to learn a language — is, it appears, part of the genetic make-up of man. By contrast, the actual learning of a particular language, though influenced by unconscious inborn needs and motives, is not a gene-regulated process and therefore not a natural process, but a cultural process" (Popper (1977:48), cited in Sampson (1997)). A similar position is adopted by some cognitive scientists working in the connectionist paradigm such as Bates & Elman (1996) who argue that domain-general learning techniques may explain more of language than was previously thought, while allowing that "[e]ven if we assume that a brain (real or artificial) contains no innate knowledge at all, we have to make crucial assumptions about the structure of the learning device, its rate and style of learning, and the kinds of input that it 'prefers' to receive" (Bates & Elman 1996:1850).

Proponents of this anti-Nativist position are often called upon to refute perhaps the strongest argument in favour of UG — the argument from the poverty of the stimulus. Pullum & Scholz (2002) investigate the evidence behind the fourth aspect of the poverty of the stimulus (given above), that children never encounter certain sentence constructions which would be required to acquire an I-language by a learning procedure unconstrained by UG. One example is the structure-dependent generalisation about subject-auxiliary inversion in question formation in English, introduced as an example of the poverty of the stimulus in Chomsky (1971).

(1.1) The dog in the corner is hungry
(1.2) Is the dog in the corner hungry?
(1.3) The dog that is in the corner is hungry
(1.4) Is the dog that is in the corner hungry?
(1.5) *Is the dog that in the corner is hungry?

The interrogative in 1.2 could be derived from the declarative in 1.1 by two possible strategies. Under the (correct) structure-dependent hypothesis, the auxiliary in the main clause in the declarative is fronted. Under the (incorrect) structure-independent hypothesis, the first auxiliary in the declarative is fronted. These two competing hypotheses are indiscriminable on the basis of exposure to sentences 1.1 and 1.2. The interrogative in 1.4, derived from the declarative in 1.3, shows that the structure-dependent hypothesis is correct, as the structure-independent movement hypothesis would generate the (ungrammatical) 1.5. Chomsky claims that “you can go over a vast amount of data of experience without ever finding such a case [as 1.4]”. Pullum & Scholz (2002) in fact found examples of the discriminating case in every corpus they examined, including newspaper text, a play, a poem, a transcription of a television program and, notably, a corpus of child-directed speech.

The assault on the empirical bases of the poverty of the stimulus argument is a relatively recent endeavour. Consequently, it is not clear to what extent the other aspects of the argument will survive closer inspection. One aspect of the argument does seem to be on fairly solid ground, however — given that the set of possible sentences of a language is infinite, we can be sure that no child will hear them all during the process of language acquisition. Does this problem of acquiring a system of infinite productivity from a finite set of data therefore force us to accept the strong Nativist concepts of UG and the LAD? Not necessarily. Experiments such as that outlined in Elman (1993) show that, given a particular maturation schedule, domain-general statistical learning techniques can extract grammatical regularities from finite exposure to word strings. These regularities could potentially allow children to produce and understand an infinite range of utterances.

1.3 Evidence for language evolution

What sources of evidence are available to guide and constrain theories of the evolution of language? Potentially there are two main types of evidence. Archaeology could provide us with evidence which directly shows the course of the evolution of language or the language faculty. However, as discussed in the next Section, such evidence proves to be of limited utility. A second possible source is in present day phenomena — the way we acquire and use language today might suggest its evolutionary origin.

1.3.1 Archaeological evidence

The main problem with archaeological evidence is that language does not fossilize. The earliest direct evidence we have for the existence of language comes from the Sumerian writing systems, the earliest examples of which occur around 3000BC (Kramer 1963). However, it is unlikely that this date reflects the origin of language — writing is essentially a technical innovation, and a significant proportion of the world's population today remains illiterate while having an intimate and subtle knowledge of a language.

Given the absence of direct evidence of the origins of language, the archaeological record has been searched for indirect evidence of language. This search can be broadly categorised into two strands: the search for evidence of the origins of biological correlates of language, and the search for cultural correlates of language. The biological correlates of language are typically taken to be in the brain and the vocal tract. These both consist of soft tissue so, somewhat ironically, neither of these fossilize either. We can, however, draw some indirect conclusions on the evolution of these indirect markers of language.

Brain size, as estimated by cranial volume, has increased down the hominid line. This trend of increasing brain size is summarised in Table 1.1, which gives the estimated brain sizes for the main recognised species of hominids. The various species of australopithecines, which were among the earliest hominids (appearing approximately 5 million years ago), had brain sizes estimated to be between 400 and 550ml, comparable with extant apes of a similar stature. Early modern *Homo sapiens* have brain volumes of 1200–1700ml, comparable with anatomically modern *Homo sapiens*.

Brain size is not the only measure we have of the evolution of brains in the hominid line — see, for example, the evidence on sulcal patterning discussed in the summary of

Species	Date of appearance	Estimated brain size
australopithecine (various)	5m BP	400–500ml
Homo habilis	2.4m BP	500–800ml
Homo erectus	1.8m BP	750–1250ml
Homo sapiens (archaic)	400,000 BP	1100–1400ml
Homo sapiens (early modern)	130,000 yrs BP	1200–1700ml

Table 1.1: The trend in brain size. Approximate date of appearance in the fossil record (in years before present) and estimated brain size for ancestors of modern humans. Data on australopithecines from Wood (1992). Data on *Homo* from Stringer (1992).

Wilkins & Wakefield's (1995) paper below. However, this kind of evidence is extremely controversial. The general trend of increasing brain size is less controversial, but probably less useful. We do not know the minimum brain size required to support a language. It is common to equate an increase in brain size with an increase in general intelligence, and for those working in the Empiricist paradigm, where language is seen primarily as a consequence of general intelligence, this suggests that the capacity for language may have emerged further down the hominid line. However, this still does not tell us when, or whether the increase in brain size was driven by selection for language abilities or selection for general intelligence. The picture of increasing intelligence is even less relevant if we follow the Nativist paradigm, where language is seen as a consequence of a specific mental organ.

The evolution of the vocal tract has also been the subject of lively debate. The soft tissue of the vocal tract does not fossilize, but the shape of the vocal tract, the height of the larynx, and the degree of respiratory and articulator control have all been estimated based on markers which do fossilize — for example, the flexion of the base of the skull, the morphology of the hyoid bone, the width of the hypoglossal canal and the width of the thoracic vertebrate canal (see Fitch (2000) for review). However, much of this evidence is controversial — as Fitch concludes, "despite an extensive and disputatious literature, most potential fossil cues to phonetic abilities appear inconclusive, suggesting that it will be difficult to reconstruct the vocal behavior of our extinct ancestors with any certainty" (Fitch 2000:263). Even if we could reconstruct the vocal behaviour of our extinct ancestors, it would perhaps not provide any direct answers as to the evolution of language. The evolution of the vocal apparatus required for speech has little relevance for those who argue that language was initially based on a gestural system (Hewes 1973; Corballis 2002). Even if we accepted that speech was the initial modality used for language, and were able to pin down a definite date when the capacity for speech began to evolve, this would still be compatible with several competing theories: speech could have been a preadaptation for language, or could have co-evolved with language (as suggested by Lieberman, discussed below), or could have evolved as a consequence of selection pressures for efficient signal production arising from the presence of fully-fledged language.

The study of the evolution of some of the supposed biological correlates of language is therefore not terribly illuminating — not only is the archaeological evidence often debatable, but the correspondence between these physical characteristics and language is not well-established. This has led some researchers to focus on the appearance of cultural artifacts in the archaeological record. Writing is one such cultural artifact. The account of Wilkins & Wakefield (1995) appeals to the relevance of tools, another cultural artifact. The fact that the inventory of tools (or at least tools which remain in the fossil record) remained fairly constant until the late Paleolithic period, when it underwent explosive diversification, might suggest that there was some biological change which lead to a cultural revolution which impacted upon the range of tools individuals were capable of learning to manufacture. However, this is not necessarily the case. It could be that the capacity for a complex tool kit was present long before such a tool kit was in place — cultural innovations do not necessarily coincide with biological ones. Secondly, and of more relevance from our point of view, there is no obvious link between tools and language. While language may be involved in the cultural transmission of tool-making skills, we cannot conclude the presence of language from the presence of tools. We certainly cannot conclude the absence of language from the absence of tools.

Others have appealed to the origins of “symbolic culture” (ochre staining or cave art) as evidence of the origins of language (e.g. Knight (1991), Knight *et al.* (1995)). While we might agree that the presence of art indicates the presence of language, we are not forced to do so — it just seems unlikely that people who were so much like ourselves as to indulge in cave paintings did not have language. Most manifestations of “symbolic culture” are so recent as to make this position fairly uncontroversial anyway. And we still cannot conclude that the absence of art reliably indicates the absence of language — art may be a fairly recent cultural innovation.

To summarise, the evidence from the archaeological record on the evolution of language is fairly sparse and of limited use. This is either due to the debatable nature of the archaeological data, or the lack of clarity as to what that data tells us about language.

1.3.2 Current-day evidence

Far better sources of evidence are available in living species in the world today. While current-day evidence cannot directly show us the time-course of the evolution of language, it can give us a valuable insight into the kinds of processes — biological or cultural — which might have been involved in the origins and evolution of language. There are 7 sources of present-day evidence which can potentially shed some light on such questions:

1. Language uniqueness
2. Language universals
3. Genetic deficits
4. Cross-species comparison
5. Language acquisition
6. Language change
7. Creolization

Explanations of the uniqueness of language and the limited range of cross-linguistic variability evidenced in the languages of the world follow naturally from the notion of genetically-encoded UG. Only humans have language because only humans have UG, without which language cannot be acquired. Language universals (see e.g. Greenberg (1966)) exist because UG constrains the range of possible I-languages to vary in limited ways. Properties which all languages share are therefore presumed to be direct reflections of the constraints imposed by UG. Uniqueness and universals can then be seen as a consequence of the biological evolution of the language faculty in our species.

Non-biological accounts of language and language evolution also have to account for the presence of linguistic universals. Typically, such universals are attributed to universal aspects of the *use* of language, rather than universal constraints on the possible forms of language — see Newmeyer (1998) for review (Chapter 3 in particular). The uniqueness of language to humans has variously been attributed to consequences of the uniquely high degree of encephalization in humans (e.g. Bickerton (2000), Christiansen & Conway (2002)), humans' uniquely sophisticated understanding of others as intentional agents (Tomasello 1999) or the unusually long developmental period in humans (Elman 1993).

Myrna Gopnik (Gopnik & Crago 1991; Gopnik 1994) presents evidence from the study of a family, known as the KE family, where approximately half of the members of the family suffer from a heritable language impairment. According to Gopnik, this impairment is specific to inflectional marking of tense and number, and other than these problems the subjects are of normal IQ, and have “no reduction in the range of movement or the tone of the mouth and tongue musculature” (Gopnik 1994:113). Furthermore, the heritable impairment seems to be a consequence of a point mutation in a single gene (Lai *et al.* 2001). This appears to indicate the presence of a “language gene” — a gene which, if mutated, has specific and deleterious consequences for a subset of the tasks involved in language acquisition and processing.

The presence of a language gene would seem to suggest that language evolution must have involved the emergence of some language-specific biological capacity in our species. However, Gopnik’s data and conclusions have been attacked. Gopnik claims that members of the family who carry the mutated gene are of normal intelligence. Vargha-Khadem *et al.* (1995) conclude that “the affected family members have both verbal and performance intelligence quotient (IQ) scores that are on average 18–19 points below those of the unaffected members” (Vargha-Khadem *et al.* 1995:930). Furthermore, empirical studies suggest that “the affected members’ disorder transcends the generation of morphosyntactic rules to include impaired processing and expression of other areas of grammar, grossly defective articulation of speech sounds, and, further, a severe extralinguistic orofacial dyspraxia” (Vargha-Khadem *et al.* 1995:930). This serves to cast doubt on the picture of the KE family mutation as a mutation in one of the genes encoding UG — the impairment may be a consequence of genes encoding for several functions, including language, or may be a consequence of a more general cognitive and physical impairment.

We can also examine how language is acquired by humans, and how languages change in human populations. I will make reference to this type of data later in the thesis, in particular in the closing sections of Chapters 3 and 5. We can also make cross-species comparisons of neural or physiological devices which have been implicated in language, and consequently hypothesise about whether such devices are evolutionary recent or not. Hauser *et al.* (2002) provide an excellent survey of this comparative approach. I will make reference to this type of evidence later in the thesis, in particular with respect to vocabulary acquisition experiments in apes.

Finally, the Nativist account of language offers a fairly straightforward account of creolization. The classic creolization scenario involves a group of individuals with no common language being brought together, typically being forced to live or work together. In such populations, a *pidgin* typically emerges — an ad-hoc system of communication, which borrows vocabulary from the various languages represented in the group (known as the substrate languages), and from the language of the ruling body (the superstrate language). Pidgins are typically highly simplified grammatically, and are tolerant of extreme ranges of grammatical variation. Creolization occurs when children begin to learn pidgins as their first language. Creoles are typically more expressive than pidgins, more elaborate grammatically and exhibit less variation across and within individuals.

Derek Bickerton (Bickerton 1981; Bickerton 1984) further claims that superficially unrelated (and geographically dispersed) creoles exhibit certain common features, and concludes that "creole similarities stem from a single substantive grammar consisting of a very restricted set of categories and processes, which will be claimed to constitute part, or all, of the human species-specific capacity for syntax" (Bickerton 1984:178). The common elements of creole languages (for example, fronting of focussed elements, certain arrangements in the determiner system, handling of tense, modality and aspect by preverbal free morphemes occurring in a certain fixed order), according to Bickerton, are a consequence of a highly specific innate prespecification of language — "The LBH [Language Bioprogram Hypothesis] claims that the innovative aspects of creole grammar are inventions on the part of the first generation of children who have a pidgin as their linguistic input, rather than features transmitted from preexisting languages . . . the LBH claims that the most cogent explanation of this similarity is that it derives from the structure of a species-specific program for language, [which is] genetically coded" (Bickerton 1984:173). In other words, the degenerate nature of the linguistic input available to children in pidgin-speaking communities forces them to fall back on the default grammar encoded in their genes. The fact that these genes are shared by all members of the species means that all creoles share certain features. This suggests that language evolution is primarily a biological process.

Bickerton's account of creolization and creole universals as reflexes of the default settings of UG is incompatible with the Empiricist position on language, and is also largely incompatible with accounts of language evolution which appeal to a significant role for cultural processes. There are two main strands of argument against it. Firstly, it has been argued (see e.g. Singler (1990)) that there are numerous exceptions to Bickerton's putative creole universals. Secondly, it has been argued that most of the correspondences between creoles can be attributed to the influence of common substrate (e.g. Holm (1988)) or superstrate (Haspelmath (1989)) languages.

In addition to these disputes over the source of "creole universals", other studies of creolization suggest that cultural, rather than genetic, processes may play an important role in language genesis. Nicaraguan Sign Language (Kegl & Iwata 1989; Kegl *et al.* 1999) gives us an insight into the birth of a language with a reduction in the confounding factors of substrate and superstrate influence and power asymmetries typically associated with creolization situations. Nicaraguan Sign Language emerged when previously isolated deaf individuals throughout Nicaragua were brought together in schools for the deaf. Prior to arriving at the school, the isolated deaf individuals communicated using idiosyncratic homesign systems. While such systems may exhibit some degree of structure (Goldin-Meadow & Mylander 1990) they are characterised as "idiosyncratic, variable even within the individual, and lacking most characteristics, particularly syntactic, of what we would recognize as a full-fledged human language" (Kegl *et al.* 1999:179). Shortly after arriving at the school these homesigners had developed a shared system of signs and grammatical devices. This shared system developed into a fully-fledged sign language after several years and several influxes of new, typically young, deaf individuals. The use of grammatical devices by deaf individuals shows significant effects for both age of entry into the school and year of entry into the school, with younger individuals entering the school in later years producing the most sophisticated signing.

Based on the emergence of Nicaraguan Sign Language and the failure of groups of deaf individuals to develop full-blown language under different social circumstances, Ragir (2002) argues that "the emergence of new languages, both signed and spoken, depends on: (1) a *critical mass* of individuals generating shared patterns of linguistic practices; (2) historical continuity maintained by a continuous influx of new participants into the language pool; and (3) an exchange of information about diverse cooperative activities".

This suggests that the Language Bioprogram Hypothesis, as simply stated above, cannot be right — if degenerate linguistic input automatically triggered the default, innate grammar then isolated individuals would effectively self-creolize, and we would expect isolated deaf individuals to use sign systems which had the structure of early creoles. The fact that creolization is dependent on a critical mass of individuals and a degree of historical continuity suggests that there must be some factor other than innate knowledge at play in creolization events, and possibly in linguistic evolution in general.

1.4 Theories of language evolution

The (somewhat limited) evidence outlined above, in conjunction with the two major explanatory paradigms for language (Nativist and Empiricist) form the basis for a number of theories of the evolution of language. These can broadly be separated into two groups — *adaptationist* and *non-adaptationist* theories. Adaptation is the process by which an organism changes to fit its environment. Adaptationist accounts emphasise the role of natural selection and what we will (for the moment) call cultural selection in the process of adaptation — pressures acting on genetic and cultural transmission favour organisms which fit, or are adapted for, their environment, such that over time a population of such organisms comes to fit its environment rather well. In the most extreme adaptationist position, every aspect of an organism's behaviour is seen as an adaptation to fit some part of the organism's environment.

Non-adaptationists deemphasise the role of adaptation in explaining characteristics of an organism, preferring to focus on the role of chance events, architectural constraints and exaptation in evolutionary processes. Architectural constraints limit and shape the ways in which organisms develop, due to laws of growth and form (Thompson 1961) or environmental influences on development. These architectural constraints can lead to *spandrels* (a term adapted from architecture by Gould & Lewontin (1979)), where architectural constraints lead to structures which appear to be designed for some purpose, but are in fact not. Exaptation (a term introduced in Gould & Vrba (1982), sometimes replaced with the somewhat more transparent term of "evolutionary reappropriation") is the process by which some aspect of an organism, either a spandrel or an adapted trait, is put to a new use by that organism.

Theories of the evolution of the linguistic capacity and language have appealed to all three of these aspects of evolutionary processes — adaptation, architectural accidents and exaptation. Non-adaptationist accounts, viewing the capacity for language as a spandrel or an exapted trait, will be reviewed first, in Section 1.4.1, in part because such accounts have found favour with influential figures in linguistics. Adaptationist accounts, which appeal to both biological adaptation of the language capacity and cultural adaptation of language, will be reviewed in Section 1.4.2.

1.4.1 Non-adaptationist accounts

The non-adaptationist accounts reviewed here cover a fairly wide spectrum. Chomsky and Piattelli-Palmarini view the language capacity as a spandrel. Bickerton views it as a trait arising either through chance events, such as a saltationary mutation or a fluke of ontogeny, or, in his later account, as an exaptation. Lieberman and Wilkins & Wakefield view the capacity for language (as distinct from the capacity for speech) as an exaptation of neural machinery for motor control, whereas Tomasello sees language as a consequence of a more general capacity for cultural transmission.

Chomsky Noam Chomsky, the founder of the theory of innate UG, has been notoriously reluctant to offer an account of the origins of this innate and language-specific organ of the human mind. According to Chomsky, "it is safe to attribute this development [of UG] to 'natural selection', so long as we realize that there is no substance to this assertion, that it amounts to nothing more than a belief that there is some naturalistic explanation for these phenomena" (Chomsky 1972:97). Chomsky's preferred explanation for the origins of UG is dependent on as-yet-unknown laws of growth and form:

> "The answers may well lie not so much in the theory of natural selection as in molecular biology, in the study of what kinds of physical systems can develop under the conditions of life on earth" (Chomsky 1988:167)
>
> "We know very little about what happens when 10^{10} neurons are crammed into something the size of a basketball, with further conditions imposed by the specific manner in which this system developed over time. It would be a serious error to suppose that all properties, or the interesting properties of the structures that evolved, can be 'explained' in terms of natural selection." (Chomsky 1975:59)
>
> "Language is off the chart. That is the basic conclusion that follows from his [Hauser's 1996] comprehensive review of comparative communication. That doesn't mean that language is not the result of biological evolution, of course we all assume it is. But what kind of result of biological evolution? ... It is simply not understood how the physical channel constrains and controls the process of selection, beyond simple cases ... there is no way to find the answer, not just for language but for cognition altogether. Others feel that they can do something. But telling stories is not very instructive. You can tell stories about insect wings, but it remains to discover how they evolved." (Chomsky 2002:Chapter 4)

While these points have some merit, taking them fully to heart essentially rules out any attempt to explain the evolution of the language faculty. Indeed, Chomsky's position has been been criticised as "a retreat to mysticism" (Jackendoff 2002:234). Recently (Hauser *et al.* 2002) Chomsky's position has softened somewhat and he acknowledges the potential role of cross-species comparison as a test on, and possible source of insight into, theories of language evolution.

Piattelli-Palmarini Piattelli-Palmarini (1989) is somewhat ambivalent between spandrel and exaptation accounts of the capacity for language — "innate, very specific, and highly abstract structures governing language and cognition may be seen as 'spandrels', that is, biological traits that have *become* central to our whole existence, but which may well have arisen for some purely architectural or structural reason ... or as a by-product of evolutionary pressures driven by other functions" (Piattelli-Palmarini 1989:19). He supports this position with an argument based on the fact that language could be other than it is, and still be adaptive:

> "Even if a trait *is* useful and actually enhances the life expectancy of individuals who possess it, this fact does *not* grant the inference that the trait is there *because* it is useful . . . Adaptive constraints are typically insufficient to discriminate between real cases and an infinity of alternative, incompatible mechanisms and traits which, although abstractly compatible with the survival of a given species, are demonstrably absent. This applies, and with bells on, to perfectly adaptive linguistic structures that could ideally have been present, but which are, as a matter of fact, wrong for *us* . . . linguistic principles are all, obviously, compatible with our survival, but they are not uniquely (nor, for that, even approximately) 'determinable' under the adaptationist constraint *alone*" (Piattelli-Palmarini 1989:18–19).

Pinker & Bloom (1990) identify two possible responses to Piattelli-Palmarini's point that other forms of language might be equally functional. Firstly, they argue that all adaptations are compromises between a variety of constraints, therefore the purpose of an adaptation may not be obvious when considered from the perspective of a single constraint. Secondly, and perhaps more tellingly, they point out that "many 'arbitrary' constraints may have been selected simply because they defined parts of a standardized communicative code in the brains of some critical mass of speakers" (Pinker & Bloom 1990:718). Certain "perfectly adaptive linguistic structures" might not be present simply due to historical contingencies, reinforced by the need for something approximating consensus among language users.

We might take a weakened version of Piattelli-Palmarini's position on board, and concede that some aspects of the design of the language faculty are constrained by considerations such as architecture, costs and so on. Indeed, not to do so would be somewhat ridiculous — as Maynard Smith has pointed out "[i]f there were no constraints on what is possible, the best phenotype would live for ever, would be impregnable to predators, would lay eggs at an infinite rate, and so on" (Maynard Smith 1978:32). Piattelli-Palmarini's point is therefore not as forceful as he obviously feels. It is weakened still further by an obvious confusion in his argument. In the very same paragraph he claims that "*Only* those who espouse an *instructivist* [transfer of structure from the environment to the organism, i.e. learning] paradigm are in need of direct adaptive constraints. If, on the contrary, one adheres to the thesis of strong innatism, adaptive constraints become *ipso facto* weak, trivial and theoretically irrelevant" (Piattelli-Palmarini 1989:19). This appears to be a non-sequitur "with bells on". Previously we were dealing with the evolutionary history of UG, either a history of exaptation or adaptation. In a strongly innatist position these can either be exapted or adapted — there is nothing that rules out adaptation of the LAD in the strongly innatist position, although Piattelli-Palmarini would prefer that we did. Piattelli-Palmarini then switches to dealing with adaptive constraints on language *acquisition*. Adaptive constraints on *learning* in the strongly innatist position are, by definition, irrelevant, as in the scenario Piattelli-Palmarini envisages there is no such thing as learning — "we would gain in clarity if the *scientific* use of the term were simply discontinued" (Piattelli-Palmarini 1989:2). However, there is no need for adaptationists to appeal to adaptive constraints on learning — some do, but others might favour Piattelli-Palmarini's discontinuation of the term altogether yet still feel that the innate UG arose through adaptation.

One more criticism of Piattelli-Palmarini's position should be noted. Piattelli-Palmarini is a strong innatist, meaning, among other things, that he views the language capacity as language-specific. However, he prefers an exaptationist account of this language capacity to an adaptationist account. Exaptation, by definition, cannot apply to an apparatus which can serve only one purpose, therefore Piattelli-Palmarini should abandon either his exaptationist account or his language-specificity constraint. Similarly, it is not clear that a spandrel could be said to be specific to one function, since spandrels are, by definition, not "for" any function at all.

Bickerton Bickerton provides an account of the evolution of the language faculty which is, at one level, an adaptationist account. However, Bickerton's early formulation of the account appeals to a single macro-mutation delivering much of the language faculty, while his later accounts become more exaptationist in flavour.

Bickerton's original account (Bickerton 1990) proposes that there was, prior to language, an agrammatical protolanguage[3], the capacity for which possibly evolved due to selection pressures arising from group foraging activities (Bickerton 2002). Mutation of a single gene then delivered, from the base of protolanguage, the capacity for full-blown modern language, including the capacity for syntax and the modern human vocal tract. This gene was then fixed in the population by means of natural selection.

[3] Bickerton claims that we see evidence of this protolanguage in children under two, pidgin speakers and signing apes. However, the content of protolanguage and the various pros and cons of the proposal are not, for our purposes, of great relevance to the remainder of Bickerton's account.

The early formulation of Bickerton's proposal is somewhat naive. As Hauser (1996) notes, relatively simple physical attributes are encoded in multiple genes. The odds seem to be against a single gene which encodes a fairly sophisticated language faculty and, entirely coincidentally, the appropriate vocal apparatus to make efficient use of this neural architecture. Depending on one's interpretation of the evidence on "language genes" provided by the studies of the KE family, this either proves that the language faculty is coded in multiple genes, or that a single "language gene" would at least require that other general intelligence or speech genes be set up appropriately.

Bickerton (1998) abandons the macro-mutation account for the transition from protolanguage to language, but his account remains saltationary. Bickerton proposes that some of the necessary neural apparatus for language — mental apparatus dealing with theta analysis[4] and motor control of the (not fully modern) vocal tract — were present in the brains of protolanguage users, but unconnected. Through ontogenetic chance (formation of connections between the two areas) these areas became connected in the brains of certain individuals, resulting in the possibility of conceptual structure being imposed on the production of protolanguage, giving us the capacity for language. Bickerton proposes that "the linkage of theta analysis with other elements involved in protolanguage would not merely have put in place the basic structure of syntax, but would also have led directly to a cascade of consequences that would, in one rapid and continuous sequence, have transformed protolanguage into language substantially as we know it today" (Bickerton 1998:353). Bickerton suggests some of these consequences might include the evolution of the modern human vocal tract and development of further neurological apparatus for language processing and storage. Bickerton therefore views the emergence of the language capacity as a process of exaptation followed by adaptation. However, the burden of the account still falls on the initial exaptation of resources used in protolanguage, facilitated by unspecified phylogenetic or ontogenetic factors which allowed the theta analysis component of the brain to connect to the vocal motor control areas. Bickerton offers no detailed explanation for this initial exaptation, although Wilkins & Wakefield (1995), discussed below, offer a possible account compatible with Bickerton's.

Bickerton (2000) retreats from this analysis. He reasons that protolanguage would be used for discussing past events, therefore we should expect it to have had access to thematic roles via memory. Secondly, he claims that all brain regions are linked to one another, therefore the tightly modularised separation of the theta analysis component and vocal motor component would not exist. Bickerton therefore suggests that these areas

[4] Which can be characterised for the purposes of Bickerton's argument as "who did what to who with what".

of the brain were linked in the protolanguage-capable brain, but could not communicate with one another due to a lack of "signal coherence" resulting from the small number of neurons in the pre-human brain. Bickerton proposes that increases in brain size overcame this signal coherence, allowing the regions of the brain responsible for theta analysis to impose structure on motor output. The syntactic brain was born, followed by the familiar avalanche of consequences, which Bickerton now proposes include canalization of learned behaviours via the Baldwin effect. Once again, this is essentially an exaptationist account (exaptation of resources resulting from increased brain size), with subsequent adaptive change.

Hurford (2002b) criticises Bickerton's 1998 and 2000 efforts, not on the vagueness of the mechanisms appealed to (although this does merit comment in passing) but on the assumption that theta analysis areas, when wired up to vocal motor areas, gives rise straightforwardly to clausal structure, and the rest of syntactic structure follows trivially. Hurford identifies several features of language (for example, duality of patterning, the morphology-syntax distinction, case marking, anaphor-antecedent relationships, and passivization) which play no role in pre-linguistic representation but could reasonably be argued to play a role in communication over a noisy serial channel. Hurford's point is that syntax, and linguistic structure more generally, is not just theta analysis with a flourish, but a system which exhibits the appearance of design for communicating complex structures over a serial channel. While Bickerton could reply that he allows time for such aspects of language to emerge during the "cascade of consequences", if there is much more to language than theta analysis then Bickerton's account has simply covered the simplest aspect of its emergence and paid lip-service to subsequent processes.

Lieberman Classifying the work of Philip Lieberman as adaptationist or exaptationist is a somewhat fraught exercise, partly due to the difference in focus of Lieberman's work from the other theories outlined here. While the other non-adaptationists are primarily concerned with the (non)evolution of UG as it pertains to syntax, Lieberman's main focus is on speech, which is a peripheral matter in the Chomskyan paradigm. Lieberman takes a classically adaptationist position on the evolution of the physiological and neurological bases of speech — "all primates are furthermore adapted for phonation at the expense of respiratory efficiency. Anatomically-modern *Homo sapiens* continues this trend toward more efficient vocal communication" (Lieberman 1984:324–325).

However, Lieberman views the capacity of the human brain for language, meaning essentially syntax, as an exapted trait: "the rules of syntax derive from a generalization of neural mechanisms that gradually evolved in the motor cortex to facilitate the automatization of motor activity [for tool use in (Lieberman 1984), bipedal locomotion and tool use in (Lieberman 2000)]. Human syntactic ability, in this view, is a product of the Darwinian mechanism of preadaptation" (Lieberman 1984:67). In this respect, taking the Chomskyan grammar-centric view of language, Lieberman is a non-adaptationist. Language only arises through the combination of the adapted speech mechanism and the exapted trait: "The synergetic effect of rapid data transmission through the medium of encoded speech and the cognitive power of the large hominid brain probably yielded the full human linguistic system" (Lieberman 1984:329)

Unlike Chomsky and Piattelli-Palmarini, but in common with Bickerton, Lieberman allows for some post-exaptation adaptation of the syntactic capacity:

> "As the central cognitive mechanism, the brain, gradually increased its power, the selective advantage of peripheral input-output mechanisms like human speech also would have increased . . . The presence of a fully encoded speech system in recent hominids may also have more directly contributed to the development of complex syntactic organization in human languages." (Lieberman 1984:325)

Lieberman's treatment of the evolution of the human vocal apparatus is very thorough, and supported by a degree of evidence not seen in the other accounts of the evolution of language. However, his treatment of syntax is somewhat threadbare. Lieberman's statement that syntax is simply a generalisation of motor control falls foul of Hurford's (2002b) criticism of Bickerton — if language is not simply slightly elaborate theta analysis, nor is it a generalisation of movement. Lieberman could mount a similar defence to Bickerton, that the aspects of language which Hurford discusses arose during the period of adaptation following the exaptation event. However, Lieberman also fails to offer more than a hand-waving explanation of how the "synergetic effect" of a large brain and apparatus capable of rapid data transmission would lead to even the rudiments of language.

Tomasello Michael Tomasello's theories on the cultural evolution of language will be expounded in more detail in Chapter 2. Tomasello's full account of the evolution of language is exaptationist, with a heavy focus on cultural processes resulting from the proposed exaptation. Tomasello suggests that the key adaptation:

> “arose at some particular point in human evolution, perhaps fairly recently, presumably because of some genetic and natural selection events. This adaptation consists [of] the ability and tendency of individuals to identify with conspecifics in ways that enable them to understand those conspecifics as intentional agents like the self, possessing their own desires and beliefs. This new mode of understanding other persons radically changed the nature of all types of social interactions, including social learning, so that a unique form of cultural evolution began to take place over historical time, as multiple generations of developing children learned various things from their forebears and then modified them in a way that led to an accumulation of these modifications” (Tomasello 1999:202)

This exaptationist account is somewhat different from the accounts propounded by Piattelli-Palmarini and Bickerton. For Piattelli-Palmarini the mental capacity for language is a spandrel, but is language specific. Cultural processes play no role in his model. For Bickerton, in his later accounts at least, the language capacity is exapted from the protolanguage capacity (possibly with additional exaptation of resources resulting from an increase in brain size), is language-specific and results in consequential evolution, some of which might involve (minimally) cultural processes. For Tomasello, the capacity to view others as intentional agents is a general cognitive ability, rather than being language-specific, and results in consequences which are driven by purely cultural processes.

Wilkins & Wakefield The final non-adaptationist account of the origins of the human capacity for language differs somewhat in perspective from the others. Wilkins & Wakefield (1995) deal with the processes which lead up to a cognitive capacity which could be reappropriated by language, rather than assuming such an apparatus and dealing with the post-exaptation scenario.

Wilkins & Wakefield present an account under which Broca’s area and the parieto-occipito-temporal junction (POT), which includes Wernicke’s area, are necessary elements of a brain supporting language. Auditory, visual and somatosensory inputs are initially processed in unimodal association areas of the cortex. These unimodal representations then converge to form multi-modal representations of sensory input. Finally, these multi-modal representations converge at the POT — “[t]he POT is, in essence, an area of integration for the three neocortical sensory association areas . . . the ‘association area of association areas’ ” (Wilkins & Wakefield 1995:163). Wilkins and Wakefield argue, following an idea originally suggested by Geschwind (1964), that this capacity to form amodal representations is crucial to the capacity to acquire lexical items.

Following Tallal & Schwartz (1980) and Greenfield (1991), Wilkins & Wakefield suggest that Broca's area is responsible for temporal sequencing and hierarchical organisation of information. In the human brain, the POT and Broca's area are connected by a large, myelinated fibre tract, allowing information to pass between them rapidly. In their account, Broca's area hierarchically structures the amodal representation of the outside world, resulting in the development of a featural, amodal, hierarchical representation — an abstract, structured semantic representation. A similar process allows auditory or visual input to become associated with these amodal concepts — arbitrary symbolic reference becomes possible.

Wilkins & Wakefield make a case that this organisation of the human brain arose, not through selection pressures related to communication, but through pressure for improved motor skills required to fashion and use tools such as throwing stones, which have obvious uses in hunting. Skilled motor control requires larger amounts of motor cortex and somatosensory cortex devoted to controlling the hand. The necessary feedback between somatosensory cortex and motor cortex resulted, in Wilkins & Wakefield's account, in selection pressure for rapid transfer between these two physically distant regions of the brain, which in turn led to the formation of the bidirectional, myelinated fibre bundle connecting sensory to motor cortex (including Broca's area). There is some neurophysiological evidence to support this association between the evolution of motor cortex and somatosensory cortex — "each architectonic subdivision of the sensory association cortex is directly and reciprocally connected to the subdivision of motor association cortex with which it shares an equivalent state of evolutionary differentiation" (Wilkins & Wakefield 1995:174). Concurrent expansion of the visual cortex and the temporal lobe resulted, due to the close proximity of the three posterior lobes, in an overlap of the information flowing through each lobe, and the POT junction formed. Evidence for this account is provided in the archaeological record, which shows that stone tools began appearing at a time coincident with the appearance of *Homo habilis* (Harris 1983), while sulcal patterning on fossil endocasts suggests that *habilis* possessed the POT and a modern Broca's area. However, the interpretation of endocasts is controversial, as Wilkins & Wakefield acknowledge. Ralph Holloway notes in his commentary on Wilkins & Wakefield's article that "there is not one single well-documented instance of paleoneurological evidence that unambiguously demonstrates a relative expansion of the parietal-ocipital-temporal junction in early *Homo*" (Holloway 1995:191).

Wilkins & Wakefield therefore present a scenario under which selectional pressures for tool use lead to the appropriate neural and cognitive apparatus for language. Their account of how language exapted these facilities is speculative, although they offer:

> "As is well known, primates are noisy animals . . . It is not unreasonable to assume that *H. habilis* too were noisy animals, with a systematic repertoire of calls. It also seems reasonable to think that a habiline child might have recruited from this call repertoire to create a linguistic sign . . . These acoustic signals, simply part of the primate call system for the adult vocaliser, might have taken the form of linguistic signs in the mind of the child" (Wilkins & Wakefield 1995:179).

1.4.2 Adaptationist accounts

In contrast to these non-adaptationist explanations of the evolutionary origins of the faculty for language in humans, several adaptationist accounts have been advanced. These accounts appeal to the biological adaptation of the human species in response to the pressure to communicate, or the cultural adaptation of languages to pressures to be expressive or learnable.

Pinker & Bloom Pinker & Bloom (1990) present the classic adaptationist account of language. Working within the Chomskyan framework of UG, Pinker & Bloom suggest that "the ability to use a natural language belongs more to the study of human biology than human culture: it is a topic like echolocation in bats" (Pinker & Bloom 1990:707). Their position is that language exhibits the appearance of design and is supported by an innate UG, therefore the origins of UG must explicable in terms of natural selection.

What is language designed for? According to Pinker & Bloom, language is adapted for the communication of propositional structures (in the internal representational "language of thought") over a serial channel. Pinker & Bloom identify a dozen features of language which they argue exhibit the appearance of design for this function. For example, they argue that major lexical categories are used to distinguish basic ontological categories, phrase structure and overt case marking provide surface cues to the structure of the underlying propositional representations, and verb affixes signal temporal aspects of events.

Pinker & Bloom go on to identify scenarios where language exhibiting these design features might offer a selective advantage. They propose that the hunter-gatherer lifestyle hypothesised to characterise our recent evolutionary predecessors would favour an ability to communicate information about toolmaking, ecology and the behaviour of plants and animals. "Devices for communicating precise information about time, space, predicate-argument relations, restrictive modification, and modality are not wasted in such efforts" (Pinker & Bloom 1990:724). Secondly, they emphasise the value of language in supporting social interactions, which put "a premium on the ability to convey such socially relevant abstract information as time, possession, beliefs, desires, tendencies, obligations, truth, probability, hypotheticals, and counterfactuals . . . Furthermore, in a group of communicators competing for attention there is a premium on the ability to engage, interest and persuade listeners. This in turn encourages the development of discourse and rhetorical skills and the pragmatically relevant grammatical devices that support them" (Pinker & Bloom 1990:725). According to Pinker & Bloom, UG evolved as a direct consequence of the payoff it provided in relation to these communicative functions.

Jackendoff Jackendoff (2002) offers what is essentially an elaboration of the position of Pinker & Bloom (1990) that the language faculty arose through natural selection. Jackendoff focuses on possible stages in the evolution of the language faculty, following Pinker & Bloom in assuming that any increase in expressive power or precision will offer selective advantages. Jackendoff is not concerned with the ecological context in which such adaptations might have arisen, or how innovations would spread through populations, although he "agree[s] with practically everyone that the 'Baldwin effect' had something to do with it" (Jackendoff 2002:237).

Jackendoff's schedule for the evolution of language is depicted in Figure 1.3 (based on Jackendoff (2002), Fig 8.1, p238). There are several points worth noting. Firstly, Jackendoff views the human language faculty as having evolved out of preexisting primate conceptual structure *and* primate call systems — his first step on the path to language is the step from stimulus-bound to stimulus-free signalling. Secondly, cultural transmission is implied in Jackendoff's second step, to use of an open vocabulary. Thirdly, Jackendoff's account is clearly influenced by Bickerton's notion of a protolanguage intermediate between primate communication and modern language, but differs from it in the respect that, for Jackendoff, protolanguage exhibits the very beginnings of syntax. The transition to using symbol position to convey basic semantic relationships occurs prior to the protolanguage stage, whereas for Bickerton at the protolanguage stage words are "randomly concatenated" (Bickerton 1998:349).

Dor & Jablonka Dor & Jablonka (2000) provide a version of the adaptationist argument which differs from that offered by Pinker & Bloom (1990) and Jackendoff (2002) in the degree of focus on cultural processes. For Pinker & Bloom and Jackendoff the evolution of language is basically due to the evolution of the language faculty via biological evolution, possibly with the Baldwin effect playing some poorly specified role. Dor & Jablonka bring the role of cultural innovation followed by genetic assimilation to the fore, in what they call "the evolutionary spiral".

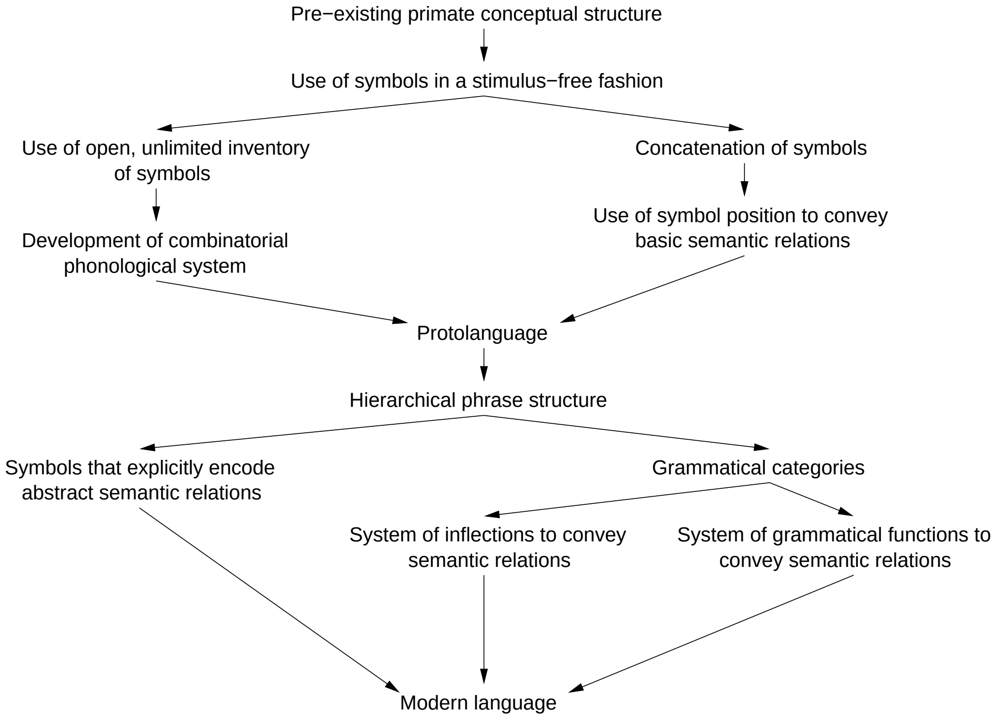

Pre−existing primate conceptual structure
Use of symbols in a stimulus−free fashion
Use of open, unlimited inventory of symbols
Concatenation of symbols
Development of combinatorial phonological system
Use of symbol position to convey basic semantic relations
Protolanguage
Hierarchical phrase structure
Symbols that explicitly encode abstract semantic relations
Grammatical categories
System of inflections to convey semantic relations
System of grammatical functions to convey semantic relations
Modern language

Figure 1.3: Jackendoff's schedule for the evolution of language. Necessary preadaptations appear on top of subsequent adaptations. Adaptations which are logically independent appear side by side.

Dor & Jablonka suggest a scenario where a hominid community uses some culturally-transmitted communication system which is subserved by a genetically-encoded propensity to acquire such a system. Note that Dor & Jablonka consider culturally-transmitted communication systems to be common in mammals, despite the view expressed by Hauser (Hauser (1996), discussed in Section 1.1.2) that this is not the case. In this population, linguistic innovations such as "new lexical items for specific referential meanings . . . some more abstract markers for existing and novel conceptual distinctions . . . new pragmatic conventions for linguistic communication" (Dor & Jablonka 2000:49) came about, through teleological processes, as a result of pressure to improve communication. Such useful innovations then spread through the population. "Crucially . . . the establishment of the innovation also raised the demands for social learning imposed on individuals in the community . . . the linguistic innovations which established themselves in the community changed the social niche, and the inhabitants of this new niche had to adapt to it" (Dor & Jablonka 2000:50). According to Dor & Jablonka this adaptation occurred through genetic assimilation, with a genetic endowment which permitted easier language acquisition spreading through the population. The process then repeats with further innovations and so on.

This scenario is opposed to that proposed by Deacon, whose account I present below. Dor & Jablonka also put a strong emphasis on the role of conscious, functionally-motivated innovation in shaping language, an issue which I will return to in Chapter 2 in connection with Tomasello's account of cultural evolution.

Deacon Deacon (1997) presents an adaptationist account which differs from the accounts of Pinker & Bloom and Jackendoff in placing relatively little emphasis on the biological evolution of the language faculty and more emphasis on cultural processes. His account is therefore superficially somewhat similar to that of Dor & Jablonka, but makes entirely different predictions.

Deacon identifies the enlarged prefrontal cortex of the human brain as the key biological component of language — "[t]he contributions of prefrontal areas [of the cortex] to learning all involve, in one way or another, the analysis of higher-order associative relationships . . . These are the most critical learning problems faced during symbol[5] acquisition" (Deacon 1997:264). Deacon regards this aspect of brain morphology as a consequence of language, rather than an exapted trait:

> "The remarkable expansion of the brain that took place in human evolution, and indirectly produced prefrontal expansion, was not the cause of symbolic language but a consequence of it. As experiments with chimpanzees demonstrate, under optimal training conditions they are capable of learning to use a simple symbol system. So, it is not inconceivable that the first step across the symbolic threshold was made by an australopithecine with roughly the cognitive capabilities of a modern chimpanzee, and that this initiated a complicated history of back-and-forth escalations in which symbol use selected for greater prefrontalization, more efficient articulatory and auditory capacities, and probably a suite of other ancillary capacities and predispositions which eased the acquisition of this new tool of communication and thought. Each assimilated change enabled even more complex symbol systems to be acquired and used, and in turn selected for greater prefrontalization, and so on." (Deacon 1997:340)

Deacon suggests that the fitness advantage of symbolic communication is due to the type of social and communicative benefits outlined in Pinker & Bloom (1990). So far, then, this is a fairly minor modification to the theory proposed by Pinker & Bloom. Deacon,

[5] Deacon's definition is somewhat different from the definition I have adopted.

however, places heavy emphasis on the cultural evolution of language itself, resulting from the pressure on language during cultural transmission:

> “The structure of a language is under intense selection because in its reproduction from generation to generation, it must pass through a narrow bottleneck: children’s minds . . . Language operations that can be learned quickly and easily by children will tend to get passed on to the next generation more effectively and intact than those that are difficult to learn. So, languages should change through history in ways that tend to conform to children’s expectations.” (Deacon 1997:110)

Under this view, the apparent innate endowment of humans to acquire language in the face of the poverty of the stimulus is merely an illusion, induced by the focus on language as a facet of individual psychology, rather than a culturally-transmitted adaptive system: “Human children appear preadapted to guess the rules of syntax correctly, precisely because languages evolve so as to embody in their syntax the most frequently guessed patterns. The brain has co-evolved with respect to language, but languages have done most of the adapting” (Deacon 1997:122).

This last sentence highlights the difference between the positions of Deacon and Dor & Jablonka — for Dor & Jablonka, the genetic makeup of a population changes to match the population’s language, whereas for Deacon the onus is on a population’s language to change to match the population’s acquisition abilities. Deacon’s position is also counter to that of Jackendoff, who states that “the limited number of possible choices to which children are constrained had better be the *right* ones, otherwise they won’t learn” (Jackendoff 2002:84). One consequence of Deacon’s view of language itself as an adaptive entity is that the choices children are constrained to are *guaranteed* to be the right ones, because language adapts to fit them.

1.5 Formal models

The brief review of theories of language evolution outlined above reveals several problems. Firstly, as indicated by the rough classification of theories as either non-adaptationist or adaptationist, there is a fairly fundamental disagreement as to whether the evolution of language can be explained in terms of properties of language itself, as the adaptationists have it, or in terms of some other properties of humans or processes of human evolution, as the non-adaptationists have it. Secondly, there tends to be a disagreement about what the evidence actually means — for example, Lieberman and Deacon both allow chimpanzees some modest language-like capacity, but while Lieberman presents this as evidence of his exaptationist account of a capacity for syntax, Deacon presents it as a gateway into an adaptationist account of language evolution. Similarly, dependent on one's interpretation of the KE family "language gene" evidence, it can be taken as indicating the presence of a genetically-encoded language faculty or the dependence of language on other, more general, cognitive abilities.

These two problems are not unique to the field of language evolution. Biologists have had a similar debate about the limits of adaptationism and the role of exaptation with respect to the evolution of traits in general, and the evidence in most fields of scientific enquiry is somewhat susceptible to interpretation. There is, however, a third problem exemplified in this review which I intend to address in this thesis.

All the theories outlined above are purely verbal — the authors of each theory present the evidence which they feel supports their theory and develop their theories in ways which seem most consistent, both internally and with the evidence. Problems arise when competing theories rely on approximately the same premise, but make rather different predictions. For example, Bickerton and Jackendoff both present theories which include a protolanguage phase. For Bickerton, once semantic relationships begin to impinge on the ordering of surface forms language emerges in a fairly straightforward, and for Bickerton fairly non-interesting, process. For Jackendoff, the ordering of constituents in protolanguage is already influenced by semantic considerations, but there is still a great deal to be explained. As another example, Dor & Jablonka and Deacon agree that cultural and genetic processes play a part in the evolution of language. However, Dor & Jablonka predict that genes will adjust to accommodate language, whereas Deacon predicts that language will accommodate the learning biases encoded in the genes. There is no way to tell which theory is more plausible, given that both are based on purely verbal reasoning and both appear to be internally consistent. The problem is that our intuitions about how complex dynamical processes, such as the gene-culture transmission of language, will unfold are notoriously poor.

Formal models can be a powerful tool for conducting this kind of "opaque thought experiment" (Di Paolo *et al.* 2000). Firstly, constructing a formal model can reveal whether the consequences of a proposed interaction are as predicted. Secondly, formal models help to identify which parts of the behaviour of a system rest on which assumptions — to give a hypothetical example, a formal model might reveal that Deacon is correct (and Dor & Jablonka are wrong) that cultures adapt to genes, given the assumption that the process of cultural evolution is ten or more times faster than genetic evolution. The extent to which these assumptions match up with the theoretical background and empirical evidence then gives a measure of the value of the theory underlying the formal model — to continue the example above, if there were good theoretical and empirical reasons for assuming that cultural evolution is always at least 50 times faster than biological evolution, the formal model would tend to support Deacon's position against that of Dor & Jablonka. Formal models allow a degree of rigour to be brought to the study of the evolution of language.

1.6 Guide to the thesis

The rest of this thesis is devoted to the discussion and development of formal models of language evolution, and the extrapolation from these models to theories of language evolution. I will focus in particular on the evolution of symbolic vocabulary and compositional structure. Much of this thesis will attempt to explain these aspects of language in terms of the cultural adaptation of language itself to pressures acting during the cultural transmission of linguistic form — symbolicism and compositionality are a consequence of cultural transmission. I will also investigate how biological evolution of learning apparatus can impact on this cultural evolution. In this respect, much of this thesis is an elaboration of Deacon's position that "[h]uman children appear preadapted to guess the rules of syntax correctly, precisely because languages evolve so as to embody in their syntax the most frequently guessed patterns. The brain has co-evolved with respect to language, but languages have done most of the adapting" (Deacon 1997:122).

In Chapter 2 I will review general and language-specific models of cultural transmission. This will form the basis for Chapter 3, in which I describe computational models which shed light on the learning biases necessary to support a culturally-transmitted symbolic vocabulary. In Chapter 4 I then investigate how these learning biases might have evolved. In Chapter 5 I turn to the cultural evolution of compositional structure, and identify two pressures which lead to the emergence of compositionality. These pressures arise from the poverty of the stimulus problem and a particular bias in learners. In Chapter 6 I model the evolution of this learning bias. Finally, in Chapter 7 I draw conclusions from the research outlined in this thesis, and highlight potential areas in which my general approach could be expanded.

CHAPTER 2

Models of cultural transmission

In this chapter I will review formal models of cultural transmission. These models fall into two groups — models which have been developed to account for cultural transmission in general, and models which have been developed to account for the cultural transmission of language in particular. The general model of cultural transmission given here is based heavily on the work of Robert Boyd and Peter Richerson, in particular Boyd & Richerson (1985) (henceforth B&R). B&R use mathematical techniques adapted from theoretical biology. The models of the cultural transmission of language have been developed by a fairly diverse group of researchers, typically (though not exclusively) working with computational models.

In Section 2.1 I review, in broad terms, two models of cultural transmission. In Section 2.2 I review in slightly more detail the mechanisms of cultural transmission acting in these models. Finally, in Section 2.3 we will see how different pressures acting on cultural transmission can drive cultural evolution and cultural adaptation, with particular reference to the cultural evolution of language.

There are two goals for this chapter. The first is to review the range of formal models which have been used to study cultural evolution in general, and the cultural evolution of language in particular. This review covers relevant techniques which have been used to address this question, and suggests areas which are worthy of further formal modelling. Secondly, some of the fundamental results of B&R's simple models will prove useful in interpreting the results of the more complex models introduced in later chapters.

2.1 General and linguistic models

2.1.1 *A general model*

A general and simple model of cultural transmission must account for three processes:

1. The cultural transmission of behaviour from a mature population to an immature population. The target of learning for the immature population is the behaviour of the mature population.
2. Individual learning by the immature population. The target of learning is determined by the environment, rather than the mature population.
3. Removal (possibly in a selective fashion) of individuals from the population.

The simplest scenario proceeds as follows. There is some set of immature individuals and some set of mature individuals. The immature individuals observe and learn from the mature individuals. The newly enculturated immature individuals then interact with the environment and adjust their behaviour according to processes of individual learning. Finally, the environment takes its toll on the population, removing some individuals and sparing others to produce a new, mature population. The process then repeats with a new immature population.

Following B&R, the simplifying assumption is made that cultural transmission, individual learning and selection can be separated out into these discrete stages. In a more realistic model these processes would be continuously modifying the population.

The distribution of phenotypes in a population at time t can be given by F_t, which specifies the proportion of each phenotype in the population. How does F_t change over time?

We will assume that F_t gives the initial phenotype distribution, prior to any social transmission, individual learning or environmental impacts. F'_t gives the distribution of phenotypes in the population after cultural transmission. This depends on three factors: 1) F_t, the distribution of phenotypes in the population prior to cultural transmission; 2) F'''_{t-1}, the distribution of mature phenotypes in the previous generation participating in cultural transmission; 3) the mechanism of cultural transmission.

F'''_t gives the distribution of phenotypes in the population once the interaction of individuals with the environment have been taken into account. These interactions can be separated into two parts: individual learning, which yields a phenotype distribution F''_t, and differential retention, which yields the final phenotype distribution F'''_t.

$$\begin{tikzcd}
F'''_{t-1} \arrow[r, "\text{cultural transmission}"] & F'_t \arrow[r, "\text{individual learning}"] & F''_t \arrow[r, "\text{selection}"] & F'''_t \\
& F_t \arrow[u] & &
\end{tikzcd}$$

Figure 2.1: A general model of cultural transmission. F_t is the original distribution of phenotypes at time t (determined by factors other than culture). F'_t is the distribution of phenotypes after cultural transmission. F''_t is the distribution of phenotypes after individual interaction with the environment (specifically, learning). F'''_t is the distribution of phenotypes after individual interaction with the environment (specifically, death).

F''_t gives the distribution of phenotypes in the population once individual learning has been taken into account. This depends on two factors: 1) F'_t, the distribution of phenotypes prior to individual learning; 2) the process of individual learning, by which individuals change their phenotype in response to the environment.

F'''_t gives the distribution of phenotypes in the population after removal of individuals through death has been taken into account. This depends on: 1) F''_t, the distribution of phenotypes prior to death; 2) the process of differential retention, by which some individuals survive and some individuals are removed due to environmental factors.

The general model of cultural transmission is illustrated in Figure 2.1.

2.1.2 Linguistic models

As discussed in Chapter 1, in the dominant Nativist paradigm language is viewed as an aspect of individual psychology. Language acquisition is seen as the "growth of cognitive structures [linguistic competence] along an internally directed course under the triggering and partially shaping effect of the environment" (Chomsky 1980:34). The environmental triggers which guide the growth of a particular linguistic competence come from the Primary Linguistic Data (PLD), the language the child observes others using. Those working within the Nativist paradigm typically emphasises the degenerate nature of the PLD, in part to offer support for the hypothesised innate UG. However, the fact that the PLD plays some role in the formation of linguistic competence suggests that language is, to some extent, culturally transmitted — an individual's linguistic competence,

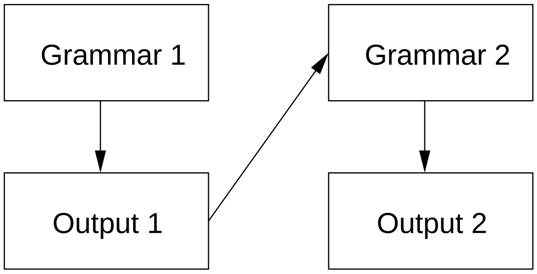


Figure 2.2: The transmission of language from generation to generation. The output of one grammar forms the PLD on which other grammars are based. Andersen (1973) points out that any attempt to make direct connections between grammar and grammar, or output and output, are spurious.

mediated by performance considerations, leads to linguistic behaviour, which forms the (degenerate) PLD for the formation of linguistic competence in other individuals.

Chomsky himself tends not to pursue this line of reasoning very far, being "concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community" (Chomsky 1965:3). However, the cultural transmission of language is of more interest to those concerned with language change, which necessarily involves a consideration of actual speaker-listeners in more or less heterogeneous speech communities.

Andersen (1973) presents an influential early account of phonological change. Andersen begins with the assertion that "[w]hat is needed is a model of phonological change which recognizes, on the one hand, that the verbal output of any speaker is determined by the grammar he has internalized, and on the other, that any speaker's internalized grammar is determined by the verbal output from which it has been inferred" (Andersen 1973:767). This scenario is sketched in Figure 2.2, adapted from Andersen's (1973) Figure 1.

Andersen applies this cultural approach to an account of phonological change in two dialects of Czech. Prior to 1300, Old Czech made a phonemic distinction between plain and palatalised ("sharped", in Andersen's terminology) dental and labial consonants. From 1300 to the end of the 1400s, most Czech dialects lost this phonemic opposition, palatalised dentals being replaced with plain dentals and palatalised labials being replaced with plain labials or plain labials plus /j/. However, in a group of dialects which Andersen terms the Teták dialects[1], palatalised labials became dentals before /i/, /e/ and /r/. The Teták dialects later lost the dental pronunciation and acquired the more

[1] This name derives from the pronunciation of the Czech word for "five" by speakers of such dialects. In more standard Czech five is /pjet/, whereas in Teták dialects it was pronounced /tet/. This difference is a consequence of the phonological change Andersen is concerned with.

standard labial pronunciation, perhaps due to the stigma associated with the Teták pronunciation — Andersen reports several standard manners for ridiculing Teták speakers, including /tiːte tiːvo ʃak je s tenou/ (meaning “Drink your beer, never mind the head”), which would be pronounced /piːte piːvo ʃak je s penou/ in non-Teták dialects. While Andersen provides an account of this later remedial change, we will focus here on his account of the first change which lead to the distinctive Teták pronunciation.

According to Andersen, the loss of a distinction between palatalised and plain labials and dentals in non-Teták dialects occurred due to errors by learners of those dialects in their analysis of the tonality of consonants. In Old Czech, the relevant consonants were either of high tonality (the dentals) or low tonality (the labials). Palatalisation further heightened the tonality of palatalised dentals or labials. This system is depicted as Stage 1 in Table 2.1a (based on Andersen’s Table 2). Heightened high tonality was reinterpreted as non-heightened high tonality by learners of the non-Teták dialects, leading to the loss of the palatalised versus non-palatalised contrast for dentals (Stage 2 in Table 2.1a). Subsequently, heightened low tonality was reinterpreted as non-heightened low tonality, yielding the final non-Teták system (Stage 3), with no phonemic distinction based on palatalisation.

The first two stages of the change in Teták dialects proceeded in the same way, as illustrated in Table 2.1b (from Andersen’s Table 3) — the palatalisation distinction was lost for dentals. However, the third stage of the change proceeded differently in the Teták dialects. The acoustic manifestation of heightened low tonality is ambiguous. In the non-Teták dialects this ambiguous tonality was interpreted by learners as representing underlying non-heightened low tonality. However, in the Teták case learners interpreted heightened low tonality as a realisation of underlying non-heightened high tonality. Adults produced linguistic behaviour which contain realisations of an underlying phoneme $/p^j/$. This behaviour constituted the PLD for learners. However, the acoustic ambiguity of the PLD led learners to interpret the consonant of interest as a manifestation of the underlying phoneme /t/.

Andersen gives examples of similar changes, induced by the ambiguity of the PLD, in the consonant system of early Latin and the vowel system of Old English. The account of the different paths taken by non-Teták and Teták dialects as sketched here is of course incomplete — it remains to be explained why learners of Teták dialects consistently reduced the phonemic distinction along different lines from learners of other dialects of Czech. This could be due to chance, or the phonological or phonetic properties of other parts of the Teták dialects. Andersen also proposes a system of *adaptive rules*, by which learners repair some of their misacquisition of the consonant system by realising certain instances

(a)

	Stage 1	Stage 2	Stage 3	
Heightened high tonality	$/t^j/$	/t/	/t/	High tonality
Non-heightened high tonality	/t/			
Heightened low tonality	$/p^j/$	$/p^j/$	/p/	Low tonality
Non-heightened low tonality	/p/	/p/		

(b)

	Stage 1	Stage 2	Stage 3	
Heightened high tonality	$/t^j/$	/t/	/t/	High tonality
Non-heightened high tonality	/t/			
Heightened low tonality	$/p^j/$	$/p^j/$		
Non-heightened low tonality	/p/	/p/	/p/	Low tonality

Table 2.1: The loss of phonemic distinctions in Czech. (a) sketches the situation in non-Teták dialects. At Stage 1 there is a distinction between palatalised and plain /t/ and /p/. The distinction between the palatalised and plain dental stops is lost at Stage 2. At Stage 3 the distinction is lost for the labial stop. This results in a reduction from four distinct levels of tonality at Stage 1 to two levels of tonality at Stage 3. (b) shows the situation in the Teták dialects. Stage 2 proceeds as in (a). However, at Stage 3 the distinction is lost in a different manner, with heightened low tonality being reinterpreted as plain high tonality.

of underlying /t/ as $/p^j/$, in line with their adult models. This additional detail is not important for our purposes — the crucial point of Andersen's work is his conclusion that misinterpretation of the PLD by language learners can lead to language change.

Lightfoot (1979) attempts to provide a fairly general account of syntactic change. Lightfoot's approach is similar to Andersen's (1973) in taking note of the cultural dimension of linguistic transmission. Under Lightfoot's account, languages gradually accumulate opacity in the derivation of surface forms from underlying syntactic structures. Once this opacity exceeds some threshold tolerance level, a therapeutic alteration to the underlying grammar is made by language learners, in order to improve the transparency of derivation. The details of Lightfoot's notions of transparency of derivation are not important here — the key point is that Lightfoot, working within the Chomskyan framework, addresses the cultural nature of language transmission:

> "An individual may be exposed to PLD which is different from the parents PLD . . . One individual may set some parameter differently from older people in her community; then it is likely that, because of the grammatical change, she will produce different utterances from other people in her community. These new expressions, in turn, affect the linguistic environment, and she will now be an agent of further change, by virtue of the fact that her younger siblings will have different PLD as a result of what she produces

> with her new grammar. As the younger siblings also set the relevant parameter in the manner of the older sister, so other people's PLD will differ. Thus a chain reaction is created" (Lightfoot 1999:101)

One of Lightfoot's central concerns is to emphasise the importance of a theory of grammar in understanding language change. For Lightfoot, the main constraint on language change derives from the restrictions placed on possible grammars by UG:

> "Each generation has to construct a grammar anew, starting from scratch. Speakers of a given grammar construct a grammar on the basis of the primary data available [the PLD] . . . A subsequent generation constructs a grammar in the same way, but if the primary data is now slightly different the grammar hypothesized will also be different, and there is no reason why it should bear any closer formal relation to that of the parent generation beyond the defining requirements of a theory of grammar; after all, small differences in output may result in large differences in the grammar, and vice versa" (Lightfoot 1979:147)

Lightfoot is particularly keen to rule out grammar-independent principles of change, which would predict how languages would change and do so in terms abstracted away from a particular theory of grammar. He criticises Traugott's (1965) position that "[t]he objectives of diachronic linguistics have always been to reconstruct the particular steps by which a language changes, and also to hypothesize about processes of language change in general", countering "[t]he neogrammarian legacy of a search for independent principles of change must be abandoned . . . the distinction outlined here between theories of grammar and change requires a change of focus, such that these predictions are derived mostly from a theory of grammar" (Lightfoot 1979:153).

Lightfoot's position is interesting for several reasons. Firstly, it shows that the Chomskyan position can amenably be applied to a study of language change, in which cultural transmission, via the PLD, plays some role. Lightfoot's objection to independent principles of change will be returned to below in connection with Hurford's conception of the Arena of Use, and also in Section 2.3.6 and Chapter 5, where research suggesting a principle of change (from less to more generalisable grammars) is presented.

Hurford (1987) presents a further elaboration of Andersen's (1973) conception of language change. Hurford closes the loop between the PLD and grammatical competence via the substrate in which communication takes place, which Hurford dubs the Arena of Use. An individual's grammatical competence is expressed in the Arena of Use, which itself imposes further communicative pressures and social and cognitive constraints. This behaviour in the Arena of Use provides the PLD upon which grammatical competence is acquired. The closing of the loop between grammatical competence and PLD via the Arena of Use is illustrated in Figure 2.3. According to Hurford, the consequences resulting from the filter imposed by the Arena between competence and linguistic output are potentially non-trivial:

> “a speaker learns the most successful ways of expressing his meanings, and the statistical shape of his output is thus influenced by his experience. It is of course conceivable that the LAD is so rich that it makes full allowance for the effect of the arena of use on linguistic output. That is, the LAD might be able to compensate fully for the ‘distorting’ factors affecting output and be able to retrieve a more or less perfect replica of the competence(s) involved in producing the output. But this strikes me as very implausible” (Hurford 1987:22)

In other words, the pressures acting on language as it passes through the Arena of Use may skew the PLD to learners so as to effect language change. The acoustic ambiguities which Andersen hypothesises lead to the change in the phoneme inventory of Czech are one possible consequence of the passage of language through the Arena of Use. Hurford proposes another, hypothetical, case, where presuppositions about the world associated with a particular culture lead to inanimate objects rarely being referred to by grammatical subjects. This consequence of the Arena of Use might lead, Hurford suggests, to children internalising a grammatical rule prohibiting the appearance of inanimate noun phrases in subject positions — the way of seeing the world, determined by non-linguistic culture, impacts via the Arena of Use on the grammar. To the extent that the consequences of the Arena are predictable, they may result in, contra Lightfoot, independent principles of change.

To summarise the positions of Andersen (1973), Lightfoot (1979) and Hurford (1987), an individual’s grammatical competence is now not solely a matter of individual psychology, but how that individual psychology applies to data provided by other individuals. Linguistic behaviour, or E-language[2], begets linguistic competence, I-languages, which beget, via the Arena of Use and the LAD, E-languages. This cultural process, which has been dubbed the Expression/Induction (E/I) cycle (Hurford 2002a) is sketched in Figure 2.4.

[2] By E-language I mean the external linguistic behaviour of individuals. The same term has sometimes been taken to refer to language as an object external to human minds. This is not the interpretation which makes most sense in this context.

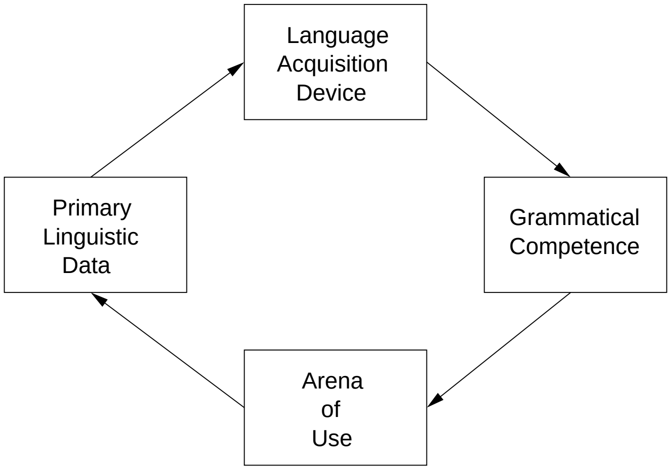


Figure 2.3: The Arena of Use in the cycle of linguistic transmission. The Arena mediates between an individual's grammatical competence and their expression of that competence as linguistic behaviour. Considerations arising from the Arena of Use may impact significantly on the PLD available to subsequent language learners, and may therefore result in language change.

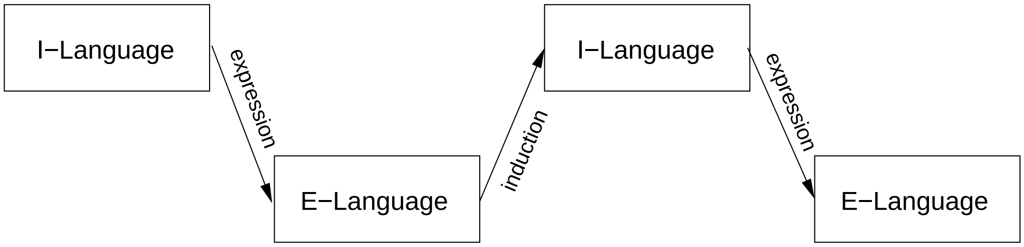


Figure 2.4: The Expression/Induction cycle. I-Language leads, via expression, to E-language. E-language leads, via induction (or acquisition, in more neutral terms) to I-language.

The E/I model constitutes a fairly general model of the cultural transmission of language. Its main shortcoming, as revealed by a comparison with B&R's general model of cultural transmission outlined in Section 2.1.1, is a lack of a formal specification of the pressures acting on language during its cultural transmission. This issue has been addressed recently with the advent of formal models of linguistic evolution, which adopt the E/I cycle as their starting point. These models can broadly be classified as either *Negotiation Models* (NMs) or *Iterated Learning Models* (ILMs). The primary methodological difference between the NM and ILM approaches is in their treatment of the verticality of cultural transmission.

B&R's general model given above is framed purely in terms of what is known as *vertical* transmission — a population is considered to consist of discrete, non-overlapping generations, with cultural transmission only taking place between generations. Intra-generational transmission is referred to as *horizontal* transmission, and is ignored in the model sketched above. However, the mathematical analyses which B&R use to investigate cultural evolution in the general model are essentially agnostic about the vertical-horizontal distinction — much of the analysis remains the same if we view a "generation" in the general model as snapshot of a monogenerational population interacting with itself.

Similarly, the E/I model is essentially agnostic regarding the vertical-horizontal distinction. In contrast, the distinction between vertical and horizontal transmission has become something of a defining characteristic in implementations of models of the cultural transmission of language. In the NM transmission is exclusively horizontal, whereas in the ILM vertical transmission is paramount. The transmission of language in the real world presumably lies somewhere between these two extremes.

2.1.2.1 The Negotiation Model

Populations in the NM consist of collections of individuals, who acquire their linguistic competence based on observations of the behaviour of other individuals. A population consists of a monogenerational collection of individuals. At each time-step members of the population produce observable linguistic behaviour, which is observed and learned from by other members of the population. This process is illustrated in Figure 2.5. There is no population turnover in the NM — new individuals do not enter the population and no individual leaves. Transmission is therefore exclusively horizontal. The NM framework is used by, for example, Hutchins & Hazelhurst (1995), Steels (1997), Batali (1998), Hazelhurst & Hutchins (1998), Steels (1998), Smith (2001a) and Batali (2002).

2.1.2.2 The Iterated Learning Model

In the ILM, as in the NM, populations consist of collections of individuals, who acquire their linguistic competence based on observations of the behaviour of other individuals. Unlike in the NM, there is population turnover in the ILM. This turnover can either be generational (see Figure 2.6 a), where the entire population is replaced at each time-step, or gradual (see Figure 2.6 b), where a single individual is replaced at each time-step. In the generational ILM individuals at generation $n + 1$ acquire their competence based on observations of the behaviour of the generation n population. In the gradual turnover ILM, each new individual acquires its competence based on observations of the behaviour of the population they enter. In either case, transmission is exclusively vertical, from fully enculturated individuals to naive individuals. The ILM framework is used in, for example, Hare & Elman (1995), Oliphant & Batali (1997), Kirby (1999), Livingstone & Fyfe

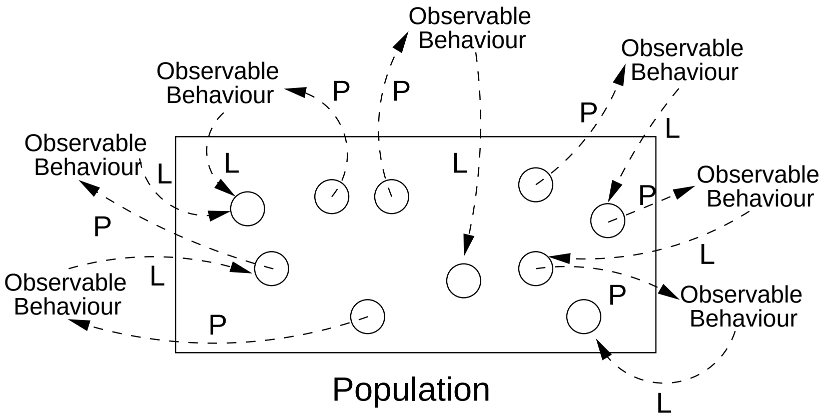


Figure 2.5: The Negotiation Model. The population consists of a collection of individuals, represented by circles. Individuals produce observable (linguistic) behaviour, which is learned from by other members of the population. There is no population turnover. Cultural transmission is therefore purely horizontal, acting within the monogenerational population.

(1999), Nowak *et al.* (1999), Oliphant (1999), Brighton (2000), Hurford (2000), Kirby (2001), Nowak *et al.* (2001), Kirby (2002), Smith (2002), Smith *et al.* (forthcoming) and Smith *et al.* (submitted).

2.2 Transmission and Cultural Traits

Given these two frameworks for studying cultural transmission (B&R's general model, and the E/I framework), several issues remain to be resolved. Firstly, how do we characterise the range of possible cultural variants? Secondly, how are these cultural traits transmitted? Section 2.2.1 describes B&R's two models of cultural traits and the transmission of such traits. Section 2.2.2 outlines the nature of cultural traits in the E/I model and the somewhat contentious issue of what information is available to learners during the transmission of linguistic structure.

2.2.1 Cultural traits and transmission in the general model

B&R provide two basic and very abstract treatments of possible cultural traits. In the *dichotomous trait* model there are two possible cultural traits, with individuals having one or the other trait. In the slightly more complex *continuous trait* model there are an unlimited number of possible cultural traits, each of which is assigned a numerical value. Each individual's cultural character is given by the numerical value of the trait they possess.

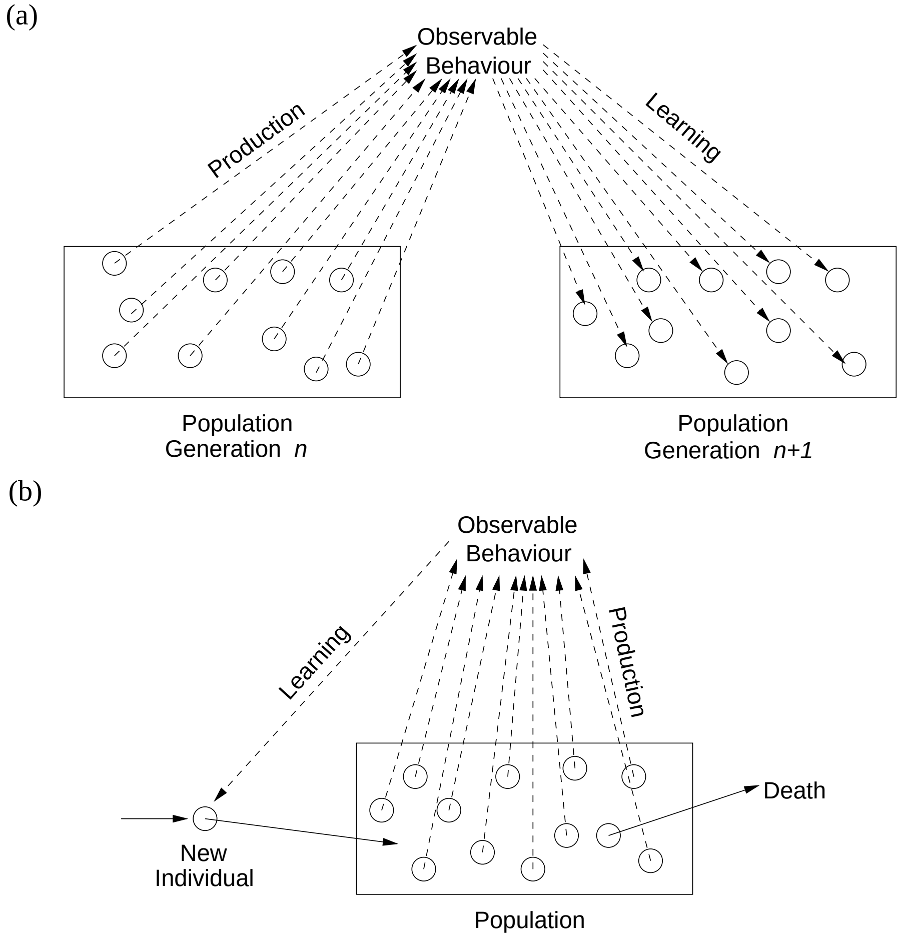


Figure 2.6: The Iterated Learning Model. In the generational version of the ILM (a), the observable behaviour produced by generation n individuals is observed and learned from by generation $n + 1$ individuals. In the gradual turnover version (b), a single individual is removed from the population at each time-step, to be replaced by a single individual. The new individual learns from the observable behaviour produced by the rest of the population. In both versions of the ILM cultural transmission is therefore purely vertical, with transmission proceeding from fully enculturated individuals to naive individuals.

B&R are not primarily concerned by how an individual's cultural trait manifests itself in that individual's behaviour. They assume that naive individuals can estimate the cultural trait possessed by mature individuals, possibly with some small error. B&R are more concerned with how the transmission of such traits can affect the cultural make-up of the population. In part, B&R can avoid being specific about how cultural traits are stored in the brain, manifested in behaviour and estimated by naive individuals because their model is very general. Firstly, their somewhat simplistic treatment of cultural traits is unlikely to specifically offend any particular specialism, or indeed offend all groups equally. Secondly, their model of cultural traits *should* be poorly specified, given their general aims of investigating how cultural transmission can result in cultural evolution, abstracted away from any particular putative cultural trait.

As shown in Section A.1.1 of Appendix A, B&R show that cultural transmission alone, of either dichotomous or continuous traits, does not result in cultural evolution — assuming that naive individuals acquire their cultural traits on the basis of a random sample of enculturated individuals, and individuals are unbiased with respect to which cultural trait they acquire, the distribution of cultural variants in the population will remain unchanged. Cultural evolution therefore does not automatically follow from cultural transmission — some pressure other than unbiased transmission is required for cultural evolution.

2.2.2 Cultural traits and transmission in linguistic models

B&R's simple characterisation of cultural variants and their assumption that the particular cultural variant an individual possesses is easily determinable from that individual's behaviour prove somewhat limiting when applied to the modelling of the transmission of language. There are three key difficulties in applying B&R's approach to models of linguistic transmission, outlined in the following three sections.

2.2.2.1 The simplicity of B&R's cultural traits

Firstly, B&R's characterisation of cultural traits (dichotomous or continuous) is too simple to capture most aspects of linguistic behaviour. Variations of their model have proved useful in certain cases. For example, Kirby (1999) (Chapter 2) considers the cultural evolution of grammars which exhibit either verb-object or object-verb order, and are either prepositional or postpositional. Kirby therefore treats grammars as pairs of dichotomous traits, and models how the distribution of the two traits impact on each other. Briscoe (2000a) characterises grammars as either head-initial or head-final, equivalent to a single dichotomous trait. Finally, Nowak *et al.* (2001) model the cultural evolution of multiple competing grammars. In B&R's terms, each grammar could be treated as a distinct integer, with grammatical competence then corresponding to a continuous trait. These models are described in more detail in Section 2.3.

However, B&R's model of cultural traits is insufficient to deal with more detailed questions of linguistic structure. Firstly, in a dichotomous or continuous trait model there is no notion of the structure *within* a particular trait. It would be possible to interpret such a model in this way, and say, for example, that trait *t* represents a language with agglutinating morphology and head-initial syntax. This is essentially the approach adopted by Kirby (1999) and Briscoe (2000a). There are two main problems with this approach:

1. There is no way to model how the internal structure of a morphological or syntactic system changes over time due to cultural transmission — a cultural trait is either present or absent in B&R's system, and the only possible change in a cultural trait is a change from presence in a particular individual to absence, or vice versa.
2. There is no way to investigate how the internal structure of a grammatical system impacts on its fecundity or fidelity during cultural transmission, short of explicitly imposing such factors. For example, Kirby (1999) defines a constant which determines to what extent verb-object order is favoured in the presence of prepositions, then sets this constant to some (theoretically well-motivated) value.

In addition to these problems relating to a lack of structure *within* a particular traits, there is an associated problem of a lack of structure *between* traits. In the dichotomous model there is no structural relationship between traits, other than one of dichotomy. In the continuous trait model, cultural traits are organised in a linear fashion, with the only structural relationship being one of numerical distance. Nowak *et al.* (2001) fall foul of this problem in modelling the degree of similarity between grammars, which they effectively treat as numerical values. Given the lack of structure within grammars in their model, it is meaningless to say to what extent users of two distinct grammars share grammatical structures, or to quantify the probability of a learner acquiring a particular grammar on exposure to expressions generated by another grammar. Nowak *et al.* circumvent this problem by assigning arbitrary (constant or random) degrees of similarity between grammars, a rather unsatisfactory solution.

Most implementations of E/I models therefore develop a more complex treatment of cultural traits. The nature of the model depends on the linguistic behaviour of interest. Broadly, models can be classified as investigating the cultural transmission of *vocabulary* systems, or *syntactic* systems.

In vocabulary models, an individual's competence typically consists of a mapping between a set of unstructured meanings and a set of unstructured signals. Such models are typically concerned with the impact of the population's vocabulary on communication within the population. These models therefore typically include some evaluative phase, where individuals attempt to communicate meanings to one another using their acquired mappings from meanings to signals. The communicative accuracy between two individuals is calculated according to a formula with the canonical form:

$$\text{communicative accuracy}\,(S,H)=\sum_i\sum_j p\left(s_j|m_i\right)\cdot r\left(m_i|s_j\right)$$

where S and H are speaker and hearer, $p\left(s_j|m_i\right)$ gives the probability of the speaker S producing signal s_j to communicate m_i and $r\left(m_i|s_j\right)$ gives the probability of the hearer H interpreting signal s_j as communicating meaning m_i. It should be noted that 1) the evaluation of communicative accuracy is typically distinct from the cultural transmission phase and 2) communicative accuracy typically plays no role in shaping the communication system of the population and 3) implicit in this measure is the assumption that meanings are functionally distinct — for example, if two meanings m_i and m_j result in the same behaviour on the part of the receiver and $r\left(p\left(m_i\right)\right)=m_j$ then the communication would be measured as a failure but could, at the behavioural level, be considered a success.

The definition of communicative accuracy above implies a definition of communication which is similar to that of Johnson-Laird (1990), who states that “the communicator [must] construct an internal representation of the external world, and then …carry out some symbolic [not necessarily in the strict sense I have used in Chapter 1] behaviour that conveys the content of that representation. The recipient must first perceive the symbolic behaviour, i.e. construct its internal representation, and then from it recover a further internal representation of the state that it signifies” (Johnson-Laird 1990:2–4). Communication systems are therefore defined as systems for mapping between meanings and signals. Communication is a success where the internal representations of communicator and recipient are the same.

In syntactic models, an individual's competence typically consists of some system, possibly principled, for mapping between structured meanings and structured signals (for example, a context free grammar with semantic operations associated with rewrite rules). Syntactic models tend to be less concerned with how cultural evolution impacts on communication within a population, and more concerned with how the system for mapping from meanings to signals changes over time. A key question tends to be the extent to which the structure of meanings and signals is exploited during the meaning-signal mapping process — do compositional mappings emerge?

Examples of these models will be described in more detail later in this Chapter. The key point here is that a much more complicated model of cultural traits is used. This has several benefits. Firstly, interesting linguistic behaviour can be modelled in a less abstract way. An explicit model of a meaning-signal mapping, be it a vocabulary-type mapping or a syntactic mapping, allows measures of structure within and between different individual's acquired cultural traits to be fairly naturally defined. However, this enriched treatment does have drawbacks. Firstly, the notion of cultural traits becomes somewhat fuzzy. For example, in a vocabulary system, is a cultural trait an individual's whole system of mapping from meanings to signals, or could an individual's cultural character be considered as consisting of several traits, with each trait specifying a single meaning-signal association? Secondly, the more complex models make the type of mathematical analysis used by B&R difficult, if not impossible. There is a tradeoff between the richness of the treatment of cultural traits and the transparency of the model's results. One of the goals of this Chapter is to relate B&R's simple models to the more complex, linguistically-flavoured models, allowing the rich results generated by the latter to be interpreted in the simple, clear terms of the former.

2.2.2.2 Is the cultural transmission of language possible?

A second major problem with applying B&R's approach to linguistic evolution is identifying what cultural traits correspond to in linguistic theory. In terms of the E/I framework, there are two possibilities — a cultural trait could characterise an individual's linguistic competence, their I-language, or their linguistic behaviour, their E-language. Our preference, implicit in the discussion above, should perhaps be towards interpreting an individual's cultural trait as corresponding to their I-language. E-language is derived from I-language, and is contingent on the set of situations which require this internal competence to be pressed into service to produce linguistic behaviour. With this interpretation, the range of possible cultural variants is circumscribed by the limitations imposed by the representation of I-language. An individual's externally visible manifestation of their cultural trait is the E-language they produce, which is determined by factors related to considerations imposed by the Arena of Use.

However, this interpretation throws up another problem. I-language, as is apparent from Figures 2.2, 2.3 and 2.4, is not directly transmitted, but is acquired via the PLD, which in turn comes from E-language. B&R assume that an individual's cultural trait can be determined from their behaviour either straightforwardly (in the dichotomous trait model) or with some normally-distributed error (in the continuous trait model). This is too simplistic an approach to dealing with the transmission of I-language. To take an extreme view, it could be argued that the filtering through E-language makes it difficult, or inaccurate, to speak of I-language being transmitted at all. This would be the position favoured by Chomsky ("it seems that a child must have the ability to 'invent' a generative grammar" (Chomsky 1965:201)) and Lightfoot ("[e]ach generation has to construct a grammar anew, starting from scratch" (Lightfoot 1979:147)). This argument could be applied to the cultural transmission of any trait. For example, how can a naive individual guess at the internal state of its cultural parent that determines political persuasion, based on behaviour?

Two points should be made to at least cast doubt on this very strong negative position. There is a large body of evidence suggesting that cultural transmission is a fact for non-linguistic traits — B&R cite 57 articles in a brief review of evidence pointing towards the reality of cultural transmission (Boyd & Richerson 1985:47–55). It is therefore possible that cultural transmission also applies to language. With particular reference to language, the fact that speakers within a speech community tend to agree on what constitute valid grammatical sentences and so on should make us doubt that they all have radically different I-languages. The most parsimonious explanation for this is that I-language is, to some extent, culturally transmitted.

If we reject the strong negative view, we are left with a wide range of possible degrees of cultural transmission. The strongest positive view would be that the filter through E-language is completely irrelevant. This too seems unreasonable. Firstly, there may be several (or indeed infinitely many) possible grammars which are consistent with a particular PLD. A learner bound to be consistent with the observed PLD might therefore converge on a different grammar from the grammar which produced that PLD. Whether or not learners are bound to be consistent with their PLD is another issue. Both Andersen (1973) and Lightfoot (1979) assume that they are. However, radical restructuring events such as creolization suggest that this constraint may be fairly weak. A second point against the strongly positive view of cultural transmission of I-language would be Hurford's (1987) comments about the possible skewing effects introduced by the Arena of Use.

Where does this leave us? B&R's assumption of unproblematic cultural transmission is too simple when it comes to language — the filtering of I-language through E-language is unlikely to be so trivial as be ignorable. However, I-language probably is transmitted to some extent — we should reject Chomsky's and Lightfoot's position that children invent their I-language anew each generation.

2.2.2.3 The nature of the PLD

A subsidiary question is: what actually constitutes the PLD? The only aspect of linguistic behaviour which is uncontroversially determinable is the acoustic (or visual) sequence produced by an individual when that individual speaks (or signs). However, most implementations of the E/I model (NMs and ILMs) assume that this observable signal is paired with the communicative intention of the individual producing the signal — the PLD consists of meaning-signal pairs. This is not an uncontroversial assumption, nor is it an assumption which is always made.

In most implementations of the E/I model there is no explicit modelling of an environment outwith the individuals that make up the population — meanings and signals in the model are arbitrary agent-internal representations. It is typically assumed that there is some shared, stable mapping between external situations and internal representations of those events — indeed, cultural transmission of linguistic structure would be impossible without an external manifestation of internal representations of signals. However, this mapping is not typically the focus of E/I models, which concentrate instead on the mapping between internal representations of meanings and signals. An account of language as a mapping from aspects of the environment to other aspects of the environment must account for two additional mappings (see Figure 2.7):

- the mapping between states of the environment representing situations to be communicated and internal representations of those states (meanings).
- the mapping between communicative alterations of the environment and internal representations of those alterations (signals).

The nature of the mapping between environment and meaning forms a key part of the symbol grounding problem (Harnad 1990) — the meaning of internal representations must, at some point, be related to the objects or situations in the world which they refer to. The mapping between environment and signal corresponds to a mapping between strings of phonemes and articulatory movements. Most E/I models simply assume that these two mappings from internal representations to the environment are shared by all simulated individuals.

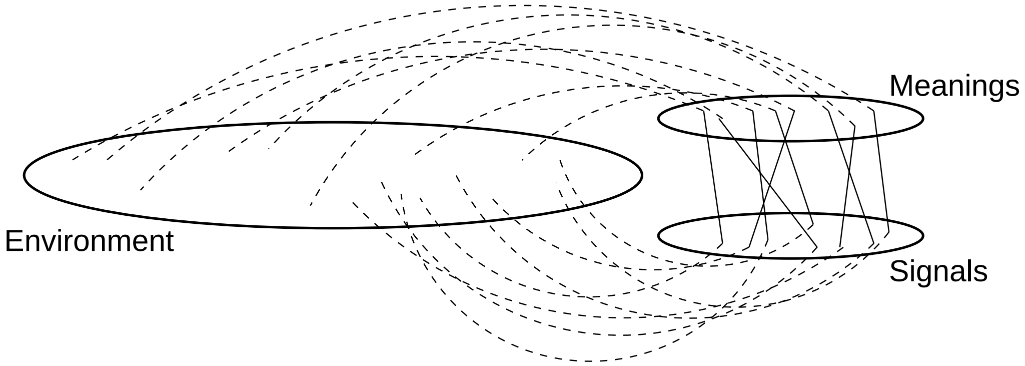


Figure 2.7: Language as a mapping between three spaces (represented as ellipses). Typically, E/I models focus on the mapping between two spaces — the internal representational spaces of meanings and signals. This mapping is given in solid lines. A complete model must account for two additional mappings — the mappings between the environment space and the internal representational spaces (dashed lines).

Given the absence of an explicit model of the environment, the assumption that learners observe meaning-signal pairs is unavoidable. In a fuller model, the more reasonable assumption could be made that learners are exposed to an environment which includes a state being communicated about and a set of articulatory gestures intended as a communicative alteration to the environment. The learner then has the task of identifying the communicatively relevant state and the communicative alteration, representing both internally and then learning the mappings between the internal representations and possibly the mappings between the internal representations and the environment.

There is a body of evidence which suggests that children have various strategies for mapping from the environment to internal representations of relevant parts of the environment. Much of this points to the importance of joint attention and intentional inference. Studies by Baldwin (Baldwin 1991; Baldwin 1993a; Baldwin 1993b) show that infants cannot learn words for toys simply by hearing the word for the toy while attending to the toy. The child must witness an intentional agent direct their attention to the toy while naming it. Under these circumstances the infant will learn the word for the toy, even if there is a delay between witnessing the intentional agent directing their attention at the toy and being able to attend to the toy directly themselves.

While most E/I models ignore the issue of the environment-internal representation mappings, some computational modelling work within the E/I framework has sought to tackle this problem. Neural network models show that genetic evolution can lead to the formation of internal representations which correspond to a categorisation of the environment (Cangelosi & Parisi 1998). These internal representations may form the basis of a (partially) culturally-transmitted communication system (Cangelosi 1999). Hazelhurst & Hutchins (1998) show that the negotiation of ritualised shifts of joint attention subserves the emergence of a learned communication system. Symbolic computational models demonstrate that shared mappings from the environment to internal representations of meanings can emerge through individual learning, both with explicit feedback (e.g. Steels (1997), Steels (1998)) and without (Smith 2001a). Finally, it has been demonstrated that repeated expression and induction of strings of words can lead to the emergence of meaning, where meaning is defined in terms of the relationship between words and other words — the emergent mesh of word-word associations constrains and guides the interpretation of signals (Hashimoto 1998). These models and the data from real language acquisition outlined above give some hope that an integrated model, of the type depicted in Figure 2.7, will be achievable. However, for the purposes of this thesis I will assume that the PLD available to learners consists of meaning-signal pairs.

2.3 Forces acting on cultural transmission

The general model of cultural transmission provided by B&R and the more language-specific E/I models raise several interesting issues on the nature of cultural traits and the manner in which these traits are transmitted. However, these questions are typically not the central concern of such models, nor are they of great importance for the purposes of this thesis. Of more interest is how the processes involved in cultural transmission outlined in Section 2.1.1 (transmission itself, individual learning by enculturated individuals, and selective removal of enculturated individuals) impact on the distribution of cultural traits in the population. Do any of these processes lead to significant cultural evolution?

The models of B&R are used to frame much of this discussion. They provide mathematical accounts of how three pressure acting on transmission can result in cultural change and cultural evolution. These are:

1. Natural selection of cultural variants, resulting from selective removal of enculturated individuals.
2. Guided variation, resulting from individual learning by enculturated individuals.
3. Biased transmission, resulting from the strategy of learners during cultural transmission. The forces of biased transmission can be further subdivided into three forms:
   (a) Directly biased transmission, resulting from a preference for learners to acquire one cultural variant over another.

(b) Indirectly biased transmission, resulting from a preference for learners to acquire cultural traits which are associated with other cultural traits.

(c) Frequency-dependent transmission, resulting from a disproportionate preference for learners to acquire the most (or least) frequent cultural trait in the population.

In Sections 2.3.1 to 2.3.5 B&R's models for these pressures are reviewed. In the interests of clarity, a separate section is devoted to each of the three subtypes of biased transmission. Mathematical details are given in Appendix A.

The main goal of this part of the thesis is to relate B&R's simple models to the more complex, linguistically-flavoured models. This allows the linguistically-interesting, somewhat complex results generated by the linguistic models to be interpreted in the simple, clear terms of B&R's models. To this end, each section describing one of B&R's pressures on cultural transmission is followed by one or more sections which introduce details of a linguistic (usually E/I) model, and discusses how that model can be interpreted in terms of B&R's model.

Finally, in Section 2.3.6, a new driving force for cultural evolution is introduced. This pressure, which arises from transmission through a bottleneck, does not feature in B&R's taxonomy, but is extremely relevant to the cultural transmission of linguistic structure.

As the structure of this Chapter is somewhat intricate, a full preview is perhaps in order.

Section 2.3.1 : Natural selection of cultural variants

- **Section 2.3.1.1** : B&R's model
- **Section 2.3.1.2** : Linguistic case study: the evolution of vocabulary under natural selection.
- **Section 2.3.1.3** : Linguistic case study: the evolution of grammar under natural selection.

Section 2.3.2 : Guided variation

- **Section 2.3.2.1** : B&R's model
- **Section 2.3.2.2** : Linguistic case study: Tomasello's cultural ratchet

Section 2.3.3 : Directly biased transmission

- **Section 2.3.3.1** : B&R's model
- **Section 2.3.3.2** : Linguistic case study: the evolution of consistent head ordering under direct bias
- **Section 2.3.3.3** : Linguistic case study: the evolution of vocabulary under direct bias

- **Section 2.3.3.4** : Linguistic case study: the evolution of morphological structure under direct bias

Section 2.3.4 : Indirectly biased transmission

- **Section 2.3.4.1** : B&R’s model
- **Section 2.3.4.2** : Linguistic case study: Croft’s utterance-based theory

Section 2.3.5 : Frequency-dependent transmission

- **Section 2.3.5.1** : B&R’s model
- **Section 2.3.5.2** : Linguistic case study: the evolution of head ordering under frequency-dependent bias

Section 2.3.6 : Transmission through a bottleneck

- **Section 2.3.6.1** : Linguistic case study: the evolution of recursive compositionality in the ILM
- **Section 2.3.6.2** : Linguistic case study: the evolution of recursive compositionality in the NM

2.3.1 Natural selection of cultural variants

Natural selection was originally conceived of by Darwin as a mechanism for explaining the appearance of design in biological organisms.

> “Owing to this struggle for life, variations, however slight and from whatever cause proceeding, if they be in any degree profitable to the individuals of a species, in their infinitely complex relations to other organic beings and to their physical conditions of life, will tend to the preservation of such individuals, and will generally be inherited by the offspring. The offspring, also, will thus have a better chance of surviving, for, of the many individuals of any species which are periodically born, but a small number can survive. I have called this principle, by which each slight variation, if useful, is preserved, by the term Natural Selection.” (Darwin 1859/1964:61)

Modern definitions appeal to three factors mentioned in Darwin’s original formulation — variation, inheritance and selective survival or propagation. While there is some ongoing debate about the precise formulation of a definition of natural selection, Futuyma (1998) concludes that “[m]ost authors agree that the definition must include the following concepts: some attribute or trait must vary among biological entities, and there must be a consistent relationship, within a defined context, between the trait and one or more components of reproductive [implying heredity] success, where ‘reproductive success’ includes both survival (a prerequisite for reproduction) and the reproductive processes themselves.” (Futuyma 1998:349).

This definition is actually rather general, excepting the single reference to “biological entities”. Donald Campbell (Campbell 1965; Campbell 1975) presents an extension of the principle of natural selection to account for cultural evolution. Campbell’s main contribution is to present an argument that the three central factors (variation, inheritance and selective retention or reproduction) required for natural selection occur in cultural systems. If cultural systems exhibit variation within or between populations, if cultural systems are in some sense heritable, and if there is selection either in the survival of cultural systems or selection in the elevation to roles allowing influence in the enculturation of others, then we should expect to see cultural evolution under natural selection. Campbell is optimistic that culture exhibits variation and selection, but acknowledges that “retention and duplication [inheritance], is also more problematic for social evolution than for biological evolution. What are required are mechanisms for loyally reproducing the selected variations.” Campbell concludes that such inheritance systems are possibly present in culture — “through social mechanisms of child socialization, reward and punishment, socially restricted learning opportunities, identification, imitation, emulation, indoctrination into tribal ideologies, language and linguistic meaning systems, conformity pressures, social authority systems, and the like, it seems reasonable to me that sufficient retention machinery exists for a social evolution of adaptive social belief systems and organizational principles to have taken place” (Campbell 1975:1107).

2.3.1.1 B&R’s model

B&R model the natural selection of cultural variants by assuming that there are a set of distinct social roles (e.g. mother, father, uncle, priest, teacher). Each naive individual acquires their cultural characteristic based on observation of a subset of these roles. In order to model natural selection we must assume that the probability that an individual attains a particular social role depends on the cultural variant that that individual possesses. As shown in Section A.1.2.1 of Appendix A, if a particular cultural variant c offers a selective advantage when averaged over social roles (i.e. individuals with variant c are, on average, more likely to attain a social role than individuals with some other cultural variant) then it will increase in frequency in the population. In other words, if possessing variant c makes an individual more likely to occupy a role which allows them to enculturate others and transmit that variant, then c will increase in frequency in the population. The rate of increase of the favoured variant is dependent on cultural variation in the population — in the extreme case, where the population exhibits no variation, natural selection of cultural variants has no impact.

2.3.1.2 Linguistic case study 1: the evolution of vocabulary under natural selection

Martin Nowak and colleagues have applied techniques from theoretical biology to the study of language evolution. With the exception of one published work (Nowak *et al.* (2000), discussed in Chapter 6), the work of Nowak *et al.* falls within what B&R describe as natural selection models of cultural evolution. Nowak *et al.* make the assumption, following the lead of Pinker & Bloom (1990), that "[i]t pays to talk. Cooperation in hunting, making plans, coordinating activities, task sharing, social bonding, manipulation and deception all benefit from an increase in expressive power. Natural selection . . . can certainly see the consequences of communication" (Nowak & Komarova 2001:288). Based on this assumption, Nowak *et al.* use essentially the same model to study the evolution of symbolic vocabulary (described here) and universal grammar (as described in Section 2.3.1.3).

In Nowak *et al.* (1999), individuals are required to communicate about n objects using m signals. Each individual is characterised by an *association matrix*, A, which is an $n \times m$ matrix where entry a_{ij} gives the number of times during learning that that individual has observed its cultural parents communicating about object i using signal j. From this A matrix P and R matrices can be derived[3]. An individual's P matrix is an $n \times m$ matrix, where entry p_{ij} gives the probability of that individual producing signal j to communicate about object i. An individual's R matrix is an $m \times n$ matrix, where entry r_{ij} gives the probability of that individual associating signal i with object j when acting as a receiver. P and R are derived from A by normalising over rows and columns respectively:

$$p_{ij} = \frac{a_{ij}}{\sum_{l=1}^{m} a_{il}}$$

$$r_{ji} = \frac{a_{ij}}{\sum_{l=1}^{n} a_{lj}}$$

The communicative payoff for two individuals I_1 and I_2, where individual I_k is characterised by A matrix A_k, is:

[3] Nowak *et al.* refer to the R matrix as the Q matrix. However, R is used here to avoid confusion with the Q matrix introduced in Nowak *et al.* (2001), which serves a completely different purpose.

$$F\left(I_1, I_2\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(p_{ij}^{(1)} r_{ji}^{(2)} + p_{ij}^{(2)} r_{ji}^{(1)}\right)$$

where $p_{ij}^{(k)}$ is an entry in the P matrix derived from A_k and $r_{ij}^{(k)}$ is an entry in the R matrix derived from A_k. This equation states that the payoff for communication between I_1 and I_2 is the average of I_1's ability to communicate about objects to I_2 and I_2's ability to communicate about objects to I_1, averaged over all objects. The communicative payoff for an individual I_I with respect to a population of N individuals I_1 to I_N is:

$$F_I = \sum_J F\left(I_I, I_J\right)$$

where $J = 1, \ldots, N$ but $J \neq I$.

An individual arrives at their A matrix by sampling the production behaviour of K cultural parents. For each cultural parent, the individual observes them produce a signal for every object k times, according to the parent's P matrix. For each observation of a cultural parent producing signal j in association with object i the learning individual increments the value of a_{ij} with probability $1 - \rho$ and increments the value of $a_{ik \neq j}$ with probability ρ. ρ therefore gives the probability of errors during learning.

Nowak *et al.* consider two models of cultural transmission. In the first case, individuals select K cultural parents at random from the preceding generation of the population. In this case there is no natural selection on cultural transmission — an individual's communicative accuracy, a consequence of their culturally-acquired communication system, does not influence the probability of that individual acting as a cultural parent. Nowak *et al.* report two results for this case:

1. when $\rho = 0$ (learning is error free) the populations converge on sub-optimal communication systems. The population converges on a shared random binary (all entries are either 0 or 1) A matrix. Typically, these random matrixes will result in some intermediate level of communicative payoff. Fairly minor effects were found for different values of k (number of exposures to object-signal pairs) and K (number of cultural parents). For $k = 1$ or $K = 1$ convergence is rapid. For $k > 1$ or $K > 1$ convergence is somewhat slower, but the populations converge to similar levels of communicative payoff.

2. when $\rho > 0$ communicative payoff decreases, with shared communication systems failing to emerge when $\rho \geq 0.01$.

In their second model of transmission, each individual selects K cultural parents from the preceding generation, with the probability of any individual I being selected as a cultural parent being $F_I / \sum_J F_J$. This amounts to natural selection during cultural transmission, with systems which result in higher communicative payoff being more likely to be transmitted. Nowak *et al.* report three results for this case:

1. when $K = 1$ (individuals have a single cultural parent) and $0 < \rho < 0.01$ the populations converge on a range of communication systems, some of which offer maximal payoff and some of which are sub-optimal. The overall level of payoff is higher than for the no-selection case. $k = 1$ (a single exposure to each meaning-signal pair) results in the fastest convergence, but the eventual level of communicative payoff is independent of k.
2. when $K > 1$ (individuals have multiple cultural parents) and $\rho = 0$ (learning is error-free) the populations converge on a range of communication systems, some of which are suboptimal. Convergence is slower than in the $K = 1$ case, but the average payoff of the final systems is higher.
3. when $K > 1$ and $\rho > 0$ (learning is subject to errors) the average payoff of the populations depends on ρ. When $0 < \rho < 0.01$ the populations behave approximately as they did when $\rho = 0$, with some populations converging to optimal systems and some converging to suboptimal systems. When $\rho > 0.01$ the populations fail to converge on any shared communication system. However, when $\rho = 0.01$ all populations converge on an optimal system. Nowak *et al.* do not explore further, to identify the width of this "sweet spot" for ρ.

These results essentially meet the predictions of B&R's dichotomous trait model. In the absence of selection, the distribution of cultural variants remains unchanged except for changes introduced by random factors, which B&R do not consider. For the selection case, cultural variants which maximise some fitness function become more frequent, although this increase in frequency is dependent on variance in the population. Both $K > 1$ (multiple cultural parents) and $\rho > 0$ (errors during learning) introduce variance in the population, although there appears to be a sweet spot for ρ. Nowak *et al.* only experiment with a relatively limited range of values of K however, so it is not possible to tell if there is a sweet spot for K.

The model does offer several advances on the very simple model provided by B&R, however. There is a far wider range of possible cultural variants (there are a potentially infinitely many A matrices, although the set of culturally-stable A matrices is much smaller, being restricted to the m^n possible binary matrixes which have a single signal associated with each meaning). Secondly, the cultural fitness function depends on the structure of the cultural variants involved, rather than being arbitrarily assigned.

2.3.1.3 Linguistic case study 2: The evolution of universal grammar

Nowak *et al.* (2001) use a similar technique to study the evolution of universal grammar. In their model, a universal grammar U consists of a set of n grammars, numbered 1 to n. Each grammar G_i could be described as a rule system that defines a mapping between syntactic representations and semantic representations. σ_1 is the finite syntactic alphabet and σ_2 is the finite semantic alphabet. σ_1^* is the countably infinite set of all possible strings of characters drawn from σ_1, representing the set of all possible syntactic representations. Similarly, σ_2^* is the countably infinite set of all possible meanings, where meanings are strings of characters drawn from σ_2. A grammar G_i specifies a (potentially infinite) subset of $\sigma_1^* \times \sigma_2^*$, a mapping between semantic and syntactic representations. However, Nowak *et al.* do not exploit this notion of a grammar, as we will see — the behaviour of grammars with respect to communication and learning is essentially arbitrary.

As in Nowak *et al.* (1999), a measure of similarity among different communication systems is required. For the case of the grammar model, this is given by the probability that a meaning-signal pair present in G_i is present in grammar G_j. This probability is denoted[4] by c_{ij}, and it is assumed that $c_{ii} = 1$. This measure of overlap between two grammars allows for a straightforward definition of the communicative payoff between two grammars, $F(G_i, G_j)$:

$$F(G_i, G_j) = \frac{1}{2}(c_{ij} + c_{ji})$$

Notice the similarity to the communicative payoff equation given in the previous section — the form is identical, but the summation and production-reception calculations are parcelled up into the measure c_{ij}. The fitness of a grammar G_i with respect to a population is given by

[4] Nowak *et al.* (2001) use a_{ij} to denote the probability that meaning-signal pairs generated by G_i are acceptable to G_j. However, the notation c_{ij} is used to avoid confusion with the A matrix notation in Nowak *et al.* (1999).

$$f_i = \sum_j x_j F\left(G_i, G_j\right)$$

where x_j is the frequency of grammar j in the population. This is exactly equivalent to the equation used by Nowak *et al.* (1999) in their model of the evolution of vocabulary.

Finally, a model of learning is required. Nowak *et al.* assume that children attempt to learn the grammar of their parents. Q_{ij} is the probability that a child whose parent uses grammar G_i will acquire grammar G_j. The dynamics of the population are then specified by:

$$\dot{x}_i = \sum_{j=1}^{n} x_j f_j Q_{ji} - \phi x_i$$

where ϕ is the average fitness of the population, $\phi = \sum_i x_i f_i$. The change in frequency of a grammar G_i therefore depends on the product of the frequency of some grammar G_j, the fitness of that grammar relative to the average fitness of the population, and the probability of learning grammar G_i based on exposure to grammar G_j, summed over all grammars. This is clearly a model of natural selection acting on cultural transmission, given that the change in frequency of a grammar depends on the fitness of that grammar. However, the frequency of a grammar also depends on its learnability — a grammar with below-average Q_{ji} will decrease in frequency due to its being difficult to learn, unless it offers above-average communicative payoff. This model can therefore, depending on the choice of values of Q, model directly-biased transmission (to be discussed in Section 2.3.3) in addition to natural selection. However, Nowak *et al.* do not pursue this avenue.

Nowak *et al.* report three results for this model.

1. For the case where learning is error-free (i.e. $Q_{ii} = 1$, $Q_{ij} = 0$ where $j \neq i$) there are n stable equilibria where one grammar completely dominates the population ($x_i = 1$ and $x_j = 0$ for all $j \neq i$).
2. For high error rates, the only stable solution is the case where each grammar occurs in the population with approximately equal frequency ($x_i \approx 1/n$).
3. For the special case where all grammars are equidistant (i.e. $c_{ii} = 1$ and $c_{ij} = c$ for all $j \neq i$, where $0 \leq c \leq 1$, and $Q_{ii} = q$ and $Q_{ij} = (1-q)/(n-1)$ for $j \neq i$) there is one symmetric solution where all grammars occur with equal frequency ($x_i = 1/n$). There are also several possible asymmetric solutions where one grammar G_i is dominant (although does not necessarily completely dominate

> the population) and all other grammars occur with equal frequency. These asymmetric solutions will be stable provided that a *coherence threshold*, q_1 is met — if $q > q_1$ (the probability of acquiring the same grammar as your parents exceeds some value given by q_1) then the asymmetric solutions will be stable. When q exceeds a second threshold q_2 the symmetric solution becomes unstable and only the asymmetric solutions are stable. The key point here is that q will tend to decrease as n increases, given that more possible grammars leads to a greater probability of selecting the wrong grammar during acquisition. Nowak *et al.* conclude that Universal Grammar must restrict the range of possible grammars such that q remains above the threshold q_1, therefore allowing the possibility of grammatical coherence within a population.

Nowak *et al.* then consider how q changes for incremental and batch learners as the number of sample sentences increases. They also consider how the population behaves where all grammars are not equidistant from one another. For this case they assume that the values of c_{ij} where $j \neq i$ are randomly selected from a normal distribution in the range 0 to 1. Given this assumption, the overall behaviour is broadly similar to the case where all grammars are equidistant.

Their model allows Nowak *et al.* to discover a fairly fundamental result regarding the relationship between the number of possible grammars, the number of sentences a learner is exposed to, the probability of acquiring the correct grammar and the behaviour of populations of such learners. However, there are several undesirable aspects to the model. The notion of a grammar is a fairly impoverished one — grammars specify allowable mappings between meanings and signals, which is a fairly standard assumption, but there is no notion of structure in a grammar or of any interplay between the sentences a grammar allows and the sentences it does not. The model of a grammar essentially boils down to a model of a very large communication system of the type modelled in Nowak *et al.* (1999). This is at odds with the commonly-held view that language is of a different type to agrammatical vocabulary, rather than simply being an extremely large vocabulary.

The other undesirable feature of their model, partially due to their simple model of a grammar, is that the utility and learnability of a grammar is entirely arbitrary. In their earlier work on communication systems, the structure of the object-signal mapping impacted on its functionality and, indirectly, on its learnability, with the most functional and stable A matrix being a binary matrix with a distinct signal for each object. In their paper on UG, values of q and c are arbitrarily assigned. While it would, in principle, be possible to calculate values of q and c for grammars which specified a non-infinite set of sentences, this would be undesirable for two reasons. Firstly, in restricting the grammars to only those which specified finite sets of sentences, the model of grammars would obviously reduce to a model of a large communication system, which as outlined above is undesirable. Secondly, the calculation would still depend on the sets of sentences allowed by grammars, rather than any internal structure of the grammar that specified the set of allowable sentences. The fact that grammars are modelled as essentially arbitrary sets of sentences, rather than collections of rules generating sets with some logical internal structure, makes any calculation of values of q and c essentially meaningless.

2.3.2 Guided variation

In B&R's taxonomy of pressures on cultural transmission, guided variation "results from the cultural transmission of the results of learning and acts to increase the frequency of traits that best satisfy the learning criteria" (B&R p 174). Individuals acquire their initial value for the phenotype through cultural transmission, then modify this value through individual, adaptive learning. The population of fully-matured phenotypes then acts as cultural parents for the next generation. If individual learning prefers one particular phenotype then that phenotype will be disproportionately represented in the distribution of phenotypes observed by learners at the next generation.

2.3.2.1 B&R's model

A model of guided variation requires a model of individual learning. B&R assume that individual learning takes as its starting point a culturally-acquired trait, and then moves this trait towards some optimal value specified by the external environment — individual learning is a process of adaptation to the environment. After cultural transmission and individual learning, an individual is considered mature, and can act as a cultural parent. Cultural parents transmit their traits, which are determined both by cultural transmission and individual learning, to naive individuals.

B&R show (see Section A.1.2.2 of Appendix A) that when individual learning is powerful (individuals are free to make large adjustments to their culturally-acquired trait during individual learning) the population moves towards the value of the phenotype favoured by the environment, due to the transmission of cultural traits favoured by individual learning. In contrast, when individual learning is weak (individual learning tends to conserve their culturally-acquired trait) the mean value of the population's cultural trait remains unchanged by individual learning — no cultural evolution takes place.

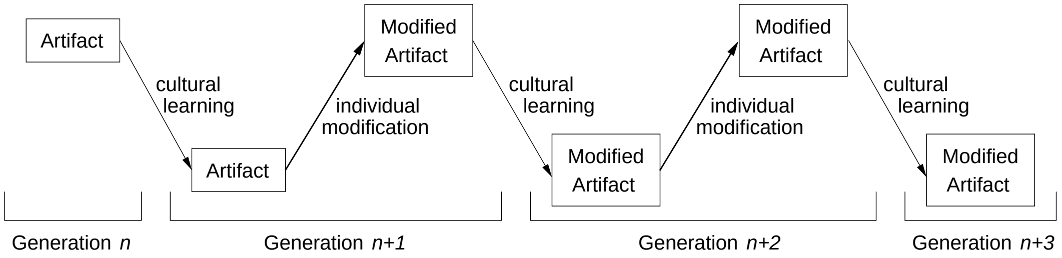


Figure 2.8: The cultural ratchet. A cultural artifact is passed from generation to generation by cultural learning. Each generation makes modifications to the artifact, which are subsequently transmitted.

2.3.2.2 Linguistic case study: Tomasello's cultural ratchet

Tomasello (Tomasello 1993; Tomasello 1999) has proposed a fairly general model of cultural transmission, which he terms "cumulative cultural evolution" or "the [cultural] ratchet effect". While this model might be equally at home in the section on general cultural models, it is included here as Tomasello's focus is primarily on the transmission of the products of individual learning.

The cultural ratchet is depicted in Figure 2.8 and proceeds as follows. Some cultural artifact is acquired, through cultural learning, by children. Those children then mature and make modifications to the cultural artifact to improve its functionality. The modified cultural artifact is then acquired, via cultural learning, by the next generation of children. This cultural transmission must be of sufficiently high fidelity to prevent the loss of earlier modifications — the ratchet must not slip backwards.

This model of the accumulation of modifications which are introduced by goal-directed adjustment of culturally-transmitted artifacts is clearly similar to B&R's model of guided variation, where goal-directed individual learning modifies culturally-transmitted traits. Tomasello's cultural ratchet theory is not supported by a formal model but his assertion that this process will result in well-adapted cultural artifacts is supported by B&R's model, which shows that a combination of social and individual learning can result in characteristics which are suited to the requirements of the environment.

Tomasello views this cultural ratchet theory as an explanation for all complex artifacts, such as tools, religious rituals, mathematics and governmental institutions. He makes particular reference to language, which in his view is established through the same processes as other cultural artifacts:

> “the way human beings have used objects as hammers has evolved significantly over human history. This is evidenced in the artifactual record by various hammer-like tools that gradually widened their functional sphere . . . it is presumably the case that some cultural conventions and rituals (e.g. human languages and religious rituals) have become more complex over time as well, as they were modified to meet novel communicative and social needs” (Tomasello 1999:37)

Tomasello suggests that speakers introduce new words or constructions in order to meet novel communicative needs, or grammaticalise loosely-structured, commonly-occurring discourse structures into syntactic constructions, to improve the functionality of the linguistic system.

The explicitly teleological source of modifications to cultural artifacts (where by “teleological” I mean “designed with purpose in mind”) is relatively unproblematic when applied to the cumulative modification of artifacts such as hammers and governmental institutions. However, the assumption that language changes through repeated, conscious, goal-directed modification is somewhat controversial. Lass (1980) rules out the possibility of purposeful modification — “[linguistic] change does not involve (conscious) human purpose (which I think can be accepted without argument)” (Lass 1980:82).

In addition to appealing to the established orthodoxy in linguistics that teleology plays little role in language change, a further criticism can be made of Tomasello’s view of linguistic evolution. The burden of this type of teleological change must rest with the speaker — a hearer cannot decide to understand innovatively, therefore new constructions must be introduced by speakers. However, unless we assume speaker altruism this may lead to innovations which are highly non-functional with respect to the hearer.

Teleology may reasonably be expected to play a role in the introduction of new words into a new language — new technological innovations, for example, typically require new words to name them. However, while teleology may account for the introduction of such terms it perhaps cannot account for their subsequent diffusion. For example, we might be tempted to explain the preference for the term “mobile phone” over “cell phone” in English speakers in the British Isles in terms of conscious choice by language users in favour of the more functional variant. However, English speakers in North American typically favour “cell phone”. It is possible that “cell phone” is more functional than “mobile phone” in the context of American English but not British English, in which case the teleological explanation of diffusion of the new compound could still apply. This would require a case-by-case analysis of the functionality of words which differed between varieties of English (e.g. "pavement" and "sidewalk" in addition to "mobile phone" and "cell phone" between British and American English, "turnip" and "swede" between Scottish English and non-Scottish English etc). However, it may be more parsimonious to allow that teleological explanations play a fairly limited role in the cultural evolution of language.

2.3.3 Directly biased transmission

Biased transmission arises when naive individuals are more likely to adopt one cultural variant than another, or when mature individuals are more likely to produce one cultural variant than another when acting as a model. In B&R's model, biased transmission is typically conceived of as "arising from the attempts of [individuals] to evaluate the adaptiveness (that is, their effects on genetic fitness) of the different cultural variants" (B&R, p134). B&R identify three subclasses of bias on transmission: direct bias, indirect bias and frequency-dependent bias. Directly-biased transmission will be discussed in this section, with models of indirect bias and frequency-dependent bias being discussed in Sections 2.3.4 and 2.3.5 respectively.

Direct bias occurs when one cultural variant is intrinsically more attractive than others. This intrinsic attractiveness makes naive individuals more likely to acquire that cultural variant. This attractiveness could derive from several sources. For example, individuals could prefer to acquire variants which they believe will be most successful, in which case the direct bias will be in favour of the cultural variants which offer the greatest fitness payoff. Alternatively, individuals could preferentially acquire one variant over another due to an arbitrary preference not necessarily related to functionality.

2.3.3.1 B&R's model

Direct bias can be simply modelled by assuming that individuals are disproportionately likely to acquire one cultural trait, to the detriment of other cultural traits. If individuals are disproportionately likely to acquire a particular cultural trait at the expense of all other cultural traits, then individuals are directly biased in favour of that trait. As shown in Section A.1.2.3, directly biased transmission will increase the frequency of the favoured variant in the population. The rate of increase depends on the strength of the bias — a weak bias will result in slow convergence of the favoured variant, whereas a strong bias will result in rapid convergence. The rate of increase also depends on the variance in the population — in the most extreme case, in a population which is culturally homogeneous, directly-biased transmission has no impact.

2.3.3.2 Linguistic case study 1: evolution of consistent head ordering

Kirby (1999) (Chapter 2 in particular) uses a simple ILM to develop an explanatory account of word-order universals. Hawkins (e.g Hawkins (1990)) notes a statistical tendency for languages of the world to be consistently head-initial or head-final across phrasal categories. Hawkins suggests that this word-order universal can be accounted for in terms of a preference for language users, when parsing utterances, to construct trees as rapidly as possible. This principle is termed Early Immediate Constituent recognition (EIC), and languages which exhibit consistent head ordering score more highly on the EIC metric and are easier to parse. However, as noted by Kirby, this does not constitute an explanation of the observed statistical universal — how does the preference of individual language users for constructions which score well on the EIC metric translate to a word-order universal?

Kirby constructs a generational ILM to attempt to answer this question. The linguistic competence of individuals in Kirby's model consists of a specification of a simple grammar which states the preferred ordering of head and complement in verb phrases (verb-initial, VO, or verb-final, OV) and adpositional phrases (prepositional, PreP, or postpositional, PostP). There are therefore 4 possible grammars — VO and PreP, VO and PostP, OV and PreP, OV and PostP. The first and last of these possibilities have consistent head ordering and therefore score more highly on the EIC metric.

Individuals produce utterances consistent with their grammars, where an utterance is not an actual sentence but a specification of the head-ordering in verb and adpositional phrases. Utterances are essentially direct externalisations of an individual's I-language. The next generation of learners take a random sample of these utterances, and according to the procedures outlined below, select their own grammar.

Selection of a grammar is based on the frequency of utterances exhibiting particular word orders in the sample pool of utterances, but also on the parsability of those utterances. The probability of an individual selecting a variant v for their grammar is given by $p\left(v\right)$ and the number of utterances exhibiting that variant in the sample pool is given by n_v. w_v gives the degree to which variant v is preferred. A learner selects their grammar features probabilistically according to the formulae:

$$\begin{aligned} p\left(PreP\right) &= \frac{w_{PreP}n_{PreP}}{w_{PreP}n_{PreP} + w_{PostP}n_{PostP}} \\ p\left(PostP\right) &= \frac{w_{PostP}n_{PostP}}{w_{PreP}n_{PreP} + w_{PostP}n_{PostP}} \end{aligned}$$

$$
\begin{aligned}
p\left(VO\right) &= \frac{w_{VO}n_{VO}}{w_{VO}n_{VO}+w_{OV}n_{OV}} \\
p\left(OV\right) &= \frac{w_{OV}n_{OV}}{w_{VO}n_{VO}+w_{OV}n_{OV}}
\end{aligned}
$$

The degree to which a particular variant is preferred depends on the frequency of other variants in the utterance pool and the EIC metric. For example, if, as mentioned above, the EIC for VO-PreP is better than VO-PostP then $w_{PreP} > w_{PostP}$ if VO order is common, and $w_{PreP} < w_{PostP}$ if OV order is common. The preferences are calculated as follows:

$$
\begin{aligned}
w_{PreP} &= \alpha n_{VO} + \left(1-\alpha\right) n_{OV} \\
w_{PostP} &= \alpha n_{OV} + \left(1-\alpha\right) n_{VO} \\
w_{VO} &= \alpha n_{PreP} + \left(1-\alpha\right) n_{PostP} \\
w_{OV} &= \alpha n_{PostP} + \left(1-\alpha\right) n_{PreP}
\end{aligned}
$$

where α is a constant reflecting the EIC metric of the various combinations. Kirby reports results for $\alpha = 0.6$, which corresponds to the situation where consistent head ordering is somewhat more parsable than inconsistent head ordering.

Kirby conducts multiple runs of the ILM and finds that there are two stable final states, corresponding to the two consistent head-orderings, VO-PreP and OV-PostP, with the simulation runs converging at random on one of these two states. The learner's bias in favour of consistent head ordering results in the emergence of word-order universals. In terms of B&R's scheme, this is a clear example of how directly biased transmission can be applied in a linguistic context to generate clear and parsimonious accounts of linguistic universals — a theoretically-well motivated bias of learners to preferentially acquire a particular cultural variant leads to that cultural variant dominating the population. Kirby also notes an S-shaped trajectory of change, as predicted by B&R's model of directly biased transmission.

2.3.3.3 Linguistic case study 2: the evolution of vocabulary

Hutchins & Hazelhurst (1995) present an early model of the negotiation of vocabulary in a population. Their key concern is to show that conventionalised symbolic vocabulary can arise through cultural processes. Hutchins & Hazelhurst model individuals using autoassociator neural networks. Autoassociator networks consist of an input layer of nodes,

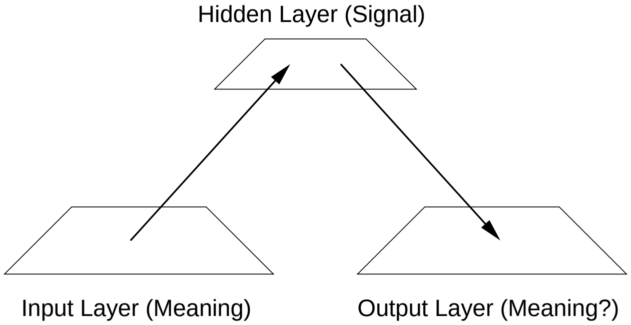


Figure 2.9: Hutchins & Hazelhurst's autoassociator network. The flow of activation (indicated by arrows) proceeds from the input layer via the hidden layer to the output layer. The goal of learning in an autoassociator network is to reproduce the input pattern of activation at the output layer. This necessitates forming a compressed representation of the input pattern at the hidden layer. Hutchins & Hazelhurst interpret the input pattern of activation (and the output pattern of activation) as a visual stimuli (meaning) and the pattern of activation at the hidden layer as an observable signal.

a smaller hidden layer and an output layer of the same size as the input layer. Autoassociator networks are trained (using the backpropagation learning algorithm in Hutchins & Hazelhurst's model) to map from an input pattern of activation, via an internal representation at the hidden layer, back to a pattern of output activation which exactly matches the input pattern — autoassociator networks essentially associate an input pattern of activation with itself.

In Hutchins & Hazelhurst's model there is a set of scenes which agents are required to communicate to one another about using a set of signals. Hutchins & Hazelhurst (1995) consider input/output patterns of activation to represent visual scenes, equivalent to meanings in the canonical E/I vocabulary model, while patterns of activation over the network's hidden layer are considered to represent observable signals. The network structure and interpretation of the various layers is illustrated in Figure 2.9.

At every time-step two individuals are selected at random from the population and one scene is chosen from the set of scenes, again at random. One individual acts as speaker and produces a signal given the selected scene. The second individual acts as a learner and is trained to associate the input scene with an identical output pattern of activation (i.e. perform the autoassociator learning task), while at the same time learning to associate the input scene with the signal produced by the speaker. The two individuals are then returned to the population. A simulation run consists of several thousand such pairwise interactions. Note that, as there is no turnover of population and therefore no separation between learners and producers, Hutchins & Hazelhurst's (1995) model is a classic implementation of the Negotiation Model. There is no explicit measurement of communicative accuracy in this model.

Hutchins & Hazelhurst report that, initially, the individuals in the population do not have distinct signals for each distinct scene and there is no consensus across the population as to which signals should be associated with which scenes. This is due to the random initialisation of the weights in each individual's network. However, after several thousand pairwise interactions, every individual in the population associates each distinct scene with a distinct signal and there is consensus between individuals as to which signal should be associated with which scene. The population converges on a communication system which would be optimal in terms of communicative payoff, as defined in Section 2.2.2.1.

What drives the emergence of these optimal communication systems? We can rule out natural selection of cultural variants, given the absence of any reward for successful communication. Guided variation can also be discounted, given the absence of any individual learning, as can indirect bias (to be discussed in Section 2.3.4). As will be discussed in Chapter 3, the autoassociator model of individuals used by Hutchins & Hazelhurst is strongly biased in favour of communication systems which are optimal in terms of communicative payoff — autoassociator agents are more likely to acquire such systems than systems which happen to be suboptimal in terms of communicative function. As predicted by B&R, the application of a direct bias on cultural transmission results in the increase in frequency of the cultural variant favoured by that bias. Hutchins & Hazelhurst observe the increase in frequency of scene-signal mappings which are optimal in terms of the biases of their chosen model of individual learners. These mappings also happen to be optimal in terms of communicative function.

The model outlined in Hutchins & Hazelhurst (1995) represents an interesting advance. The predictions of B&R are shown to hold with a far less abstract model of cultural variants and a bias which arises naturally from the chosen model of a learner. They also illustrate that natural selection is not the only possible pressure on cultural transmission that can lead to the emergence of communicatively optimal vocabulary systems — in their model there is no direct pressure for communicative function, nor even any evaluation of communicative accuracy, yet communicatively optimal systems still emerge. However, it is unclear how general their results are. Will any model of a learner result in the emergence of shared vocabulary when placed in the context of the Negotiation Model? What properties must the learner have to ensure the emergence of shared vocabulary? These questions will be returned to in depth in Chapter 3.

2.3.3.4 Linguistic case study 3: the evolution of morphology

Batali (1998) presents a computational implementation of the Negotiation Model where a small population of agents negotiate a vocabulary to communicate a set of (fairly minimally) structured meanings. The structure of the meanings and the biases of the learners results in the emergence of partially regular morphology.

In Batali's simulation individuals are modelled using simple recurrent neural networks capable of mapping a temporal sequence of input patterns to an output pattern of activation, with the eventual output pattern of activation depending on both the actual input patterns presented and the sequence in which input patterns are presented. Input patterns are representations of characters, with sequences of such input patterns being words. The developed output pattern of activation is considered to be a meaning. Batali's networks therefore take as input a sequence of characters forming a word and arrive at a meaning, their interpretation of that word. Meanings are analysed as consisting of a predicate component and a referent component. There are 10 possible predicates, represented by distinct but overlapping activation patterns over part of the output layer, and 10 possible referents, once again represented by distinct but overlapping activation patterns over the remainder of the output layer.

The simulation model follows the classic pattern of the Negotiation Model, with members of the population in turn producing meaning-signal pairs and learning from the produced meaning-signal pairs of other agents. Learning is carried out using the backpropagation algorithm. During each exposure the learner is presented with a meaning and a character sequence constituting the observed signal and attempts, using the backpropagation algorithm, to learn the association.

In addition to acting as learners, individuals act as producers (for communication and for producing observable behaviour for other agents to learn from) and as receivers (for the purpose of evaluating communicative accuracy). When acting as a receiver, the network is presented with a character sequence and constructs a meaning. Production is somewhat more complex, given that recurrent neural networks are non-reversible and Batali's networks map from input signals to output meanings. To produce a signal for a given meaning, an agent considers all possible characters and selects the character which produces the lowest output error in its own network with respect to the target meaning. That character is sent to the learner as the first character of the word associated with the target meaning. The producer continues producing characters until the error with respect to the target meaning drops below some threshold value or the number of characters in the word exceeds some arbitrary limit. This process of an individual using themselves as a model of the hearer and sending the signal which they themselves would interpret as the target meaning is known as the *obverter* strategy and is encountered elsewhere.

The accuracy of a communicative episode between two such individuals can be computed by comparing the speaker's given meaning vector with a hearer's output meaning vector after being exposed to the signal produced by the speaker. The communicative accuracy of a single episode is the number of positions on the meaning vector for which speaker and hearer agree on the value, within some tolerance. This corresponds to a variant of the canonical measure of communicative accuracy, with a distance metric over meanings and partial payoff for partially correct meanings. However, as with most Negotiation Models, there is no actual payoff for being a successful communicator. Cultural evolution is not driven by natural selection.

In the initial population of agents the accuracy of communicative episodes is at chance levels and distinct meanings are not necessarily communicated using distinct signals. As with the model of Hutchins & Hazelhurst (1995), this is due to the random starting values in each agent's initial network. Batali (1998) presents two results for this simulation model:

1. After several thousand learning interactions between members of the population, the population arrives at a near-optimal communication system. The accuracy of communicative episodes is high, each distinct meaning is communicated using a distinct signal and different individuals largely agree on the meaning associated with each signal.
2. The population converges on a semi-regular morphological system, with the predicate component of a meaning typically being associated with a root morpheme portion of a signal, and the referent component of a meaning typically being associated with a suffix component of the signal. For example, the pattern of activation corresponding to the predicate *happy* is usually, but not always, communicated using the root sequence "ba-", while the pattern of activation corresponding to the first person plural referent *we* is typically communicated with a suffix "-d" or some variant (e.g. "-dc" or "-ddc").

As with the model described in Hutchins & Hazelhurst (1995), the convergence of the population on a near-optimal communication system can be accounted for in terms of a direct bias on cultural transmission resulting from the learner model. This arbitrary bias of the learner results in the emergence of systems of meaning-signal mappings which happen to be optimal in terms of communicative function. An in-depth discussion of the bias is postponed till Chapter 3.

The emergence of semi-regular morphology is an interesting and novel aspect of the model. Batali also attributes this emergent structure to the biases of the agent architectures and the negotiation task. The negotiation task forces agents to arrive at shared mappings from signals to meanings. Sharing such mappings makes it likely that agents will share mappings from partially-presented signals (partial sequences of characters) to output patterns of activation, given that any large divergence in output patterns of activation midway through processing will be difficult to remedy later in the string. Agents will come to use fairly regular sequences to guide other agents (or themselves) into appropriate regions of output vector space. This is reflected in the semi-regularity of the final morphological system — there is a tendency to systematically use a particular sequence when expressing a particular portion of meaning, but this tendency is not so strong as to force a completely regular morphological system.

This model represents an interesting and significant development in the formal models of the cultural evolution of linguistic structure. Moderately sophisticated linguistic structure is investigated, using a methodology that has been applied to the cultural evolution of simpler, unstructured communication systems. The first result, that communicatively-useful communication systems can arise in the absence of natural selection, provides support for Hutchins & Hazelhurst's (1995) earlier conclusion. The use of a different model of a learner shows that neither Batali's nor Hutchins & Hazelhurst's results are dependent on a particular model of a learner. However, there is no exploration of the nature of the bias that leads, via negotiation, to the emergence of optimal communication and no attempt to relate the biases embodied in this model to those in earlier models. Furthermore, the explanation of the emergence of semi-regular morphological structure is perhaps somewhat underdeveloped, appealing to vague tendencies for agents to coordinate their movement through output vector space.

2.3.4 Indirectly-biased transmission

In directly-biased transmission one cultural variant is preferred over alternatives due to its own intrinsic properties. In indirectly biased transmission individuals must acquire two cultural traits — an *indicator* trait and an *indirectly biased* trait. Certain variants of the indicator trait are, as in the direct bias case, intrinsically preferable. No variant of the indirectly biased trait is intrinsically more preferable than any other. However, individuals prefer to acquire the indirectly biased trait of individuals who have an attractive indicator trait. For example, an indicator trait might be clothing styles (assuming an intrinsically-preferable style) and an indirectly biased trait might be political persuasion. Individuals will preferentially acquire the preferred clothing style (the indicator trait) and will preferentially acquire the political viewpoint of individuals who wear the preferred style of clothes — people will prefer to dress like smart dressers, and tend to vote like them too.

2.3.4.1 B&R's model

B&R assume that each individual is characterised by two cultural traits — an indicator trait and an indirectly-biased trait. They further assume that there is a direct bias in favour of a particular variant of the indicator trait — individuals preferentially acquire the preferred variant of the indicator trait. As discussed above, this will lead to the population converging on that variant of the indicator trait. It is furthermore assumed that individuals acquire the indirectly-biased trait of individuals who are have the preferred variant of the indicator trait — if an individual has the preferred indicator trait, they are disproportionately likely to transmit this, but also disproportionately likely to transmit their other cultural trait.

As shown in Section A.1.2.4, this results in the convergence of the population on values of the indirectly-biased trait which happen to be correlated with the preferred variant of the indicator trait — "variants of the indirectly biased trait that are positively correlated with the admired variants of the indicator trait will increase in frequency" (B&R p254). The rate of increase of the correlated trait depends on the strength of the correlation — indirectly-biased traits which are only weakly correlated with the preferred indicator trait will increase in frequency slowly, whereas indirectly-biased traits which are strongly correlated with the preferred indicator trait will rapidly come to dominate the population.

2.3.4.2 Linguistic case study: language evolution through acts of identity

Croft (2000) introduces an "utterance-based selectional theory" of linguistic evolution. In Croft's theory, differential replication of "linguemes" results in linguistic evolution, where a lingueme is a linguistic structure embodied in an utterance. Croft proposes that differential reproduction arises not through natural selection, where the functionality of a cultural variant determines access to roles which yield opportunities to transmit culturally, but through an evaluation of the social affordances of a particular lingueme by language users — biased transmission.

There is an established tradition in the sociolinguistic literature in accounting for the linguistic behaviour of individuals in terms of *prestige* (Labov 1966) and *covert prestige* (Trudgill 1972). In these terms, the choice of a particular linguistic form in preference to alternative forms constitutes an act of identity (LePage & Tabouret-Keller 1985) on the part of the speaker. A population consists of several, possibly overlapping, social groups — for example, a population can be subdivided according to gender, religion, social class or geographic region. Different social groups may use different linguistic systems. For example, residents of Japan, by and large, use Japanese whereas residents of the Korean Republic typically use Korean. Middle-class residents of Edinburgh typically have a different accent from working-class residents of Edinburgh. If an individual wishes to identify themselves with a particular social group they will adjust their linguistic behaviour to conform more closely with the linguistic behaviour associated with that group. In Croft's model it is such acts of identification that determine the selective advantage of one lingueme over another:

> "the factor in language use granting selective advantage to an individual speaker (and thus to the way she talks) is the desire of hearers who interact with her to identify with the community to which she belongs. This selective advantage ensures the differential perpetuation of the replicators she produces, that is, the propagation of the linguistic variants associated with the linguistic community to which she belongs" (Croft 2000:183).

To put it in the terminology Croft himself introduces, linguemes which are associated with social groups which individuals wish to identify themselves with will have a selective advantage. The social groups with which individuals wish to associate themselves are contingent on context and the notions of prestige and covert prestige. For example, in situations where covert prestige (say "working-classness") was important, linguemes associated with groups which rate highly on the covert prestige scale (e.g. linguemes associated with "working-classness") would have a selective advantage.

This informal model is highly reminiscent of the indirectly-biased transmission model of B&R, and Croft's conclusions are equivalent to the predictions of B&R's simple mathematical model — "variants of the indirectly biased trait [=linguemes] that are positively correlated with the admired variants of the indicator trait [= admired social groups] will increase in frequency [= have a selective advantage]". Croft's model is of course more complex, given that his notion of cultural traits is rather more complex and his equivalent of the biasing function on indicator traits, groups with which individuals wish to identify themselves, changes from occasion to occasion. It is also somewhat debatable as to whether an individual's social group can truly be treated as a cultural trait. An acceptable circumlocution might be to say that an individual's social group is determined by a complex of traits, many of which are culturally acquired. In this case, the comparison with B&R's model passes at least a cursory inspection.

There is a further complication, however. In B&R's model the correlation between an indicator trait and an indirectly biased trait can be presupposed, or can arise through cultural transmission of indicator and indirectly biased traits, assuming that errors during cultural transmission increase the correlation between indicator and indirectly biased traits. Croft has nothing to say on how this correlation between a particular social group and a particular set of linguistic behaviours gets off the ground in his model. For groups defined in terms of regional boundaries this might be unproblematic, given that spatial separation leads over time to linguistic separation. However, it is less clear how social groups defined in terms of religion or social class might come to be associated with particular linguistic behaviours. It is perhaps sufficient to simply presuppose this, or to assume it can be introduced by chance factors, or correlated errors between the cultural transmission of measures of prestige and linguemes.

2.3.5 Frequency-dependent bias

Cultural transmission is of course frequency dependent in all of the models considered so far — in the case of unbiased transmission, the probability of acquiring a particular cultural trait depends linearly on the frequency of that trait in the population, whereas in the direct and indirect bias cases, the relationship between the frequency of a trait and the probability of acquiring it is non-linear, with the non-linearity being introduced by the bias in favour of a particular trait. B&R reserve the term "frequency-dependent bias" for the case where the probability of acquiring the *most common* cultural variant in the population, regardless of which variant it is, is greater than (in the case of conformist frequency-dependent bias) or less than (in the case of non-conformist frequency-dependent bias) the probability of acquiring that variant in the unbiased transmission scenario.

2.3.5.1 B&R's model

B&R assume that individuals are disproportionately likely to acquire the majority cultural variant in the population (*conformist* frequency dependent bias). Assume that there are two possible cultural variants, c and d. As shown in Section A.1.2.5 of Appendix A, under conformist transmission, variant c will increase in frequency if the frequency of c is greater than the frequency of variant d. Conversely, c will decrease in frequency if its frequency is less than that of d. The rate of change of the most common variant is at its lowest as its frequency approaches 1 (where the population is converged on the most frequent variant) or 0.5 (the point where the population is perfectly split between the two variants). Conformist transmission results in the spread of the most common cultural variant.

2.3.5.2 Linguistic case study: the spread of head order

Briscoe (2000a) considers a model, similar in some aspects to that of Kirby (1999) described in Section 2.3.3.2, of the spread of a particular head ordering parameter within a population. Briscoe's (2000a) model is rather more complex than Kirby's model — learners acquire categorial grammars based on exposure to sets of triggers. However, the details of this model are not particularly relevant for the purpose of this section. It will suffice to say that learners acquire a grammar which is either head-initial (which we shall call G_I) or head-final (G_F). Triggers are specified as being of type I or F. Individuals using head-initial grammars (G_I) produces triggers of type I, and individuals using head-final grammars (G_F) produces triggers of type F. Learners observe a set of n triggers, and decide on whether to acquire G_I or G_F. Briscoe considers three possible learning procedures:

- The learner acquires G_I if the *first* of the n triggers it observes is of type I, and G_F if the first of the n triggers is of type F.
- The learner acquires G_I if the *last* of the n triggers is of type I, and G_F if the last of the n triggers is of type F.
- The learner acquires G_I if the *majority* of the n triggers are of type I, and G_F is the majority of the n triggers are of type F.

Learning procedures 1 and 2 are in fact equivalent and lead to a probability of acquiring grammar G_x of approximately $p\,(G_x)$, where $p\,(G_x)$ is the proportion of grammar G_x in the population. This is equivalent to the linear transmission rule for a dichotomous character as discussed in Section A.1.1.1 of Appendix A, where there is a single cultural parent. In the context of Briscoe's simulation model, where populations are finite, this leads to random fluctuation in the frequencies of G_I and G_F until one grammar reaches fixation.

Learning procedure 3 has rather different behaviour. Assuming that n is odd (learners observe an odd number of triggers), and that k is the first integer such that $j > n/2$ (i.e. k is the lowest value such that k represents more than half the number of models n), the probability that an agent will acquire grammar G_I is:

$$Prob\left(G_I\right) = \sum_{j=k}^{n} \binom{n}{j} p\left(G_I\right)^j \, p\left(G_F\right)^{n-j}$$

In other words, the probability of acquiring grammar G_I is equal to the probability of picking a set of n models from a population characterised by $p\left(G_I\right)$ and $p\left(G_F\right)$ such that individuals of type G_I are in the majority. The probability of acquiring G_F can be similarly defined.

These equations results in population dynamics equivalent to B&R's equation describing the dynamics of a population acquiring a dichotomous character under frequency-dependent bias, in the case of a strongly conformist bias function.

Briscoe's results confirm those of B&R — the frequency of the initially most frequent cultural variant increases, with the rate of change decreasing as the population approaches saturation. In Briscoe's simulations, where the population is finite, random sampling errors can move the population away from the case where both variants are equally frequent, which in B&R's infinite population model would be an unstable fixed point.

2.3.6 Transmission through a bottleneck

The final pressure acting on cultural transmission to be discussed in this chapter is a novel discovery, arising from computational models of the cultural transmission of language. The character of this bias suggests that it may be a specific bias acting on the transmission of infinite systems, such as language.

2.3.6.1 The transmission bottleneck in the Iterated Learning Model

Kirby (2002) presents a generational ILM which deals with the cultural evolution of compositionality and recursiveness, two of the design features of language discussed in Chapter 1. This model also reveals a new mechanism of cultural evolution, not covered in B&R's set of possible mechanisms.

In Kirby's model an individual's linguistic competence consists of a definite-clause grammar with attached semantic arguments. These definite-clause grammars consist of a set of rules, where the left hand side consists of a non-terminal category and the semantic label for that category and the right hand side consists of zero or more non-terminal categories, with semantic labels, and zero or more strings of characters, which correspond to phonetically-realised components of a signal. Semantic representations are predicate-argument structures, which may have hierarchical structure. Two example grammars might be:

Grammar 1:

S / sees′(mark′,loves′(lynne′,garry′)) → `markseeslynnelovesgarry`

Grammar 2

S / p(x,y) → *N/x V/p S/y*
S / p(x,y) → *N/x V/p N/y*
V / sees′ → `sees`
V / loves′ → `loves`
N / mark′ → `mark`
N / lynne′ → `lynne`
N / garry′ → `garry`

Atomic semantic elements are marked with primes, characters are represented in `typewriter font`, upper case italics represent non-terminal categories and lower case italics represent variables over semantic elements.

Both grammars would produce the string `markseeslynnelovesgarry` meaning sees′(mark′,loves′(lynne′,garry′)), but clearly do so in rather different ways — grammar 1 would do so in a holistic manner, whereas grammar 2 would do so in a compositional (each subpart of the meaning corresponds to a subpart of the signal) and recursive (the first *S* rule rewrites as a string of non-terminals including an *S*) manner.

Learners in Kirby's model are presented with a set of utterances, where utterances consist of meaning-signal pairs, and induce a grammar based on these utterances. Grammar induction consists of two main processes — rule incorporation and rule subsumption. In an incorporation event a learner is presented with a meaning-signal pair $\langle m, s\rangle$ and forms a rule: *S*/$m \rightarrow s$. This amounts to simply memorising an observed utterance.

Subsumption involves two main sub-processes, chunking and merging. During chunking, pairs of rules are examined in the search for meaningful chunks, which are then separated out into new syntactic categories. For example, if a grammar contains two incorporated rules:

S / loves′(lynne′,garry′) → `lynnelovesgarry`
S / loves′(lynne′,beppe′) → `lynnelovesbeppe`

chunking would identify that there is a single difference in semantics between the two rules (the second argument) and a single difference in signal (the string-final `garry` and `beppe`). The two sentence level rules are then replaced with a single rule

S / loves$'$(lynne$'$,x) $\rightarrow$ `lynneloves` N/x

and two new rules are added referring to the newly-introduced syntactic category:

N / garry$'$ $\rightarrow$ `garry`
N / beppe$'$ $\rightarrow$ `beppe`

Merging compares pairs of rules and attempts to reduce the number of distinct syntactic categories in the grammar. For example, suppose that a learner with the grammar just given makes a third observation of the meaning-signal pair $\langle$loves$'$(garry$'$, beppe$'$), `garrylovesbeppe`$\rangle$. This will lead, via incorporation and chunking, to the grammar:

S / loves$'$(y,x) $\rightarrow$ M/y `loves` N/x
N / garry$'$ $\rightarrow$ `garry`
N / beppe$'$ $\rightarrow$ `beppe`
M /garry$'$ $\rightarrow$ `garry`
M /lynne$'$ $\rightarrow$ `lynne`

The merging operation will notice that the first example of category N and the first example of category M are identical, and will therefore replace all mentions of M with N. After removal of redundant rules, this leads to the grammar

S / loves$'$(y,x) $\rightarrow$ N/y `loves` N/x
N / garry$'$ $\rightarrow$ `garry`
N / beppe$'$ $\rightarrow$ `beppe`
N /lynne$'$ $\rightarrow$ `lynne`

Merging therefore leads to generalisation — the learner with this grammar can now parse or produce utterances for nine meanings, based on three observations.

When called upon to produce an utterance for a meaning, Kirby's agents attempt to find a combination of rewrite rules which will cover the given meaning. If such a set of rules exist then the producer will produce an utterance consisting of the meaning and the string of terminals produced by the application of these rules. However, if such a set of rules does not exist then the producer will be forced to apply an invention procedure. Invention involves using existing rules as much as possible, with parts of the meaning which are not expressible using the grammar being expressed with random strings. During invention, the inventor learns from its own invention via incorporation, subsumption and so on.

Using these models of production and learning, Kirby conducts a series of generational ILM simulations, with a population size of 1 at each generation. At each generation individuals produce utterances for 50 randomly selected meanings involving no embedded predicates (e.g. loves′(lynne′,garry′)), then 50 meanings involving a single embedded predicate (e.g. sees′(mark′, loves′(lynne′,garry′))) then 50 meanings involving two embeddings (e.g. thinks′(beppe′, sees′(mark′, loves′(lynne′,garry′)))). These utterances then constitute the learning data for the learner at the next generation. Given that there are 5 embedding predicates, 5 non-embedding predicates and 5 possible atomic arguments each individual sees only a tiny fraction of the possible language of the previous generation — learners suffer from one aspect of the poverty of the stimulus problem. In Section 1.2.1, one aspect of the poverty of the stimulus problem was identified as the fact that children's data-exposure histories are finite, yet they acquire the ability to produce or understand an infinite number of sentences. Kirby's learners are exposed to a small number of utterances drawn from a system which is at least very large (covering all possible meanings involving two embeddings) and in principle infinite — embedding could be continued to an arbitrary depth. Kirby terms this aspect of the poverty of the stimulus the *transmission bottleneck.*

The language in the early stages of a simulation is almost entirely holistic, including rules such as:

S/ thinks′(beppe′, sees′(mark′, loves′(lynne′,garry′))) $\rightarrow$ `im`

Such grammars are obviously fairly large, as each distinct meaning is represented by a single rule. However, after 1000 generations of the ILM the grammars reduce to maximally compressed, fully compositional grammars with a single non-recursive sentence rule (for the non-embedded predicates) and a single recursive sentence rule (for the embedded predicates), and three further non-terminals (one for arguments, one for non-embedding predicates and one for embedding predicates). What drives this evolution of recursive compositionality?

One might be tempted to attribute the emergence of compositionality to directly biased transmission — learners compress their grammars wherever possible and are therefore biased in favour of acquiring cultural variants corresponding to compressed grammars. While this is almost certainly true to a degree, directly biased transmission is only part of the story. Were each learner exposed to utterances for the full set of meanings, the learner bias in favour of compressed grammars would lead to a partially compressed grammar, but not the radically compressed grammars that do emerge.

This radical compression is driven by the transmission bottleneck — utterances produced using holistic rules must be observed to be reproduced, whereas utterances produced using general rules can be reproduced even if they have not been seen, provided some other utterances produced using those same general rules have been observed. More general rules are more likely to be represented in the data presented to the learner, and are therefore more likely to survive cultural transmission. The maximally stable grammar is maximally general, and the bottleneck on transmission filters out non-stable grammars until the maximally stable, maximally general grammar is found.

This therefore represents a new mechanism for cultural change, one which is particularly relevant to the cultural transmission of language. Batali (2002) (discussed below) argues that such a dynamic is also present in his model.

This mechanism by which languages change from less to more generalisable due to the pressure introduced by the bottleneck might be seen as an independent principle of language change. Recall Lightfoot's assertion that the "search for independent principles of change must be abandoned . . . predictions are derived mostly from a theory of grammar" (Lightfoot 1979:153). This does not seem to be the case in Kirby's model — the initial, idiosyncratic, holistic systems in Kirby's simulations are as compatible with Kirby's theory of grammar (such as it is) as the final, systematic, compositional languages. Of course the representational restrictions of the grammar and the inductive biases of the learner, which could be seen as part of the theory of grammar, play a role in shaping language emergence in Kirby's simulations. But his results are not predictable solely from these factors — the bottleneck on transmission imposes a key pressure. As such, the change from holism to generalizability could be seen as an independent principle of change.

While Kirby's findings are obviously significant, one or two minor criticisms might be made. Firstly, the transmission bottleneck is enforced in all of his simulation runs. Some runs with no bottleneck on transmission would allow us to identify how much linguistic evolution would take place in the absence of the bottleneck, which would in turn allow us to identify the strength of the learning bias in his model and the strength of the dynamic arising from the bottleneck. Secondly, it is not clear to what extent the incremental increase in complexity of embedding during production (no embedding first, then degree-1 embedding, then degree-2 embedding) assists the emergence of recursive compositionality — to what extent would these results still hold if the "starting small" (Elman 1993) procedure was not applied. Finally, how important is the assumption that producers learn from their own invention? This may be a reasonably plausible assumption, but would we still see the evolution of recursive compositionality if we weakened this assumption and assumed that producers only occasionally learned from their inventions, or never did?

2.3.6.2 The transmission bottleneck in the Negotiation Model

Batali (2002) presents another computational implementation of the NM. Unlike in his earlier implementation (Batali 1998), Batali takes a symbolic, rather than connectionist, approach to modelling agents and the emergent communication systems in his model are striking in their intricacy.

In his symbolic model Batali treats meanings as formula sets, equivalent to predicate logic constructions. Formula sets consist of conjunctions of zero or more predicates, each of which take one or two arguments, where the arguments are variables represented numerically. For example, the feature set $\{(lizard\ 1)\,(likes\ 1\ 2))\}$ is equivalent to the logical formula $\exists x \exists y\, [lizard\,(x) \wedge likes\,(x, y)]$. Signals are simply strings of characters, of unbounded length.

Linguistic structures are mappings between formula sets and signals. The simplest type of linguistic structure simply has a formula set paired with a signal. Batali terms these basic structures “tokens”, and they are equivalent to lexical items. More complex structures consist of combined sets of tokens. Tokens are combined in binary branching structures, with the formula set for the whole structure being the union of the formula sets for its constituent tokens and the signal for the whole structure being the concatenated signals associated with each token. Argument maps rename variables during the unification of the formula sets for two distinct tokens. Some example structures are given in Figure 2.10.

Agents learn by observing meaning-signal pairs produced by other members of the population, in the classic Negotiation Model framework. Individuals acquire linguistic knowledge by building up a set of exemplars, where each exemplar is a structure of the type shown in Figure 2.10 with an associated cost. When presented with a meaning-signal pair a learner has several choices.

- Store the observed meaning-signal pair in their memory as an unanalysed token.
- Search among existing exemplars and find an exemplar which can be modified so that it matches the observed meaning and signal. Exemplars can be modified by replacing one subpart of a complex exemplar with another exemplar, which may

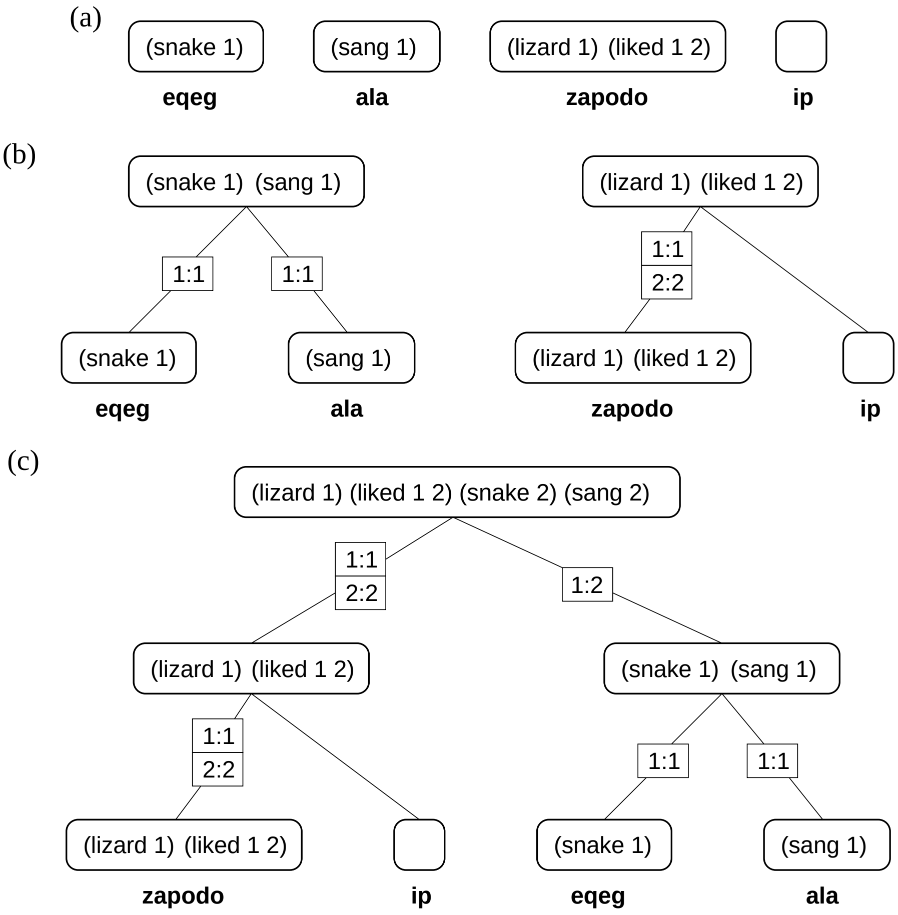


Figure 2.10: Exemplars. The exemplars in (a) are tokens, corresponding to lexical items. The formula set is enclosed in a box, while the string is given in sans serif font. The exemplars in (b) are constructed by combining the tokens from (a), with argument maps (in square boxes) specifying how to relabel variables from the tokens in the formula set for the complex exemplars. The exemplar in (c) is constructed by combining the exemplars in (b). Note that one argument map rewrites occurrences of 1 as 2.

have to be created from scratch. The modified exemplar and any newly created subparts are stored in memory.

- Combine two or more exemplars to form a new complex exemplar, which has the same meaning and signal as the observed utterance. This may involve creating new exemplars. The new, complex exemplar and any newly-created subparts are added to the store of exemplars.

Every newly-created exemplar has a initial cost of 1. A learner searches for the cheapest exemplar or combination of exemplars with the lowest total cost which matches the given meaning and signal. Exemplars which are used during learning have their cost reduced. During learning, agents also perform a search for inconsistent exemplars. If two exemplars or combinations of exemplars have matching signals but non-matching formula sets, they have their costs increased by a fixed amount. This amounts to a penalisation of homonymy. Finally, exemplars which have not been used in the last 200 episodes of production, reception or learning are removed from an agent's memory.

Agents are called upon to produce a signal for a given meaning, either when producing behaviour for another individual to observe or when taking part in a communicative interaction (as described below). Production involves searching through an agent's set of exemplars for an exemplar or combination of exemplars which will have as its formula set the meaning to be conveyed.

Agents have several options during production:

- Exemplars can be retrieved as a whole from the agent's set of exemplars. Such exemplars have an associated cost.
- Exemplars can be modified, by replacing one subpart with another exemplar. Such exemplars have a cost equal to the sum of the costs of the two exemplars used, plus a fixed modification cost.
- Exemplars can be combined, in which case the combined exemplar has a total cost equal to the sum of the costs of the component exemplars plus a fixed combination cost.

During any one of these processes, new unstructured token can be created. Such tokens consist of a formula set and a random string, and have a cost proportional to the sum of the length of the string and the number of predicates in the formula set. The producer searches for the cheapest combination of exemplars. Reception (the search for an exemplar which matches a given signal) proceeds in the same fashion. Any new exemplars which are created during production or reception are *not* stored in an individual's memory — unlike in Kirby's model, individuals do not learn from their own inventions.

90% of the interactions an individual participates in are of the learning type, with one individual called upon to produce a signal for a randomly-selected formula set containing between 2 and 7 predicates and 1 and 3 different variables, and the other individual learning from the produced behaviour. 10% of an agent's interactions are of a purely communicative nature, with one individual producing a signal and the other individual receiving the signal and arriving at a formula set. The success of such interactions are evaluated according to the formula:

$$\text{Communicative Accuracy} = \frac{1}{2}\left(\frac{c}{s} + \frac{c}{r}\right)$$

where s is the number of formulae in the sender's formula set, r is the number of formulae in the receiver's formula set and c is the number of formulae common to both sets. This is similar to Batali's (1998) communicative accuracy measure, again with partial payoff for partial communicative success. Communicative accuracy varies between 0 and 1, with a value of 1 representing perfect communication.

Batali reports several interesting results relating to the level of communicative accuracy within the population and the structure of the emergent communication systems:

1. The population's communicative accuracy increases from 0 to close to 1 over tens of thousands of rounds of negotiation. Individuals settle on fairly stable sets of exemplars, and typical exemplar cost is low. In the final stable system individuals rarely need to produce new structures during learning. The majority of tokens (unanalysed pairings between a signal and a formula set) contain a single formula in the formula set. The emergent systems are compositional — the meanings of complex exemplars depend of the meanings of their constituent exemplars, and the way those constituents are combined.
2. In some emergent systems empty tokens (unanalysed pairings of a string and an empty formula set) play an important role in the types of argument map applied to a complex exemplar. Some examples are given in Figure 2.11, which involve empty tokens which introduce a collapsing map and an inverting map.
3. Systems which do not use empty tokens rely on word order to construct argument maps. For example, in a system with no empty tokens, predicates relating solely

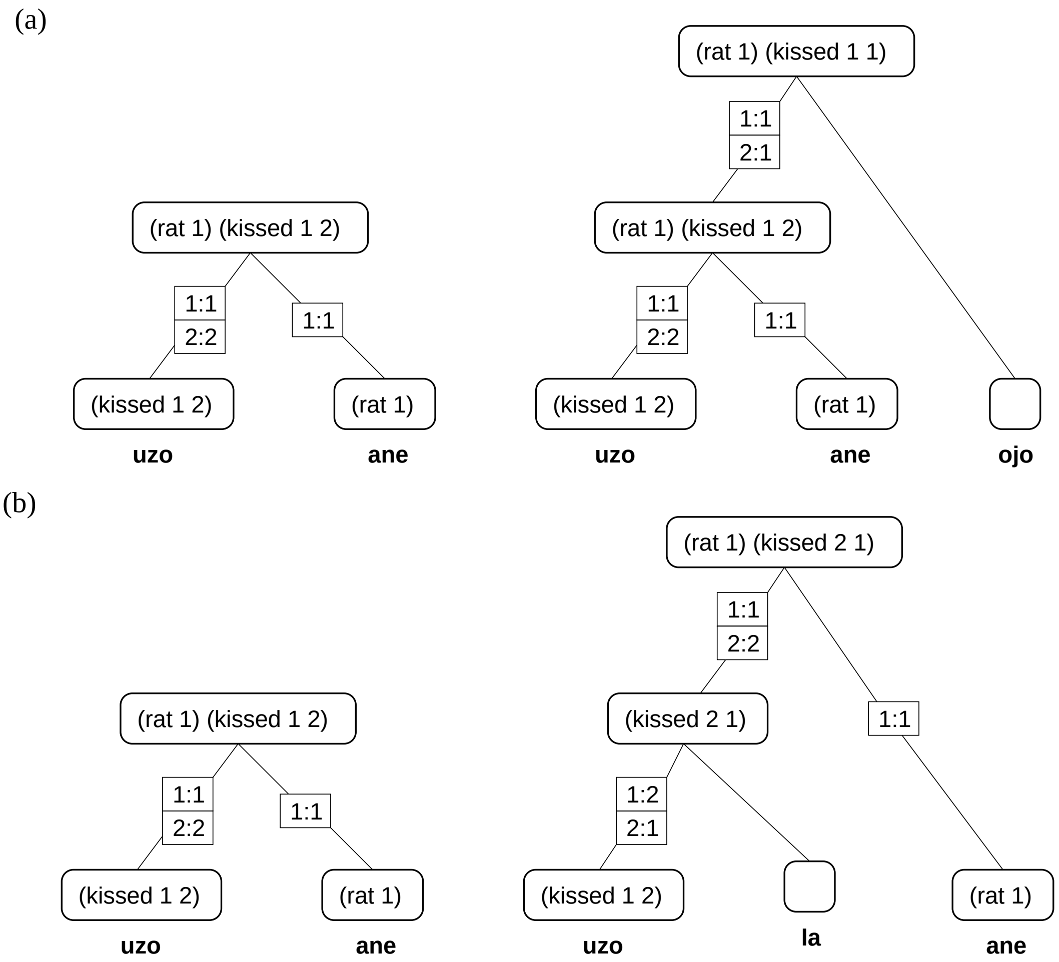


Figure 2.11: Argument maps in Batali's model. The larger complex exemplar in (a) is constructed by combining the smaller exemplar with the token with the string ojo. While this token contributes nothing to the formula set of the utterance as a whole, it does introduce another layer of structure. In the argument map introduced with this top layer of structure, the variable 2 is collapsed with variable 1. The ojo token could be said to function in this case as some kind of marker of a reflexive. (b) shows two more exemplars. The smaller of the two exemplars yields the formula set which could be glossed as meaning "rat kissed someone". In the second complex exemplar, the insertion of the semantically-empty token with the string la allows an intermediate layer of structure to be built. The associated argument map swaps the variables around. This exemplar, when combined with a further exemplar, yields the formula set which would be glossed as "someone kissed rat" — the la token could be considered a marker of a passive-like construction.

> to individual x will appear first, followed by a two-place predicate involving x and another individual y, followed by all predicates relating solely to individual y.

What drives the emergence of coordinated communication and linguistic structure? In B&R's terms, the emergence of linguistic structure in Batali's model appears to be driven by a direct bias. Systems encodable using a small number of exemplars which are frequently reused and recombined are favoured by this bias. The exemplars encoding these systems will have a low cost and will therefore be unlikely to be replaced by random, less structured inventions, which have a high cost.

This bias will apply both to production — agents will prefer to reuse cheap exemplars, which will therefore become cheaper and more reusable — and to learning — analyses of novel meaning-signal pairs which involve existing exemplars will be preferred by learners, reinforcing the reusable components and forcing out analyses which involve large, non-reusable chunks. The division between systems which use empty tokens to control argument maps and those which use a rigid word order is an example of two alternative solutions to the problem of arriving at a particular formula set when combining exemplars. Presumably the state of the final system with respect to this choice is contingent on coincidences in the early rounds of the negotiation process.

We should also expect the transmission bottleneck to have an impact on the population's emerging language — learners acquire the ability to communicate approximately 10^{13} meanings after a few thousand exposures. However, the NM framework makes it difficult to tell how much cultural evolution is due to direct bias and how much is due to the transmission bottleneck. The sharp discontinuity between generations in the ILM highlights the importance the transmission bottleneck. In the NM, there is no notion of a generation, no sharp discontinuity and no clear indication of when the bottleneck applies. In fact, we would expect the impact of the bottleneck to be at its most severe at the start of an NM simulation — at this point learners have made few observations and are still called upon to produce utterances. The strength of the bottleneck decreases as observations are accumulated. Is the structure of the population's language determined early on, when the bottleneck is very tight, or later, after its severity is diminished?

Batali's model represents a significant development of the very simple models of cultural transmission proposed by B&R. Cultural variants are highly structured aggregations of smaller culturally-transmitted variants. However, as discussed above, the framework of the Negotiation Model makes the role of bias during cultural transmission difficult to separate from forces arising from cultural transmission itself.

2.4 Summary of the Chapter

In this Chapter I have introduced B&R's general model of cultural transmission, and the more language-specific Expression/Induction Model. The latter is typically implemented as either an Iterated Learning Model, where vertical transmission is paramount, or a Negotiation Model, where horizontal transmission is paramount. Treating language as a culturally transmitted trait requires us to question to what extent language is indeed culturally transmitted, what constitute the units of cultural replication, and what form the Primary Linguistic Data available to learners takes.

I then provided a summary of B&R's taxonomy of pressures acting on cultural transmission — natural selection of cultural variants, guided variation, and biased transmission, which can be further subdivided into directly biased transmission, indirectly biased transmission and frequency-dependent bias. Each of these pressures has been implicated in the cultural evolution of some aspect of language, and I have given examples of models which demonstrate how these pressures can drive linguistic evolution. Finally, I described a further pressure, which does not appear in B&R's taxonomy, which has been hypothesised to play a role is the evolution of linguistic structure — the pressure to generalise arising from the bottleneck on cultural transmission.

These causes of cultural evolution, in particular the pressures arising from direct bias and a transmission bottleneck, will play a recurring role in the thesis. I will show that some of the fundamental structural properties of language are a consequence of the interaction of these pressures during the cultural transmission of language.

CHAPTER 3

The cultural evolution of communication

The first step in an investigation into the evolution of the distinguishing design features of language — cultural transmission, symbolicism and compositionality — is to consider the cultural evolution of simple, unstructured communication systems. Such systems are the equivalent of a symbolic vocabulary. In particular, I will search for the circumstances under which optimal communication emerges. My line of reasoning, for the purpose of this chapter, is as follows. Humans have a culturally transmitted, symbolic vocabulary. Therefore, humans must have the necessary mental apparatus to support such a symbolic vocabulary. I will make the default adaptationist assumption that this capacity must have provided a fitness payoff at some point in evolutionary history, and this payoff must have been due to the communicative benefits of symbolic vocabulary. Therefore, humans have the necessary mental apparatus to support a *communicatively useful*, culturally transmitted symbolic vocabulary. The aim of this chapter is therefore to identify the learning bias required to support a communicatively useful vocabulary, in the strongest case an optimal communication system, and equate this with the mental capacity of humans.

The assumption that the human capacity for symbolic vocabulary evolved due to fitness payoff arising from communication will be reexamined in Chapter 4.

In Section 3.1 I review previous computational models which tackle this issue. In Section 3.2 a simple model of communication is developed. In Section 3.3 a new ILM is introduced. This model shows that the key determinant of the population's communicative behaviour is the direct bias on cultural transmission resulting from the learner's bias. In light of this, a more sophisticated model is developed in Section 3.4 to investigate a wider range of learning biases. In Section 3.5 the key bias for the cultural evolution of vocabulary is identified and defined. Finally, parallels are drawn between this learning bias and the learning bias applied by human language learners to the task of vocabulary acquisition. This comparison suggests that the human learning bias (and perhaps not the learning bias of other, closely related species) exhibits the appearance of design for the cultural evolution of communicatively-optimal symbolic vocabulary.

3.1 Models of the evolution of vocabulary

In the review carried out in Chapter 2 I covered the models of the cultural evolution of vocabulary systems described in Hutchins & Hazelhurst (1995) and Nowak *et al.* (1999). To summarise briefly, Hutchins & Hazelhurst (working within the NM framework) report that communicatively-optimal symbolic vocabulary evolves culturally. I attributed this to a direct bias acting on cultural transmission, a consequence of the autoassociator network architecture used in their model. Nowak *et al.* also demonstrate (working within the ILM framework) that communicatively-optimal symbolic vocabulary can evolve culturally. However, in their model this is a consequence of the natural selection of cultural variants — successful communicators are more likely to act as cultural parents, therefore successful communication systems are preferentially retained in the population.

Hurford (1989) describes possibly the first computational investigation into the evolution of communication. One of the central concerns of Hurford's paper is the biological evolution of learning strategies, and his work in this area will be returned to in Chapter 4. However, Hurford does cover purely cultural evolution as well.

The communicative behaviour of individuals in Hurford's model is represented with two probability matrices — a production matrix, which gives the probability of producing a particular signal given a certain meaning, and a reception matrix, which gives the probabilistic reception behaviour of the individual. A generational ILM is used. Learners form their production and reception matrices based on a sample of the observable behaviour produced by the previous generation. Each learner samples once from the population's production and reception behaviour, yielding a set of observed meaning-signal pairs (based on a stochastic sample of the population's production behaviour) and a set of observed signal-meaning pairs (based on a stochastic sample of the population's reception behaviour). Hurford considers three learning strategies. *Imitator* agents form their transmission matrix based on observed transmission behaviour, and their reception matrix based on observed reception behaviour. *Calculators* base their production behaviour on observed reception and their reception on observed production. *Saussureans* base their production behaviour on observed production, then derive their reception behaviour from this matrix.

Imitator learners form their production and reception matrixes on the basis of direct observation of production and reception behaviour — if, for example, an Imitator observes signal $s1$ being produced for meaning $m1$, it will set the probability of producing $s1$ for $m1$ to 1 in its own production matrix.

Saussurean learners derive their production matrix from production behaviour in a similar manner, then design their reception matrix so as to make it optimally coordinated with their own production behaviour — for example, if a Saussurean learner arrives at a production matrix where $s1$ is produced for meanings $m1$ and $m2$, then it will interpret $s1$ as meaning $m1$ with probability 0.5 and $m2$ with probability 0.5.

Calculators form their reception matrix on the basis of observed production behaviour, in the same way that a Saussurean learner forms its reception matrix on the basis of its own production behaviour. For example, if a Calculator learner observes a population where $s1$ is produced for meanings $m1$ and $m2$, then it will interpret $s1$ as meaning $m1$ with probability 0.5 and $m2$ with probability 0.5. By the same optimisation process, Calculators calculate their production matrix on the basis of observed reception behaviour.

Hurford reports two results with relation to this ILM. Firstly, populations of Calculator agents are unable to preserve an optimal communication system over time. Secondly, populations of Imitator and Saussurean learners are capable of creating communication systems which lead to intermediate levels of communicative accuracy through purely cultural processes.

There are two candidate pressures acting on cultural transmission in Hurford's model. Firstly, the behaviour of the populations could be explained by natural selection of cultural variants — more successful communicators are more likely to act as cultural parents in Hurford's model. Secondly, the different learning strategies could result in different direct biases on cultural transmission. I will demonstrate in Section 3.5.3 that this later pressure is probably the key one.

Oliphant & Batali (1997) introduce another Iterated Learning Model of vocabulary. Individuals are required to communicate about a small set of meanings using a small set of signals. As in Hurford's (1989) model, individuals are modelled using probabilistic functions, with each individual being characterised by a production function, which gives the probability of each signal being sent for a given meaning, and a reception function, which gives the probability of a given signal being interpreted as a particular meaning.

Oliphant & Batali use a gradual population turnover model. At each time-step a single individual is removed from the population and replaced by a new individual. This individual estimates the average production and reception functions in use in the population, by making a number of observations of the population's production and reception behaviour. Based on these estimated functions, the new individual then creates its own production and reception functions according to one of two learning procedures, termed *Imitate-Choose* (a slight variation on Hurford's Imitator) and *Obverter* (related to Hurford's Calculator).

The learner's estimation of the probability with which the population produces signal σ_j for meaning μ_i is given by $P\left(\mu_i, \sigma_j\right)$ and the learner's estimation of the probability with which the population interprets σ_j as meaning μ_i is given by $R\left(\sigma_j, \mu_i\right)$. The learner must choose their own production and reception probabilities, given by $p\left(\mu_i, \sigma_j\right)$ and $r\left(\sigma_j, \mu_i\right)$.

The Imitate-Choose learner proceeds as follows:

```
For each meaning $\mu_i$:
  – Find the signal $\sigma_j$ for which $P\left(\mu_i, \sigma_j\right)$ is maximum.
  – Set $p\left(\mu_i, \sigma_j\right) = 1$ and $p\left(\mu_i, \sigma_k\right) = 0$ for all $k \neq j$.
For each signal $\sigma_j$:
  – Find the meaning $\mu_i$ for which $R\left(\sigma_j, \mu_i\right)$ is maximum.
  – Set $r\left(\sigma_j, \mu_i\right) = 1$ and $r\left(\sigma_j, \mu_k\right) = 0$ for all $k \neq i$.
```

The Imitate-Choose learner therefore bases its production behaviour on the average production behaviour of the population, selecting the most frequently used signal for each meaning. Similarly, reception behaviour is based on the population's reception behaviour, with the most frequent interpretation of a given signal being learned as the *only* interpretation of that signal. Note that there is no coupling between production and reception behaviour.

The Obverter learning procedure proceeds as follows:

```
For each meaning $\mu_i$:
  – Find the signal $\sigma_j$ for which $R\left(\sigma_j, \mu_i\right)$ is maximum.
  – Set $p\left(\mu_i, \sigma_j\right) = 1$ and $p\left(\mu_i, \sigma_k\right) = 0$ for all $k \neq j$.
For each signal $\sigma_j$:
  – Find the meaning $\mu_i$ for which $P\left(\mu_i, \sigma_j\right)$ is maximum.
  – Set $r\left(\sigma_j, \mu_i\right) = 1$ and $r\left(\sigma_j, \mu_k\right) = 0$ for all $k \neq i$.
```

Obverter learners will produce the signal which is most commonly interpreted by the rest of the population as conveying the meaning they wish to convey. Similarly, Obverter learners will interpret a signal as meaning the meaning it is most frequently produced for. The Obverter learner therefore bases its production behaviour on the population's reception behaviour and its reception behaviour on the population's production behaviour. Note that, unlike in the Imitate-Choose strategy, this results in the coupling of production and reception behaviour — the population's production behaviour at time t will influence its reception behaviour at time $t + 1$.

Oliphant & Batali define a measure of communicative accuracy for individuals using these probabilistic send and receive functions, and measure how the communicative accuracy of a population changes over time as new individuals are introduced and learn according to one of the two strategies. Communicative accuracy within the population is measured according to a variant of the canonical formula given in Chapter 2, Section 2.2.2.1 — simply put, a communicative episode between two individuals is a success if the hearer interprets the form produced by the speaker as conveying the meaning that the speaker intended.

Oliphant & Batali report that the Imitate-Choose strategy can increase communicative accuracy among a population where communicative accuracy is already high. However, in poorly coordinated populations the use of Imitate-Choose can result in further degradation. In contrast, use of the Obverter strategy always results in a steady increase in communicative accuracy until optimal levels are reached.

Why does the Obverter learning strategy result in optimal communication, but the Imitate-Choose strategy does not? Oliphant & Batali attribute the success of the Obverter strategy to its implicitly communicative aims — during learning, signals are selected so as to maximise their probability of being understood. Imitator agents do not have this built-in understanding of the communicative task. They suggest that both strategies build in ambiguity-avoiding measures — in both cases the most popular meaning-signal combinations are selected to the exclusion of other possible combinations. As we will see in the remainder of this Chapter, these comments are somewhat wide of the mark. Firstly, it is not necessary to build in an implicit understanding of the communicative task — in Section 3.4 I will demonstrate that optimal communication can emerge in a population of learners who do not select signals so as to maximise the probability of being understood. Secondly, building in an understanding of the communicative task does not necessarily lead to optimal communication — Hurford's Calculators are similar to Oliphant & Batali's Obverters, but cannot even preserve an optimal system. Finally, it will be shown in Section 3.5.3 that the Obverter and Imitate-Choose strategies respond differently to different types of ambiguity, and that this difference is crucial in understanding the behaviour of populations of such learners.

What is clear, however, is that the emergence or non-emergence of optimal communication in these populations is driven by what B&R term directly biased transmission — the Imitate-Choose and Obverter strategies have different biases as to how they acquire communication systems, and over the course of repeated cultural transmission the cultural variants which most closely match these biases come to dominate the population. We can surmise that the Imitate-Choose and Obverter strategies have different biases, with only the Obverter strategy being biased in favour of communication systems which maximise communicative accuracy.

Livingstone & Fyfe (1999) investigate the evolution of vocabulary using a model which is a hybrid NM-ILM. Livingstone and Fyfe's main concern is the emergence of diversity of vocabulary, but they do make some observations of the overall structure of the vocabularies in their populations. Individuals are modelled using neural networks, mapping from input signals to output meanings[1]. The neural network model of an agent has N input units and M output units, where these units can take values of ± 1. Meanings are represented by patterns of activation over the M nodes where a single node has an activation of $+1$. This yields M distinct meanings. Signals are represented by arbitrary patterns of activation over the N signal nodes, yielding 2^N possible signals. The network's behaviour while producing signals for a given meaning and arriving at the interpretation of a particular received signal are determined by the single layer of connection weights in the network, connecting all nodes in N with all nodes in M.

At each generation each individual in the generation $g + 1$ receives t exposures to the communicative behaviour of generation g individuals, where each exposure consists of an observation of a single meaning-signal pair. Generation $g + 1$ individuals then receive a further $t/2$ exposures to the communicative behaviour of other generation $g + 1$ individuals. This model therefore exhibits a degree of hybridization between the Iterated Learning and Negotiation models. However, it is more appropriate to classify the model as of the Iterated Learning type, as the exposures to the previous generation's communicative behaviour occur first and will have the greatest impact. The model outlined in Section 3.3 also suggests that the behaviour of the model would be qualitatively similar if the negotiation portion of learning were omitted.

[1] Livingstone & Fyfe (1999) actually present the network as one which maps from input meanings to output signals. However, during learning signals are treated as input and meanings as output. This turns out to be the key factor in understanding the behaviour of the model.

At each training episode the learner is presented with a signal-meaning pair. The learner takes the signal as input and produces a pattern of activation over the output meaning nodes, x', representing that individual's interpretation of that signal. The teacher's meaning x is then used to perform weight adjustment according to:

$$\Delta w_{ij} = \eta \left(x_i - x'_i \right) y_i$$

where w_{ij} is the weight of the connection between input node i and output node j, y_i and x_i gives the activation levels of the ith input and output unit respectively and η is the learning rate. This type of network model is typically referred to as an Obverter network (as is the network described in Batali (1998), discussed in Section 2.3.3.4), by analogy with the Obverter learning strategy of Oliphant & Batali (1997) — observed production behaviour is used to acquire reception behaviour.

Livingstone and Fyfe report that, over time, populations of such agents converge on shared, stable mappings between meanings and signals which would be optimal in terms of the communicative accuracy measures used by Oliphant & Batali. What drives the emergence of this optimal vocabulary system? As with the models of Hutchins & Hazelhurst (1995), Oliphant & Batali (1997) and Batali (1998), the learning bias of these agents results in directly biased transmission, with the learners happening to favour communication systems which are optimal in terms of communicative accuracy. A discussion of the nature of this bias is postponed until later in this Chapter.

3.2 The communication model

A communication system C consists of a *production* function $p\left(m\right)$, mapping from unstructured meanings m to unstructured signals s, and a *reception* function $r\left(s\right)$, mapping from signals s to meanings m. m and s are selected such that $m \in \mathcal{M}$ and $s \in \mathcal{S}$ where $\mathcal{M} = \left\{ m_1, m_2 \ldots m_{|\mathcal{M}|} \right\}$ and $\mathcal{S} = \left\{ s_1, s_2 \ldots s_{|\mathcal{S}|} \right\}$. This simple model is suitable for studying the emergence of conventionalised symbolic vocabulary.

How can we evaluate the communicative accuracy of a population using such a communication system? The accuracy of a single communicative event involving a producer P with production function $p\left(m\right)$, a receiver R with reception function $r\left(s\right)$ and a meaning $m_i \in \mathcal{M}$, $ca\left(P, R, m_i\right)$, is defined as:

$$ca\left(P, R, m_i\right) = \left\{ \begin{array}{ll} 1 & \text{if } r\left(p\left(m_i\right)\right) = m_i \\ 0 & \text{otherwise} \end{array} \right\}$$

When $ca\left(P, R, m_i\right) = 1$ the communication is successful. A population's communicative accuracy can be estimated by taking the average $ca\left(P, R, m_i\right)$ for a random sample of P, R and m_i. In a population possessing an *optimal communication system* $ca\left(P, R, m_i\right) = 1$ for any choice of P, R and m_i. This method of measuring communicative accuracy is adopted in Section 3.3.

Equivalently, if the production function $p(m)$ is viewed as a probabilistic function $p(s_j|m_i)$, which gives the probability of producing signal s_j given meaning m_i, and the reception function $r(s)$ is similarly viewed as a probabilistic function $r(m_i|s_j)$ then the communicative accuracy between two individuals with respect to a single meaning, $ca\left(P, R, m_i\right)$, is given by:

$$ca(P, R, m_i) = \sum_{j=1}^{j=|\mathcal{S}|} p(s_j|m_i) \cdot r(m_i|s_j)$$

The communicative accuracy of P and R over all meanings, $ca(P, R)$ can then be defined as the average of their communicative accuracy over each meaning $m_i \in \mathcal{M}$ e.g.

$$ca\left(P, R\right) = \frac{\sum_{i=1}^{i=|\mathcal{M}|} \sum_{j=1}^{j=|\mathcal{S}|} p\left(s_j|m_i\right) \cdot r\left(m_i|s_j\right)}{|\mathcal{M}|}$$

In a population possessing an optimal communication system $ca(P, R) = 1$ for any choice of P and R. This method of evaluating communicative accuracy is more appropriate for the model outlined in Section 3.4.

3.3 Model 1: a feedforward network model

In this Section a simple ILM is described, which is designed to allow the investigation of the impact of learner bias and natural selection of cultural variants on emergent communication systems. This model is based on my undergraduate dissertation (Smith 1998), and has been published in more recent form in Smith (in press).

In this ILM, communicative agents are modelled using feedforward neural networks. Neural networks were chosen for several reasons. Firstly, there is some tradition of using neural networks in research on the evolution of communication — neural networks of some form are used by Batali (1994), Hutchins & Hazelhurst (1995), Batali (1998), Cangelosi & Parisi (1998), Cangelosi (1999), Livingstone & Fyfe (1999) and Kirby & Hurford (2002). Continuing this tradition provides several benefits. In particular, using a similar model allows the results of this research to be more easily related to previous research and the generality of the results of earlier simulations to be tested.

Secondly, well-established mechanisms exist for training neural networks to learn input-output mappings (i.e. backpropagation). Using an established learning mechanism reduces the amount of novel elements contained in the model, as well as allowing our understanding of that mechanism to be expanded.

Finally, using neural networks allows both genetically-transmitted and culturally-transmitted information to influence, in principle, the eventual behaviour of agents in the model. This will prove useful in Chapter 4, when I will consider dual-transmission models.

3.3.1 The communicative agent

The model of a communication system is as described above in Section 3.2. Communicative agents must be capable of representing, using and learning such systems.

3.3.1.1 Representation

Feedforward neural networks are used to model communicative agents. Each individual is modelled using a single network mapping between meanings and signals. There are two possible types of networks: one which takes a representation of a meaning as input and produces a representation of a signal as output, and one which takes a signal as input and produces a meaning as output. The structure of the two networks are shown in Figure 3.1. Feedforward networks mapping from input meanings to output signals will be termed *imitator* networks, whereas networks mapping from input signals to output meanings will be referred to as *obverter* networks. The precise nature of the meaning-signal mapping in these networks is determined by the network connection weights.

Given that the input and output layers in these networks have three nodes, communication systems are mappings between three-dimensional meaning vectors and three-dimensional signal vectors. Binary vectors are used, giving 2^3 possible meanings and 2^3 possible signals. A subset of the set of possible meaning vectors, $\mathcal{M}_{CRS}$, are considered

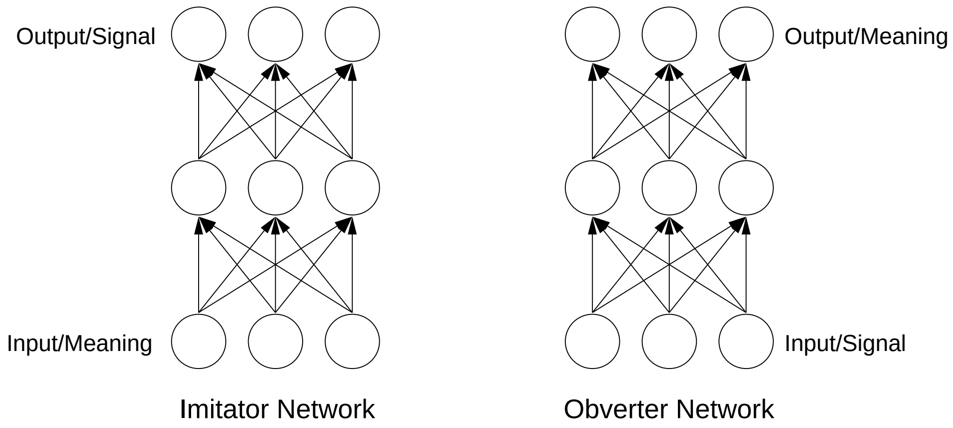


Figure 3.1: The two network architectures. Imitator networks map from input meanings to output signals. Obverter networks map from input signals to output meanings.

to be communicatively relevant situations, meaning the agents are required to communicate about them. For all simulations outlined in this section, $\mathcal{M}_{CRS}$ consists of the unit meanings represented by the vectors $(1\ 0\ 0)$, $(0\ 1\ 0)$ and $(0\ 0\ 1)$. The full set of possible signals are allowed. Therefore $|\mathcal{M}| = 8$, $|\mathcal{M}_{CRS}| = 3$ and $|\mathcal{S}| = 8$.

3.3.1.2 Production and reception

These neural network models of agents embody the production and reception functions $p\,(m)$ and $r\,(s)$. Deriving $p\,(m)$ from an imitator network or $r\,(s)$ from an obverter network is straightforward. For imitator networks $p\,(m)$ is derived by presenting the pattern of activation corresponding to each $m \in \mathcal{M}_{CRS}$ to the network, propagating activations forward through the network and thresholding the resultant real-valued output pattern of activation to give the signal $s \in \mathcal{S}$ associated with the given meaning. Similarly, for obverter networks $r\,(s)$ is derived by presenting the pattern of activation corresponding to each $s \in \mathcal{S}$ to the network, propagating activations forward through the network and thresholding the resultant real-valued output pattern of activation to give the meaning $m \in \mathcal{M}$ associated with the received signal.

Reception for imitator networks and production for obverter networks is slightly more complex, given that the networks are not bidirectional. To derive $r\,(s)$ for an imitator network each signal $s \in \mathcal{S}$ is considered in turn. All $m \in \mathcal{M}_{CRS}$ are propagated through a given agent's network to produce a real-numbered output pattern of activation for each meaning. Each output pattern is given a confidence rating, corresponding to how closely that pattern matches the signal currently under consideration, s. The meaning which

Process	Network Type	
	Imitator	Obverter
Production	propagate	confidence measure
Reception	confidence measure	propagate

Table 3.1: A summary of the production and reception procedures for the two types of networks. Imitators produce a signal for a given meaning by propagating activations forward through their network, and arrive at a meaning given a received signal using the confidence measure. Obverter networks produce using the confidence measure process, and receive by propagating activations.

yields the real-numbered output vector closest to s, according to the confidence measure, is chosen as the interpretation of s. This method is based on the method used by Batali (1998) and Kirby & Hurford (2002) for producing outputs for similar networks. Similarly, to derive an obverter network's $p\left(m\right)$ each meaning $m \in \mathcal{M}_{CRS}$ is considered in turn. All $s \in \mathcal{S}$ are propagated through a given agent's network to produce a real-numbered output pattern of activation for each signal. Each output pattern is given a confidence rating, corresponding to how closely that pattern matches the meaning currently under consideration, m. The signal which yields the real-numbered output closest to m, according to the confidence measure, is chosen as the network's production for m.

The confidence measure that a given real-numbered output vector, o, of length n matches a target binary vector t of length n is given by $C\left(t, o\right)$. $C\left(t, o\right)$ is simply the product of the confidence scores for each individual node $1...n$ in the output vector i.e.

$$C\left(t\left[1\ldots n\right], o\left[1\ldots n\right]\right) = \prod_{i=1}^{n} C\left(t\left[i\right], o\left[i\right]\right)$$

where the confidence measure for node i is

$$C\left(t\left[i\right], o\left[i\right]\right) = \begin{cases} o\left[i\right] & \text{if } t\left[i\right] = 1, \\ \left(1 - o\left[i\right]\right) & \text{if } t\left[i\right] = 0. \end{cases}$$

(Equations adapted from Kirby & Hurford (2002))

The production and reception processes for both types of networks are summarised in Table 3.1.

The deterministic nature of these networks during production means that a definition of ambiguity for communication systems can be formally stated. Communication systems used by neural networks will be termed:

- *Unambiguous* if $p(m)$ is a one-to-one function.
- *Partially ambiguous* if $p(m)$ is a many-to-one function, but the range of $p(m)$ is not a singleton set.
- *Fully ambiguous* if the range of $p(m)$ is a singleton set.

3.3.1.3 Learning

In common with most implementations of the NM and ILM, I assume here that individuals learn from observed meaning-signal pairs. Well-established procedures exist for training feedforward networks to associate pairs of input-output pairs — here I use the backpropagation method (Rumelhart *et al.* 1986). For imitator agents, the training process involves attempting to associate an input meaning with an output signal. Imitators are therefore learning their production function on the basis of observed production behaviour. Obverter networks learn to associate input signals with output meanings — obverters learn their reception function on the basis of observed production behaviour, as in Batali (1998), Livingstone & Fyfe (1999) and others.

3.3.2 The Iterated Learning Model

As discussed in Chapter 2, the results of repeated cultural transmission can be investigated using an Iterated Learning Model. In an ILM agents acquire their competence through learning from observations of the behaviour of other agents. This competence is then used to generate behaviour which is observed in turn by other agents. In the case of this model, the culturally-transmitted behaviour of interest is a communication system.

The process of iterated learning requires a model of population turnover. In this model I use a generational population turnover model, illustrated in Figure 2.6 (a) in Chapter 2. At every time-step a new population of a certain size is created. The pre-existing population produces some observable behaviour and the members of the new population observe and learn from that behaviour. The pre-existing population is then removed and replaced by the newly-created population and the process repeats.

More formally, the generational ILM consists of an initialisation process and an iteration process:

Initialisation Create a population $population_{g=0}$ of N agents[2]. Each agent is either an imitator or obverter, as described above, with populations being homogeneous in this

[2] $N = 100$ for all ILMs outlined in this section.

respect. Each agent has random initial connection weights in the range $[-1, 1]$. Each agent's communication system is determined by these random initial connection weights.

Iteration

1. Evaluate the communicative accuracy of every member of $population_g$ by evaluating every individual's communicative accuracy as both producer and receiver with two randomly selected partners according to the measure $ca\,(P, R, m)$, for every $m \in \mathcal{M}_{CRS}$.
2. For every member of the population $population_g$, generate a set of meaning-signal pairs by applying the network production process to every $m \in \mathcal{M}_{CRS}$. Noise is added to each meaning-signal pair[3] with probability p_n.
3. Create a new population $population_{g+1}$ of N agents of the same type (imitator or obverter) as $population_g$, where each member of $population_{g+1}$ has random initial connection weights in the range $[-1, 1]$.
4. Each member of $population_{g+1}$ receives e exposures to the observable behaviour generated by $population_g$. During each of these e exposures the new agent observes the complete set of meaning-signal pairs generated by a member of $population_g$ selected randomly from among the t most successful communicators in $population_g$. For each exposure the learner updates their connection weights according to the observed meaning-signal pairs using the backpropagation learning algorithm[4].
5. $population_g$ is removed and replaced with $population_{g+1}$. Return to 1.

Each pass through the iteration process will be termed a *generation*. Note that the selection of individuals to observe depends on t and therefore allows the possibility of *natural selection* of cultural variants, as described by B&R. If $t = N$ then selection of individuals to act as cultural parents is independent of the communicative success of those individuals and there is no natural selection of cultural variants. When $t < N$ the probability of an individual being observed and learned from will depend on their evaluated communicative success, and there will be natural selection, acting on cultural transmission, in favour of communication systems which result in successful communication.

The fact that every individual in a population begins their life with a particular network type (imitator or obverter) and a particular set of connection weights (randomly

[3] In order to add noise to a meaning-signal pair $\langle m_i, s_j \rangle$, s_j is replaced with a randomly-selected $s_k \in \mathcal{S}$, where $k \neq j$.

[4] A learning rate of 0.5 is used

distributed within some range) suggests some kind of innate endowment of these components. It is our goal to investigate the impact of this innate endowment on the communicative behaviour of the population. However, every agent begins life with the *same* endowment – there is no possibility of genetic variation within the population. The emergent behaviour of the population will therefore be determined by the dynamics resulting from the iterated cultural transmission of communication systems among individuals with a common genetic endowment. In Chapter 4 I will investigate how the biological evolution of this innate endowment in a genetically heterogeneous population can impact on the evolution of communication systems.

3.3.3 Network architecture, learning bias and natural selection

The goal of this Chapter of the thesis is to identify the learning mechanisms necessary to create, through cultural processes, a communicatively useful vocabulary. The ILM described above can be used to investigate whether imitator and obverter agents construct an optimal, unambiguous communication system from random initial behaviour, and under what circumstances. To this end, runs of the iterated learning model were carried out. In these simulations, the communication system used by agents in the initial population is dependent on their random connection weights, and is therefore random. 10 runs were carried out for each set of experimental conditions, with runs proceeding for 1000 generations. We are primarily interested in the *end states* of these runs, rather than their progress through time. In order to evaluate the end state communication system in use in the populations, the average communicative accuracy of the population is recorded for the last 10 generations of each run. Each point in the plots that follow therefore represents the average communicative accuracy of 10 populations over a period of 10 generations.

3.3.3.1 Learning bias and no natural selection

Figure 3.2 shows the results for simulation runs for imitator and obverter populations where $t = N$, (every member of the population is a potential cultural parent) for various numbers of learning exposures (e), in the absence of noise on cultural transmission ($p_n = 0$).

The different network architectures clearly result in very different behaviour, when placed in the context of the ILM. For imitator networks, the populations converge on communication systems which result in communicative accuracy of 0.33. This level of communicative accuracy is a consequence of the population using a shared, fully ambiguous

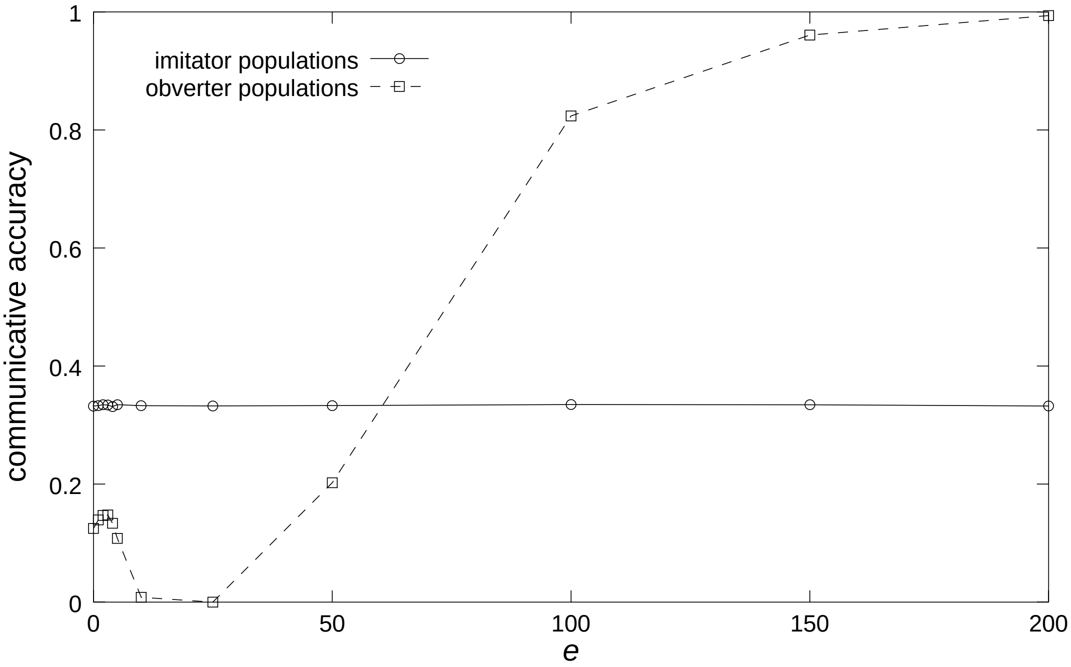


Figure 3.2: The average final communicative accuracy of imitator (solid line) and obverter (dashed line) populations as a function of the number of learning exposures (e) used during the simulation runs. These results are for the case where there is no natural selection of cultural variants ($t = N$). Imitator populations converge on systems which yield chance levels of communicative accuracy, regardless of e. In contrast, given high enough e, obverter populations converge on communication systems which give high levels of communicative accuracy.

communication system. In contrast, obverter populations converge on levels of communicative accuracy close to optimal when e is large — given large enough e, obverter populations converge on a shared, unambiguous vocabulary.

Why do the two different network architectures display this behaviour when placed in the context of the ILM? We can rule out natural selection of cultural variants (because $t = N$). This behaviour therefore must be due to direct bias pressure operating on cultural transmission.

In order to understand the source of this bias, it is necessary to assess the ability of individual agents, in isolation, to acquire systems of various levels of ambiguity, with varying levels of exposure to such systems. The possible range of meaning-signal mappings is actually rather small — $|\mathcal{M}_{CRS}| = 3$ and$|\mathcal{S}| = 8$ gives $|\mathcal{S}|^{|\mathcal{M}_{CRS}|} = 512$ possible meaning-signal mappings. It is therefore possible to examine the ability of agents to acquire every possible system. Of the 512 possible systems, 8 are fully ambiguous, 168 are partially ambiguous and 336 are unambiguous. For each system, 100 networks with random initial weights in the range $[-1, 1]$ were given e exposures to that system of meaning-signal

e	System Type		
	Fully Ambiguous	Partially Ambiguous	Unambiguous
1	25.4	0.3	0.0
2	47.1	0.8	0.0
3	74.0	1.4	0.0
4	91.8	1.4	0.1
5	98.0	1.7	0.0
10	100.0	0.5	0.0
25	100.0	1.6	0.1
50	100.0	34.2	13.1
100	100.0	92.8	82.9
150	100.0	99.4	98.3
200	100.0	100.0	99.8

Table 3.2: The imitator learning bias. The table shows the percentage of imitator networks which succeed acquiring languages of the various classifications, according to e, the number of exposures to the system. For imitator networks, fully ambiguous systems are easier to learn.

e	System Type		
	Fully Ambiguous	Partially Ambiguous	Unambiguous
1	0.1	0.2	0.3
2	0.0	0.3	0.3
3	0.1	0.3	0.5
4	0.0	0.5	0.5
5	0.1	0.4	0.7
10	0.0	0.9	1.6
25	0.0	3.7	7.8
50	0.0	10.1	26.2
100	0.0	14.9	49.1
150	0.0	15.2	53.8
200	0.0	15.2	54.8

Table 3.3: The obverter learning bias. The table shows the percentage of obverter networks which succeed acquiring languages of the various classifications, according to e, the number of exposures to the system. Obverter networks find unambiguous systems easier to learn.

mappings. Learning proceeds via the backpropagation process, with the same learning rate as used in the ILM. A network was judged to have learned a system successfully if the observed system could be reproduced in production — for every meaning-signal pair $\langle m_i, s_j \rangle$ production of the signal associated with m_i resulted in s_j being produced. The results are summarised in Tables 3.2 and 3.3 by communication system type.

As can be seen from Table 3.2, for imitator agents systems exhibiting a higher degree of ambiguity are easier to acquire than systems exhibiting a lower degree of ambiguity, for

System Type	% population
Unambiguous	2
Partially Ambiguous	25
Fully Ambiguous	73

Table 3.4: The behaviour of imitator agents with random connection weights. The table shows the percentage (based on 1000 test networks) of imitator networks with random connection weights (in the range [-1,1]) who use a communication system of the given type. Random imitator networks tend to produce fully ambiguous systems.

all values of e. Table 3.3 shows that obverter agents have the opposite learning bias — systems exhibiting lower degrees of ambiguity are easier to acquire, for all values of e. Learnability never reaches 100%, even for unambiguous communication systems. It appears that certain unambiguous systems are unlearnable by obverter agents, while certain unambiguous systems are 100% learnable. The key point is that certain unambiguous systems are highly learnable whereas partially ambiguous and fully ambiguous systems are less learnable.

Returning to the results for ILM runs involving imitator populations, for low values of e, no communication system can reliably be learned. Populations essentially behave in a random fashion. The typical random behaviour of imitator agents is shown in Table 3.4. The majority of individuals use fully ambiguous systems, resulting in chance levels of communicative accuracy. As e increases, fully ambiguous systems rapidly become highly learnable, and are always more learnable than less ambiguous systems. Less ambiguous systems are less likely to be successfully learned than fully ambiguous systems, and are unstable over time. The populations therefore converge on fully ambiguous systems, resulting in low levels of communicative accuracy.

In contrast, the communicative accuracy in obverter populations increases as e increases. For low values of e all systems are unlearnable, and individuals use a random system. As shown in Table 3.5, obverter agents with random connection weights tend to use a unambiguous systems. The communicative accuracy of the population is therefore low, as there are a large number of uncoordinated unambiguous systems present. As e increases, the learnability of unambiguous systems increases, and is always higher than the learnability of more ambiguous systems. Unambiguous systems become increasingly stable relative to more ambiguous systems and the populations converge on shared unambiguous communication systems, resulting in high levels of communicative accuracy for high values of e.

System Type	% population
Unambiguous	65
Partially Ambiguous	33
Fully Ambiguous	2

Table 3.5: The behaviour of obverter agents with random connection weights. The table shows the percentage (based on 1000 test networks) of obverter networks with random connection weights (in the range [-1,1]) who use a communication system of the given type. Random obverter networks tend to produce unambiguous systems.

The behaviour of these populations in the ILM is therefore determined by the learning biases of the two network architectures. These learning biases result in direct bias acting on cultural transmission, with the cultural variants favoured by the bias eventually reaching fixation.

3.3.3.2 Learning bias and natural selection

Figures 3.3 and 3.4 show the results for simulation runs for imitator and obverter agents where $t < N$ (only the top t individuals act as cultural parents, and t is less than the population size N, therefore the less able communicators may not act as cultural parents), for various values of t and e (learning exposures), again in the absence of noise ($p_n = 0$). In these simulations there are two pressures operating on the communication systems in the populations:

1. *Selection for learnability*, driven by the agents' learning bias, favouring either more ambiguous communication systems (in the case of imitator agents) or less ambiguous systems (in the case of obverter agents).
2. *Selection for communicative success* driven by natural selection of communication systems, favouring systems which result in successful communication.

In populations of imitator agents pressures 1 and 2 are in conflict, with selection for learnability favouring fully ambiguous systems (as discussed in the previous section), while natural selection favours shared unambiguous systems. For obverter populations these pressures are not in conflict, with both favouring the development of shared unambiguous systems.

The addition of natural selection of cultural variants has little impact on the emergent communication systems — as with the case where there is no natural selection, imitator populations converge on fully ambiguous communication systems and communicative accuracy remains uniformly low, while obverter populations converge on unambiguous

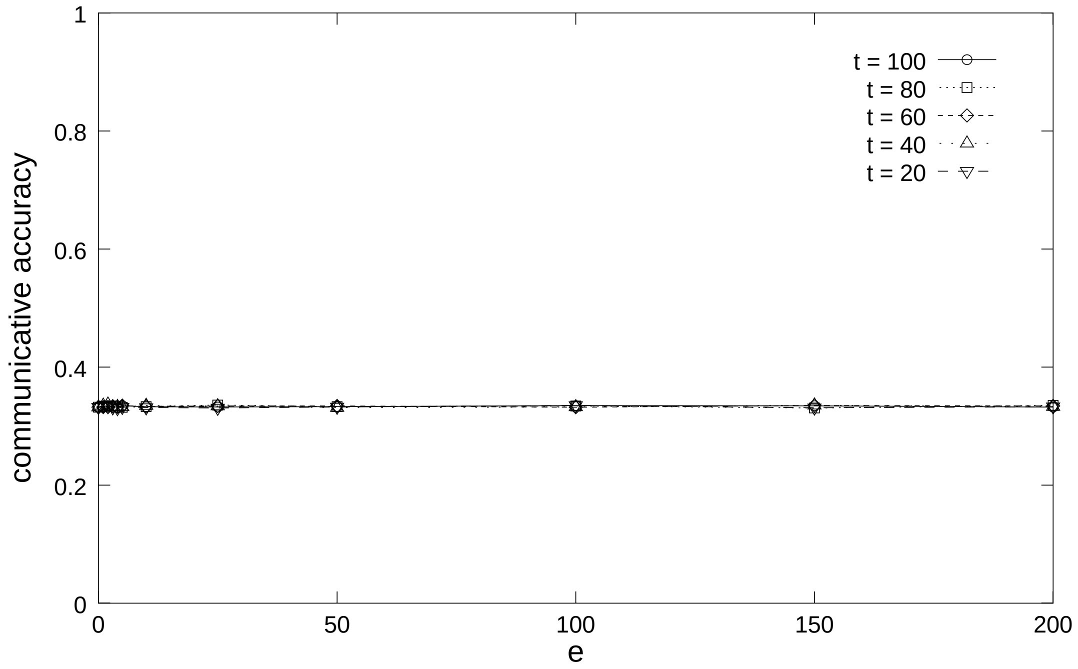


Figure 3.3: The average final communicative accuracy of imitator populations where there is natural selection of cultural variants ($t \leq N$), and no noise on cultural transmission ($p_n = 0$), as a function of e. Natural selection of cultural variants clearly has no impact.

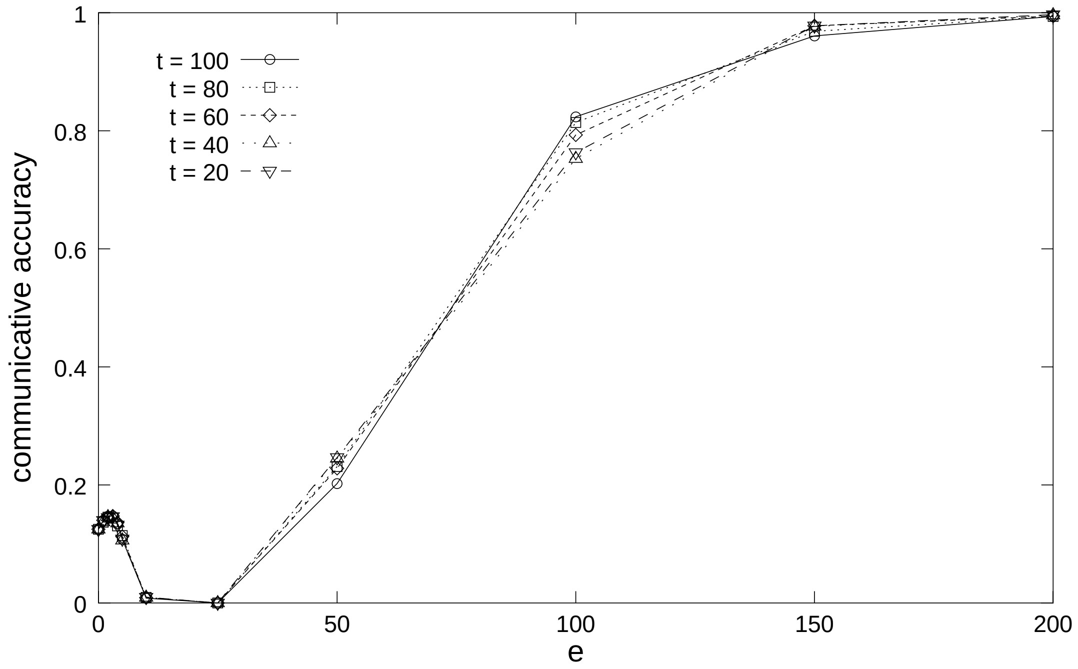


Figure 3.4: The average final communicative accuracy of obverter populations where there is natural selection of cultural variants ($t \leq N$), and no noise on cultural transmission ($p_n = 0$), as a function of e. Natural selection of cultural variants has little impact — there are very slight differences in final communicative accuracy, dependent on t.

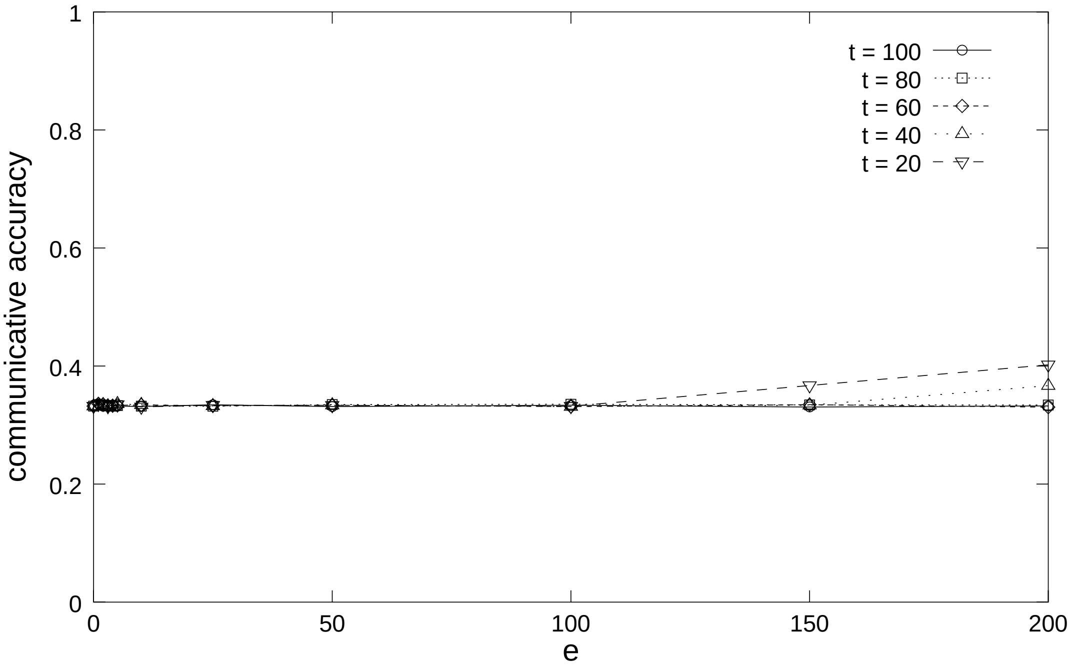


Figure 3.5: The average final communicative accuracy of imitator populations where there is natural selection of cultural variants ($t \leq N$), and noise on cultural transmission ($p_n = 0.05$), as a function of e. Natural selection of cultural variants clearly a slight impact for high e.

communication systems given sufficiently high e, with consequently high communicative accuracy. The behaviour of the populations is still dominated by the intrinsic learning bias of the agents.

B&R highlight the importance of cultural variation in populations where cultural transmission is undergoing natural selection — where there is no variation, natural selection is powerless. It is possible that we are not seeing any impact from natural selection due to a lack of cultural variability in the populations. While the initial populations exhibit variability (see Tables 3.4 and 3.5), which direct bias clearly feeds off, it could be that biased cultural transmission eliminates this variability too quickly, preventing natural selection of cultural variants from functioning. In order to investigate this possibility, the experiments outlined above were repeated with noise on cultural transmission ($p_n = 0.05$). This noise will potentially introduce cultural variation, which natural selection can then feed off. The results are plotted in Figures 3.5 and 3.6.

The introduction of noise has a slight impact. In imitator populations, when e is very high, there is a slight increase in average communicative accuracy for the case where $t = 20$. This is in fact due to one (when $e = 150$) or two (when $e = 200$) of the ten runs converging on partially ambiguous communication systems. This only occurs when

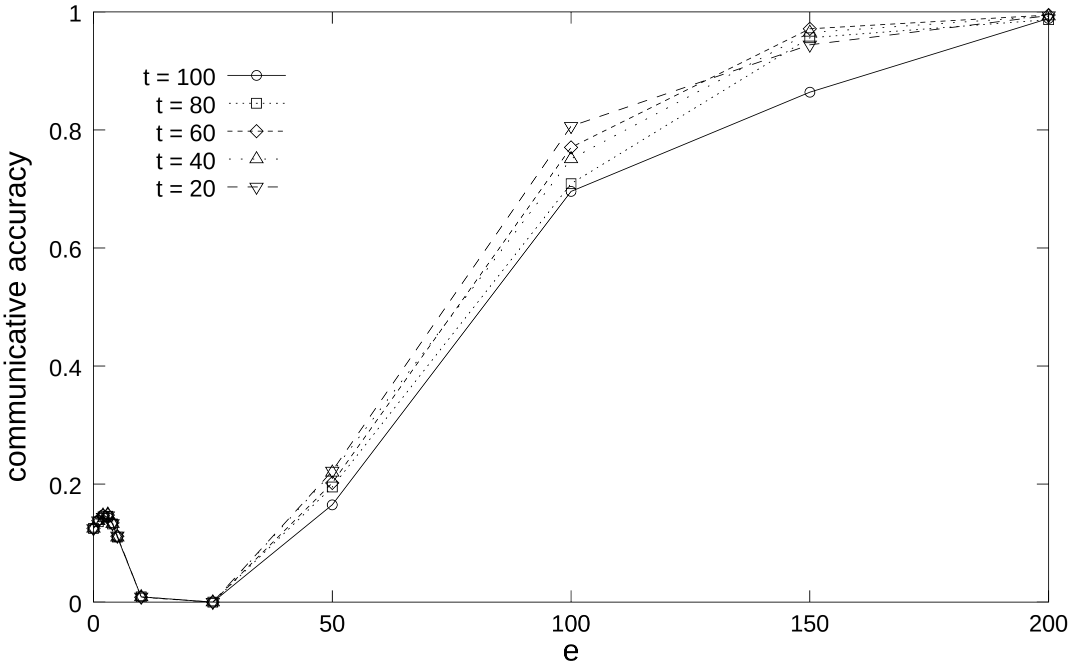


Figure 3.6: The average final communicative accuracy of obverter populations where there is natural selection of cultural variants ($t \leq N$), and noise on cultural transmission ($p_n = 0.05$), as a function of e. Natural selection of cultural variants has a noticeable impact for $e = 100$ or 150.

e is very high as this is the point where the individual's learning bias is weakest — as can be seen from Table 3.2, partially ambiguous systems seem to be as learnable as fully ambiguous systems where $e \geq 150$. However, the natural selection of cultural variants needs to be very severe to produce this slight effect.

In obverter populations there is, similarly, a slight impact, with natural selection of cultural variants improving the populations' communicative accuracy somewhat for certain amounts of learning ($e = 100$ or 150). The effect is still fairly minor.

3.3.4 Summary

The two distinct models of communicative agents have different learning biases, as highlighted by the acquisition tests outlined in Section 3.3.3.1. Placing these agents within an ILM allows the consequences of the iterated application of these learning biases to be explored. In the case where the direct bias on cultural transmission introduced by the agents' learning bias is the sole factor at play, the populations converge on the type of communication system favoured by that bias, as predicted by B+R. Natural selection of cultural variants in addition to this biased transmission has a very minor impact, and even then only when noise is injected into the system to provide variation.

Why do imitator and obverter agents have different biases and how can this bias best be described? This issue will be returned to in Section 3.5. However, it is clear from the relatively simple experiments outlined in this section that the behaviour of populations of individuals is strongly determined by their learning bias, which can override other pressures acting on cultural transmission, such as natural selection. Comparison of more than two alternative learning biases remains desirable, and a model allowing the comparison of a much wider range of biases is outlined in the next section. The feedforward neural network model is returned to in Chapter 4, where it is used to investigate the interactions between genetic and cultural transmission of communication.

3.4 Model 2: an associative network model

The feedforward network model described above is limited in the sense that the learning bias of individual agents is a consequence of their network architecture, and there are only two such possible architectures — the imitator architecture, and the obverter architecture. Ideally we would like to be able to experiment with a wider range of learning biases, in order to isolate the elements of bias which drive the cultural evolution of symbolic vocabulary.

A promising approach to addressing precisely this question is outlined in Oliphant (1999). Oliphant investigates how different learning rules influence the development of a vocabulary system through cultural processes within a population of associative networks. While the approach described in this paper is promising, its execution suffers from several shortcomings. Firstly, only three possible learning rules are considered. Secondly, while it is shown that certain learning rules result in the emergence of optimal communication, the properties of the learning rules that result in this behaviour are not explicitly identified. Thirdly, the results for those three learning rules are not related to other results in the field.

In this section I introduce a model, based on Oliphant's, which allows a wide range of learning rules, and associated learning biases, to be explored. This exploration allows me, in Section 3.5, to identify the key bias leading to the emergence of communicatively optimal, symbolic vocabulary through cultural processes. This bias can also be identified in the feedforward network model described above, and in most other models of the emergence of vocabulary via cultural evolution.

3.4.1 Communicative agents

The model of communication is as outlined in Section 3.2. Given the nature of the communicative agent model, the probabilistic interpretation of the communicative accuracy function is more natural, and is used in this section.

An associative network is used to model communicative agents. Since this model is less standard than the feedforward network model outlined in Section 3.3 a detailed and somewhat formal description is given here.

3.4.1.1 Representation

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

Patterns of activation over $\mathcal{N}_M$ are considered to represent meanings, whereas patterns of activation over $\mathcal{N}_S$ are considered to be signals. Restricting these patterns of activation to contain a single active unit yields $|\mathcal{N}_M|$ orthogonal meaning representations and $|\mathcal{N}_S|$ orthogonal signal representations, suitable for representing sets of unstructured meanings and unstructured signals such as those described in Section 3.2. If Gi is the ith node from the set $\mathcal{N}_G$ and the activation of node Gi is a_{Gi} then the meaning m_i corresponds to a pattern of activation over $\mathcal{N}_M$ where $a_{Mi} = 1$ and $a_{M(j \neq i)} = 0$. Similarly, the signal s_i corresponds to a pattern of activation over $\mathcal{N}_S$ where $a_{Si} = 1$ and $a_{S(j \neq i)} = 0$. This representational scheme is illustrated in Figure 3.7.

3.4.1.2 Production and reception

Patterns are retrieved from the network using a k-winners-take-all strategy. In order to retrieve a pattern of activation over nodes in $\mathcal{N}_S$ based on an input pattern of activation over nodes in $\mathcal{N}_M$ the weighted sum of inputs to node Si, q_{Si}, for each $Si \in \mathcal{N}_S$ is calculated according to the formula:

$$q_{Si} = \sum_{j=1}^{j=|\mathcal{N}_M|} a_{Mj} \cdot w_{Mj,Si}$$

where $w_{a,b} \in \mathcal{W}$ is the weight of the connection between nodes a and b. The k nodes in $\mathcal{N}_S$ with the highest values of q then have their activations set to 1, while all other nodes in $\mathcal{N}_S$ have their activations set to 0. If several nodes have equal q a random winner is selected from among them. Patterns of activation over the nodes in $\mathcal{N}_M$ are retrieved

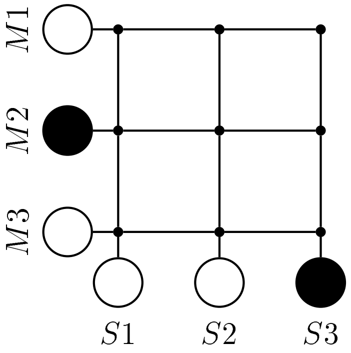


Figure 3.7: A neural network where $|\mathcal{N}_M| = |\mathcal{N}_S| = 3$. Large filled circles represent nodes with activation of 1, large empty circles represent nodes with activation of 0. The pattern of activation over $\mathcal{N}_M$ therefore represents the meaning m_2 ($a_{M2} = 1$, $a_{M1} = a_{M3} = 0$). Similarly, the pattern of activation over $\mathcal{N}_S$ represents the signal s_3

based on input patterns of activation over $\mathcal{N}_S$ in exactly the same way. For all simulations outlined in this paper, $k = 1$ — retrieved patterns of activation only ever consist of a single active node and $(|\mathcal{N}| - 1)$ non-active nodes. This ensures that retrieved patterns of activation conform to our representation of meanings and signals outlined above. This retrieval process is illustrated in Figure 3.8.

Retrieving a pattern of activation over $\mathcal{N}_S$ given an input pattern of activation over $\mathcal{N}_M$ corresponds to retrieving the signal associated with a given meaning — *production* of a signal associated with a given meaning. Retrieving a pattern of activation over $\mathcal{N}_M$ given an input pattern of activation over $\mathcal{N}_S$ corresponds to retrieving the meaning associated with a given signal — *reception* of a given signal and interpretation of that signal to yield a meaning. Note that the production and reception behaviour of such networks are not necessarily closely related — for example, the network in Figure 3.8 would produce $s2$ when prompted with $m2$, but would interpret $s2$ as meaning $m3$. Using a single network for both production and reception, as opposed to two separate networks, does however allow the possibility of a coupling of production and reception.

3.4.1.3 Learning

In order to store the association between patterns of activation over $\mathcal{N}_M$ and $\mathcal{N}_S$ the activations of the nodes in $\mathcal{N}_M$ and $\mathcal{N}_S$ are set to the required values and the weights of the connections in $\mathcal{W}$ are adjusted according to some weight-update rule W. If we assume that W must only adjust connection weights based on local information and that all patterns of activation will be binary, W can be specified by the 4-tuple $(\alpha\ \beta\ \gamma\ \delta)$,

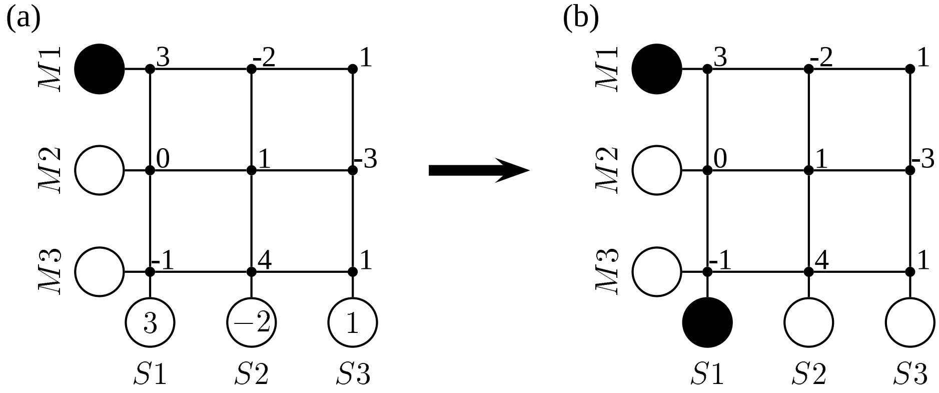


Figure 3.8: Retrieval of a pattern of activation over $\mathcal{N}_S$ based on a pattern of activation over $\mathcal{N}_M$. As before, large filled circles represent nodes with activation of 1. Connections between nodes are represented by the intersections of connecting lines and have an associated weight. In (a), the nodes in $\mathcal{N}_M$ have been set to a pattern of activation, resulting in a pattern of weighted sums of inputs over the nodes in $\mathcal{N}_S$ — the q values for those nodes The numbers in the centre of the nodes in $\mathcal{N}_S$ represent the weighted sums to those nodes. In (b) the result of the application of the winner-take-all process is shown — q_{S1} is greater than q_{S2} or q_{S3}, therefore node $S1$ has its activation set to 1 while nodes $S2$ and $S3$ have their activations set to 0.

where the value in α specifies how the weight of connection $w_{i,j}$ should be adjusted when $a_i = a_j = 1$, the value in β specifies how $w_{i,j}$ should be adjusted when $a_i = 1$ and $a_j = 0$, the value in γ specifies how $w_{i,j}$ should be adjusted when $a_i = 0$ and $a_j = 1$ and the value in δ specifies how $w_{i,j}$ should be adjusted when $a_i = a_j = 0$. While weights could be adjusted in many ways we will restrict ourselves here to the simplest case where α, β, γ and δ must take integer values in the range $[-1, 1]$. This yields a range of $3^4 = 81$ possible weight-update rules.

Given our interpretations of patterns of activations of $\mathcal{N}_M$ and $\mathcal{N}_S$ this storage process represents the process of learning the association between a meaning and a signal in a meaning-signal pair $\langle m, s\rangle$ according to some rule W. The learning process is illustrated in Figure 3.9.

3.4.2 Acquisition of an optimal system

We now have a model of communication, a model of an agent and processes of production, reception and learning. The feedforward neural network model highlighted the importance of the learning biases of agents when accounting for the behaviour of populations of such agents. The first question to be addressed here is therefore to ask whether individual agents, in isolation, can acquire an optimal communication system. To this end an unambiguous set of meaning-signal pairs $\mathcal{A} = \{\langle m_1, s_1\rangle, \langle m_2, s_2\rangle \ldots \langle m_{10}, s_{10}\rangle\}$

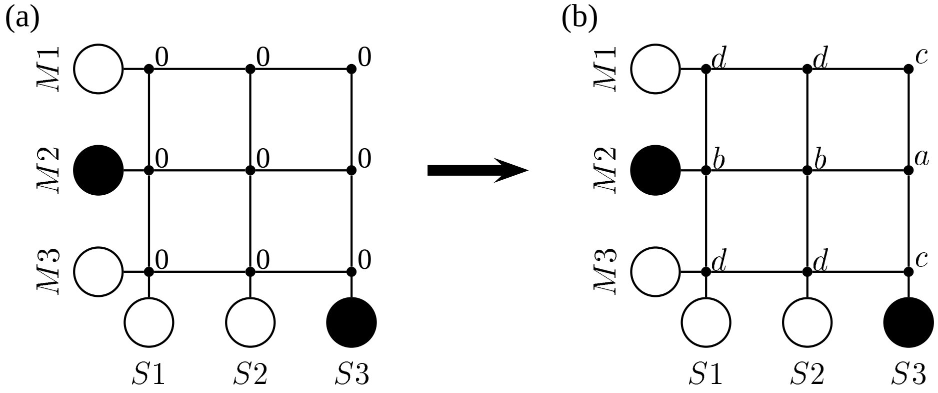


Figure 3.9: Storage of the meaning-signal pair $\langle m_2, s_3 \rangle$ using the weight-update rule $W = (a\ b\ c\ d)$. In (a), the nodes in $\mathcal{N}_M$ and $\mathcal{N}_S$ have been set to the patterns of activation representing m_2 and s_3. All connections have weight 0. In (b) the result of the application of the storage process is shown — all connections now have weights of a, b, c or d, depending on the activations of the nodes they connect.

was constructed. Agents using each of the 81 possible weight-update rules were then trained on $\mathcal{A}$, by storing each meaning-signal pair in $\mathcal{A}$ in their network. The agents were then evaluated to see if they had successfully acquired an optimal communication system based on exposure to the unambiguous set of meaning-signal pairs $\mathcal{A}$. Agents are judged to have acquired an optimal system, if, for every $\langle m_i, s_i \rangle \in \mathcal{A}$ both:

1. Production of the signal associated with m_i always[5] results in s_i being produced, i.e. $\langle m_i, s_i \rangle$ can be reproduced in production *and*
2. reception of s_i always results in the interpretation m_i, i.e. $\langle m_i, s_i \rangle$ can be reproduced in reception, meaning that the agent would communicate optimally with itself or another agent using the same weight-update rule exposed to $\mathcal{A}$.

The 81 weight-update rules can therefore be classified according to a [±learner] feature. 31 of the 81 possible weight-update rules were judged to be capable of acquiring the optimal communication system and were classified as [+learner]. The remaining 50 weight-update rules were classified [−learner].

[5] The term "always" has to be introduced to account for the stochastic nature of the behaviour of some networks, resulting from multiple nodes in the network receiving the same weighted sum of inputs on presentation of a pattern. In practice, "always" was reduced to "for every one of 1000 trials".

3.4.3 The Iterated Learning Model

As with the feedforward network model discussed above in Section 3.3, this model of a communicative agent can be slotted into an Iterated Learning Model to evaluate how the different weight-update rules influence the development of communication over time in a population. Unlike the feedforward network model, a *gradual*, rather than generational, population turnover model is used. The gradual population turnover model was preferred to counter the possibility in the new model of “inverting” learners, who learn the opposite communication system to their cultural parents. In a generational ILM populations of such agents would score highly on the intra-generational communicate accuracy measures, but could not be said to have learned the communication system of their cultural parents.

In the gradual population turnover model (see Figure 2.6 (b) in Chapter 2) at every timestep a single agent is selected at random and removed from the population. The remaining members of the population produce some observable behaviour, in the case of this model sets of meaning-signal pairs. A new individual arrives and learns based on observations of the population’s observable behaviour, then enters the population. The process then repeats.

More formally, the ILM consists of an initialisation process and an iteration process:

Initialisation Create a population of N agents[6], each using the weight-update rule W and possessing communication system L.

Iteration

1. Select an agent at random from the population and remove it.
2. For every remaining member of the population, generate a set of meaning-signal pairs by applying the network production process to every $m \in \mathcal{M}$. Noise is added to each meaning-signal pair[7] with probability p_n.
3. Create a new agent with connection weights of 0 who uses weight-update rule W.

[6] $N = 100$ for all ILMs outlined in this section.

[7] In order to add noise to a meaning-signal pair $\langle m_i, s_j \rangle$, s_j is replaced with a randomly-selected $s_k \in \mathcal{S}$, where $k \neq j$.

4. The new agent receives e exposures to the population's observable behaviour. During each of these e exposures the new agent observes the complete set of meaning-signal pairs of a randomly selected member of the population and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W.
5. The new agent joins the population. Return to 1.

Each pass through the iteration process will be termed a *cohort*. Note that the random removal of agents from the population means there is no selection based on communicative ability. As with the feedforward network ILM, the fact that every individual begins its life with a weight-update rule and initial set of connection weights suggests some kind of innate endowment of these components. In the simulations outlined in this section populations are homogeneous with respect to this endowment, and we restrict ourselves to investigating the impact of cultural transmission factors on the emergent communication systems. In Chapter 4 the biological evolution of these innate endowments in a genetically heterogeneous population will be investigated.

3.4.4 Maintenance of an optimal system

The first question to be addressed using the ILM is whether a population of agents possessing a weight-update rule W can maintain an optimal system over time in the presence of a small degree of noise. Recall from the description of the ILM given above that the agents in the initial population use some predefined communication system L. For the experiments outlined in this section, the initial population's set of weights $\mathcal{W}$ were constructed such that the $p(m)$ of the initial L generates the set of meaning-signal pairs $\mathcal{L} = \{\langle m_1, s_1\rangle, \langle m_2, s_2\rangle \ldots \langle m_{10}, s_{10}\rangle\}$ — the initial population shares an unambiguous meaning-signal mapping. ILMs were run with each of the 81 possible learning rules, with noise introduced with probability $p_n = 0.05$ and each individual receiving exposures to the communication systems of three randomly-selected members of the population ($e = 3$). Populations were defined as having *maintained* the initial optimal system if the population's communicative accuracy remained above 0.95 for every cohort of a run.[8] Weight-update rules were classified as [$+$maintainer] if the optimal system was maintained for each of ten 2000-cohort runs.

The populations exhibited four typical patterns of behaviour, illustrated in Figure 3.10. Populations (a), (b) and (c) in Figure 3.10 have failed to maintain the optimal system

[8] The population's communicative accuracy was estimated by evaluating every individual's average communicative accuracy as both producer and receiver with two randomly selected partners according to the measure $ca(P, R)$ given in Section 3.2, averaging over all individuals in the population.

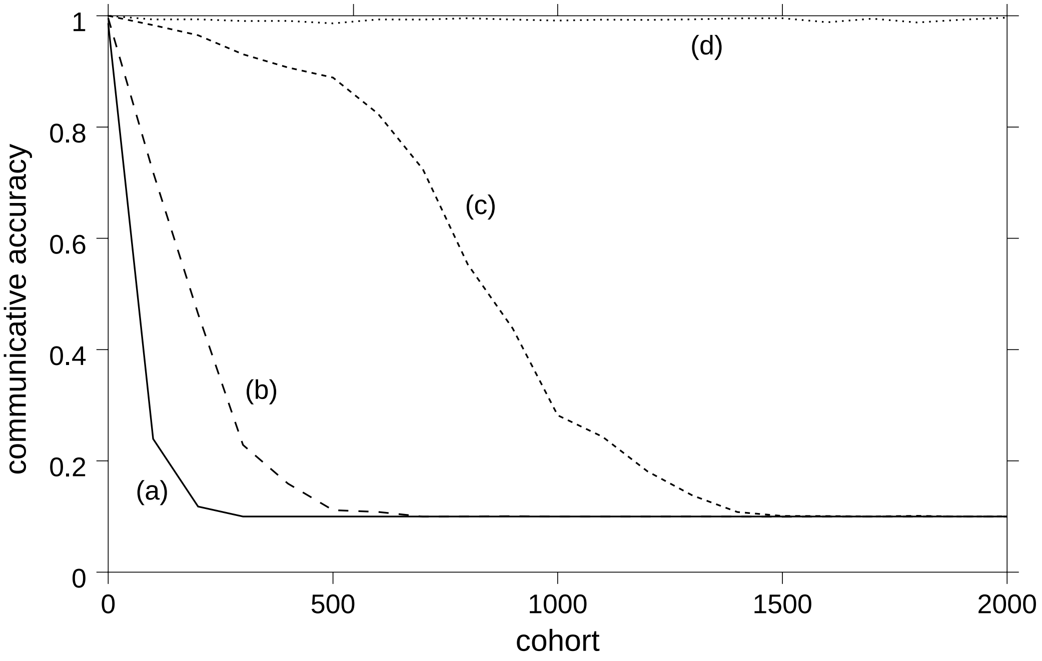


Figure 3.10: Populations of agents using the 81 learning rules exhibit four patterns of behaviour when attempting to maintain an optimal system. This figure plots the communicative accuracy over time of single populations exhibiting these patterns of behaviour: rapid collapse to chance levels of communicative accuracy, as in (a); less rapid collapse to chance levels of communicative accuracy, as in (b) and (c); maintenance of the optimal system, as in (d).

and can therefore be classified as [$-$maintainer], although population (a) in Figure 3.10 exhibits a more rapid decrease in communicative accuracy than populations (b) and (c). Unsurprisingly, all 50 populations using weight-update rules with the [$-$learner] feature followed the pattern of (a) and can therefore be classified [$-$learner, $-$maintainer]. Of the remaining 31 weight-update rules, 13 resulted in the type of pattern exemplified by populations (b) and (c) and can be classified as [$+$learner, $-$maintainer] and 18 resulted in patterns similar to that of population (d) in Figure 3.10 and can be classified as [$+$learner, $+$maintainer].

3.4.5 Construction of an optimal system

Finally, the 81 weight-update rules were examined to see whether they resulted in the emergence of optimal communication systems from random behaviour when placed in the context of the ILM. In the previous section the initial population's communication system, L, was optimal. In the models outlined in this section L has maximum entropy — every $m \in \mathcal{M}$ is associated with every $s \in \mathcal{S}$ with equal probability, $|\mathcal{M}| = |\mathcal{S}| = 10$. This was achieved by setting the connection weights of every individual in the initial population to 0. Unlike in the previous section, cultural transmission is noise-free —

$p_n = 0$ (although results show that similar behaviour occurs with $p_n > 0$). Simulations were run for each of the 81 possible learning rules. A population was defined as having *constructed* an optimal system if the population's communicative accuracy reached 1.0. Weight-update rules were classified [+constructor] if optimal systems were constructed in each of ten 5000-cohort runs.

The populations exhibit three typical patterns of behaviour, of which populations (a), (b) and (c) in Figure 3.11 are representative examples. The populations which fit the pattern exemplified by (a) in Figure 3.11 have clearly failed to construct an optimal system and in fact persist at the random level of performance for $|\mathcal{M}| = |\mathcal{S}| = 10$. All of the weight-update rules which were classified as [$-$maintainer] follow this pattern and can be classified as [$-$constructor].

Populations behaving similarly to population (b) in Figure 3.11 are performing above the random level, but have not constructed an optimal system as defined above. In fact, as suggested for a more limited case by Oliphant (1999), the level of communicative accuracy in these populations hovers around the level we would expect given a random assignment of signals from $\mathcal{S}$ to meanings from $\mathcal{M}$ with replacement:

$$\text{communicative accuracy} \approx 1 - \left(1 - \frac{1}{|\mathcal{S}|}\right)^{|\mathcal{M}|}$$

The reason for this level of performance will be made clear in section 3.5.1. Nine of the 18 weight-update rules which were classified [+maintainer] fit this pattern and can be classified as [$-$constructor].

Populations fitting the pattern exemplified by population (c) in Figure 3.11 have succeeded in constructing an optimal system from random behaviour and can be classified as [+constructor]. Nine of the 18 weight-update rules which were classified as [+maintainer] fit this pattern.

3.4.6 Summary: The classification hierarchy

The three tests outlined above divide the 81 weight-update rules into four groups, summarised in Table 3.6.

The fact that all weight-update rules which are [+constructor] are [+maintainer] and all rules which are [+maintainer] are [+learner] suggests a hierarchy of weight-update rules, summarised in Figure 3.12.

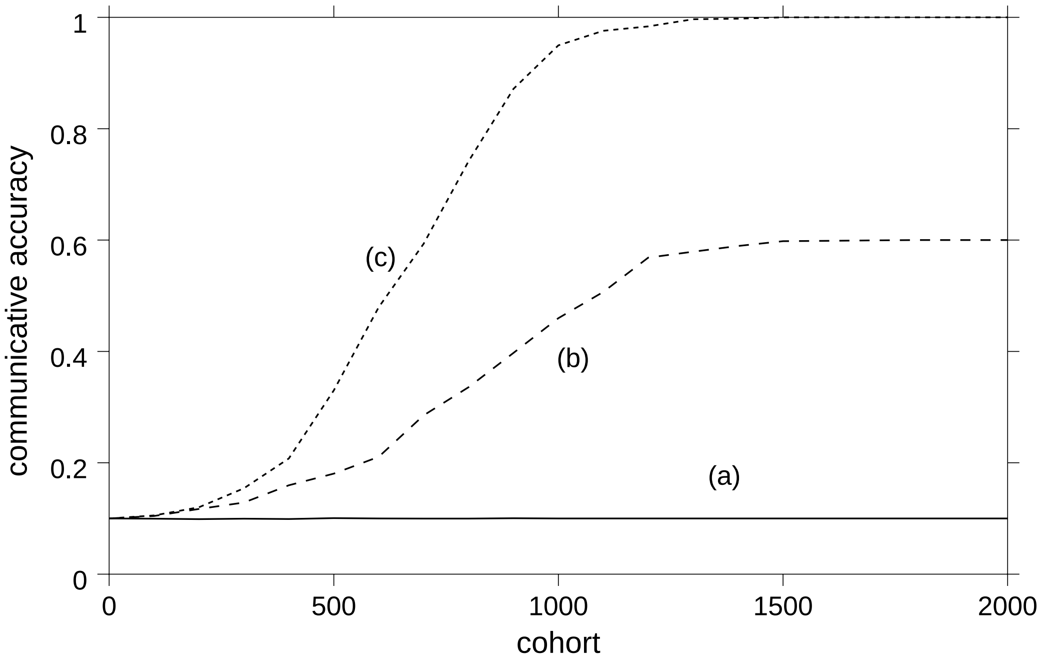


Figure 3.11: Populations of agents using the 81 learning rules exhibit three patterns of behaviour when attempting to construct an optimal system: failure to construct an optimal system and chance-level communicative accuracy, as in (a); failure to construct an optimal system, but levels of communicative accuracy significantly above chance, as in (b); construction of an optimal system, as in (c).

Classification	Number
[−learner, −maintainer, −constructor]	50
[+learner, −maintainer, −constructor]	13
[+learner, +maintainer, −constructor]	9
[+learner, +maintainer, +constructor]	9

Table 3.6: The number of weight-update rules of each particular complete classification, from the sample of 81.

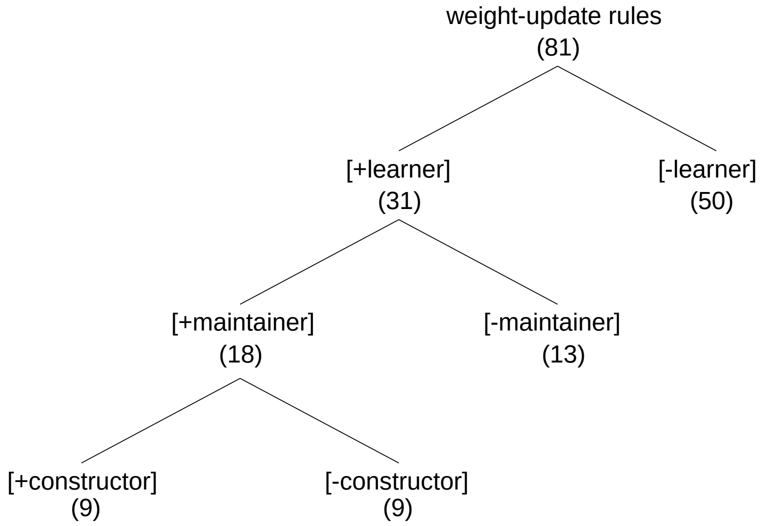


Figure 3.12: The hierarchy of weight-update rules. Read from the top, each node places additional restrictions on the properties of the weight-update rules. The numbers possessing each feature are given in parentheses at each point in the tree.

3.5 The Key Bias

Why are obverter networks in the feedforward network model described in Section 3.3 biased in favour of acquiring unambiguous systems but imitator agents, with a slightly different architecture, are biased in favour of acquiring fully ambiguous systems? Similarly, what is it about the particular assignment of -1s, 0s and 1s to the four conditions α, β, γ and δ in the associative network model[9] (described in the previous Section) that makes one weight-update rule incapable of learning an optimal communication system whereas another weight-update rule is capable of constructing such a system from random behaviour in the context of iterated cultural transmission?

The learning biases of the different network architectures or weight-update rules are best described in terms of the one-to-one nature of mappings between meanings and signals. As defined in Section 3.2, in an optimal communication system $r(p(m)) = m$ for all $m \in \mathcal{M}$. This requires that:

1. Each $m \in \mathcal{M}$ should be expressed by a distinct $s \in \mathcal{S}$, i.e. $p(m)$ should be a one-to-one function.
2. Each $s \in \mathcal{S}$ should map back to a single $m \in \mathcal{M}$ such that $p(m) = s$, i.e. $r(s)$ should be a superset of the inverse of $p(m)$.

[9] See Section 3.4.1.3. To recap, the value in α specifies how to change the connection weight between coactive units, β specifies how to change the connection weight between an active meaning node and an inactive signal node, γ specifies how to change the connection weight between an inactive meaning node and an active signal node, and δ specifies how to change the connection weight between two inactive units.

3.5.1 The key bias in the associative network model

There is a clear pattern relating the properties of weight-update rules to the assignment of actions to values in the $(\alpha \ \ \beta \ \ \gamma \ \ \delta)$ 4-tuple. Given the (approximately) bidirectional nature of the networks and assuming $|\mathcal{S}| \geq |\mathcal{M}|$, point 1 above ($p\,(m)$ should be a one-to-one function) proves to be crucial in determining which weight-update rules are [$+$constructor], which are [$+$maintainer, $-$constructor] and which are [$+$learner, $-$maintainer, $-$constructor]. Weight-update rules which are [$+$constructor] are biased in favour of a one-to-one $p(m)$, those which are [$+$maintainer, $-$constructor] are neutral with respect to the one-to-one nature of $p(m)$ and those which are [$+$learner, $-$maintainer, $-$constructor] are biased in favour of a many-to-one $p(m)$.

3.5.1.1 The [$+$constructor] bias

Is there any pattern of assignment of values to conditions in the weight-update rule specification $(\alpha \ \beta \ \gamma \ \delta)$ that characterises rules which are [$+$constructor] but not rules which are [$-$constructor]? Yes.

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

Why does this pattern of weight changes result in the construction of optimal systems from random behaviour? Consider a network where $|\mathcal{N}_M| = |\mathcal{N}_S| = 2$ using the weight-update rule $(a \ b \ c \ d)$. Prior to learning, all the connection weights in $\mathcal{W}$ are 0. If we represent $\mathcal{W}$ as a matrix with the value in row i and column j representing the weight of the connection between nodes Mi and Sj then its initial weights will be:

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

If this network is exposed once to the meaning m_1 (recall from Section 3.4.1.1 that for this meaning $a_{M1} = 1$, $a_{M2} = 0$), paired with the signal s_1 (similarly, $a_{S1} = 1$, $a_{S2} = 0$), its weight matrix will be:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

For rules which are [$+$constructor] $a > b$. This means that if our simple network uses a [$+$constructor] rule it will correctly produce s_1 to communicate m_1, due to the winner-take-all retrieval procedure.

For [+constructor] rules, $d > c$. In the context of our simple network, this means that if the network uses a constructor rule it will automatically prefer to use the signal s_2 to communicate meaning m_2, despite the fact it has only been trained to associate m_1 with s_1. This is the crucial property of [+constructor] rules — they are biased in favour of acquiring one-to-one mappings between meanings and signals. What consequences does this bias have in the context of iterated cultural transmission?

Only communication systems which conform completely to the biases of learners will be stable over iterated cultural transmission — communication systems which partially conform to learner biases will be less likely to be acquired than systems which conform more fully to the learner biases, and will therefore be filtered out of the population over time. This differential retention of communication systems resulting from learner biases results in direct bias on cultural transmission, as defined by B&R. The [+constructor] bias in favour of one-to-one mappings between meanings and signals results in many-to-one mappings being filtered out of the population. Eventually, through the process of iterated learning, the population converges on a shared one-to-one mapping between meanings and signals — an optimal communication system is constructed.

3.5.1.2 The [+maintainer] bias

Can the [+maintainer] property also be explained in terms of allocations of actions to the $(\alpha\ \beta\ \gamma\ \delta)$ weight-update rule specification? First, is there any pattern which uniquely identifies the [+maintainer, $-$constructor] rules? Yes.

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

Once again consider a network where $|\mathcal{N}_M| = |\mathcal{N}_S| = 2$ using the rule $(a\ b\ c\ d)$ exposed once to m_1 paired with s_1. As before, the resultant weight matrix is:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

As for [+constructor] rules, for [+maintainer, $-$constructor] rules $a > b$. This means that if our simple network uses a [+maintainer, $-$constructor] rule it will correctly produce s_1 to communicate m_1.

For [+maintainer, $-$constructor] rules $d = c$. This means that, unlike [+constructor] rules, the network using a [+maintainer, $-$constructor] rule will be equally likely to express m_2 using s_1 or s_2, due to their equal weights in the network. [+maintainer,

$-$constructor] rules are therefore neutral with respect to one-to-one mappings. This explains both the ability of populations of agents using such rules to maintain optimal systems in the context of the ILM and the behaviour of these populations as they attempt to construct optimal systems.

[$+$maintainer, $-$constructor] rules can maintain an optimal system in the presence of noise. The initial optimal system is, by definition, a one-to-one mapping between meanings and signals. Given the neutrality of [$+$maintainer, $-$constructor] rules to the one-to-one nature of mappings, such optimal systems can be acquired in the presence of noise, provided the noise is not sufficient to drown out the one-to-one mapping.

Recall from Section 3.4.5 and Figure 3.11 that, when provided with an initially random system, populations of agents using [$+$maintainer, $-$constructor] rules converge on the level of communicative accuracy one would expect given a random assignment, with replacement, of signals to meanings. This can be explained in terms of the neutrality of [$+$maintainer, $-$constructor] rules to the one-to-one nature of mappings. The initial population's random behaviour, when taken as a whole, will embody a random assignment of signals to meanings. This random assignment will become shared among the population through the process of cultural transmission. While [$+$constructor] agents remove the many-to-one elements of the initial random system, [$+$maintainer, $-$constructor] agents do not — the population's eventual communication system will embody the same number of many-to-one mappings as the initial random behaviour.

What then of the [$+$maintainer] property in isolation from the [$\pm$constructor] feature? This can be captured thus:

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

The fact that rules which are [$+$constructor] are always [$+$maintainer] is captured by this statement, as is the fact that it is possible to be [$+$maintainer, $-$constructor].

3.5.1.3 The [$+$learner] bias

The pattern of assignments of actions to the weight-update rule specification $(\alpha\ \beta\ \gamma\ \delta)$ that characterises rules which are [$+$learner] is:

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

or, in simple terms, in order to be able to acquire an optimal communication system you must make stronger associations between units which tend to have matching activations

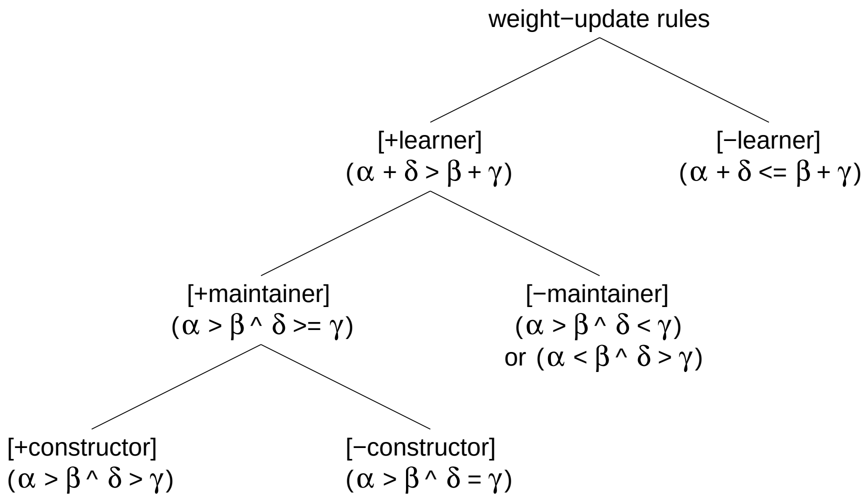


Figure 3.13: The hierarchy given in Figure 3.12, expressed in terms of restrictions on possible values in each condition of weight-update rules.

than between units which tend to have conflicting activations. Note that the $\alpha > \beta \wedge \delta \geq \gamma$ constraint on [+maintainer] rules guarantees that all such rules are also [+learner].

Why are rules which are [+learner, −maintainer, −constructor] unable to maintain or construct optimal communication systems? As we might expect, such weight-update rules are biased *against* one-to-one mappings between meanings and signals and in favour of many-to-one mappings. This immediately rules out construction of the one-to-one mappings characterising optimal systems, and also maintenance of such systems. Any many-to-one mappings introduced by noise will be preferentially acquired by [+learner, −maintainer, −constructor] agents and will spread through populations of such agents, resulting in the type of decrease in communicative accuracy seen in Figure 3.10.

3.5.1.4 Summary of the key bias in the associative network model

The weight-update rule hierarchy given in Figure 3.12 is re-presented in Figure 3.13 in terms of the constraints on the values of the weight-update rules. Each terminal node of the tree has a bias, summarised in Table 3.7.

3.5.2 The key bias in the feedforward network model

In the feedforward network model both imitator and obverter agents learn using the backpropagation algorithm. The bias is therefore introduced by the architecture of these networks, rather than the particular learning rule used. Imitator networks map from input

Classification	Bias
[$-$learner, $-$maintainer, $-$constructor]	NA
[$+$learner, $-$maintainer, $-$constructor]	favours many-to-one mappings
[$+$learner, $+$maintainer, $-$constructor]	neutral
[$+$learner, $+$maintainer, $+$constructor]	favours one-to-one mappings

Table 3.7: A summary of the learning biases of each particular combination of features. Weight-update rules which are classified as [$-$learner, $-$maintainer, $-$constructor] cannot be said to have a learning bias as they cannot learn.

meanings to output signals, whereas obverter networks map from input signals to output meanings. This turns out to be crucial in understanding the bias of these networks.

Feedforward neural networks learn many-to-one functions. Due to the deterministic nature of the feedforward propagation of activation values they cannot learn one-to-many mappings. The easiest function for a network to acquire is an all-to-one mapping from inputs to outputs, the hardest learnable function is an injective (one-to-one) function and one-to-many mappings are unlearnable. The reversal process used to model reception behaviour for imitators and production behaviour for obverters is similarly biased — it generates a function, which may be injective or many-to-one, based on the function the feedforward network has acquired. In general, if the network has acquired a function $f(x)$ which has a range y, then the reversal process ensures that element $y_i \in y$ will map onto a single element $x_i \in x$ such that $f(x_i) = y_i$ — in simple terms, the reversal process deterministically reverses the function acquired by the network.

In imitator agents the feedforward network learns functions from meanings to signals — it learns $p(m)$. Since it is a feedforward network it will be biased towards acquiring a many-to-one or all-to-one $p(m)$. As illustrated in Figure 3.14 and discussed in the caption, the maximally stable $p(m)$ for imitator agents is therefore an all-to-one fully ambiguous function. Imitators are therefore biased against one-to-one mappings from meanings to signals. Reception in imitators will be based on their acquired $p(m)$ — as shown in Figure 3.14, in the case of an all-to-one $p(m)$, in $r(s)$ the signal s_i that constitutes the range of $p(m)$ will map onto a single element from m. Therefore a population of imitators agents will tend to produce the same signal for every meaning and interpret the ambiguous signal as communicating one arbitrary selected meaning. This situation results in performance equivalent to random guessing.

In obverter agents the feedforward network learns functions from signals to meanings — it learns $r(s)$. As illustrated in Figures 3.15 and 3.16 the only culturally stable system has a one-to-one $p(m)$ and an $r(s)$ which includes at least the inverse of $p(m)$. Obverter

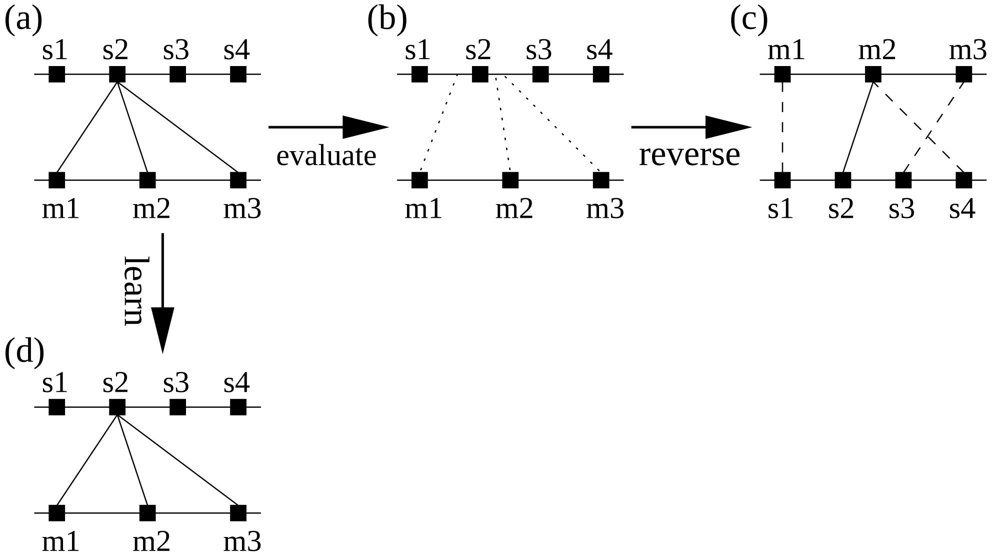


Figure 3.14: (a) is a representation of an imitator agent's feedforward network encoding an all-to-one $p(m)$ mapping three meanings onto a single signal, $s2$. The function from a domain of real numbers (input unit activations) to a codomain of real numbers (output unit activations) is represented by two lines, the lower line representing the domain, the upper representing the codomain. Squares represent particular points on the line corresponding to binary meanings or signals. Associations are shown with solid lines between elements in the domain and elements in the codomain. (b) represents the confidence-measuring step of the reversal process for the network underlying (a). In order to decide $r(s2)$, the real-number values of $p(m1)$, $p(m2)$ and $p(m3)$ are calculated. These real-numbered mappings are represented by dotted lines in (b). (c) represents the $r(s)$ derived from applying the reversal process to (a). $r(s2) = m2$ because $m2$ mapped closer to $s2$ than any other m in (b). The other associations are effectively random. The random nature of these mappings is represented by dashed lines. (d) represents the function acquired by an imitator network exposed to behaviour generated by (a) — as it is an all-to-one function between meanings and signals it is easily learned by imitator agents. This is in fact the only stable function for imitators.

agents are therefore strongly biased in favour of acquiring systems with the properties of optimal communication systems.

How can we relate these feedforward network biases to the classification hierarchy developed for the associative network weight-update rules? Obverter networks, biased in favour of one-to-one mappings between meanings and signals, should clearly be classified as [+constructor]. The classification of imitator agents is less clear. Imitator agents are capable, given sufficiently high e, of acquiring an optimal, unambiguous communication systems, and should therefore be classified as at least [+learner]. Populations of such agents cannot construct an optimal system, and should therefore be classified [$-$constructor]. Their status with respect to the [$\pm$maintainer] feature is less clear. We would expect, given their bias in favour of many-to-one functions, that they should be classified as [$-$maintainer]. However, simulation runs were carried out to measure the

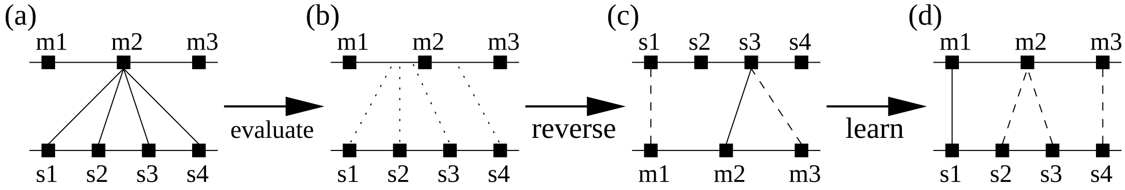


Figure 3.15: (a) represents an all-to-one $r(s)$ encoded in an obverter agent's feedforward network. As obverters map from signals to meanings this is the most learnable $r(s)$. (b) represents the confidence-measuring step of reversing this $r(s)$ to generate a $p(m)$ — as before, real-number mappings are shown as dotted lines. (c) shows the $p(m)$ derived from (a). $p(m2) = s3$ as $s3$ mapped closest to $m2$ in (b). The other associations are essentially random. The $p(m)$ in (c) produces the meaning-signal pairs $\{(m1, s1), (m2, s3), (m3, s3), \}$. Meanings and signals in these pairs are transposed (yielding $\{(s1, m1), (s3, m2), (s3, m3), \}$) to train the next generation of obverter networks. (d) shows the $r(s)$ resulting from training an obverter network on the signal-meaning pairs $\{(s1, m1), (s3, m2), (s3, m3), \}$. $r(s1) = m1$, as expected. However, feedforward networks cannot learn one-to-many mappings so $r(s3)$ is effectively randomly assigned to a signal, in this case $m2$. As $s2$ and $s4$ are not represented in the training set they are effectively randomly assigned mappings. Notice that the mapping in (a) has been destroyed in (d) — the many-to-one mapping in (a) is not culturally stable.

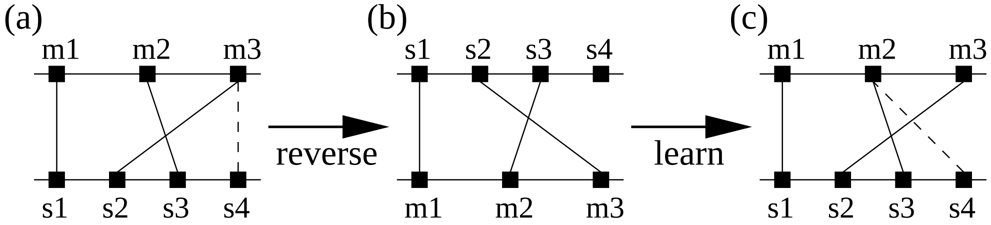


Figure 3.16: Only an unambiguous $p(m)$ is stable for obverter agents. (a) represents an obverter agent's $r(s)$. (b) is the $p(m)$ derived from reversal of (a) — it is a one-to-one function. (c) illustrates that the $r(s)$ resulting from training the next generation of agents on data produced by (b) is effectively similar to (a) and will therefore lead to (b) once again — (a) and (b) are culturally stable. The only unstable aspect is the floating synonym $s4$. This synonym is highly unlikely to interfere with the mapping in (b) and the floating synonym phenomenon can be observed in other obverter models.

ability of populations of such agents to maintain an optimal system in the presence of noise ($e = 200$, $p_n = 0.05$) and no runs were found for which they failed to do so. However, this appears to be due to the large value of e, which reduces the impact of noise. As we will see in Chapter 4, injection of a slightly different form of noise does result in failure to maintain an optimal system. We will therefore classify imitator networks as [$+$learner,$-$maintainer,$-$constructor].

3.5.3 The key bias in other models

Can we understand the behaviour of other models of the cultural evolution of vocabulary in terms of this key bias? Specifically, in the models where cultural evolution is driven by direct bias, does the direct bias result from a learning bias similar to that of [$+$constructor] agents? Dealing with other neural network models first, does this key bias appear in the neural network models of Hutchins & Hazelhurst (1995), Batali (1998) and Livingstone & Fyfe (1999) (discussed in Section 3.1) and Kvasnička & Pospíchal (1999) (which will be discussed in Chapter 4), and Hare & Elman (1995) and Kirby & Hurford (2002) (discussed in Chapter 5)?

Hutchins & Hazelhurst's (1995) model can be treated separately from the other models, which all share a common model of a learner. Hutchins and Hazelhurst use autoassociator networks to model communicative agents, with patterns of activation over the hidden layer being interpreted as signals. Autoassociator networks must develop a distinct pattern of activation over the hidden layer for every input-output pair (input-output pairs are equivalent to meanings as defined here) in order to succeed in the autoassociator task. Interpreting the hidden-layer patterns of activation as signals therefore builds in a one-to-one bias of the type identified as crucial for developing an optimal communication system.

Batali (1998), Kvasnička & Pospíchal (1999), Livingstone & Fyfe (1999) and Kirby & Hurford (2002) all use the obverter feedforward network configuration, with networks mapping from input signals to output meanings. As discussed in Section 3.5.2, such a configuration results in a learning bias in favour of one-to-one meaning-signal mappings. The obverter network configuration, which is quite common in the literature, therefore builds in a strong bias in favour of optimal communication systems.

Hare & Elman's (1995) model of morphological change deserves a brief mention here. This network maps from semantic representations of verbs to representations of the phonological realisation of those verbs, and Hare & Elman observe a simplification of the phonological system, with increasing numbers of verbs being expressed with similar affixes. While this pattern exhibits some complex interactions between phonological regularity and frequency of tokens, the general pattern of convergence to many-to-one mappings is what we should expect to see from an imitator network architecture. We can speculate that, had Hare & Elman allowed their simulations to continue for several hundred generations, all verbs would end up being expressed with a single phonological form. The learning bias of the imitator network would essentially destroy the morphological system. Contrast this with Batali's (1998) results, where the obverter network architecture results in the emergence of a morphological system.

Non-neural network models of the evolution of vocabulary, where that evolution is driven by direct bias, are actually rather scarce, the only clear examples being the models of Hurford (1989) and Oliphant & Batali (1997). As discussed in Section 3.1, Hurford considers three learning strategies — Calculators, Imitators and Saussureans. Populations of individuals using the first of these strategies cannot maintain optimal systems over time, even when there is no noise on cultural transmission, while Imitator and Saussurean populations can construct communication systems which yield intermediate levels of communicative accuracy, with Saussureans being somewhat more successful than Imitators in this respect.

Based on these results, we would expect that Calculators can be classified as [$\pm$learner, $-$maintainer, $-$constructor] and Imitators and Saussureans can be classified as [$+$learner, $+$maintainer, $-$constructor]. Saussureans have the additional advantage over Imitators that their production and reception behaviour are necessarily closely coupled — while Imitator learners acquire their production and reception matrixes completely independently from one another, Saussureans construct their reception matrix on the basis of their own production matrix.

Do these learning strategies have the biases we would expect with respect to the one-to-one nature of the meaning-signal mapping? In other words, are Imitators and Saussureans neutral with respect to the one-to-one nature of mappings, and how are Calculators biased with respect to this property?

It should be fairly obvious that Imitators and Saussureans *are* neutral with respect to many-to-one mappings from meanings to signals — they acquire their production matrix straightforwardly, by memorising the meaning-signal pairs they observe being produced. The story with Calculators is somewhat more complicated. Consider a Calculator trying to learn a system involving 2 meanings and 3 signals. Assume that the learner observes a system where $m1$ is communicated using $s1$ and $m2$ is communicated using $s2$, with

$s3$ being unused. This is a one-to-one system. On the basis of this observed production behaviour the Calculator will arrive at the reception matrix

R	m1	m2
s1	1	0
s2	0	1
s3	0.5	0.5

$s3$ is interpreted as meaning $m1$ and $m2$ with equal probability, given that these signals are unused in the observed production system. This individual's reception behaviour will now be used by the next generation of learners to form their production behaviour. What happens?

The Calculator above will produce two possible sets of reception behaviour. In both, $s1$ is interpreted as $m1$, $s2$ interpreted as $m2$, with $s3$ being interpreted as $m1$ or $m2$ at random. What consequence does this have for the production matrices of the next generation? If the learner at the next generation observes $s3$ being interpreted as $m1$ they will arrive at the production matrix

P	s1	s2	s3
m1	0.5	0	0.5
m2	0	1	0

In other words, they will now produce either $s1$ or $s3$ to communicate $m1$. If the learner instead observes $s3$ being interpreted as $m2$ they will arrive at the production matrix where $s2$ and $s3$ are produced with equal probability for $m2$. In other words, the spare signal leads to the creation of one-to-many mappings between meanings and signals. The inability of Calculators to acquire an optimal system (albeit over the course of two learning episodes) therefore indicates they should be classified as [$-$learner]. This classification explains the fact that populations of such individuals immediately lose an initially perfect system in an ILM where there is no noise on cultural transmission.

As discussed in Section 3.1, Oliphant & Batali contrast two learning strategies (Imitate-Choose, henceforth imitator, and obverter). Populations of obverter learners can construct an optimal system, whereas populations of imitators cannot. Oliphant & Batali attribute this to the fact that obverter agents base their production behaviour on the population's reception behaviour, therefore explicitly designing their communication systems so as to be understood, whereas imitator agents do not. However, the [+constructor] agents described in Section 3.4 base their production behaviour on production behaviour, yet still arrive at an optimal communication system. Do the obverter agents described by

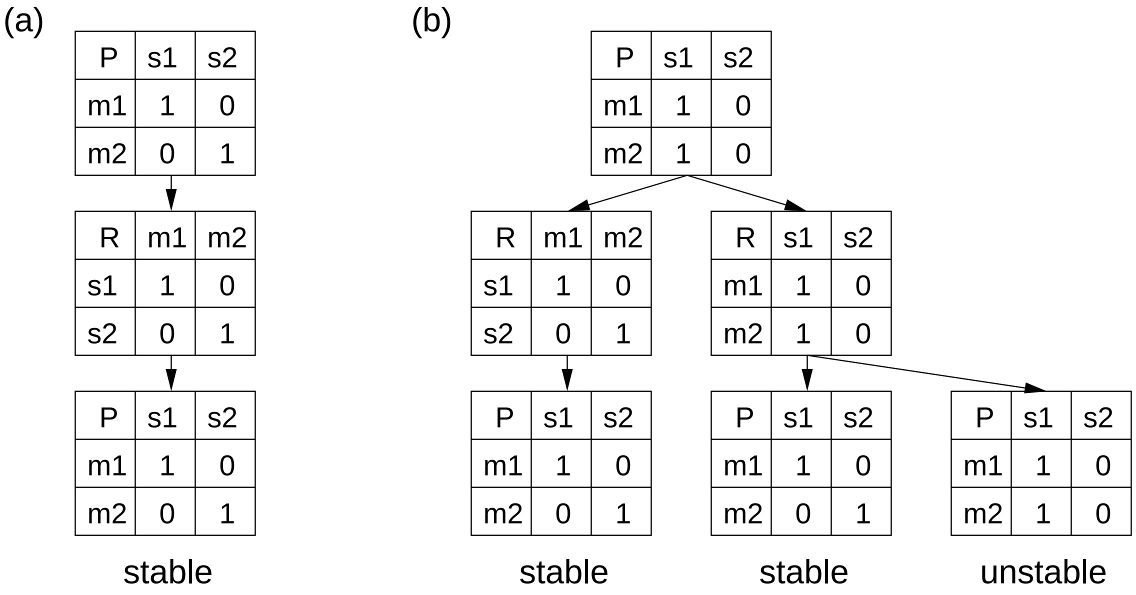
Oliphant & Batali arrive at an optimal system in a different way, or can the behaviour of populations of such individuals be better explained in terms of a learning bias in favour of one-to-one mappings between meanings and signals?

Figure 3.17 analyses how a population's production (P) and reception (R) functions change over time, for two initial starting conditions — an initial one-to-one mapping, and an initial many-to-one mapping. Both these initial mappings are learnable by imitator agents. For obverter agents, the one-to-one mapping remains stable over time — the one-to-one P matrix leads to a one-to-one R matrix, which in turn leads back to the one-to-one P matrix. In contrast, the many-to-one P matrix is unstable. The obverter procedure attempts to derive an R matrix from this P matrix by finding the meaning m_x for which s_1 and s_2 is at a maximum. Since m_1 and m_2 are equally likely in both these contexts, a random choice is made, which leads to four possible R matrixes, all equally probable. For the sake of convenience, only two are shown in Figure 3.17. One of these is a one-to-one R matrix, which leads to a one-to-one P matrix, which, as we have seen, is stable. The other is a many-to-one R matrix, which leads, again through random selection, to 4 possible P matrixes, two of which are shown. Only the one-to-one P matrix is stable — the other, as we have seen, is unstable. The learning bias of the obverter agents favours one-to-one mappings between meanings and signals.

3.6 Biases in vocabulary acquisition in humans and non-humans

Is there any evidence that language acquisition in humans is guided by biases in favour of one-to-one mappings between meanings and signals? If so, then the result of the computational models shown here would suggest that optimal, or at least communicatively useful, communication systems could arise in human populations through purely cultural processes.

The one-to-one bias described here is typically broken down into two subcomponents when talking about vocabulary acquisition in humans. The one-to-one bias consists of both a bias against homonymy (many-to-one mappings from meanings to signals) and a bias against synonymy (one-to-many mappings from meanings to signals). In Sections 3.6.1 and 3.6.2 I will present evidence from the language acquisition literature that suggests that children apply both these biases to the learning of vocabulary. I will start with the proposed bias against synonymy, as this is perhaps slightly less controversial, then move on to homonymy. Finally, in Section 3.6.3 I will briefly review some evidence from ape 'language' learning experiments which suggest that apes may not possess similar learning biases to human infants. The postulated uniqueness of these learning biases



(a)

P	s1	s2
m1	1	0
m2	0	1

R	m1	m2
s1	1	0
s2	0	1

P	s1	s2
m1	1	0
m2	0	1

stable

(b)

P	s1	s2
m1	1	0
m2	1	0

R	m1	m2
s1	1	0
s2	0	1

R	s1	s2
m1	1	0
m2	1	0

P	s1	s2
m1	1	0
m2	0	1

stable

P	s1	s2
m1	1	0
m2	0	1

stable

P	s1	s2
m1	1	0
m2	1	0

unstable

Figure 3.17: The learning bias of Oliphant & Batali's obverter learner. As illustrated in (a), one-to-one mappings are stable over time — a one-to-one production matrix (marked by a P) leads to a one-to-one reception matrix (R), which leads back to a one-to-one production matrix. Many-to-one mappings are unstable, as illustrated in (b). The many-to-one production matrix leads to four possible reception matrixes (only two are shown here). The one-to-one reception matrix leads to a one-to-one production matrix which, as shown in (a), is stable. The many-to-one reception matrix leads either to a one-to-one production matrix, which is stable, or a many-to-one production matrix, which as we have seen is unstable. In other words, only one-to-one matrixes are stable over time in populations of such learners.

to humans forms the motivation for Chapter 4, where investigations into the biological evolution of the one-to-one learning bias are discussed. I will return to the role of one-to-one biases in the acquisition of linguistic structure (both syntactic and morphological, as opposed to essentially unstructured vocabulary which I discuss here) in Chapter 5.

3.6.1 Biases against synonymy in humans

Eve Clark and Ellen Markman have proposed that children have word-learning biases which make synonyms (mappings from one meaning to several distinct words) either unlearnable or difficult to learn. Both authors claim that these biases help children in the rapid acquisition of vocabulary — the acquisition process in children has been termed "Fast Mapping" (see Bloom (2000), Chapter 2, for review). While their conception of this bias is in fact rather different, both base their theories on a set of experimental studies carried out by Kagan (1981), replicated by, among others, Markman & Wachtel (1988).

In Kagan's original study, children were shown three objects, two of which were familiar (a doll and a dog) and one of which was unfamiliar (for example, a lemon zester). Children were allowed to play with the objects, and were then asked to "Give me the zoob" (or some other nonsense word) by the experimenter. The children showed a strong preference for giving the novel item.

In Markman & Wachtel's (1988) replication, children were shown a single familiar object (for example, a plate) and an unfamiliar object (e.g. a radish rosette maker) and asked by a puppet frog to "Show me the fendle" or some other nonsense word. Children reliably respond by giving or showing the unfamiliar object. Results from a control group study, where children were asked simply to "Show me one", indicated that this preference was not due to a preference on the part of children to respond with the unfamiliar object — children only exhibit such a preference when prompted with a novel word.

Clark (Clark 1988; Clark 1990; Clark 1993) proposes two pragmatic principles guiding vocabulary acquisition. The first, which she terms the Principle of Contrast (henceforth Contrast), states that "different words have different meanings". The second, the Principle of Conventionality (henceforth Conventionality), states that "for certain meanings, there is a form that speakers expect to be used in the language community".

How does this relate to the child's behaviour in the experiment outlined above? The child knows, through Contrast, that contrasting words contrast in meaning. The child knows, through Conventionality, that established words have priority. It is assumed that the child already knows the established word for the familiar object. The child then deduces, via Conventionality, that if the experimenter wished to refer to the familiar object they would use the conventional word for that object. However, notes the child, the experimenter used a novel word. The child reasons that that word cannot refer to the familiar object, because different words have different meanings. The child therefore concludes that the new word must refer to the unfamiliar object, responds appropriately by giving the experimenter the unfamiliar object, and learns the (nonsense) name for the unfamiliar object.

Contrast rigidly rules out synonyms — Clark emphasises that any difference in the form of a word indicates a difference in meaning. To put it another way, according to Clark synonyms do not exist. Clark is keen to point out that it does not rule out homonymy (many-to-one mappings from meanings to signals), a point which I will return to below.

Markman (Markman & Wachtel 1988; Markman 1989; Markman 1992) proposes a Mutual Exclusivity (ME) bias in children — "children should be biased to assume, especially at first, that terms [words] are mutually exclusive" (Markman 1989:188) and "each object will have only one label". Note that, unlike Contrast, this is not an inviolable principle, but a tendency or bias that can be overridden given sufficient evidence. Like Contrast, ME discourages synonymy — each object ideally has only one label, therefore there should be no one-to-many mappings between meanings and signals. Markman is less clear on the status of ME with respect to homonymy (for example, the term does not appear in the index to her 1989 book), although she makes frequent reference to "one-to-one" biases in vocabulary acquisition. Like Contrast, ME enables the child in the experiment above to deduce that the novel word should refer to the unfamiliar object and learn that labelling — ME dictates that the novel word cannot refer to the object which is already labelled, therefore the novel word must refer to the unfamiliar object.

The main difference between Clark's position and Markman's is in the severity of the bias. Markman (1989) presents what she considers to be evidence that children *can* learn to violate ME. Specifically, they can learn super-ordinate terms ("poodle" but also "dog" and "animal"). However, children find such super-ordinate terms difficult to learn, with children as old as 14 making errors. Typically, the error involves mistaking a super-ordinate term for a term expressing a collection of the subordinate items. Macnamara (1972) gives the example of a child who will accept that a particular plaything is a "truck" or a "train", but will simultaneously deny that it is a "toy", with the term "toy" being reserved for a group of trucks, trains, Teddy bears and so on. This type of evidence leads Markman to conclude that ME is violable. In contrast, Clark would say that super-ordinate terms have a different meaning from their subordinate terms, therefore Contrast is preserved, although this offers no explanation as to why children find super-ordinates hard to learn. However, regardless of the debate between Contrast and ME, the experimental evidence suggests that children have a bias against acquiring synonyms, one-to-many mappings from meanings (objects in the experiment) to signals.

3.6.2 Biases against homonymy in humans

The status of biases with respect to homonymy is somewhat problematic. One does not have to look very far through any language to find homonyms, mappings from several meanings to a single surface form. At first blush, this seems to indicate that children do not have any bias against acquiring many-to-one mappings from meanings to signals. Indeed, Clark is keen to point out that Contrast and Conventionality do not in any way bias against homonymy. Markman remains silent on the subject of homonymy, although in several places she refers to "one-to-one" biases in vocabulary acquisition.

Briefly considering a possible experiment should, however, serve to cast doubt on the first intuition that children are unbiased with respect to the acquisition of homonyms. Imagine a slight variation in the experiment outlined above, where, rather than the experimenter asking the child to “show me the zoob”, the experimenter asks the child “Show me the shoe” or whatever the familiar object was. I suspect that the child would respond by showing the shoe, at least with the same level of reliability as children prompted with the novel word would respond by showing the unfamiliar object. The fact that nobody, to my knowledge, has carried out this experiment indicates that this is probably not a very controversial hypothesis — our everyday experience indicates that people know the names for things and if you request those things then they don’t assume you are talking about something else.

However, this experiment, were it to proceed as I expect, would illustrate that children must be biased against homonymy. If children are unbiased with respect to homonymy then under such experimental circumstances they should show the shoe or the unfamiliar object with equal probability — if many-to-one mappings between meanings and signals are as possible as one-to-one mappings, then the child cannot tell whether “shoe” means shoe$_1$, the familiar sense of shoe, or shoe$_2$, a new use of homophonous “shoe” to refer to the unfamiliar object. It could be argued that the child would prefer the familiar sense of “shoe” due to the fact that they have frequently encountered this use of “shoe” (this is essentially Clark’s Conventionality principle). However, if we agree that children prefer not to learn new meanings for established words then we are accepting a bias against homonymy. It could be argued that this bias against homonymy only comes into play once a well-established convention is in place — for example, once the child has experienced several hundred utterances of “shoe”, all with the same reference, they will be resistant to learning a new meaning for “shoe”. However, it cannot simply be sheer weight of numbers which performs this function. Bloom & Markson (1998) describe an experiment where children are presented with two novel objects, given a single nonsense name for one object (“bem”, for example), then asked to “Show me the jop”. Under these circumstances children still reliably show the unnamed novel object. Given that neither object has been encountered before, and the word “bem” is also novel, it seems that a single exposure to a word paired with an object biases children against interpreting or acquiring that word as conveying a different meaning.

It in fact seems that, with a learner unbiased with respect to homonymy, word learning would become all but impossible — every possible utterance of a familiar word could refer to any object at all. If we accept Contrast or ME, then the problem is fractionally reduced — any word can refer to any object which has not already been labelled with a different word. However, Contrast/ME cannot really resolve the issue as learning even a single label becomes an intractable task. This problem becomes worse when we consider that children can learn labels for subparts of objects. Joint attention might narrow down the focus of possible objects, ruling out other whole objects as the referent of "shoe", but when I say "Show me the shoe", do I mean shoe$_1$, the whole object, or shoe$_3$, the string that ties the shoes up, or shoe$_4$, the man-made fibre which the shoe uppers are made of?

I believe that we are forced to conclude, on logical grounds, that children must have some bias against acquiring homonyms, many-to-one mappings from meanings to words. Without such a bias, word learning would become impossible or at best extremely arduous. This is in fact not a terribly new position, although the argument given above may be a novel one. McMahon (1994) briefly discusses work carried out by Jules Gilliéron in France from 1896. Gilliéron, with the aid of several fieldworkers, compiled the Linguistic Atlas of France. One of Gilliéron's primary concerns was to construct "phonetic etymologies", by comparing the expected forms of modern French words (based on a set of hypothesised changes occurring between Latin and modern French) to the words actually attested. Where the predictions and the data did not match up Gilliéron attempted to explain the discrepancy.

Gilliéron's theory predicted that the modern French word for cockerel should be derived from the Latin "gallus". In the Gascony region these hypothesised changes should have lead to the form "gat". However, this is also the predicted form for the word for cat, "cattus" in Latin. Gilliéron's survey revealed that "gat" in Gascony does in fact refer to cat, with cockerel being referred to by another word. Gilliéron appealed to an avoidance of homonymy to explain this mismatch — cats and cockerels are both farmyard animals (or presumably were in turn-of-the-century France), and homophony involving meanings from the same semantic field is avoided, therefore "gat" was restricted to meaning cat and an alternative form was employed for cockerel.

This early example does not indicate where the bias against homonymy resides — is it in the language learner or the language user? Martinet (1972) proposes the second of these alternatives. Martinet's primary concern is an account of phonemic change, but he works within a functionalist framework: "The basic assumption of functionalists in such matters is that sound shifts do not proceed irrespective of communicative needs, and that one of the factors which may determine their direction and even their appearance is the basic necessity of securing mutual understanding" (Martinet 1972:144). The imperative to preserve mutual understanding should discourage, among other things, homonymy — homonymy leads to ambiguity.

Lass (1980) provides a concrete example of an irregular phonemic change which appears to result in the avoidance of homonymy. During the change from Old English to modern English, the Old English vowel /y/ appears to change in two distinct ways. In the first, regular path, Old English /y/ changes to modern English /ɪ/, via Middle English /i/. However, in some lexical items Old English /y/ appears to change to Middle English /u/ and then to modern English /ʌ/. The vowel of modern English "shut" proceeded according to the less frequent, irregular path. Had it proceeded according to the more regular path, the result would have been "shit", as Lass puts it "[t]his particular homophony would be, I would think, about as 'pernicious' as any" (Lass 1980:76). Those subscribing to Martinet's line of reasoning would perhaps argue that "shut" avoided the more regular change in order to avoid homonymy, which would lead to a decrease in communicative function, although Lass argues strongly against this interpretation, as we will see below.

Turning briefly to cross-categorial homonymy, Macnamara (1982) reports two pieces of evidence, based on the acquisition behaviour of two young subjects, that children prefer not to acquire homophonous terms which refer both to an action and an object. Macnamara's first subject preferred to use ambiguous terms such as "comb" to refer to either the action or object, but not both, even when the child's parents used the term as both noun and verb in the child's presence. Macnamara's second subject, his son, went so far as to invent a new word to avoid this type of homonymy.

To summarise, the logical argument and proposed experiment outlined at the beginning of this Section suggest that children must be biased to some extent against acquiring many-to-one mappings from meanings to signals. This is supported by some concrete examples of where this type of homonymy avoidance might be observed empirically. The empirical evidence at this point is somewhat weak, and will be considerably strengthened in Chapter 5, where more evidence for a bias against homonymy is presented. However, for now it is time to return to the two main arguments against a bias with respect to homonymy — the fact that languages contain numerous homonymous mappings, and that teleological mechanisms do not exist for homonymy avoidance.

Firstly, if children do need to be biased against homonymy, why is homonymy so frequent in language? Doesn't this prove that children are in fact not biased against homonymy? There are two possible responses to this position, both similar to Clark and Markman's respective defences of their proposed biases relating to synonymy.

We could imitate Clark's rather rigid line of argument, and insist that, just as difference in signal reflects difference in meaning, congruence of signal represents similarity of meaning. This line, in a rather strong form, is pursued by Haiman (1980). While this argument can probably be used to deflect some cases of homonymy, such as polysemous uses of “mouth” (mouth of an animal, mouth of a cave, mouth of a river etc), it perhaps does not do to push it too far.

Alternatively, we could appeal to the kind of explanation that Markman makes to explain violations of ME — perhaps homonyms are simply somewhat harder to learn than non-homonyms, but still learnable. This testable hypothesis therefore allows that we should indeed expect to see homonymy in language, particularly when we consider that a learner bias against homonymy is not the only pressure acting on language and language acquisition — not only do phonological shifts and borrowing continually bring homonymy into a language, but there are other possible pressures:

> “[These other pressures might] pertain to the number of fixed expressions, patterns, and locutions that a speaker must master, remember, and manipulate in language use. The impracticality of having a separate lexical item for every conceivable object, event, or situation a speaker is likely to encounter is of course a truism. Languages never provide a lexical inventory that is vast enough to label with uniqueness and precision the elements of every conceivable contingency; rather they depend on the speaker to use creatively a more restricted inventory of lexical units in conjunction with the resources of the grammatical system.” (Langacker 1977:114)

It is interesting that this tendency to minimise the number of lexical items which have to be learned impacts differently on synonyms and homonyms. Synonyms will be disfavoured — memorising two words for a single object increases (perhaps unnecessarily) the learning burden. The preference for smaller vocabularies then reinforces the child’s bias against synonyms. However, a pressure to minimise vocabulary size *favours* homonyms — if a single word can be used for two meanings, then the total number of words which must be learned is reduced. The tendency to minimise vocabulary size therefore fights against the postulated learner bias against homonyms. This is one possible explanation for the apparently contradictory fact that there are few (if any) true synonyms in language and numerous homonyms, while children are biased against both synonyms and homonyms.

The second objection to proposed anti-homonymy biases is the suggested location of such biases. Lass (1980) deals rather forcefully with functionalist explanations. Perhaps his most telling criticism is that, according to the functionalists, “it seems that speakers avoid homophones by prolepsis, i.e. by taking avoiding action in advance . . . [b]ut this is surely absurd . . . the only mechanism left is for speakers actually to produce the offending articles, and then, having discovered what they've done, to remove them ('My God, I've just said "please *shit* the door"; better change it to *shut*')" (Lass 1980:79). If we accept this line of argument, this leaves us with the homonymy avoidance residing in the language learner.

Croft (2000) deals with teleological explanations for homonymy avoidance:

> "perhaps the greatest objection is that there is no plausible theory motivating a teleological mechanism [whereby language users change the linguistic system for the sake of the linguistic system] . . . innovations must be brought about ultimately as a result of actions by speakers. Yet there is no obvious motivation for speakers to innovate to make the grammatical system more symmetrical, or to preserve distinctions for the sake of preserving distinctions. Speakers have many goals when they use language, but changing the linguistic system is not one of them" (Croft 2000:70)

I could not agree more. Locating a bias against homonymy in the language learner avoids the distasteful aspects of teleological and functional explanations — children learning language avoid homonyms not because they're worried that they'll say "Please shit the door mother", but simply because they can't help it — the bias against homonymy is a component of the innate device which determines the way children acquire language. Of course functionalist explanations of the origin of this language learning bias still have to be explained, which will be the role of Chapter 4.

3.6.3 Biases in non-human animals?

Clark and Markman present evidence that children are biased against acquiring one-to-many mappings between meanings and signals. I have also presented an argument that children must be biased against acquiring many-to-one mappings between meanings and signals. We should therefore, if we accept both these factors, expect children to be biased in favour of one-to-one mappings between meanings and signals. This is precisely the bias that we found to be necessary for the cultural evolution of functional communication systems. This provides partial support for our theory that only humans have the mental capacity to support learned symbolic vocabulary — we have established that humans have the necessary bias. The second supporting strut for this argument would be to fill in the “only” piece — to show that no non-human animals have a learning bias which favours one-to-one mappings between meanings and signals.

Do non-human animals have a bias in favour of one-to-one mappings between meanings and signals? Some evidence on this front comes from the ape language-learning experiments, but this evidence is somewhat sketchy. This is largely due to the fact that Kanzi, possibly the most prominent and successful ape language learning subject, essentially learned what he learned without the researchers noticing — while the focus had been on teaching Kanzi’s mother to communicate “Kanzi had been keeping a secret. He had been learning these words all along . . . We thought he did not know how to talk with the keyboard, but he did.” (Savage-Rumbaugh *et al.* 1998:22). While this set of circumstances may shed some light on whether or not apes need explicit reinforcement to learn a communication system, it is rather disappointing from our current perspective — nobody noticed how Kanzi went about learning lexical items.

There is some empirical evidence on this point, however. David Premack reports (in the discussion section following Premack (1983)) on an experiment where chimpanzees requested a previously unnamed object using a “new word”, a newly introduced piece of plastic. This suggests a bias against homonymy — if the apes were unbiased with respect to homonymy, they could happily refer to the unnamed object using a known word, therefore introducing a many-to-one mapping between meanings and signals.

Other experimental work casts doubt on this conclusion, however. Lyn & Savage-Rumbaugh (2000) describe a fairly rigorous set of experiments into the ability of two pygmy chimpanzees (Kanzi and his younger half-sister Panbanisha) to learn new words for novel items. Their overall experimental setup is inspired by the experimental setup used to investigate the Contrast/ME principle in human infants.

Lyn & Savage-Rumbaugh tested the ability of the two apes to acquire new words (lexigrams) for ten novel items. Before an ape was tested on an item, it received pre-exposure to that item being named by two human experimenters in a naturalistic dialogue. These dialogues took place outside the ape’s living enclosure, but in clear view of the ape. During each presentation session, the two experimenters played with and discussed the item, naming it between six and 19 times. Presentation sessions were grouped in threes, with the experimenters discussing one novel item, going away, returning with another novel item, and so on.

Within one hour of these groups of three presentation sessions the apes were tested on their ability to name the novel items. During each test session the ape was presented with five familiar items and the three novel items it had just seen named. The apes were allowed to play with all the items until their initial curiosity waned. 11 tests were then conducted, during which the ape was asked to give a particular item to the experimenter.

During the first three such tests the apes were instructed to hand over a familiar item. Both apes responded correctly in at least two thirds of these initial tests. For the remaining eight tests the apes were told to give the experimenter each of the eight items, in random order. If the ape made a mistake during any one of these eight tests, the request was repeated a single time. During testing, food and praise were given freely to the apes, but "[i]ndication of the incorrectness of a response was kept to a minimum" (Lyn & Savage-Rumbaugh 2000:261), notwithstanding the repetition of the request.

If an ape failed to correctly name all three novel items the presentation and test sessions were repeated within 48 hours. The repeat presentation sessions proceeded exactly as before, with all three novel items being named. However, during the test sessions, the apes were only prompted to give the novel items which they had earlier failed on.

Lyn & Savage-Rumbaugh report that Kanzi required a mean of 2 presentation/test sessions before he correctly responded to the name of a novel item, which amounted to a mean 22.8 exposures to that novel item being named by experimenters. Panbanisha fared rather less well, requiring a mean of 4.1 presentation/test sequences and 42.5 exposures per item. Furthermore, on 70% of the test incidents when the apes were asked to give a novel item they failed to do so, either choosing a familiar item, choosing more than one item or refusing to answer. As Lyn & Savage-Rumbaugh point out, this is a below-chance level of performance, indicating a preference by the apes to select familiar items when asked for a novel item. Finally, while Kanzi's level of performance remained fairly constant over the series of experiments, Panbanisha's performance gradually improved — it took her a greater number of cycles to learn the name for the first novel item than it did for the last novel item.

How should we interpret these results with respect to homonymy and synonymy biases in apes? Firstly, it should be noted that apes clearly have a rather more difficult time with the task than children, requiring pre-exposure and multiple tests to select 'novel' items when presented with 'novel' words. Secondly, the apes' performance at some aspects of the task indicate a lack of a bias against homonymy, or at least a very weak bias. The apes sometimes failed when asked to give a familiar item during the first three trials of each test session, although not more than 33% of the time. This suggests that they are not strongly biased against associating a familiar word with a novel object or the wrong familiar object (a many-to-one mapping from objects to words). Their performance on requests for familiar items during the remaining eight tests is not reported. Thirdly, the apes' preference for selecting familiar, already-named items when prompted with a novel word indicates the absence of a bias against synonymy — the apes fail, or at least take time to pass, the Contrast/ME test. Finally, Panbanisha apparently has to learn how to perform the task, while it comes naturally to human infants. Whereas the experimenters interpret this as Panbanisha coming to terms with the test environment (Heidi Lyn, personal communication), it could be taken to indicate that she takes time to come to terms with the idea of naming novel items with novel words. What is clear from this set of experiments is that apes find the acquisition of words much more difficult than humans, and the process, which is known as Fast Mapping in human infants, takes time. The experiments also suggest that apes are either unbiased with respect to homonymy and synonymy, or at least much less biased than human infants, although this conclusion is more a matter of interpretation.

To further muddy the waters, it could be noted that, in a detailed report on a separate set of ape language experiments, Savage-Rumbaugh *et al.* (1986) report that Mulika (Kanzi's younger sister) "began by using the lexigram *milk* for many different things, including requests to be picked up, requests for attention, requests to travel to different places, requests for food and requests for milk" (Savage-Rumbaugh *et al.* 1986:219). Matata, Kanzi and Mulika's mother, "did not develop an adequate concept of one-to-one correspondence between a given symbol and a given referent" (Savage-Rumbaugh *et al.* 1986:215). These (admittedly circumstantial) examples of a lack of any obvious one-to-one bias in ape vocabulary acquisition should throw further doubt on whether the biases humans bring to this task are present in a closely related species.

3.7 Summary of the Chapter

In this Chapter I have presented two models of the cultural evolution of unstructured communication systems — one revolving around a feedforward network model of a learner, the other based on an associative network model. In both models, a learning bias in favour of one-to-one mappings between meanings and signals was found to be key in driving the cultural evolution of communication. In the feedforward network model, obverter agents have this bias whereas imitator agents do not. Consequently, populations of obverter agents converge on optimal, unambiguous communication systems whereas populations of imitator agents do not. These results were found to hold even in the face of fairly strong natural selection of cultural variants. In the associative network model, only certain weight-update rules (those classified as [+learner, +maintainer, +constructor]) possess the one-to-one bias. Only those weight-update rules, when placed in the context of the ILM, result in the emergence of communicatively optimal systems of meaning-signal mappings.

An examination of the learning biases involved in other models of the cultural evolution of communication shows that this one-to-one learning bias is paramount. More significantly from the point of view of understanding language evolution in human populations, one-to-one biases seem to be brought to bear by human infants when acquiring vocabulary. I have presented arguments that human infants are biased against acquiring synonyms and homonyms. Furthermore, non-human primates appear not to have this bias. This suggests that the one-to-one biases applied by humans to the vocabulary learning task may be unique among primates, and may explain the uniqueness (among the primates) of language as a culturally transmitted, symbolic communication system.

CHAPTER 4

The genetic and cultural evolution of communication

In the previous Chapter I equated the properties of the learning bias required for the cultural evolution of optimal communication with properties of the human language acquisition device, in particular the apparent human biases against synonymy and homonymy during vocabulary acquisition. I made the assumption, for the purpose of Chapter 3, that this human capacity must have evolved through natural selection for communication.

It is time to return to this assumption and investigate whether the learning bias required to support a learned symbolic vocabulary can, in fact, evolve under selection pressure for communicative success. If this is the case then we can form an argument that this aspect of the human cognitive apparatus evolved specifically to support vocabulary acquisition.

Chapters 2 and 3 focussed on models of cultural transmission. In Section 4.1 I introduce a simple mathematical model of genetic transmission. This is synthesised with the general model of cultural transmission discussed in Chapter 2 to form B&R's *dual transmission model* (Section 4.2). The Iterated Learning equivalent of the dual transmission model is the Evolutionary Iterated Learning Model, outlined in Section 4.3. In Sections 4.4 and 4.5 I describe how the two ILMs which form the basis of Chapter 3 can be expanded to Evolutionary Iterated Learning Models, to investigate the possibility of coevolution between a culturally-transmitted communication system and the genetically-encoded learning bias facilitating the acquisition of such a system.

4.1 Modelling genetic transmission

The most basic models of genetic transmission are based around three processes:

$$\begin{tikzcd}
G_t \arrow[r, "\text{ontogeny}"] & \begin{pmatrix} F_t \\ G_t \end{pmatrix} \arrow[r, "\text{selection}"] & \begin{pmatrix} F_t' \\ G_t' \end{pmatrix} \arrow[r, "\text{transmission}"] & G_{t+1}
\end{tikzcd}$$

Figure 4.1: A simple model of genetic transmission. G_t gives the distribution of genotypes at time t. This leads, via developmental processes, to an associated distribution of phenotypes, F_t. The environment then takes its toll, removing some phenotypes and their associated genotypes. This gives a new distribution of phenotypes ${F_t}'$ and genotypes G_t'. The remaining genotypes then lead, via reproduction, to the distribution of genotypes at time $t + 1$.

- Ontogeny, where the genotype[1] is translated, in the context of the environment, into the phenotype. This depends on 1) the distribution of genotypes population, G_t, and 2) the process of ontogeny.
- Selection, where the environment takes its toll on the population by removing individuals based on the performance of their phenotypes. This depends on 1) the distribution of phenotypes, F_t, and 2) the selection procedure for removing phenotypes from the population.
- Genetic transmission, where the surviving individuals reproduce and transmit their genotypes to their offspring. This depends on 1) the distribution of phenotypes after selection, F_t', which implicitly gives G_t', the distribution of genotypes of surviving individuals, and 2) the procedure by which new genotypes are formed during reproduction.

This cycle is illustrated in Figure 4.1.

4.1.1 Natural selection on genetic transmission

The simplest models of natural selection acting on genetic transmission deal with the changes in frequency of alleles of a single gene in asexually-reproducing haploid populations — each individual has a single gene drawn from a set of n alleles and each individual inherits the allele of their single parent. In sexually-reproducing diploid populations the equations are complicated by the fact that each individual has two alleles for each gene and receives one allele from each of their two parents.

[1] A clarification of terminology is perhaps useful here:
Genotype: The particular genetic makeup of an individual.
Phenotype: The particular physical, non-genetic makeup of an individual, which will be a consequence of an individual's genotype in interaction with the environment.
Genome: The specification of possible genotypes, which limits the range of possible genotypes.
Phenome: The specification of possible phenotypes, which limits the range of possible phenotypes.

Ontogeny is typically treated in a very simplistic manner in mathematical models of population genetics. In the haploid organism, single gene case there are n distinct alleles and therefore n distinct genotypes $G_1 \dots G_n$. It is typically assumed that there are n distinct phenotypes $F_1 \dots F_n$ and ontogeny maps genotype G_i onto phenotype F_i. Selection then acts on the phenotype, but since there is a one-to-one correspondence between genotypes and phenotypes we can talk of selection acting on genotypes and effectively ignore ontogeny.

The *fitness* of such a genotype is the average probability of individuals with that genotype surviving to reproductive age, multiplied by the average number of offspring that each genotype of reproductive age produces. Now consider a population with two genotypes G_a and G_b with fitness f_a and f_b respectively. Evolution by natural selection takes place in such a population where the two genotypes do not reproduce at equal rates — $f_a \neq f_b$. As shown in Section A.2 in Appendix A, if $f_a > f_b$ then genotype G_a will increase in frequency, and if $f_a < f_b$ then it will decrease in frequency — the fitter genotype comes to dominate the population. The rate of change is at a maximum when genetic diversity is at a maximum, which occurs when both genotypes occur with equal frequency — in other words, natural selection depends on genetic diversity, and the rate of evolution is higher when the population exhibits more diversity.

4.1.2 Models of the genetic transmission of communication

Models of the genetic evolution of communication can be roughly divided into two main groups — those which address the question “when should we expect to see communication?” and those which ask “what structure should we expect communication to exhibit?”.

4.1.2.1 When should we expect to see communication?

The first question has typically been addressed by theoretical biologists but has recently been tackled by researchers using agent-based modelling techniques. Researchers in this area are concerned with the interlocking issues of signal honesty, altruism and signal costs.

Much of this work has been inspired by, or relates to, Krebs & Dawkins’s (1984) conception of signalling as manipulation. Krebs & Dawkins define signals as “means by which one animal makes use of another animal’s muscle power” (Krebs & Dawkins 1984:382) — a signal emitted by one animal causes another animal to act in a certain way. This is bound up with their definition of communication (given in Dawkins & Krebs (1978)) that communication systems are systems where the signaller has evolved to induce a response in the receiver which benefits the signaller in terms of inclusive fitness.

Where the interests of the signaller and the receiver are coincident (both signaller and receiver get some benefit), the evolution of signalling seems fairly unsurprising — individuals who signal will be selected for, as will individuals who respond to signalling. One common thread in the formal modelling approach revolves around identifying common interests between signallers and receivers, particularly in situations where no common interest is obvious. For example, alarm call systems obviously offer a benefit to receivers of alarm calls (they can evade a predator they may not have noticed), but the signallers receive no obvious benefit (they have already seen the predator), and may in fact incur a cost (their signalling behaviour may attract the predator). So why do alarm call systems exist? Kin selection, whereby signallers receive an indirect payoff via the genes they share with receivers, has been advanced as an explanation for this apparently altruistic behaviour, based on both empirical evidence from studies of alarm-calling species (Sherman 1977) and evidence from computational modelling (Ackley & Littman (1994) and Oliphant (1996), but see Noble (1999) for a negative result).

Krebs & Dawkins (1984) make the subsidiary prediction that, where the interest of signallers and receivers coincide, there should be coevolution between signaller and receiver so that the signal is cheap for the signaller to produce and the receiver is sensitive enough to pick up on these “conspiratorial whispers”. Analytic and simulation models (Noble 1998) support a slightly modified version of Krebs & Dawkins’s (1984) conspiratorial whispers theory — where both sender and receiver receive a clear payoff then signals should be cheap, whereas if the signaller can expect to receive a large payoff but the receiver only receives a marginal benefit, receivers should be resistant to cheap signals and signallers should therefore evolve to use costly signals.

Matters are more complicated where there is a direct conflict of interest between signaller and receiver, which cannot be resolved by appeals to mechanisms such as kin selection. Consider the case of males who produce signals to indicate to females their quality as a mate. Males, typically, will wish to mate as often as possible whereas females should be choosy, preferring to bear offspring only for high-quality males. In this scenario there is a direct conflict of interest and we would predict that signalling would be useless — any population of “honest” signallers, where the signals of males reliably reflected their quality and the females took the males at their word, would be prone to invasion by “liars”, males who exaggerated their own quality.

One potential solution to this problem is the handicap principle (Zahavi 1975; Zahavi 1977). If the signals males produce are costly, with a cost proportional to their claims of quality, then females can trust the signals — only a male of high quality could afford to bear the cost of producing the signal indicating high quality, therefore individuals producing the high quality signal must be telling the truth. This result has been verified by analytic and simulation models, with one or two qualifications (see, for example, Grafen (1990) and Bullock (1997)).

Alternatively, Krebs & Dawkins (1984) suggest that signalling could persist in the face of a conflict of interests, but result in an evolutionary arms race between signallers and receivers — signallers will be selected to produce increasingly effective manipulative signals, and receivers will be selected for increased resistance to these manipulative signals. This position has received less support from formal models — Noble (1999) presents results from both analytic and simulation models which suggest that communication is never more than a transitory phenomenon when a conflict of interests exists.

While these issues are obviously important, they are somewhat outwith the main area of inquiry of this thesis — I assume that signallers and receivers have coincident interests, and both receive a payoff from successful communication. In part this is due to my focus on the *structure* of communication systems and the learning bias required to bring it about, but also to the observation that issues of honesty, altruism and so on don't seem to have a great impact on the structure of language — people tell the truth and tell lies, and altruistically give and selfishly withhold information all the time, and it's not clear if this has any lasting structural consequences.

4.1.2.2 How should communication be structured?

There are several models which show that communicatively-optimal, innate communication systems (i.e. shared one-to-one mappings) can evolve under natural selection for communicative success (e.g. Werner & Dyer (1992), MacLennan & Burghardt (1993), Levin (1995)). Such models typically assume a pre-existing dedicated communication channel, although it has been shown recently that communication can evolve in the absence of such a dedicated channel, through evolutionary reappropriation of initially non-communicative behaviours (Quinn 2001). Functional communication systems have also been shown to emerge as a consequence of the evolution of internal representations (Cangelosi & Parisi 1998).

Turkel (2002)[2] presents an interesting model, based on Hinton & Nowlan's (1987) influential paper, which demonstrates the occurrence of the Baldwin effect. Turkel interprets his model as a model of language evolution. This is perhaps the weakest point of the model, for reasons which will become obvious. However, the results show interesting behaviour with respect to the evolution of coordinated behaviour, and as such it is relevant to models of the evolution of communication.

In the model, an individual's genotype is a string of 1s, 0s and ?s. The process of genotype-phenotype mapping simply directly maps this genotype to a phenotype string of 1s, 0s and ?s. Individuals then participate in pairwise coordination interactions, which could be considered communicative interactions, where coordination is simply phenotype matching. During each interaction the two individuals involved temporarily[3] replace all the ?s in their phenotype with a 0 or a 1, selected at random. If the two phenotypes then match then the interaction is considered to be a success. Otherwise, a different temporary replacement of ?s with 1s and 0s is attempted. This process repeats until coordination is established or a certain prespecified number of replacement trials is reached, in which case the interaction is considered to be a failure. The level of success of an episode of interaction is inversely proportional to the number of replacement trials required to achieve a phenotype match.

There are several points to note here. Firstly, if, at any given point on the phenotype, one individual has a 1 and the other individual has a 0 then coordination is impossible. Secondly, individuals with large numbers of ?s are likely to take longer to successfully match their phenotypes. Finally, Turkel explicitly penalises individuals who require 0 trials to match their phenotypes (i.e. both individuals have no ?s in their phenotypes and their phenotypes match) — this apparently capricious step is taken to ensure that ?s never disappear from the population, a consequence of Turkel's interpretation of the model as a model of P&P UG.

Turkel's main contribution is to show that aspects of phenotypes which were originally plastic (?s) become fixed (to 1s or 0s), via the Baldwin effect, although plasticity is never fully eliminated, in part due to Turkel's quirky fitness function. The Baldwin effect can be summarised thus: behaviour which is initially plastic becomes innately fixed when plasticity has some cost. In Turkel's model, excess ?s have a cost as they reduce the

[2] This paper has in fact been in circulation, and receiving citations, since 1994. However, for a variety of reasons, it has only recently been published. This explains the fact that, for example, Kirby & Hurford present an extension of a model in 1997 that was not published till 2002.

[3] The replacement is only temporary, therefore there is no true cultural transmission in the population — an individual's phenotype after a series of interactions is exactly the same as its phenotype before those interactions.

coordination payoff individuals can expect to receive. Parts of the genotype which began as ?s therefore tend to be replaced by 1s or 0s, depending on the genetic makeup of the rest of the population. This model therefore demonstrates that the Baldwin effect can potentially lead to the transfer from a learned coordinated system of behaviour (perhaps a communication system) to a largely innate, coordinated system of behaviour. Yamauchi (2001) demonstrates that this can only happen when the relationship between genotypes and phenotypes is fairly simple — when a single phenotypic trait is determined by several genes, or when a gene determines several phenotypic traits, the impact of the Baldwin effect is reduced.

4.2 The dual inheritance model

How can models of genetic transmission be combined with models of cultural transmission to produce a unified model of transmission? B&R propose the *dual inheritance model*, sketched in Figure 4.2.

The population at time t is defined by F_t, a distribution over phenotypes, and G_t, a distribution over genotypes. G_t is derived from G'_{t-1}, the distribution of genotypes of individuals of reproductive age at the previous generation. The process of ontogeny and cultural transmission then give F_t, the phenotype distribution, based on:

- the mechanisms of ontogeny,
- G_t, the distribution of genotypes in the population,
- the mechanisms of cultural transmission and
- F'_{t-1}, the distribution of phenotypes of individuals of reproductive age at the previous generation.

In a dual transmission model the distribution of genotypes in the population is relevant because genetic variants are taken to have an influence on the process of cultural transmission — for example, individuals with a particular genotype may not take part in cultural transmission, or may prefer to acquire a particular cultural trait.

The dual inheritance model then proceeds in a similar fashion to the general model of cultural transmission given in Chapter 2. F_t^l gives the distribution of phenotypes in the population once individual learning has been taken into account. This depends on two factors: 1) F_t, the distribution of phenotypes prior to individual learning; 2) the process of individual learning, by which individuals change their phenotype in response to the

$$\begin{tikzcd}
\begin{pmatrix} F'_{t-1} \\ G'_{t-1} \end{pmatrix} \arrow[rr, "\text{cultural transmission}"] \arrow[dr, "\text{mating}"'] & & \begin{pmatrix} F_t \\ G_t \end{pmatrix} \arrow[r, "\text{individual learning}"] & \begin{pmatrix} F^l_t \\ G_t \end{pmatrix} \arrow[r, "\text{selection}"] & \begin{pmatrix} F'_t \\ G'_t \end{pmatrix} \\
& G_t \arrow[ur] & & &
\end{tikzcd}$$

Figure 4.2: The dual transmission model. The genotype distribution of the population at time t, G_t is determined by the the reproduction process and the distribution of genotypes in the population at time $t-1$. Cultural transmission, constrained by the genetic makeup of the population (G_t) then yields the distribution of cultural variants F_t. This process is dependent on the mechanisms of cultural transmission, the influence of genes on the cultural transmission process, and the distribution of cultural variants in the population at time $t-1$.

environment. F'_t and G'_t give the distribution of phenotypes in the population after removal of individuals through death has been taken into account. These depend on: 1) F^l_t and G_t, the distribution of phenotypes and genotypes prior to death; 2) the process of differential retention, by which some individuals survive and some individuals are removed due to environmental factors. Note that the distribution of genotypes remains unaffected by cultural transmission and individual learning.

4.2.1 The genetic transmission of direct bias

B&R consider the circumstances under which a biological capacity for individual learning and biased and unbiased cultural transmission will be favoured by natural selection. In general, their technique is to assume that the capacity for the behaviour of interest is genetically encoded, and that there are two possible alleles. Individuals with one allele will participate in the relevant behaviour, whereas individuals with the other allele will not. B&R construct equations to see under what circumstances the allele for the particular behaviour will be favoured by natural selection on genetic transmission.

For the purpose of this thesis it is sufficient to review their model of the genetic evolution of direct bias. Recall from Chapter 2 that direct bias on cultural transmission will increase the frequency of the favoured variant in a population, with the rate of increase depending on the strength of the direct bias and the cultural variance in the population. B&R expand this model, following their general technique outlined above, to consider the case where an individual's genotype determines their preference for cultural variants.

Let us assume that there are two cultural variants, c and d, and two genetic variants, e and f. e is the unbiased genotype, and f is the biased genotype. In other words, individuals with genotype e will happily acquire cultural variants c or d, whereas individuals with genotype f will prefer to acquire one particular cultural variant, let's say c.

As shown in Section A.3 of Appendix A, if genetic parents are selected at random (there is no natural selection acting on genetic transmission) then, as we would expect, the frequency distribution of genotypes in the population remains unchanged. Among individuals with the biased allele f cultural variant c increases in frequency according to the strength of the bias and the cultural variance in the parent population, and variant d decreases by a similar factor. Among individuals with the unbiased allele e the frequency of the two cultural variants remains unchanged. In other words, individuals with the unbiased genotype acquire either cultural variant, whereas individuals with the biased genotype preferentially acquire variant c at the expense of variant d.

B&R then go on to add natural selection to the model. Natural selection weeds individuals out after cultural transmission and prior to breeding. Let us assume that individuals in possession of cultural variant c receive some fitness payoff s (and are therefore more likely to reproduce genetically and act as cultural parents), and individuals with the biased genotype f incur some cost z for that bias (and are therefore penalised if $z > 0$).

Assume for a moment that the proportion of individuals with cultural variant c, is fixed at some arbitrary value. What happens to the frequency of individuals with the biased genotype? As shown in Section A.3 of Appendix A:

- if the biased genotype has no associated cost ($z = 0$):
  - if the population exhibits no cultural variation then the biased allele has no fitness advantage over the unbiased allele and does not change in frequency.
  - if the population exhibits cultural variation then the biased allele will increase in frequency.
- if the biased genotype has a cost ($z > 0$):
  - if the population exhibits no cultural variation then the biased genotype will decrease in frequency — the biased allele will suffer a fitness penalty due to its cost and no fitness benefit over the unbiased allele due to the lack of cultural variation.

– if the population exhibits cultural variation then the biased genotype will either be selected for or against, dependent on the relative strength of the direct bias, the cultural makeup of the population, the advantage of possessing cultural variant c and the cost associated with the bias.

To summarise, in a population which is completely converged culturally (on either variant) the frequency of the biased variant should either remain constant (if biased learning is costless relative to unbiased learning), or decrease (if biased learning has a cost).

What can we predict about the frequency of cultural variant c? As discussed in Section A.3 of Appendix A, variant c is always favoured by selection, and by biased transmission when there are individuals with the biased genotype in the population. Therefore variant c will increase in frequency until the population completely converges on c. As discussed above, at this equilibrium state the biased genotype either has no advantage over the unbiased genotype or is at a disadvantage (where $z > 0$). Therefore, at equilibrium we should expect selection to either be neutral with respect to bias, or to see only the unbiased allele — directly biased transmission pushes the population to converge on the favoured cultural variant, at which point selection pressure on the population's genotypes either stops, or acts to reduce the frequency of the biased allele which drove cultural convergence in the first place.

4.3 The Evolutionary Iterated Learning Model

The Evolutionary Iterated Learning Model (EILM) is an extension of the ILM to model the dual transmission of cultural and biological traits — it is the Iterated Learning equivalent of B&R's dual transmission model. As in the ILM, populations consist of collections of individuals, who acquire their linguistic competence on the basis of the observed behaviour of other individuals. Individuals also inherit their genetic endowment from other members of the population, their biological parent(s). As with the ILM, population turnover can either be generational, as depicted in Figure 4.3, or gradual, as in Figure 4.4.

4.3.1 The evolutionary iterated learning of vocabulary

Hurford (1989) presents the original EILM. Hurford assumes three possible genetically-encoded learning strategies, which he calls Imitators, Calculators and Saussureans. As discussed in Chapter 3, Section 3.5.3, based on experiments outlined in Hurford (1989) and an analysis of these three learning strategies, Calculators can be classified as [$-$learner, $-$maintainer, $-$constructor] and Imitators and Saussureans can be classified as [$+$learner,

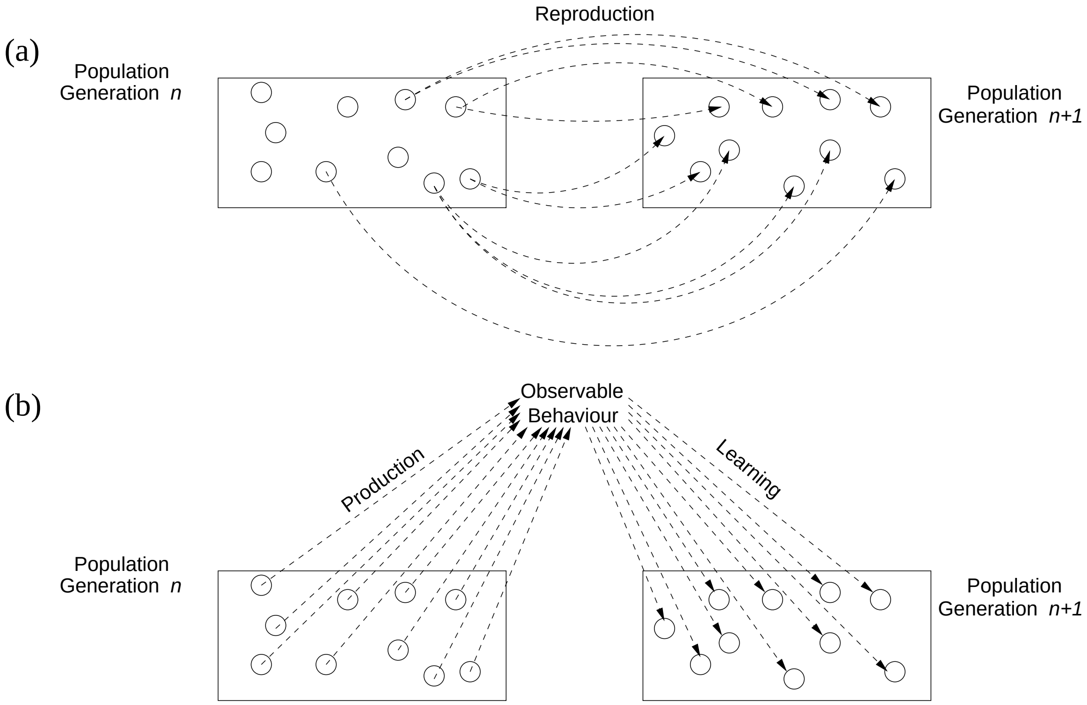


Figure 4.3: The generational EILM. (a) illustrates the process of genetic transmission — individuals at generation $n + 1$ (represented by circles) inherit their genotypes via genetic transmission (arcs) from individual at generation n. (b) illustrates the process of cultural transmission — the generation n individuals produce a set of observable behaviour which is observed and learned from by generation $n + 1$ individuals.

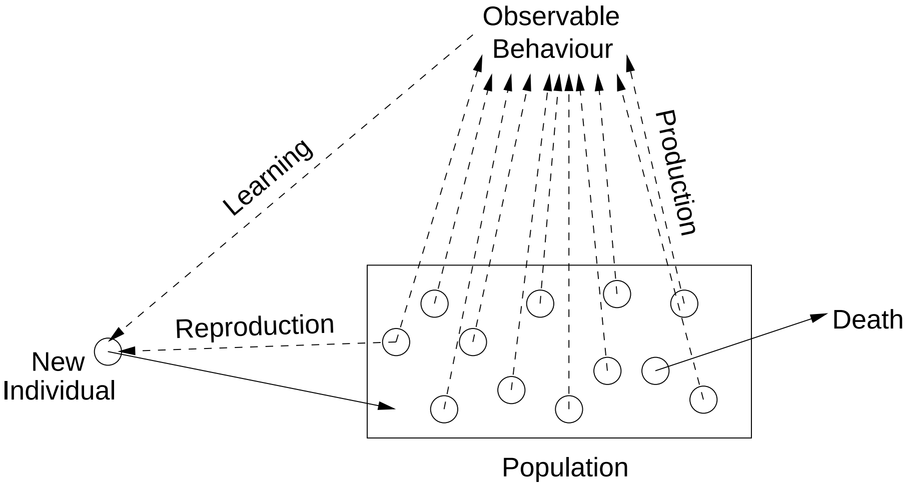


Figure 4.4: The gradual EILM. A single individual is removed from the population, representing death. The remaining individuals then reproduce to produce a new individual, who inherits its genotype from its parent(s). The new individual then learns based on the observable behaviour produced by the population, and enters the population.

+maintainer, −constructor]. Saussureans have the additional advantage over Imitators that their production and reception behaviour are necessarily closely coupled.

Hurford conducts three sets of experiments, the first two of which have been discussed in Section 3.5.3 and will not concern us here. In the third set of experiments, using a generational model of population turnover, Hurford investigates how natural selection during genetic transmission of learning strategies affects genetically heterogeneous populations. Hurford's simulation runs start with each meaning–signal pair being equiprobable, with an equal distribution of the three genetically-encoded learning strategies. There is strong selection during genetic transmission, with individuals who are more successful at communicating with other members of the population being more likely to breed. Under such circumstances, the Saussurean learning strategy comes to dominate each of 20 simulated populations — natural selection during genetic transmission leads to the evolution of the 'optimal' learning bias. However, Hurford (1989) does not report the frequencies of the various genotypes over the course of the runs, and does not retain the original data (James Hurford, personal communication). This is a pity, as we will see in Section 4.5 that the manner in which the population arrives at the final state is in fact more telling than the final state itself.

Kvasnička & Pospíchal (1999) present an EILM similar to the model which I will describe in Section 4.4. Briefly, Kvasnička & Pospíchal use a feedforward network model of a communicative agent, with the obverter architecture. Individuals acquire their communication system based on the observable behaviour produced by the previous generation of the population. The connection weights in an individual's network, together with some limited aspects of network topology, are specified genetically, and opportunity to breed is determined by communicative accuracy. Kvasnička & Pospíchal report that optimal communication emerges from initially random behaviour, and conclude that this is a consequence of gene-culture coevolution. However, it is not clear from their model that there is actually any gene-culture coevolution occurring. While the network topology may indeed evolve to facilitate communication (a similar result is reported by Livingstone & Fyfe (2000)), it is not clear to what extent the transmission of initial connection weights plays a role in biasing the learners towards acquiring optimal communication systems. As discussed in Chapter 3, the obverter network architecture builds in a bias in favour of one-to-one mappings from meanings to signals — optimal communication can emerge in populations of obverter networks in the absence of any genetic transmission of connection weights. The relative importance of pressures acting on genetic and cultural transmission is not adequately decoupled in Kvasnička & Pospíchal's (1999) model.

EILMs have also been used to investigate gene-culture interactions in the evolution of structured communication — for example, the models described in Kirby & Hurford (1997) and Briscoe (2000b). Discussion of these models will be delayed until Chapter 6.

4.4 Model 1: Adding genetic transmission to the feedforward network model

In this Section, I describe an extension of the feedforward network model outlined in Section 3.3 which can be used to investigate possible interactions between the genetic transmission of learning bias and the cultural evolution of communication. This research appears in Smith (in press).

Feedforward neural networks are sensitive to their initial connection weights (Kolen & Pollack 1990) — certain combinations of initial connection weights can make certain input-output mappings unlearnable, whereas other initial weights can reduce the number of exposures required to learn particular mappings. The initial connection weights effectively bias a feedforward neural network towards acquiring certain input-output mappings. This suggests a first step towards investigating the dual transmission of vocabulary systems — if we assume that the initial connection weights in an individual's neural network are genetically determined, then we have a model where genetic information affects that individual's learning bias.

The feedforward network model outlined in Section 3.3 can be straightforwardly extended to include the genetic transmission of initial connection weights, in order to investigate how differential genetic transmission of learning bias impacts on the evolution of unstructured communication. Some of the assumptions arising from this treatment are somewhat dubious, in particular the assumptions that there is a one-to-one correspondence between genotype and phenotype and that genetic information merely provides a starting point for unconstrained learning in the phenotype. Two points can be made in defence of this model. Firstly, the technique of evolving initial connection weights for neural networks is not a new one, and has been adopted for non-linguistic problem domains with some success by, for example, Montana & Davis (1989), Belew *et al.* (1992) and Nolfi *et al.* (1994). Batali (1994) adopts a similar technique to a study of the learnability of structured languages. Secondly, the feedforward network model merely provides a starting point — while it yields some useful results, it will be superseded by the more plausible model outlined in Section 4.5.

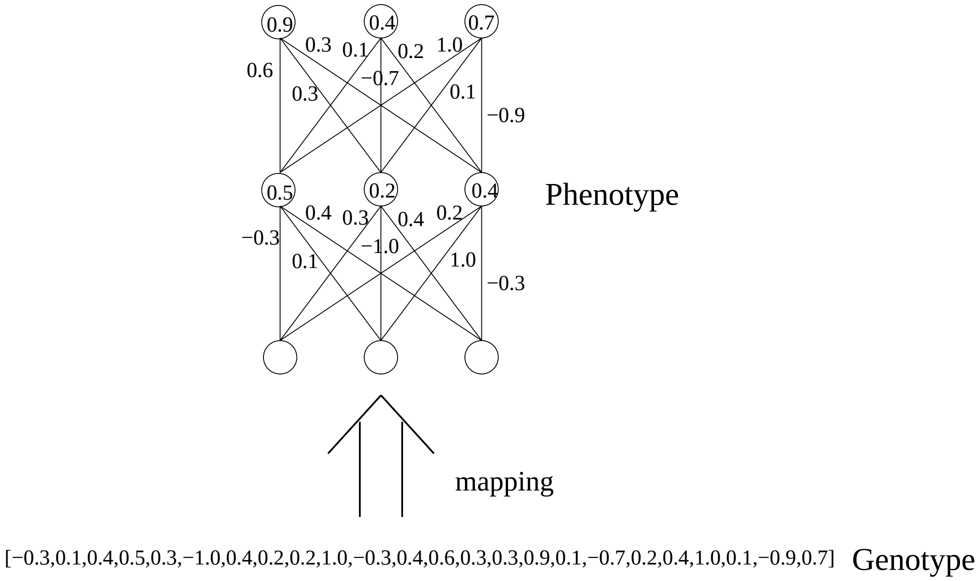


Figure 4.5: Genotype-phenotype mapping, the process of translating a genotype into a phenotype. Each locus on the genome specifies the weight of a connection in the phenome. In the genotype-phenotype mapping process, the allele at each locus in an individual's genotype specifies the initial weight of the associated connection in the individual's phenotype network, as illustrated here. The weights from bias nodes to nodes in the hidden and output layer are shown in the centre of the associated node.

4.4.1 Genotypes, phenotypes and the genotype-phenotype mapping

The EILM requires a model of the genome and a model of the phenome. The phenome model is simply the feedforward network model of a communicative agent outlined in Section 3.3. The genome is a chromosome with 24 loci. The possible alleles for each locus are the set of real numbers.

Ontogeny, in the terms outlined in Section 4.1, involves the mapping from a genotype to a phenotype. Each individual consists of a genotype-phenotype pair. Each locus in the genome corresponds to a connection in the phenome network. In the process of genotype-phenotype mapping the allele at each locus in an individual's genotype specifies the initial weight of the associated connection in the individual's phenotype network. This mapping process is illustrated in Figure 4.5.

4.4.2 Reproduction

Individuals inherit their genotype from their parents. In this model the genome is haploid, so asexual reproduction would be a reasonable and simple option. However, previous research suggests that asexual reproduction in small populations tends to result in a rapid loss of genetic diversity and a corresponding weakening of evolutionary pressures (Mitchell 1996). A sexual model of reproduction is therefore used, to attempt to maintain genetic diversity in the population.

Each individual has two parents. The parental genotypes are combined to form a single offspring genotype using one-point crossover[4]. Mutation[5] during reproduction is also included, once again to maintain genetic diversity in the population.

4.4.3 The EILM algorithm

As with the ILM described in Section 3.3, in the feedforward network EILM a generational model of population turnover is used. Populations consist exclusively of either imitator or obverter networks. The EILM consists of an initialisation process and an iteration process.

Initialisation Create a population $population_{g=0}$ of N agents[6]. Each agent is either an imitator or obverter, with populations being homogeneous in this respect. Each initial agent has a random genotype, with the allele at each locus selected randomly from the range $[-1, 1]$. The population is heterogeneous in this respect. Each initial agent's phenotype is determined by their genotype and the genotype-phenotype mapping process discussed in Section 4.4.1 .

Iteration

1. Evaluate the communicative accuracy of every member of $population_g$ by evaluating every individual's communicative accuracy as both producer and receiver with two randomly selected partners according to the measure $ca\,(P, R, m_i)$, for every communicatively relevant meaning $m_i \in \mathcal{M}_{CRS}$.
2. For every member of $population_g$, generate a set of meaning-signal pairs by applying the network production process to every $m \in \mathcal{M}_{CRS}$.

[4] Crossover occurs with probability p_x ($p_x = 0.95$ in all simulations outlined in this section). When crossover takes place, the alleles for the first n loci of the offspring's genotype are inherited from the first parent and the remaining alleles are inherited from the second parent. n is randomly selected to be between 1 and $l_g - 1$, where l_g is the length of the genome. When crossover does not take place the whole genotype is inherited from the first parent.

[5] Point mutations occur on the newly-formed genotype with probability $p\,_m$ ($p_m = \frac{0.1}{l_g}$ for all simulations outlined in this section). Mutation involves replacing the allele at the mutated locus, which has a value a, with an allele with the value $a+r$, where r is a random number in the range $[-1, 1]$. This mutation operator, in conjunction with the unrestricted range of alleles, allows the possibility of the emergence of extremely large-valued alleles. However, in practice such alleles do not occur. In the simulations outlined in this section all alleles remain within the range $[-5.41, 5.29]$.

[6] $N = 100$ for all the EILMs outlined in this section.

3. Create a new population $population_{g+1}$ of N agents of the same type (imitator or obverter) as $population_g$. Each member of $population_{g+1}$ inherits their genotype from two parents from $population_g$, via the reproduction process outlined above. Parents are selected randomly from amongst the fittest b members of $population_g$.
4. Each member of $population_{g+1}$ receives e exposures to the observable behaviour generated by $population_g$. During each of these e exposures the new agent observes the complete set of meaning-signal pairs generated by a member of $population_g$ selected randomly from among the t most successful communicators in $population_g$. For each exposure the learner updates their connection weights according to the observed meaning-signal pairs using the backpropagation learning algorithm[7].
5. $population_g$ is removed and replaced with $population_{g+1}$. Return to 1.

In this model there are three possible pressures operating on the population:

1. *Selection for communicative success operating on genetic transmission* (when $b < N$), driven by natural selection, favouring genes whose phenotype realizations are successful communicators.
2. *Selection for communicative success operating on cultural transmission* (when $t < N$), driven by natural selection of communication systems, favouring systems which result in successful communication.
3. *Selection for learnability operating on cultural transmission*, driven by the agents' learning bias, favouring either more ambiguous communication systems (in the case of imitator agents) or less ambiguous systems (in the case of obverter agents).

How do these pressures interact? The behaviour of populations attempting to construct and maintain an optimal system, for various simulation parameter settings, are discussed in the next two sections.

4.4.4 Emergence of a communication system

Simulation runs were carried out to evaluate the behaviour of imitator and obverter populations with initially random communication systems (given by the random initial weights in the generation 0 population), for various levels of natural selection pressure acting on genetic transmission (dependent on b) and cultural transmission (dependent on t). Runs were allowed to proceed for 1000 generations, with 10 runs being carried out for each set of simulation parameters. In the figures that follow, the measurement of interest is

[7] As before, a learning rate of 0.5 is used.

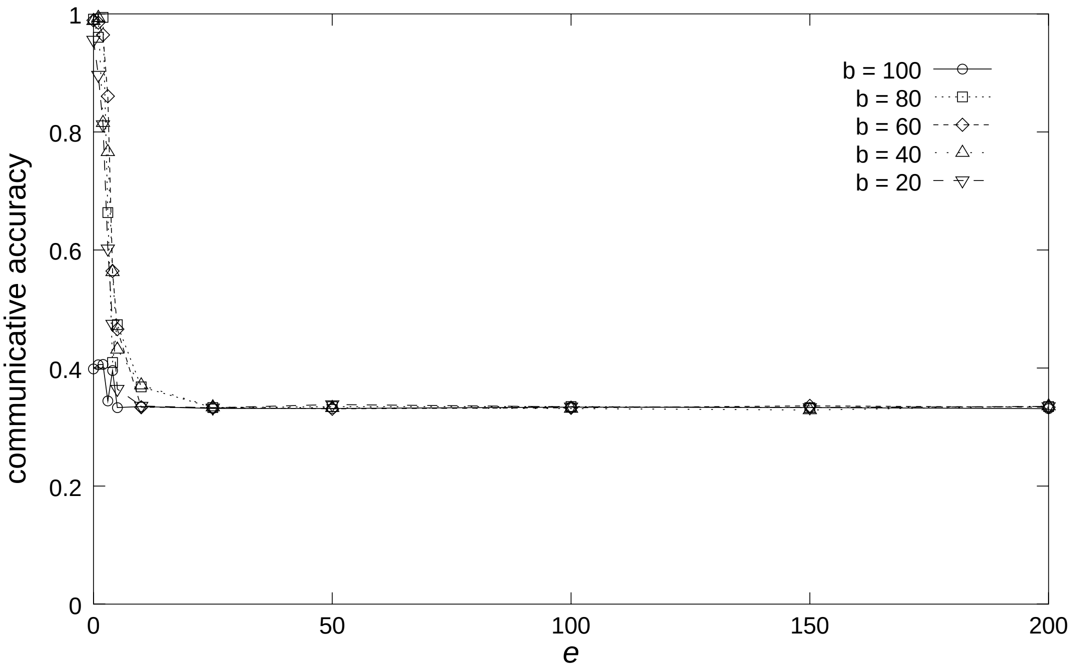


Figure 4.6: The final communicative accuracy of imitator populations in the EILM, in the case where there are varying degrees of natural selection acting on genetic transmission ($b \leq N$) and no natural selection of cultural variants ($t = N$), as a function of e. Communicative accuracy drops rapidly as e increases — cultural transmission clearly has a detrimental effect.

the final average communicative accuracy of the populations. This was obtained by measuring the communicative accuracy of each individual in the final 10 generations of the population, according to the process outlined above, averaging over the population then averaging over the 10 simulation runs carried out with that experimental setting.

4.4.4.1 Imitator populations

Figure 4.6 shows the communicative accuracy of the final systems in imitator populations, for the case where $b \leq N$ and $t = N$, for various values of e (numbers of learning exposures). Results are not shown for the case where $t < N$ — as we might expect given the results discussed in Section 3.3, differential access to cultural parent roles had virtually no impact on the behaviour of the populations.

For the case where $b = N = 100$ (breeding is independent of fitness), the behaviour of the populations is essentially as it was when there was no genetic transmission of initial connection weights — communicative accuracy remains uniformly low. As can be seen from Figure 4.7, for very low values of e and $b < N$ (natural selection on genetic transmission) communicative accuracy does rise significantly above chance levels. When

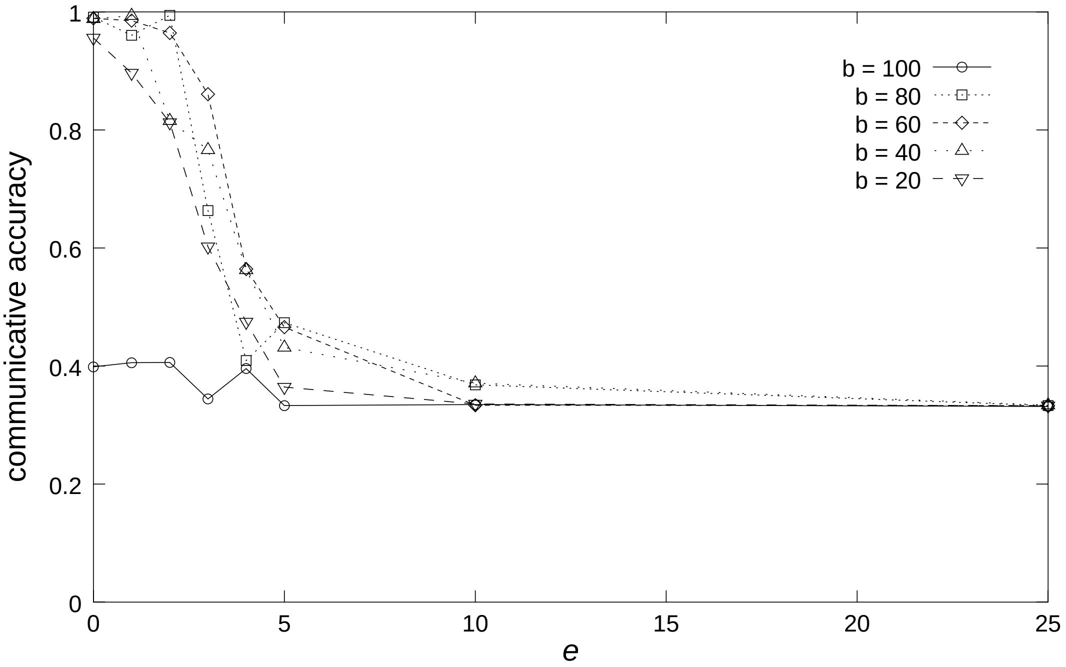


Figure 4.7: The final communicative accuracy of imitator populations where $b \leq N$, $t = N$, for low values of e.

$e = 0$ communicative accuracy reaches optimal levels. This is the result we would expect, given the mathematical and computational models of genetic transmission given above — in the absence of learning, natural selection acts to increase the frequency of fitter genotypes. In the model considered here, fitness is dependent on communicative accuracy, and natural selection therefore acts to increase the frequency of genotypes encoding successfully communicating individuals.

However, as e increases communicative accuracy rapidly drops. For $b < N$ (natural selection on genetic transmission) and $e > 10$ the populations converge on maximally ambiguous communication systems, leading to chance levels of communication. The importance of cultural transmission increases as e increases, and consequently the importance of directly biased cultural transmission increases. As discussed in Section 3.3, imitator individuals are biased in favour of acquiring ambiguous communication systems. When $e = 10$ fully ambiguous systems become completely learnable, and remain more learnable than less ambiguous systems. When this value of e is reached, the pressure for genotypes encoding successful communicators is completely overridden by the pressure on cultural transmission introduced by learner bias, and the populations converge on fully ambiguous communication systems. To put it another way, as e increases, cultural transmission begins to *shield* (Ackley & Littman 1992) genetic information from natural selection — learning makes individuals with distinct genetic makeups appear phenotypically identical, therefore selection cannot identify beneficial genotypes. This is also the situation predicted by B&R — when the population exhibits no cultural variation, selection acting on genetic transmission is neutral with respect to the genetically-encoded direct bias.

4.4.4.2 Obverter populations

Figure 4.8 plots the communicative accuracy of obverter populations, for $b \leq N$ and $t = N$, for various values of e. Again, no significant effect was found when natural selection of cultural variants was introduced — setting $t < N$ leads to similar results.

For the case where $b < N$, natural selection on genetic transmission clearly has a positive effect — the average communicative accuracy of the populations increases above the level for $b = N$. Natural selection fine-tunes initial connection weights so as to increase the learnability of unambiguous systems. For low to intermediate values of e a direct bias on cultural transmission has clearly evolved. However, for high e the obverter network learning bias alone leads to optimal communication. Under these circumstances, there is no selection pressure in favour of an additional direct bias introduced by the networks' initial connection weights — the population's initial connection weights undergo genetic drift. As discussed above, this situation matches the predictions of both B&R and Ackley & Littman (1992) — when the population exhibits no cultural variation, selection acting on genetic transmission is neutral with respect to the genetically-encoded direct bias.

4.4.5 Maintenance of an optimal system

In Chapter 3, I briefly alluded to the behaviour of imitator populations when attempting to maintain an optimal system. I predicted that imitator agents would be unable to do so, due to their bias in favour of many-to-one mappings between meanings and signals. However, no results from the ILM were found to support this prediction. It turns out that some support for this position can be found by considering populations of imitator agents attempting to maintain an optimal system in the EILM.

Figure 4.9 shows the average communicative accuracy of populations of imitator agents who start out with a shared, optimal, innate communication system — all the agents in the initial population have a set of genes which encode an unambiguous communication system. $e = 200$ for all simulations in Figure 4.9. Various amounts of selection pressure (b and t) are used on genetic and cultural transmission. As can be seen from Figure 4.9,

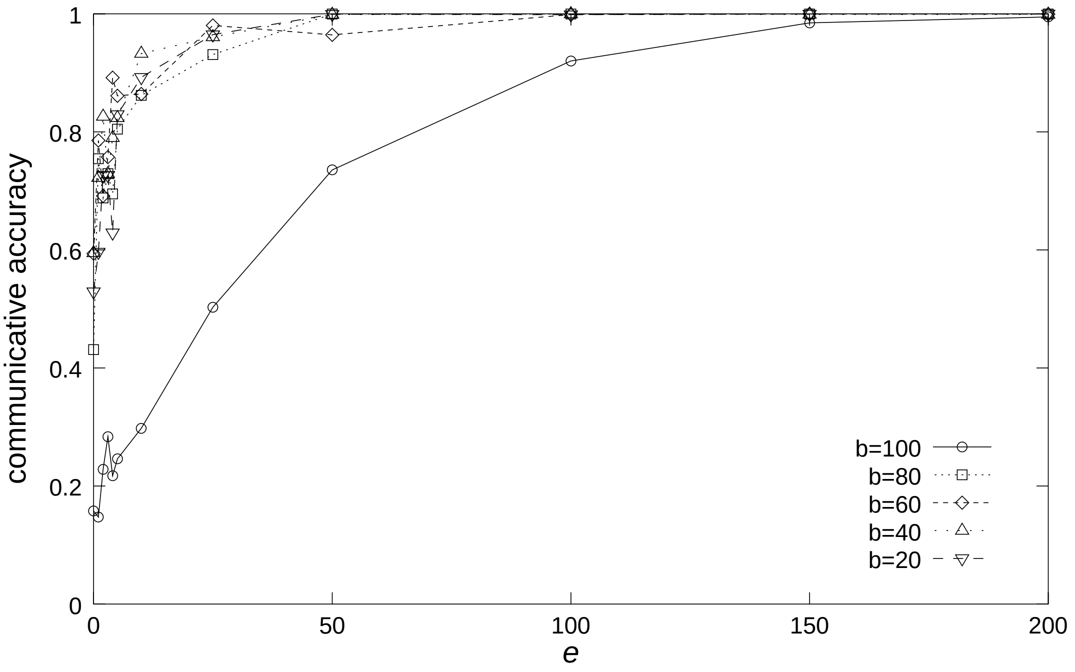


Figure 4.8: The final communicative accuracy of obverter populations where $b \leq N$, $t = N$, for various values of e. Communicative accuracy increases as e increases — cultural transmission is beneficial. Overall levels of communicative accuracy are also higher when $b < N$ — natural selection acting on genetic transmission is beneficial, although the benefit decreases as e increases.

all populations collapse from using an unambiguous communication system to using a partially or fully ambiguous communication system within 25000 generations.

As discussed above, learning in the phenotype almost completely masks an agent's genes but there are certain combinations of genes which make learning a particular communication system impossible. Given the absence of cultural variation in populations with an optimal initial system, genetic transmission occurs at random — there is no selection pressure in favour of a direct bias which results in successful communication. The population's genotypes, and therefore the population's initial connection weights, will undergo drift. In each simulation shown in Figure 4.9 an agent is eventually born whose genes are so bad that they cannot learn the unambiguous communication system in use by the rest of the population. This individual learns a partially ambiguous or fully ambiguous communication system instead. Such agents are unlikely to breed, given that their fitness is usually lower than other agents in the population. Suboptimal communicators do have a negative effect on the fitness of optimally-communicating agents, given that those optimally-communicating agents suffer a penalty for not understanding or being understood by suboptimal communicators, although this will not usually depress the population's fitness enough to allow a suboptimal communicator to breed. However,

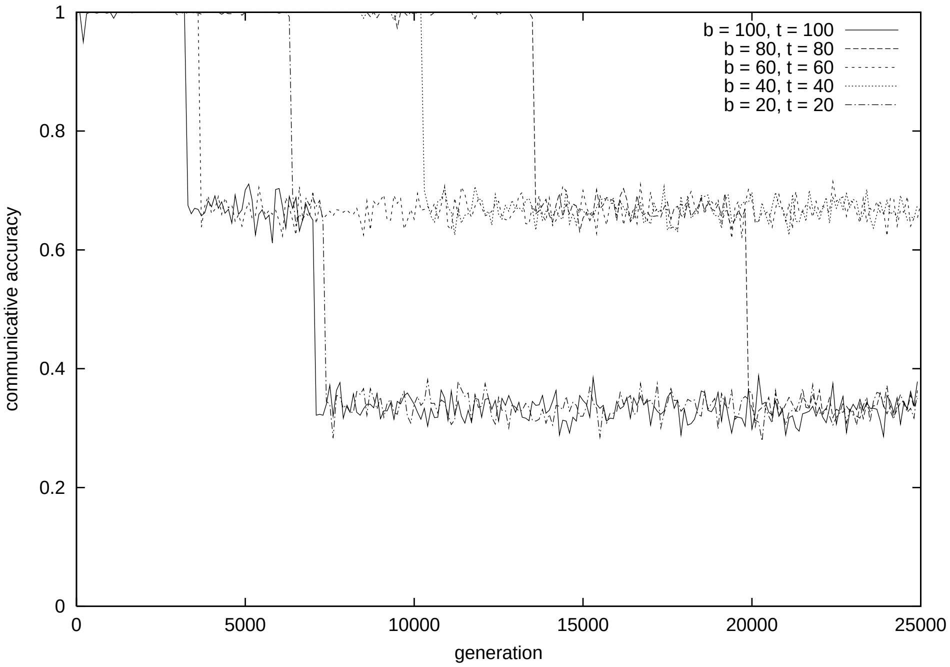


Figure 4.9: Slumping from optimal to suboptimal communication systems in imitator populations in the EILM.

while such individuals are unlikely to breed, their communication systems may be observed and learned from by agents in the next generation, depending on t. Genetic drift therefore introduces cultural variation into the population. This cultural variation then results in the reintroduction of directly biased cultural transmission, resulting from learning bias, which leads to the spread of ambiguous communication systems.

Note that $e = 200$ represents the best-case scenario for imitator agents. $e = 200$ results in the highest level of learnability for unambiguous communication systems and also ensures that, at the early stages of collapse, suboptimal communication systems will constitute only a small part of a learner's observations. In populations where $e \leq 2$ the collapse phenomenon does not occur, but as discussed above the behaviour of these populations is entirely determined by natural selection — they cannot truly be called learning populations.

4.4.6 Summary

The addition of natural selection acting on the genetic transmission of initial connection weights has rather different effects in imitator and obverter populations. In both cases, the learning bias inherent in the two network architectures is still the best predictor of the populations' eventual communication system. In imitator populations, this learning bias favours many-to-one mappings from meanings to signals. The imitator populations therefore converge on fully ambiguous mappings, unless cultural transmission has very little impact (e is low). Furthermore, imitator populations cannot maintain optimal systems in this scenario, due to phenotypic noise introduced by genetic drift.

In obverter populations, in contrast, the addition of natural selection acting on the genetic transmission of initial connection weights has positive effects. When cultural transmission is weak, the addition of natural selection leads to the emergence of initial connection weightings which help establish an optimal, shared communication system. However, this effect diminishes as the strength of cultural transmission increases. Furthermore, comparison of the behaviour of imitator and obverter populations reveals that the correct overall learning bias has to be in place before such beneficial effects are observed.

4.5 Model 2: Adding genetic transmission to the associative network model

While the results outlined in the previous section give an interesting insight into the interaction between directly biased cultural transmission and genetic transmission, they are somewhat unsatisfactory for a number of reasons. There are two forms of direct bias in the feedforward network EILM — one introduced by the network architecture (imitator or obverter), which is externally determined, and one introduced by the genetic transmission of initial connection weights. This second, and much weaker, bias is the only element manipulable by natural selection. We would like the primary bias to be genetically transmitted, to investigate whether biases favouring optimal communication can evolve.

To this end, the associative network model outlined in Section 3.4 is extended to include the genetic transmission of weight-update rules. A small part of this research (Section 4.5.5 in particular) appears in Smith (2001b). Rolls & Stringer (2000) take a similar approach to investigating the evolution of learning, although in their model learning is individual, rather than cultural.

4.5.1 Genotypes, phenotypes and the genotype-phenotype mapping

The phenome is the associative network model of a communicative agent outlined in Section 3.4 of Chapter 3, with $|\mathcal{N}_M| = |\mathcal{N}_S| = 10$. This phenome is defined by the pair $\langle \mathcal{W}, W \rangle$, where $\mathcal{W}$ is an initial set of connection weights and W is a weight-update rule. The genome is a 4-locus chromosome. A genotype is specified by the 4-tuple $(a_\alpha\ a_\beta\ a_\gamma\ a_\delta)$ where a_x is an allele drawn from the set $\{-1, 0, 1\}$.

The process of mapping from a genotype to a phenotype involves converting such a 4-locus chromosome into a $\langle \mathcal{W}, W \rangle$ phenotype. As discussed in Section 3.4.1.3, each weight-update rule W is specified by a 4-tuple $(\alpha\ \beta\ \gamma\ \delta)$. During genotype-phenotype mapping α is set to the value of allele a_α, β is set to the value of allele a_β and so on. The genotype therefore specifies the phenotype's weight-update rule. All $w_{i,j} \in \mathcal{W}$ are set to 0 — every agent has all their initial connection weights set to 0. It would be possible to specify $\mathcal{W}$ in the genotype, as was done in the feedforward network model. However, that option was not explored — the results from the feedforward network model suggest that the genetic evolution of initial connection weights will be of fairly minor importance in comparison to the evolution of the learning bias itself.

To recap, there are 81 possible genotypes, which encode the 81 possible weight-update rules discussed in Chapter 3, Section 3.4. Of these 81 rules, 50 are classified as [$-$learner, $-$maintainer, $-$constructor], 13 are classified as [$+$learner, $-$maintainer, $-$constructor], 9 are classified as [$+$learner, $+$maintainer, $-$constructor] and 9 are classified as [$+$learner, $+$maintainer, $+$constructor].

4.5.2 Reproduction

Individuals inherit their genes from their parents. As with the feedforward network model, organisms are haploid but sexual recombination (involving crossover, in an identical fashion to that outlined for the previous model) is used. Newly-formed genotypes are also subject to mutation.[8] Given the short chromosome it is not clear whether sexual reproduction is necessary to maintain genetic diversity in the population, or whether mutation is sufficient. Simulation runs similar to those described here were carried out with asexual reproduction (each individual has a single genetic parent) and the results were found to be qualitatively similar.

4.5.3 The EILM algorithm

As with the associative network ILM discussed in Chapter 3, a gradual population turnover model is used in the associative network EILM. The EILM consists of initialisation and iteration processes.

[8] Point mutations occur on the newly-formed genotype with probability p_m ($p_m = \frac{0.04}{l_g}$ for all simulations outlined in this section, where l_g is the length of the genome.) Mutation involves replacing the allele a_i at the mutated locus with another allele $a_{j \neq i}$, where a_j is selected from the set of possible alleles.

Initialisation Create a population of N agents[9]. Each initial agent has a random genotype, with the allele at each locus selected randomly from the range of possible alleles. Each initial individual's phenotype is determined by their genotype and the genotype-phenotype mapping.

Iteration

1. Select an individual from the population according to the death procedure outlined below and remove it.
2. For every remaining member of the population, generate a set of meaning-signal pairs by applying the network production process to every meaning $m \in \mathcal{M}$.
3. Create a new agent. The new agent inherits its genotype from its parents, who are selected from the population according to the reproduction procedure outlined below.
4. The new agent receives e exposures to the population's observable behaviour. During each of these e exposures the new agent observes the complete set of meaning-signal pairs of a randomly selected member of the population and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W.
5. The new agent joins the population. Return to 1.

The iteration process requires procedures for selecting which individuals die and which individuals reproduce. In the ILM version of the model death was random, and reproduction was effectively ignored. In the feedforward network EILM reproduction occurred with equal probability among the fittest b members of the population. This rank selection procedure would be acceptable for the gradual EILM described here. However, for reasons of computational efficiency [10] a *tournament selection* procedure was used. During a tournament, T individuals are selected from the population at random and evaluated. Each individual is scored according to their average communicative accuracy when acting as both producer and receiver with two randomly selected partners according to the measure $ca\left(P, R\right)$ given in Section 3.2 in Chapter 3. During selection to decide death, the individual with the *lowest* communicative accuracy from among the T selected individuals 'wins' the tournament and is removed from the population. During selection to decide reproduction, the individual with the *highest* communicative accuracy wins the tournament and reproduces.

[9] $N = 100$ for all models outlined in this section

[10] As a new individual is introduced at every cohort, communicative accuracy for the entire population would have to be reevaluated at every cohort if the rank procedure was used.

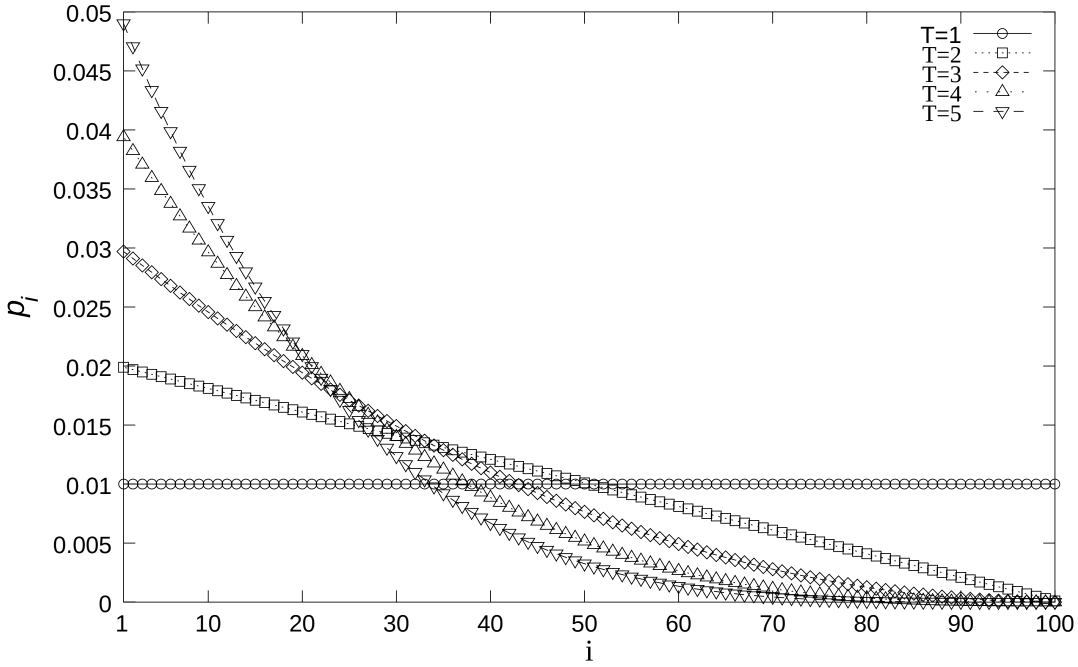


Figure 4.10: Probability of breeding or dying according to rank fitness, for various values of T. i gives the rank fitness of an individual and p_i gives the probability of that individual winning a tournament. For death tournaments, the *least* fit individual has rank 1 and the most fit individual has rank 100. For breeding tournaments, the *most* fit individual has rank 1. Based on the equation from Bäck (1994): $p_i = N^{-T}\left((N-i+1)^T - (N-i)^T\right)$

The strength of selection pressure in the population therefore depends on T. Figure 4.10 shows the probability of individuals of different rank fitness reproducing for various values of T. $T = 3$ was selected for the simulations outlined in this section — it offers a reasonable selection pressure, while still allowing less fit individuals to reproduce, therefore helping to maintain genetic diversity in the population.

One note should be made here regarding the natural selection of cultural variants. The individual who is removed from the population at each cohort does not produce observable behaviour. This was due to reasons of continuity with previous work (e.g. Oliphant (1999)) and simplicity of coding. However, given that removal is dependent on communicative accuracy in the EILM, this introduces natural selection of cultural variants during cultural transmission — less successful communicators are more likely to be removed prior to cultural transmission, and therefore unsuccessful communication systems are selectively removed from the population's culture. It turns out that this does not affect the results reported here. Firstly, it is only a very weak selection pressure (see Figure 4.11). Secondly, as we saw in Chapter 3 with respect to the feedforward network model,

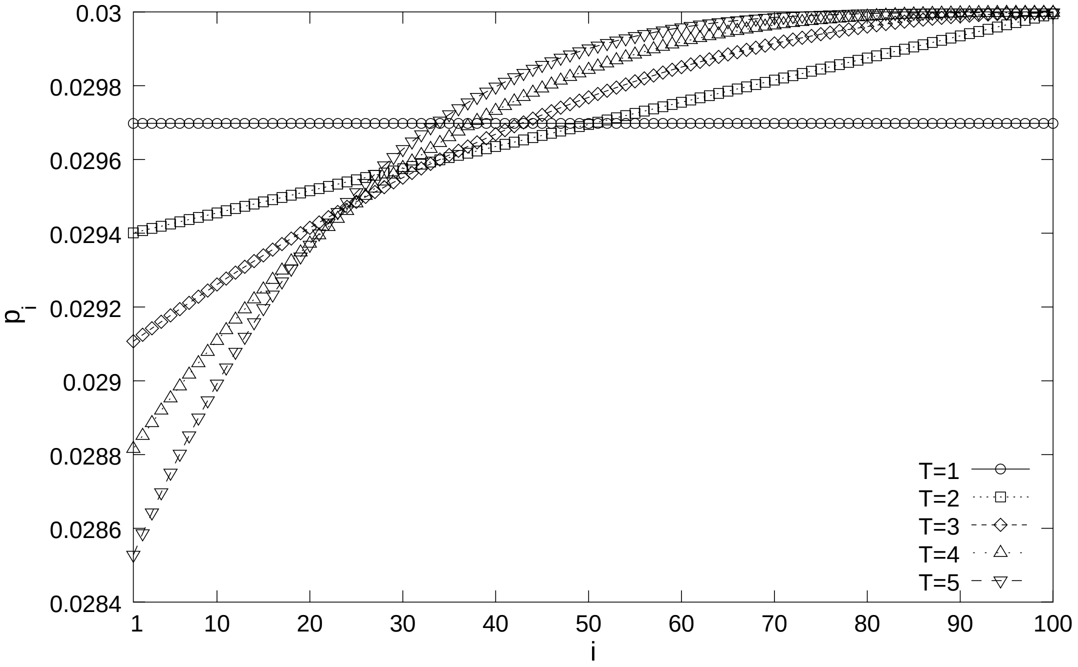


Figure 4.11: Probability of acting as a teacher according to rank fitness, for various values of T. Note the narrow range of the y-axis. i gives the rank fitness of an individual and p_i gives the probability of that acting as a cultural parent. Here the *least* fit individual has rank 1 and the most fit individual has rank 100. The probability of an individual acting as a cultural parent is given by the probability of them not dying multiplied by the probability of being selected to act as a cultural parent — $p_i = \left(1 - \left(N^{-T}\left((N-i+1)^T - (N-i)^T\right)\right)\right) \cdot \left(1 - \left(1 - \frac{1}{N-1}\right)^e\right)$. For the plots shown here, $N = 100$, $e = 3$. There is little difference between the probabilities of the least fit and most fit individuals acting as cultural parents — there is very little natural selection of cultural variants.

natural selection of cultural variants tends to be drowned out by the pressures introduced by learning bias.

4.5.4 Main result: optimal communication rarely emerges

100 runs of the EILM were carried out for $e = 3$ (the same amount of exposures as used in the ILM runs outlined in Section 3.4). Each run proceeded for 5000 cohorts. Figure 4.12 shows the progress of a run of the simulation where the population constructed an optimal[11] communication system. This is a typical example of a successful run. However, only 3 of 100 runs were successful in this respect. Natural selection does not reliably lead

[11] The definition of "optimal" was weakened in comparison to the ILM construction tests, to allow for the fact that populations would be genetically heterogeneous and therefore might suffer reduced communicative accuracy due to the presence of non-learners. A population was considered as having converged on an optimal system if communicative accuracy exceeded 0.95 within the 5000 cohorts, and once communicative accuracy exceeded this threshold it remained above it. Weaker definitions of optimal systems yield qualitatively similar results.

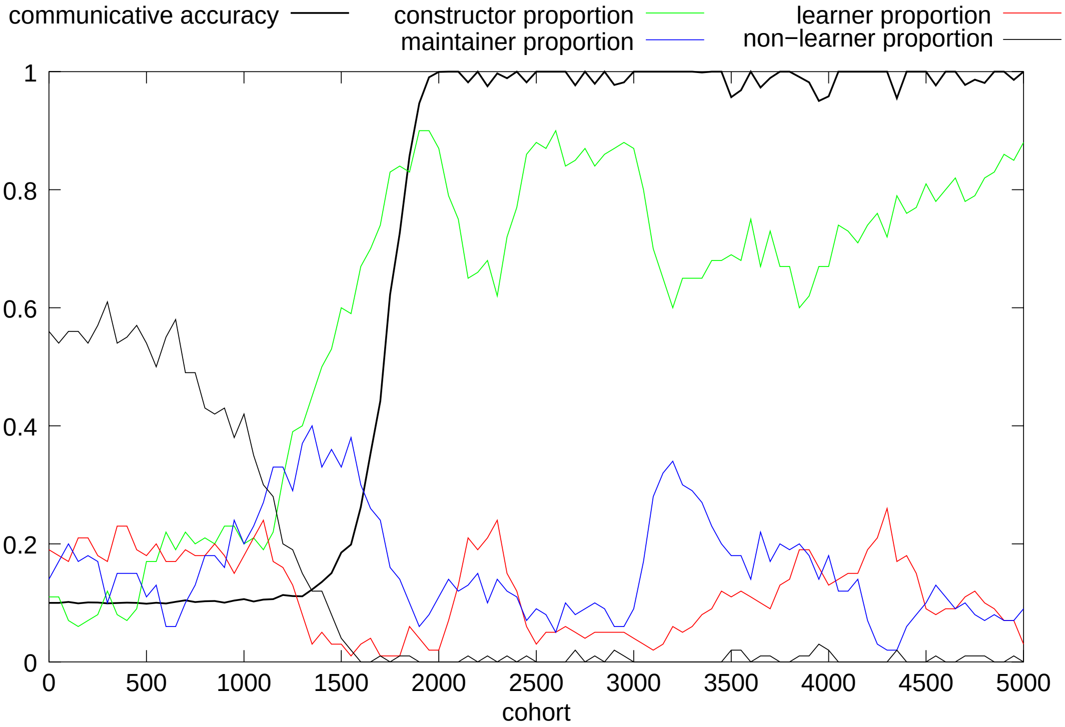


Figure 4.12: The evolution of learning biases leading to optimal communication. Proportions of all the classes of genotypes are given (for example, the black line gives the proportion of the population who use one of the 50 weight-update rules classified as [−learner]). This is a plot of a successful run of the EILM. However, such successful runs are rare.

to the evolution of optimal communication — optimal systems were found to be unlikely to emerge.

Why does natural selection during genetic transmission have such difficulty in identifying weight-update rules which lead to optimal communication? The problem is that the cultural construction of a communication system takes time. Individuals with [+constructor][12] weight-update rules gradually converge, over hundreds of cohorts, on increasingly unambiguous areas of communication system space. In the early stages

[12] I will be frequently referring to the various classifications of weight-update rules for the remainder of this Chapter, and writing out the full [±learner, ±maintainer, ±constructor] specification becomes somewhat laborious. I will therefore adopt two simplifying conventions, one which exploits the hierarchical arrangement of the classification system, and one which is more intuitive. The conventions are summarised below:

Full specification	Abbreviated specification	Name
−learner, −maintainer, −constructor	−learner	non-learner
+learner, −maintainer, −constructor	+learner, −maintainer	learner
+learner, +maintainer, −constructor	+maintainer, −constructor	maintainer
+learner, +maintainer, +constructor	+constructor	constructor

of this construction process individuals whose genotypes encode [+constructor] weight-update rules will have little fitness advantage over other individuals. As a consequence the genetic transmission process will be essentially random — the population will undergo genetic drift. In successful runs, such as that shown in Figure 4.12, genetic drift preserves genotypes encoding [+constructor] rules, by chance, in sufficient numbers for sufficient time to allow the construction process to get well under way. Individuals with genotypes encoding [+constructor] rules then show increased communicative accuracy, which leads to steady selection for such genotypes, which increase in number. In the more common, unsuccessful runs, drift does not provide constructor agents in sufficient numbers for sufficient time for the construction process to get under way.

Interestingly, when the population's communicative accuracy nears optimal levels, selection for genotypes encoding [+constructor] weight-update rules stops. The population enters a second stage of genetic drift, where constructor numbers fluctuate randomly. This is due to the fact that maintainers are capable of putting the finishing touches on a communication system and maintaining that system once it is established. Learner agents are also able to acquire the population's optimal system once it is established. Constructor agents lose their fitness advantage over maintainers and learners and genetic transmission becomes semi-random once more. However, non-learners never drift back into the population — they are incapable of learning an optimal system and suffer a severe fitness penalty.

This three-stage drift-selection-drift pattern is common over all successful runs and is illustrated schematically in Figure 4.13. In Figures 4.14–4.17 the relative communicative accuracy[13] of the four classes of genotypes are plotted against their change in number, with the population's average communicative accuracy and the proportion of constructor agents plotted for reference. As can be seen from Figure 4.14, the relative communicative accuracy (henceforth *rca*) of constructor agents fluctuates around 1 till approximately 1000 cohorts through the simulation. We might be tempted to think, given that constructor *rca* is sometimes slightly above 1, that constructors are being selected for. However, this is confounded by the fact that increases in constructor numbers brought about by drift lead to an increase in the communicative accuracy of constructor agents — drift is the cause of the fluctuations in constructor *rca*. More reliable indicators are given by the *rca* values of maintainers, learners and non-learners, for whom this effect is weaker

[13] The relative communicative accuracy of a particular weight-update rule is the communicative accuracy of individuals with that weight-update rule divided by the average communicative accuracy of the population. The relative communicative accuracy of a classification (constructors, maintainers etc) is the average of the relative communicative accuracies for all weight-update rules which fall into that classification. Relative communicative accuracy of greater than 1 indicates above-average performance.

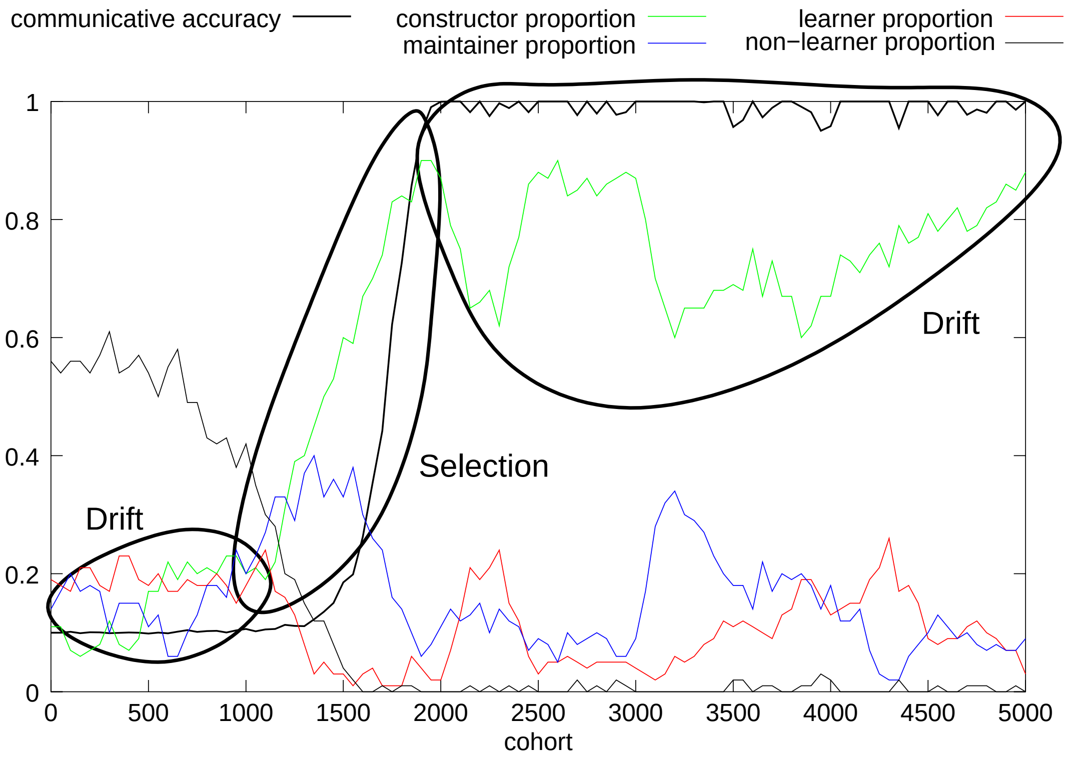


Figure 4.13: The three phases of a successful run. Drift establishes constructor agents in sufficient numbers for the construction process to get underway. Constructor genotypes are then preferentially selected, and communicative accuracy increases rapidly. When an optimal system is established, constructors loose their selective advantage over maintainers (and, to a lesser extent, learners), and a second period of drift occurs.

or non-existent. Learners and non-learners in particular show *rca* of around 1 until 1000 cohorts into the run, by which point the mini-plateau of constructor agents has been established by drift. At this point the *rca* of these individuals plunges, overall communicative accuracy increases markedly and constructor numbers increase due to natural selection for constructor genotypes. After communicative accuracy reaches 1, the *rca*s of learner, maintainer and constructor individuals hover at 1, with non-learners having very low *rca* whenever they are reintroduced by crossover or mutation.

The second period of genetic drift follows the predictions of B&R — once the population reaches cultural convergence, the selection pressure for biases disappears. However, in this model there is still strong selection pressure against [−learner] individuals, who cannot acquire the cultural variant in use in the population.

The first period of drift is not predicted by B&R's mathematical model, which suggests that, given costless learning, the biased allele should increase in frequency when the population is culturally heterogeneous. This turns out not to be the case in this model, for

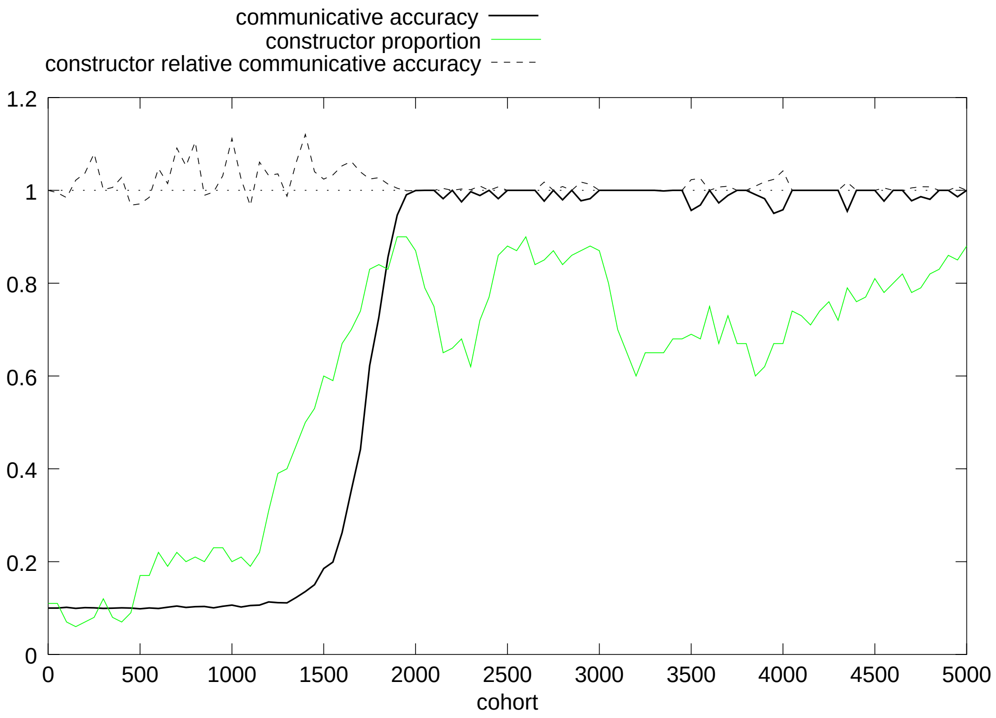


Figure 4.14: The relative communicative accuracy of agents with constructor genotypes. This value fluctuates around 1, although it is most obviously above 1 from 1300 to 1900 cohorts.

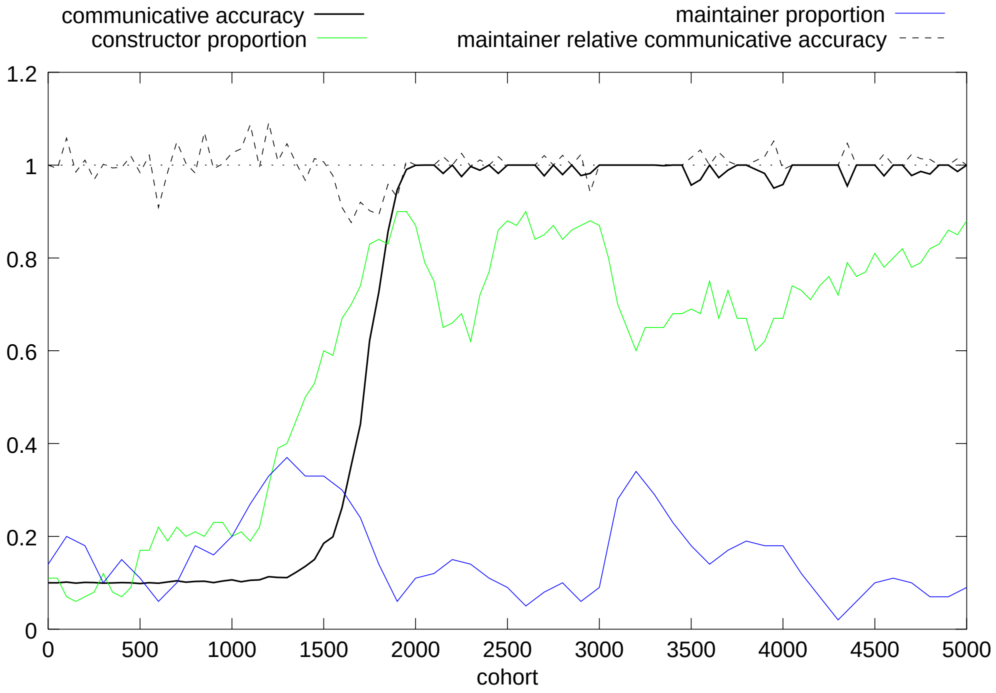


Figure 4.15: The relative communicative accuracy of agents with maintainer genotypes. This value fluctuates around 1, and is obviously below 1 from 1500 to 1900 cohorts, during which time maintainer numbers drop sharply.

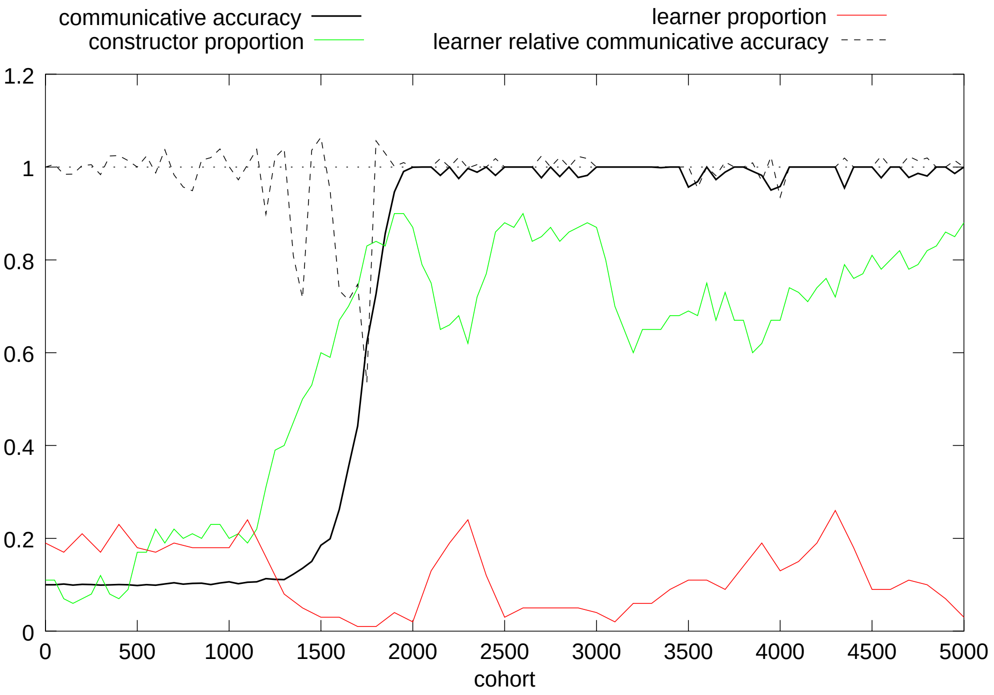


Figure 4.16: The relative communicative accuracy of agents with learner genotypes. This value fluctuates around 1, and drops most markedly below 1 from 1100 to 1800 cohorts, during which time learner numbers drop sharply.

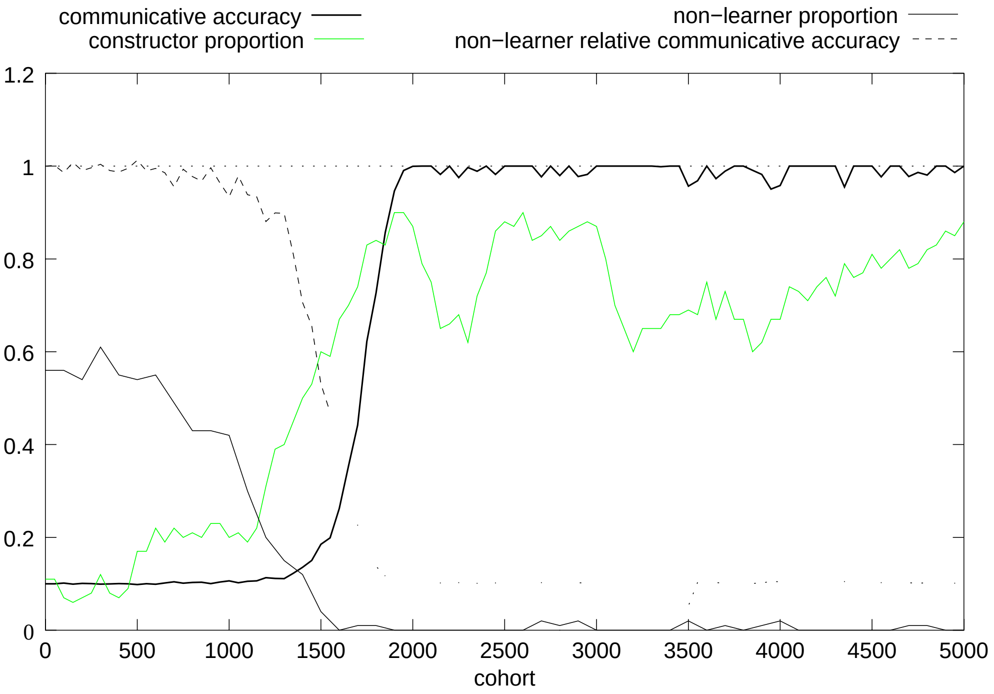


Figure 4.17: The relative communicative accuracy of agents with non-learner genotypes. This value initially fluctuates around 1, and appears to be below 1 from 600 cohorts onwards. By this time, drift has already delivered a significant number of constructor agents. The *rca* for non-learners drops markedly after about 1000 cohorts, at which point constructor numbers shoot up. After the optimal system is established, non-learner numbers and *rca* remain very low.

the initial period until drift establishes [+constructor] individuals in sufficient number. This is due to the fact that the fitness payoff in this model, unlike in B&R's model, is frequency-dependent — if no other individuals have the same cultural variant as you, you receive no payoff.

These simulation results emphasise a fairly basic logical point — if there is no established useful communication system present in the population, natural selection will not favour individuals who are predisposed to learn such a system. In the simulation results shown here, genetic drift is required to break the cycle of non-communication, at which point being biased to acquire an optimal communication system becomes advantageous. However, the point remains that, in a non-communicating population, being able to learn an optimal communication system confers no advantage.

4.5.5 Varying the speed of convergence by varying e

The speed of convergence of [+constructor] populations on optimal systems depends on e, the number of exposures learners receive to the observable behaviour of the population. In general, larger values of e lead to more rapid cultural convergence in the ILM. The results outlined above suggest that larger values of e will result in a higher proportion of runs of the EILM converging on optimal systems. For higher values of e there will be a shorter time-lag between emergence of genotypes encoding [+constructor] rules and an increase in communicative accuracy for individuals using such rules, therefore the populations will be less vulnerable to genetic drift during the abbreviated period of construction.

The experiments outlined in the previous section were repeated for all integer values of e in the range $[1, 30]$. The proportion of runs (out of 100) converging on optimal systems are plotted in Figure 4.18, along with the average speed of construction for those values of e in pure ILM runs. As can be seen from this Figure, there is a clear relationship between speed of construction in the ILM and probability of convergence on an optimal system in the EILM — as e increases, speed of construction decreases and probability of convergence in the EILM increases as a consequence. The correlation between time to convergence in the ILM and proportion of convergent runs in the EILM is statistically significant ($r = -0.56$, $p = 0.01$).

4.5.6 Varying the speed of convergence by varying cultural population size

Oliphant (1999) reports that population size impacts on the speed of cultural convergence, with larger populations taking longer to converge culturally on optimal systems. More

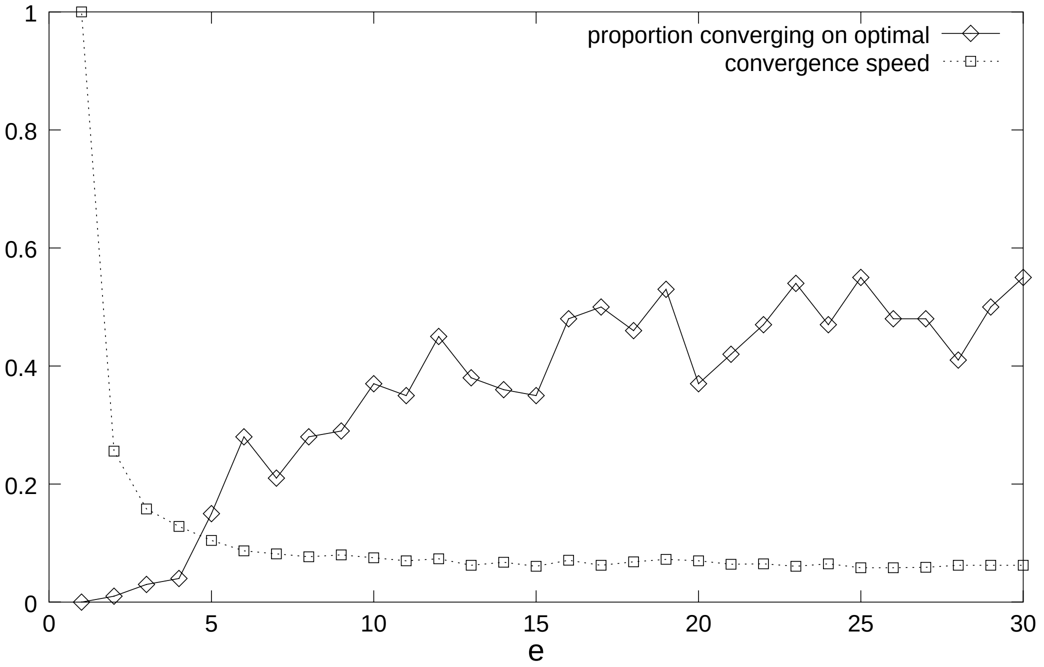


Figure 4.18: The relationship between e, convergence speed in the ILM and probability of convergence on an optimal system in the EILM. As e increases, the time taken by runs of an ILM to converge on a stable system decreases (speed of convergence is averaged over 10 runs and shown as a proportion of the time taken by the slowest converging runs). In the EILM, increased speed of convergence reduce the reliance on genetic drift, and consequently as e increases the proportion of EILM runs converging on an optimal communication system increases.

generally, we would expect cultural innovations to take longer to diffuse through larger populations, assuming that the rate of diffusion is invariant with respect to population size. Given the relationship between speed of convergence in the ILM and probability of converging on an optimal system in the EILM, we might expect smaller population sizes to lead to a higher probability of convergence in the EILM. However, this prediction is complicated by the genetic drift phase of the EILM described above. Genetic drift is more pronounced and its effects more dramatic in smaller populations, which would be a confounding factor in repeating the EILM experiments outlined above for smaller populations.

In order to decouple the effects of genetic population size and cultural population size, a model of spatial organisation was introduced to the EILM. The model of spatial organisation outlined in Oliphant (1997) is adopted for our purposes. The population of N agents is organised in a ring. This allows us to define a distance measure between individuals in the population. In the case where the population is unorganised spatially the probability of picking a particular individual is independent of their position in the population. However, in the spatial organisation case this is a function of their distance from a particular focal position on the ring.

There are four processes which involve picking individuals from the population:

1. picking individuals to participate in death tournaments.
2. picking individuals to participate in breeding tournaments.
3. picking individuals to communicate with.
4. picking individuals to observe and learn from.

Individuals are picked at random to participate in death tournaments, irrespective of their position in the population. Each of the remaining three processes is given a spatialisation parameter, d_b, d_c and d_l. The processes then proceed as follows:

Breeding Tournaments: Breeding produces a single individual to replace the single dead individual, who occupied slot i. The T individuals who will participate in the tournament are selected from the population probabilistically, according to their distance from position i. f is a normal distribution with a mean of 0 and standard deviation d_b. The probability of an individual with distance j from individual i being picked is equal to $f(j)$ — if d_b is low, only individuals close to i will be picked, whereas if d_b is high, individuals who are distant from i may be picked. The offspring produced by breeding occupies slot i.

Communication: Individual i is scored according to their average communicative accuracy when acting as both producer and receiver with two partners according to the measure $ca(P, R)$ given in Section 3.2 in Chapter 3. The partners are selected according to a normal distribution around individual i with standard deviation d_c.

Learning: Individual i receives e exposures to the population's communicative behaviour. During each of these e exposures the learner observes the complete set of meaning-signal pairs of a member of the population, where that member of the population is selected according to a normal distribution around individual i with standard deviation d_l.

For low values of d the population will be highly organised spatially, with individuals interacting only with individuals who are close by in terms of the distance measure. As d increases the population becomes less spatially organised, with the probability of interacting with distant individuals increasing. For $N = 100$ and $d \approx 50$ the population is effectively completely unorganised spatially with respect to the process of interest (breeding, communication or learning).

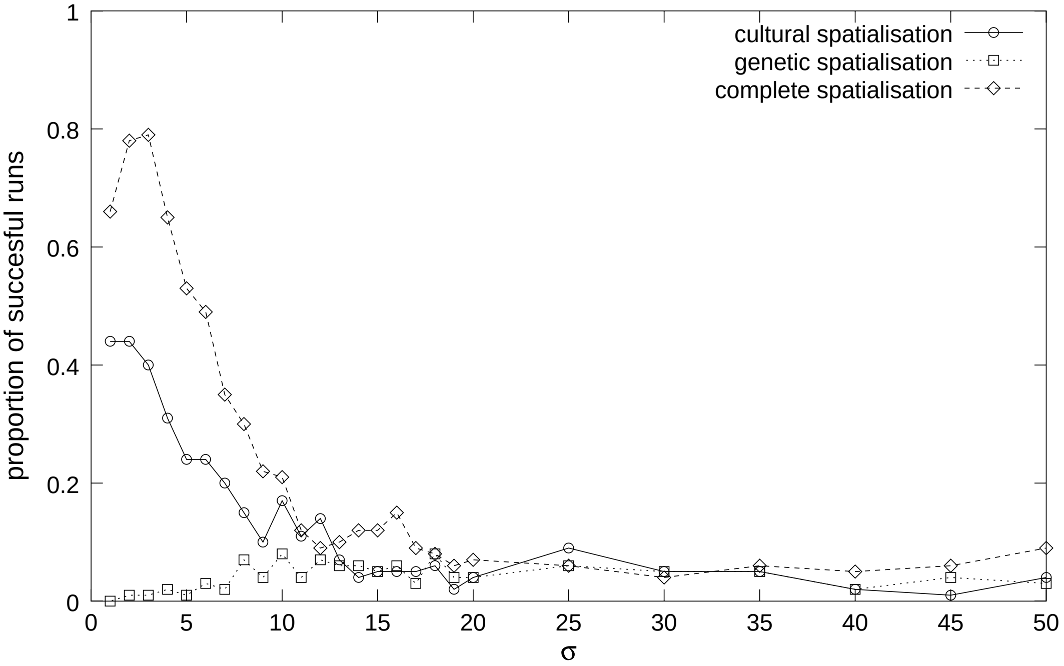


Figure 4.19: The impact of spatial organisation. σ gives the spatial organisation parameter, for those processes which are spatially organised. As σ increases, spatial organisation decreases. When cultural transmission is spatially organised, the proportion of EILM runs converging on an optimal system decreases, due to the increased time to cultural convergence. Conversely, when genetic transmission but not cultural transmission is spatially organised, the increase in σ results in a slight increase in the proportion of successful EILM runs, due to less fierce genetic drift. When genetic and cultural transmission are both spatially organised, a cultural kin-selection effect is observed, which tails off as σ increases.

I will consider three combinations of parameters:

Cultural spatialisation: $d_l = d_c = \sigma$, $\sigma \in [1, 50]$,$d_b = 50$
Genetic spatialisation: $d_b = d_c = \sigma$, $\sigma \in [1, 50]$,$d_l = 50$
Complete spatialisation: $d_l = d_b = d_c = \sigma$, $\sigma \in [1, 50]$

Figure 4.19 plots the proportion (out of 100) of populations which converge on optimal systems[14], for the three parameter conditions for various values of σ (as with the earlier ILMs, $N = 100$, $e = 3$).

Three patterns are clear here. Firstly, the number of convergent runs in the genetic spatialisation condition is very low, and is at its lowest when genetic spatialisation is at its most extreme. This is what we would expect — genetic drift in the sub-populations will

[14] Runs were classified as optimal if communicative accuracy reached and subsequently remained above 0.85. This value was lower than the previous classification values to allow for the reduction in communicative accuracy resulting from boundaries between communication systems.

be more pronounced, and cultural convergence will still take a long time. Secondly, the proportion of convergent runs in the cultural spatialisation condition decreases as σ increases — as σ increases, effective population size increases, resulting in slower cultural convergence and hence greater vulnerability to genetic drift.

Finally, the runs for the complete spatialisation case exhibit a similar pattern to the cultural spatialisation case, but with a higher overall proportion of convergence and a hump for very low σ. The hump is due to extreme genetic drift when $d_b = 1$. The overall higher level of convergence suggests a positive interaction between genetic and cultural spatialisation. Neighbouring individuals will typically have similar genotypes, will have learned from similar observable behaviour and will have similar communication systems. If individuals are of the [$+$constructor] classification, this is likely to result in very quick convergence on shared meaning-signal mappings. Sharing meaning-signal mappings leads to increased communicative accuracy, increased probability of breeding, and increased chances of adding yet more constructor agents into the constructor sub-population, who will then learn from similarly biased individuals, and subsequently communicate successfully with them.

Figures 4.20, 4.21 and 4.22 give sample convergent runs in more detail. As can be seen from Figure 4.20, tight cultural spatial organisation leads to the rapid formation of dialects within the population, which yield above-average communicative accuracy. Figure 4.21 shows that these spatially-organised dialects do not form when the population is not spatially organised for cultural transmission. A similar relationship between cultural spatial organisation and dialect formation is reported by Livingstone & Fyfe (1999). Finally, Figure 4.22 shows that, when both genetic and cultural transmission are spatially organised, the emergent, useful dialects are associated with constructor genotypes, which then spread to dominate the population.

4.5.7 Summary

The simulation results outlined in this Section show that natural selection cannot reliably identify weight-update rules which lead to the cultural evolution of optimal communication. This is due to the time-lag between the emergence of such weight-update rules and any communicative payoff to individuals possessing them. In simulation runs which do converge on optimal communication systems, genetic drift provides appropriate weight-update rules in sufficient numbers for the cultural construction process to get under way, at which point constructor weight-update rules are identified and selected for during genetic transmission. Increasing the number of observations each learner makes (increasing

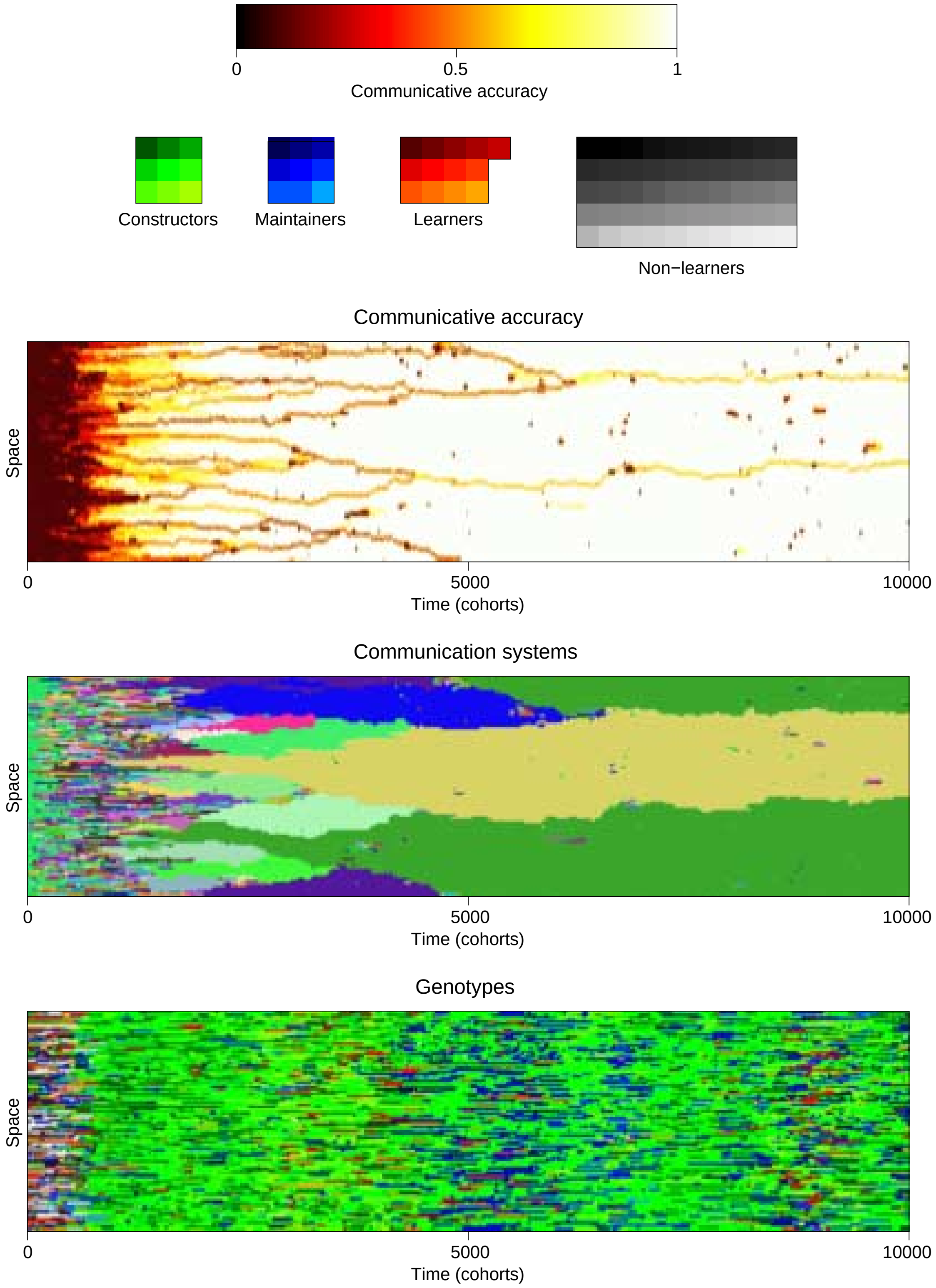


Figure 4.20: Cultural spatialisation only ($d_l = d_c = 1$, $d_b = 50$). Time-space diagrams for communicative accuracy (top), communication systems (middle) and genotypes (bottom). Each vertical strip of each diagram represents the complete population at a particular point in time (in cohorts). In the communicative accuracy diagram, different colours correspond to individuals with different levels of communicative accuracy (see key). In the communication systems diagram, each distinct $p(m)$ is assigned a distinct (random) colour. In the genotypes diagram, different colours correspond to genotypes encoding different weight-update rules (see key).

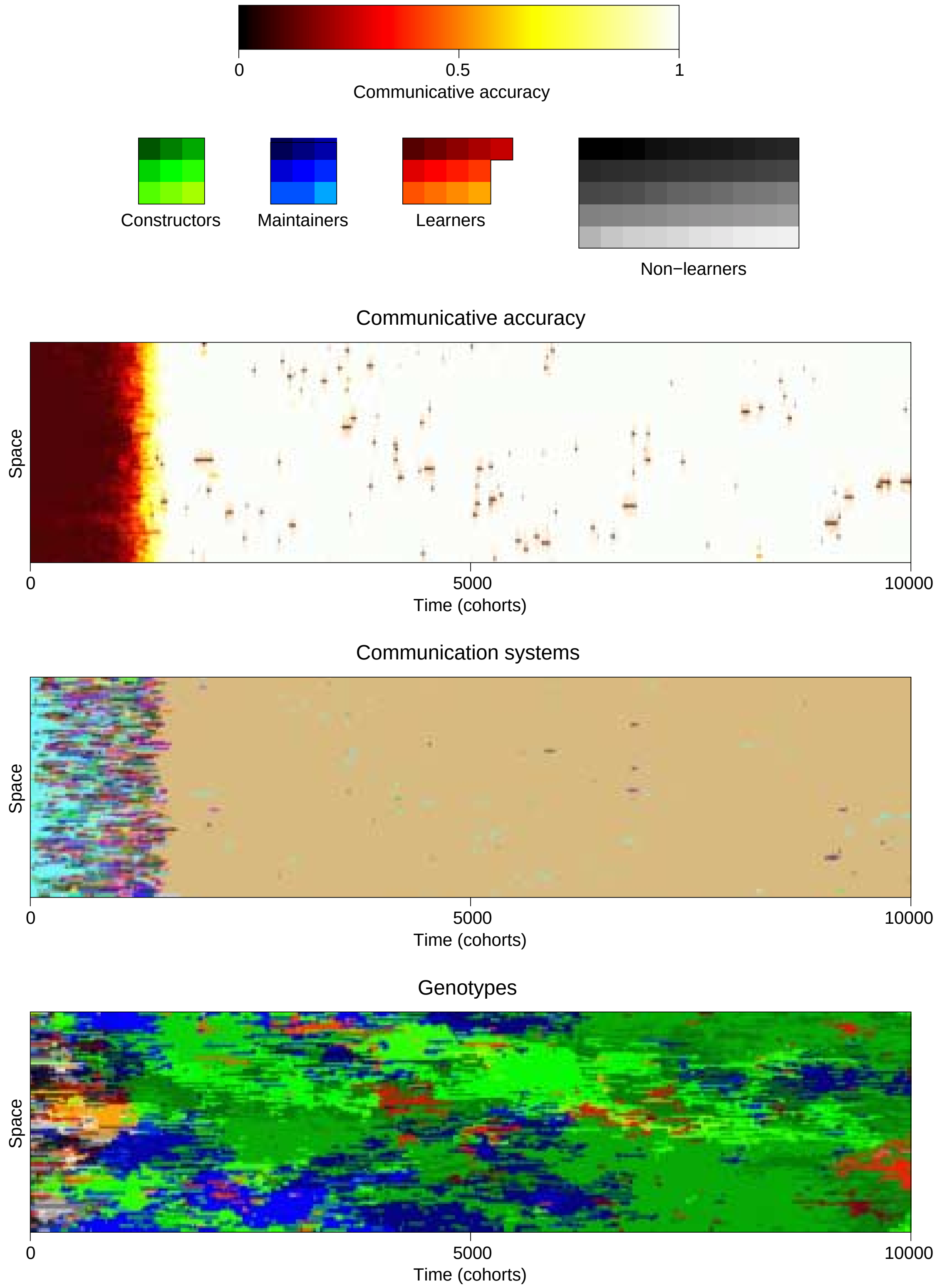


Figure 4.21: Genetic spatialisation only ($d_b = d_c = 3$, $d_l = 50$). While genetic subpopulations exist, diversity of communication systems is rapidly lost.

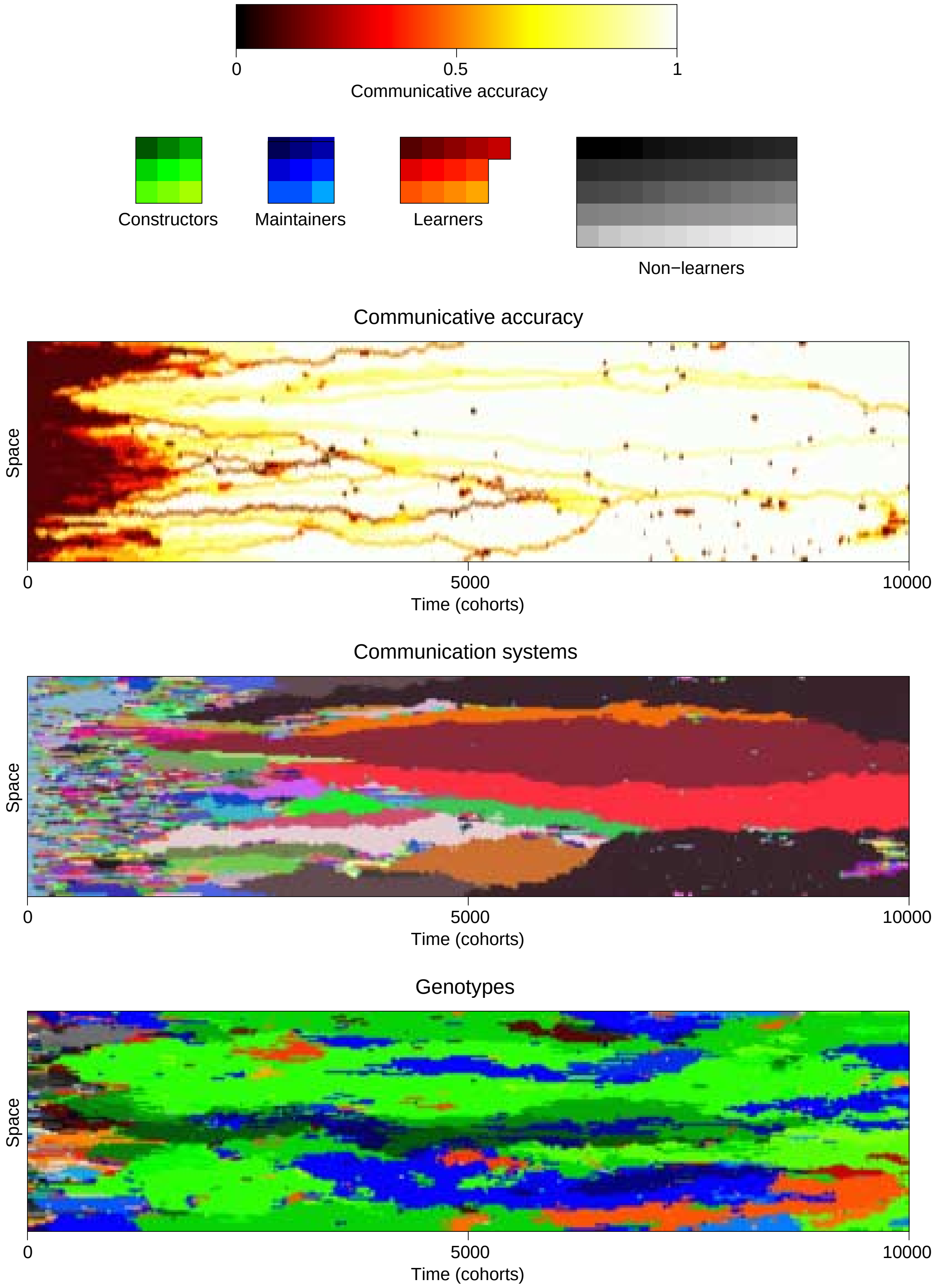


Figure 4.22: Combined spatialisation ($d_l = d_b = d_c = 1$). Both cultural and genetic diversity is maintained. Note that the early regions of increased communicative accuracy coincide with regions dominated by constructor genotypes. Later on, dialects are less associated with a particular genetic makeup. Note also the small subpopulation of learner agents, emerging at around 6000 cohorts and disappearing at 10000 cohorts. This build-up of [+learner, −maintainer] individuals leads to a drop in communicative accuracy, visible in the top diagram, and consequently the elimination of the learner genotypes.

e) or reducing the cultural population size (by introducing cultural spatialisation) speeds up the cultural convergence process, resulting in more frequent convergence on good genes and optimal communication systems. In successful runs, a kind of cultural niche construction (Odling-Smee 1988; Laland *et al.* 2000) takes place — drift provides constructor genotypes in sufficient numbers for sufficient time to construct a cultural niche, which constructor agents are then selected to occupy. However, the logical point remains that in a non-communicating population there can be no communicative advantage in being biased to acquire an optimal communication system.

4.5.8 Discussion

In Chapter 3 I argued that humans are biased in favour of acquiring one-to-one mappings between meanings and vocabulary items, and that this learning bias could result in the emergence of effective communication in human populations through purely cultural processes. The fairly natural assumption would be that this learning bias in human infants has evolved through natural selection in favour of communication. However, the results outlined in this Chapter cast doubt on this assumption — in the simulation model, such biases are unlikely to evolve, in spite of strong selection pressure in favour of successful communication.

What can these results tell us about the evolution of vocabulary acquisition biases in humans? There are two possible interpretations. I will take as the starting-point for both interpretations the assumption that pre-linguistic human populations resembled current-day primate populations in terms of communicative behaviour — they possessed some fairly limited set of stimulus-bound, innate signalling behaviours, possibly augmented by a few idiosyncratic, ontogenetically-ritualised signals shared between pairs of individuals. It should of course be remembered that non-human primates have undergone a significant period of evolution since we last shared a common ancestor, and do not constitute some kind of living fossil. However, modern humans still possess a small range of universal, innate signals (facial expressions for disgust, anger and so on) and it is reasonable to suggest that the basis of this common characteristic existed in the last common ancestor of chimpanzees and humans, and therefore in the first non-common ancestor on our side of the family tree. I will further assume that any emergent, culturally-transmitted symbolic communication system did not replace this innate, non-linguistic signalling system — the persistence of human facial expressions and the like shows that it did not.

Under the first, positive interpretation, the simulation results outlined in this Chapter can be seen as highlighting the conditions under which the learning bias in humans must have evolved. We can then draw parallels between the simulation parameters which lead most reliably to the evolution of appropriate learning biases and the hypothesised ecological setting of human evolution. Alternatively, we could emphasise the negative aspect of the simulation results, and mount an argument that the human vocabulary acquisition biases, apparently so well-designed for communication, are better understood as an exapted trait. These two contrasting interpretations are considered in the following two Sections.

4.5.8.1 The positive interpretation

The weakest positive interpretation of the simulation results is to conclude that the evolution of the human vocabulary acquisition bias was something of an evolutionary fluke — the simulation results indicate that such learning biases are unlikely to evolve but, given a large dose of luck, can do so. Under this interpretation, the uniqueness of human language as a learned symbolic communication system (at least among the primates) can be seen as a consequence of the fortuitous emergence of a significant number of individuals possessing a language-specific learning bias with the correct properties. After the cultural construction process got underway, these individuals reproduced with disproportionate success, due to the payoff of communication, and the genes encoding the bias became fixed in the population.

This kind of account is rather unsatisfactory, relying as it does on a double saltation. Firstly, the appropriate learning bias must emerge de novo in a population whose communication system was previously innate. Secondly, this learning bias must be maintained in a significant proportion of the population for a significant period of time, despite a lack of fitness advantage for those individuals possessing it. Of course, assuming that the appropriate bias is costless relative to inappropriate biases, genetic drift may preserve it, or even increase its numbers. However, given that the genes encoding the new bias will initially occur in small numbers, drift is more likely to result in their disappearance than their proliferation.

As a second possible interpretation, the simulation results could be taken as indicating that the emergent learning bias in human populations immediately provided a fitness advantage to those individuals who possessed it — in this case, the time-lag problem would not occur. It has been suggested (for example, in Chomsky (2002)) that an important function of language is individual-internal, allowing the formulation of plans and so on using “inner speech”. If this was the case, then it could be argued that the initial individuals possessing the human vocabulary acquisition bias could use their acquired, sub-optimal but non-random, communication system internally, thereby immediately gaining a fitness advantage over individuals who could not communicate with themselves.

I have never found this line of argument particularly convincing with respect to full-blown language, and it is even less convincing when stretched to cover unstructured communication. Firstly, it is not clear to what extent inner speech is a widespread phenomenon, or even a real phenomenon. Individuals who claim to think in speech could be imposing some kind of post-hoc rationalisation on their thought processes. Even if some individuals do think in speech, it is not known if this is common, or a matter of personal cognitive style. Even if all individuals do think in speech sometimes, it is also not clear that this mode of thought is applied to a wide range of tasks — while we might be happy to agree that we "think in language" when performing tasks like rehearsing a talk or writing a thesis, it is less clear to what extent we think in language when performing less wordy tasks like painting a shed, going to the shops, or hunting a wildebeest. Ignoring the empirical evidence (or lack thereof) regarding the degree to which we actually do think in language, if this inner speech provided the main impetus for the evolution of the linguistic capacity then we might expect language to be of a rather different form — see Hurford's points in Chapter 1, Section 1.4.1, on the apparent adaptation of language for the externalisation of propositional structures. With respect to thinking in an unstructured communication system, it is even difficult to see how thinking, for example, wildebeest$'$, signalling "wildebeest", and re-arriving at wildebeest$'$ offers any improvement in clarity of thought.

An alternative, more appealing possibility is that some aspect of the emergent learning bias in pre-linguistic hominids or the social structure of pre-linguistic hominids resulted in an immediate, or at least very rapid, fitness advantage to individuals possessing the appropriate bias. The simulation results in Sections 4.5.5 and 4.5.6 show that increased speed of cultural convergence, either by increasing the number of exposures learners receive or by reducing cultural population size, leads to more reliable emergence of learning biases supporting optimal communication. If a plausible argument could be presented that one or both of these factors were at play in early stages of the emergence of human language-learning biases then we might be more optimistic that these learning biases evolved specifically for communication in the hominid line.

Human infants have an unusually long period of immaturity, a consequence of the reduced size of the pelvic opening in humans, which requires that babies be born at a relatively immature stage, before the head is too big to fit through the birth canal (Martin 1992). This extended period of immaturity and plasticity could be very roughly equated with large e in the simulation models, which speeds up cultural convergence and leads to more reliable evolution of the one-to-one bias. Alternatively, and perhaps more plausibly, we could point to the speed of cultural convergence in creolization situations, arguably a consequence of a strong human learning bias. Bickerton argues that creolization takes place in a single generation (Bickerton (1990), but see, for example, (McWhorter 1997) for a contrasting view). If we accept for a moment that the emergent learning bias in the hominid line led to significant convergence on a shared communication system within one generation, or even a few generations, then natural selection would be more likely to quickly identify and select for this bias. This does, however, require that the bias in its initial form was extremely strong, ruling out initial emergence of a weak bias which was gradually selected for in stronger and stronger forms.

Reduced cultural population size leads to rapid cultural convergence in the simulation model. We could argue that this is characteristic of the population structure of prelinguistic hominids, but not other primates, facilitating the evolution of a unique learning bias in the hominid line. Arguments based on group size have been made before in relation to the evolution of language in humans (Dunbar 1996). However, two main problems exist with this approach. Firstly, Dunbar's account is centred on the assumption that group size increased down the hominid line. This produces exactly the wrong prediction in terms of our model — larger group size should lead to slower cultural convergence and therefore a lower likelihood of evolving the correct bias. It could be argued that, while overall group size increased in hominids, groups became increasingly internally structured, which would produce the correct prediction. However, this brings us to the second problem with group-size arguments. In a recent review of the archaeological evidence underpinning theories of the evolution of language, Buckley & Steele (2002) conclude that there is no good evidence that group size changed during the Pleistocene period (from 1.6 million to 10,000 years ago), and that there is little to no direct evidence at all from earlier times. Making inferences from the fairly sparse archaeological evidence on group structure, a more subtle phenomenon, is likely to be even more fraught.

As a final, (semi-)positive interpretation, it could be concluded from these simulation results that a bias for the acquisition of vocabulary did not evolve under natural selection for communication in hominid populations, but a bias for the acquisition of *structured* communication did. This possibility will be investigated in Chapter 6. We should, however, remain sceptical — we would expect the logical point that there is no advantage to learning if there is nothing interesting to learn to pertain regardless of the degree of structure of the object of learning.

4.5.8.2 The negative interpretation

The contrasting position would be to interpret the results in this Chapter as indicating that the human capacity for vocabulary acquisition, and perhaps by extension language acquisition, did not evolve under natural selection for communication. This would mean that this capacity was initially not language-specific, although subsequent modifications may have refined the scope of the application of the bias.

How then should this learning bias in humans be explained? A first possibility is that this piece of cognitive apparatus is an exapted trait — the learning bias evolved under reliable selection pressure for some other task, and was later reappropriated for the acquisition of communication systems. A slight modification of this theory would be that it evolved initially for some general function, all the while being incidentally applied to the acquisition of vocabulary. The cultural construction of effective communication systems got underway, at which point the bias was selected directly for the communicative payoff it offered.

The main problem with these accounts is that it is not clear what the original function of the learning bias was — what other learning tasks require a one-to-one bias? As discussed in Chapter 1, some theories of the evolution language have suggested that the capacity for tool use and language are intertwined. It could be argued that the explosion in specialisation in tool function which occurred towards the end of the Paleolithic period (approximately 40,000 years ago) may have reflected a new bias in favour of one-to-one mappings between functions (meanings?) and forms (signals?). However, this is wildly speculative. An alternative, more plausible possibility is that the one-to-one bias was initially a side-effect of the particular method of learning which was selected in the population, for general or specific unknown purposes — the one-to-one bias was a spandrel of the selected learning mechanism.

An interesting variation on this possibility is that the one-to-one bias is a consequence of the evolution of a sophisticated theory of mind, an understanding of other individuals as intentional agents. Tomasello (e.g. Tomasello (1997) and Tomasello (1999)) argues that the human capacity for this type of mental gymnastics is uniquely sophisticated among the primates. Tomasello argues that this cognitive capacity could have evolved under an array of selection pressures, including selection for communication, cultural learning in general, cooperation and tool use.

How would this understanding of others as intentional agents lead to a learning bias in favour of the acquisition of one-to-one mappings between meanings and signals? Clark (1990) and Bloom & Markson (1998) present the vocabulary acquisition biases of humans in essentially pragmatic terms — children understand the nature of communicative acts, which is basically to draw their attention to particular items, and infer from that understanding that ambiguity (introduced by synonymy or homonymy) is undesirable. I have tended to shy away from this interpretation here, preferring to discuss learning bias as a consequence of weight-changing procedures. However, the two interpretations are roughly equivalent for my purposes — a one-to-one bias should either be imposed on the learning act itself (as I have it), or on the process by which the learner deduces the object which is being referred to (as Bloom & Markson (1998) have it). My account has the advantage of extending, as will be discussed in Chapter 5, to the evolution of structured communication, whereas Bloom & Markson see their account as being restricted to lexical learning. Bloom & Markson's (1998) account has the advantage that two processes which are distinct under my model (identifying the meaning of an utterance and learning with a one-to-one bias) are shown to be aspects of a single cognitive process.

4.6 Summary of the Chapter

At the beginning of this Chapter I briefly outlined approaches to modelling genetic transmission, and discussed some of the issues involved in applying such models to the investigation of the evolution of innate signalling systems. My primary concern, however, was to introduce the dual transmission model, which unifies models of biological and cultural evolution. The dual transmission model forms the basis of the Evolutionary Iterated Learning Model.

I then outlined two EILMs, based on extensions of the feedforward network and associative network ILMs discussed in Chapter 3. These EILMs allow us to investigate the interactions between the biological evolution of learning bias and the cultural evolution of communication systems acquired using this biased learning apparatus.

In the feedforward network EILM, there was found to be little significant interaction — the learning bias associated with the imitator or obverter network architectures proved to be the decisive factor in determining the behaviour of populations in the EILM. In the associative network EILM, the picture was more complex. One-to-one learning biases can evolve under natural selection pressure for communicative success. However, the evolution of such biases was dependent on an initial period of genetic drift, due to the lack of an immediate benefit to individuals with the appropriate learning bias. This dependence on genetic drift can be reduced by speeding up cultural convergence within populations, either by increasing the amount of learning each individual does or by introducing spatial organisation and effectively dividing a single population up into several smaller subpopulations. However, the point remains that there is no immediate benefit to being biased to acquire an optimal communication system in a non-communicating population.

This result has implications for our understanding of the evolution of vocabulary-learning biases in human populations. We could take the simulation results as indicative of the conditions under which such a bias evolved in hominid populations. The alternative, possibly stronger view, is to conclude that these learning biases did not evolve directly for the purpose of communication — human vocabulary learning biases are a spandrel, or an exapted trait.

CHAPTER 5

The cultural evolution of compositionality

In Chapter 3 I described an investigation into the properties of a learning bias required to construct an optimal vocabulary system, and equated these properties with learning biases of humans. This investigation was carried out under the assumption that the human learning biases evolved under natural selection for communication. However, the results outlined in Chapter 4 forced a re-evaluation of this position — such learning biases are unlikely to evolve through natural selection for communication, due to the time-lag between the emergence of an appropriate bias and a communicative payoff to those individuals possessing it.

In this Chapter I will pursue a similar route to that taken in Chapter 3, investigating the conditions under which compositional language can emerge through purely cultural processes, and attempting to draw analogies between these conditions and the factors at play in human language acquisition. I will return to the issue of the evolution of the learning biases necessary to support compositional communication in Chapter 6, while for the meantime remaining sceptical as to whether these biases are communication-specific.

The review of models of the cultural evolution of linguistic structure carried out in Section 5.1 highlights three possible factors impacting on linguistic structure: the severity of the transmission bottleneck, the biases of language learners and the degree of structure shared by communicatively-relevant situations. In Section 5.2 an Iterated Learning Model is presented, based on an extension of the associative network model outlined in Chapter 3, which is designed to shed light on these issues. In Section 5.3 the impact of transmission bottleneck and environment structure on emergent communication systems is investigated. The results outlined in this Section broadly support the conclusions of other models — a transmission bottleneck leads to a pressure for generalisation, and the advantage of compositionality is at a maximum when the environment is structured. In Section 5.4 I conduct a thorough examination of the learning biases required to support compositional language in this model. A bias in favour of one-to-one mappings between elements of meaning and elements of signal is found to be key, in combination with a preference for exploiting internally-compositional representations. Similar biases are present in several other ILMs dealing with the evolution of linguistic structure. Finally, in Section 5.6, I argue that human language learners are also characterised by this type of bias.

5.1 Models of the cultural evolution of linguistic structure

In Chapter 2, Section 2.3.6, I outlined models by Kirby (Kirby 2002) and Batali (Batali 2002) which demonstrated that compositional and recursive syntax can emerge from holistic, non-recursive communication through purely cultural processes. Both authors attribute this emergent linguistic structure to the bottleneck on cultural transmission, which forces language to be generalisable.

Armed with this basic result, we might ask two further questions. Firstly, is the transmission bottleneck alone sufficient to lead to the emergence of compositional language, or are there some particular learner biases which must also be in place? Secondly, how do changes in the severity of the bottleneck impact on the structure of the evolving languages? In Sections 5.1.1 and 5.1.2 below I present other Expression/Induction models which shed some light on these two issues. Finally, in Section 5.1.3 I describe a model by Brighton which introduces another factor which impacts on the cultural evolution of compositionality — the structure of the meaning space.

5.1.1 Varying the learner model

No direct within-model manipulation of learner biases have been carried out. However, a cross-model comparison yields some insight into the types of learning biases required to support the cultural evolution of linguistic structure. Kirby and Batali's initial findings have been replicated both with connectionist and symbolic models of learners.

Kirby & Hurford (2002) present an ILM based around a feedforward neural network model. Their network maps from input signals to output meanings. Meanings and signals are modelled by binary patterns of activation over output and input nodes in the network. Individuals observe meaning-signal pairs and, based on this observed production behaviour, acquire their reception competence, which in turn determines their own production competence. This is therefore the classic obverter neural network paradigm. Production and reception in the network proceed by processes which are essentially identical to those described for obverter networks in Chapter 3.

Kirby & Hurford report that, given a certain level of bottleneck on cultural transmission (as discussed in the next Section), perfectly compositional systems emerge, where the value of a particular bit in the meaning is expressed by the value of a particular bit of the signal. Their results support the earlier finding that a bottleneck on cultural transmission leads to the emergence of a compositional mapping between meanings and signals, and demonstrate that this phenomenon is not specific to the symbolic models used by Kirby and Batali.

Hare & Elman (1995) also use a feedforward network model of a learner in an early ILM dealing with morphological change. However, Hare & Elman's network follows the imitator architecture, and as discussed in Chapter 3, Section 3.5.3, this network architecture, with the associated many-to-one bias, probably leads to the *loss* of linguistic structure. The comparison of this result with that of Kirby & Hurford (2002) demonstrates that the differences in the learning biases associated with the obverter and imitator networks may be significant when considering the cultural evolution of linguistic structure.

Hurford (2000) describes a symbolic model of a learner which is based on Kirby's (2002) context free grammar-inducing agent, but has a rather stronger bias towards compositionality. The main difference between Kirby's model and Hurford's is that, in the latter, structural generalisations can be formed on the basis of *single* exposures to structured meaning signal pairs. In Kirby's model, several such exposures are required. Hurford's results support Kirby's findings on the impact of the transmission bottleneck on linguistic structure. This is perhaps not surprising, given the strengthening of the agents' inductive bias. However, Hurford also experimentally varies the frequency with which agents apply their strong inductive bias — rather than making structural generalisations with every observation, learners in the modified experiment make such generalisations only 25% of the time, with most learning episodes simply involving memorisation. Under these circumstances, Hurford finds that the population's language, based on an analysis of produced meaning-signal pairs, appears to be completely compositional. However, examination of the agents' internal rule systems reveals that individual agents are in fact producing utterances in a partially compositional, partially holistic fashion. While this type of experimental variation of learning bias is desirable, it is not clear that the infrequent application of a strong, appropriate bias is the same as the constant application of a weak or inappropriate bias.

5.1.2 Varying the transmission bottleneck

Kirby & Hurford (2002) use their obverter network ILM to investigate the impact of various severities of transmission bottleneck. When the bottleneck is very severe (20 exposures to utterances produced for randomly selected meanings from the space of 256 possible meanings) no stable systems emerge. When there is virtually no bottleneck (2000 exposures) stable holistic vocabularies emerge. However, when the bottleneck is at some intermediate level (50 exposures) perfectly compositional systems emerge. The severity of transmission bottleneck clearly plays a role in shaping the emergent linguistic system in this model.

Both Kirby (2001) (using a model of a learner similar to that described in Kirby (2002)) and Hurford (2000) (discussed above) present results which indirectly indicate the effects of bottleneck severity on linguistic evolution. In both these models, certain meanings occur with disproportionately high frequency. Variation in frequency of particular meanings is equivalent to varying the bottleneck on the cultural transmission of different meaning-signal pairs. Given a fixed number of exposures, assuming uniform frequency distributions over meanings, we can calculate the probability of a learner observing a particular meaning-signal pair, with this probability being equal for all meaning-form pairs. However, if one meaning is disproportionately frequent then learners are more likely to observe an utterance being produced for that meaning — the frequent meaning is more likely to pass through the bottleneck than the less frequent meanings. Kirby (2001) and Hurford (2000) both report that more frequent meanings are more likely to be expressed in a non-compositional manner. This suggests that we should expect an increase in bottleneck size to lead to a reduced pressure for compositionality.

Finally, Brighton (2002) presents results for several levels of bottleneck severity. I will discuss his results in the next Section.

5.1.3 Varying the structure of the meaning space

A recent paper by Henry Brighton (Brighton 2002) addresses the impact of transmission bottleneck and meaning space structure on the relative stability of holistic and compositional language. Brighton adopts a mathematical modelling approach, and restricts himself to examining *stability*, rather than a full-blown emergence model.

Brighton models meanings as points in an F-dimensional space where each dimension has V discrete values. F and V therefore define the structure of a meaning space — high F and V give a highly structured meaning space, with the possibility of a large number of fine-grained semantic distinctions, whereas low F and V give a meaning space capable of only a few distinctions.

Brighton compares the stability of perfectly compositional and completely holistic communication systems over a single generation of cultural transmission. This allows us to predict the likely outcome of a multi-iteration ILM — over cultural time, highly stable systems will persist whereas unstable systems will disappear or change radically. Brighton varies both the structure of the meaning space and the severity of the transmission bottleneck, to investigate how these factors impact on the relative stability of compositional and holistic language.

Brighton assumes the strongest possible capacity to generalise. A learner faced with a completely holistic language cannot make any generalisations — by definition, in a holistic language there is no way to reconstruct a signal which you have not seen paired with a meaning, due to the arbitrary associations between meanings and signals. A learner of a holistic language simply has to memorise observed meaning-signal mappings. In contrast, a learner confronted with a perfectly compositional *can* generalise, due to the structure in the system of meaning-signal mappings. Brighton assumes that a learner who has observed some subset of a perfectly compositional language will be able to express a meaning if it has observed all the feature values that make up that meaning, paired with signal substrings — in other words, if they know how to express each part of the meaning individually, they know how to express the whole meaning in a compositional manner.

If an individual (who has either memorised some portion of a holistic language or learned from some subset of a compositional language) is called upon to express a meaning they have not observed being expressed and cannot generalise to, they have two options. Firstly, they could simply not express. Alternatively, they could produce some random signal. In either case, any association between meaning and signal that was present in the previous individual's hypothesis will be lost — part of the meaning-signal mapping will change. A shortfall in expressivity therefore results in instability over cultural time. Brighton uses mathematical techniques to compare relative expressivity (and therefore relative stability) of a learner exposed to some subset of a holistic or compositional language. Where these two languages have equal stability we should expect them to emerge with equal frequency over cultural time. When compositional languages are more stable than holistic languages we should expect them to emerge more frequently, and persist for longer, than holistic languages, and vice versa.

Brighton reports two main results:

1. The stability advantage of compositional language is at a maximum when learners only observe a small subset of the language of their cultural parents. Holistic languages will not persist over time when the bottleneck on cultural transmission is tight. In contrast, compositional languages are generalisable, due to their structure, and remain relatively stable even when a learner only observes a small subset of the language of the previous generation.
2. A large stability advantage for compositional language only occurs when the meaning space exhibits a certain degree of structure, suggesting that structure in the conceptual space of language learners is a requirement for the evolution of compositionality. In such meaning spaces, distinct meanings tend to share feature values. A compositional system in such a meaning space will be highly generalisable — the signal associated with a meaning can be deduced from observation of other meanings paired with signals, due to the shared feature values. However, if the meaning space is too highly structured the compositional language has little stability advantage, as few distinct meanings will share feature values and the advantage of generalisation is lost.

This first result is a verification of the results reported by Kirby, Batali and others, outlined in the previous Section. The second result is a novel one, and indicates a possible precondition for the cultural evolution of compositionality.

5.2 An Iterated Learning Model

Where are we now, based on the review? We have seen that learning bias, bottleneck, and meaning-space structure can all impact on the cultural evolution of compositionality. The ILM outlined in this Section is an extension of the associative network ILM described in Chapter 3 and is designed to allow investigation into the importance and interaction of these three factors. This ILM is also described in Smith *et al.* (forthcoming) and Smith *et al.* (submitted).

5.2.1 Languages and communication

The model of a structured language is based on the model of unstructured communication systems outlined in Chapter 3, Section 3.2. A language L consists of a production function $p\left(m\right)$, mapping from meanings m to signals s, and a reception function $r\left(s\right)$, mapping from signals s to meanings m. m and s are selected such that $m \in \mathcal{M}$ and $s \in \mathcal{S}$ where $\mathcal{M} = \left\{m_1, m_2 \ldots m_{|\mathcal{M}|}\right\}$ and $\mathcal{S} = \left\{s_1, s_2 \ldots s_{|\mathcal{S}|}\right\}$.

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

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

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

In the simulations outlined in Chapters 3 and 4, the set of meanings agents were required to produce utterances for was equivalent to the meaning space $\mathcal{M}$. However, this need not be the case. I will introduce a new term, the *environment*, $\mathcal{E}$, appealing to the notion that the agents' external environment provides the situations which they are required to produce signals for. The meanings in $\mathcal{E}$ constitute a *subset* of $\mathcal{M}$.

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

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

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

In previous Chapters I defined measures of communicative accuracy, and measured the change in communicative accuracy within populations over time. However, in the models outlined in this chapter I will consider the case where the population at each generation consists of a single individual. The concept of communication therefore becomes essentially meaningless, given that individuals have no-one to communicate with. It would be possible to measure inter-generational communicative accuracy by similar techniques to those used to measure intra-population communicative accuracy in previous Chapters. However, I will focus instead on the evolution of linguistic structure, and postpone the consideration of communication until Chapter 6, where the model outlined in this Chapter is extended to deal with larger populations.

5.2.2 Communicative agents

Communicative agents in the model must be capable of representing such communication systems, modelling production and reception functions of the type outlined above and modifying their behaviour based on observations of systems of the type outlined above. The associative network model outlined in Chapter 3 is used as a basis for the model of a communicative agent. The basic associative network model is altered to allow the manipulation of structured meanings and signals. The most fundamental changes are in the processes of production and reception, while the process of learning remains largely unchanged.

5.2.2.1 Representation

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

What do these nodes represent? In the associative network outlined in Chapter 3, meanings and signals are discrete atomic items, therefore each node represented a particular meaning or signal. However, in the case of structured languages each distinct meaning or signal has structure, some of which may be shared with other meanings or signals. In order to represent systems of this type, it is necessary to introduce two new pieces of terminology — the notion of *components* of meanings and signals, and *analyses* of meanings and signals. I will define components now, and postpone the definition of analyses until the section on production and reception.

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

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

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

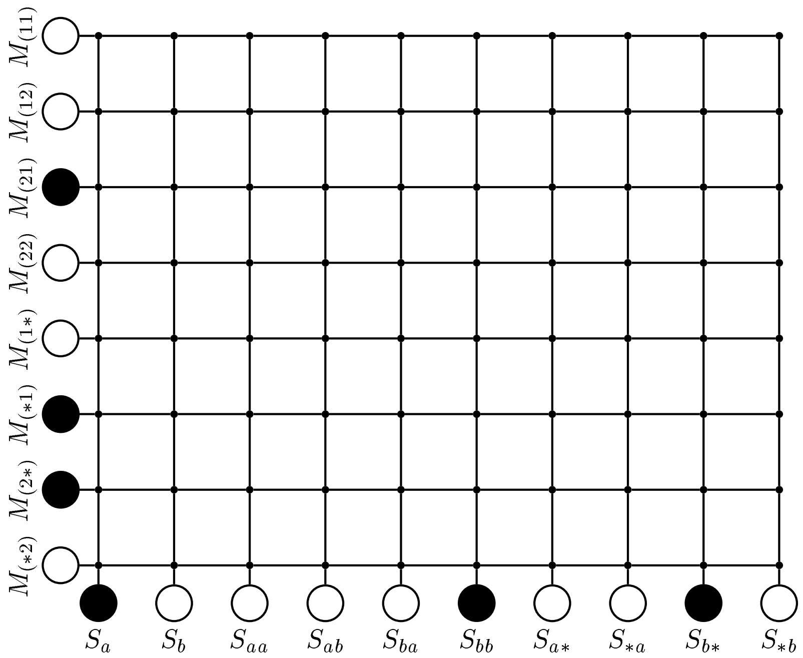


Figure 5.1: A network where $F = V = 2$, $l_{max} = 2$, $\Sigma = \{a, b\}$. Large filled circles represent nodes with activation of 1, large empty circles represent nodes with activation of 0. The pattern of activation over $\mathcal{N}_M$ therefore represents the meaning components $(2\ 1)$, $(2\ *)$ and $(*\ 1)$, which all happen to be components of the meaning $(2\ 1)$. The pattern of activation over $\mathcal{N}_S$ represents the signal components a, bb and $b*$.

the signal represented by the string bd are bd, $b*$ and $*d$, but not $e*$ (first character doesn't match), $**$ (no specified characters) or a (not of correct length).

Each node Mi in $\mathcal{N}_M$ represents a component of a meaning, and there is a single node in $\mathcal{N}_M$ for each component of every possible meaning. Similarly, each node Si in $\mathcal{N}_S$ represents a component of a signal and there is a single node in $\mathcal{N}_S$ for each component of every possible signal. In order to represent the meaning component c_m the activation, a_{Mc_m}, of node Mc_m is set to one. In order to represent the signal component c_s, a_{Sc_s} is set to 1. This representational scheme is illustrated in Figure 5.1. Note that, unlike in the associative network model outlined in Chapter 3, there is no restriction that only a single node in $\mathcal{N}_M$ and $\mathcal{N}_S$ can be active.

5.2.2.2 Learning

During a learning event, a learner observes a meaning-signal pair $\langle m, s\rangle$. The activations of the nodes corresponding to all possible components of m and all possible components of s are set to 1. The activations of all other nodes are set to 0. The weights of the connections in $\mathcal{W}$ are adjusted according to some weight-update rule W where, as before, W is specified as a 4-tuple $(\alpha\ \beta\ \gamma\ \delta)$, where α, β, γ and δ take integer values in the range $[-1, 1]$. The storage process is illustrated in Figure 5.2.

5.2.2.3 Production and reception

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

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

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

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

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

where n is the number of components in the analysis of meaning and signal, $w_{c_m^i, c_s^i}$ gives the weight of the connection between the nodes representing the ith component of the meaning analysis and the ith component of the signal analysis and $\omega\,(x)$ is a weighting function which gives the non-wildcard proportion of x. The production process is illustrated in Figure 5.3.

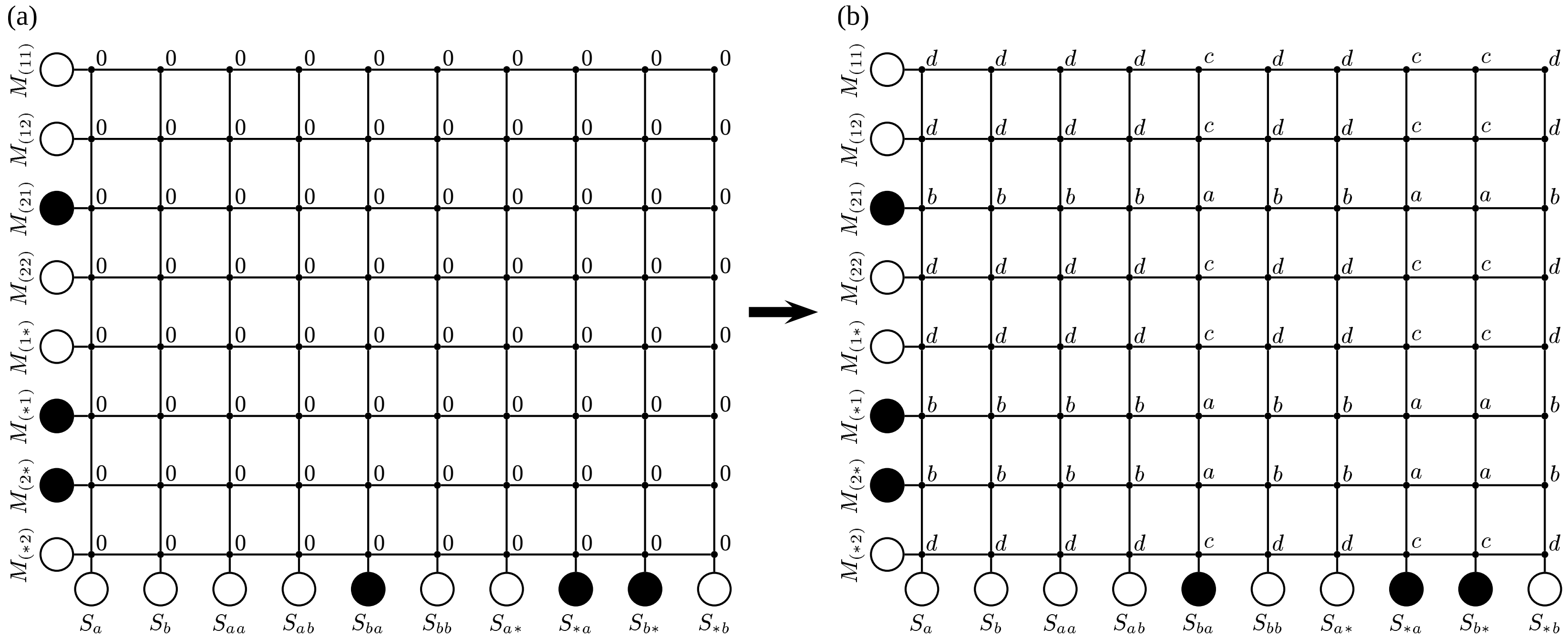
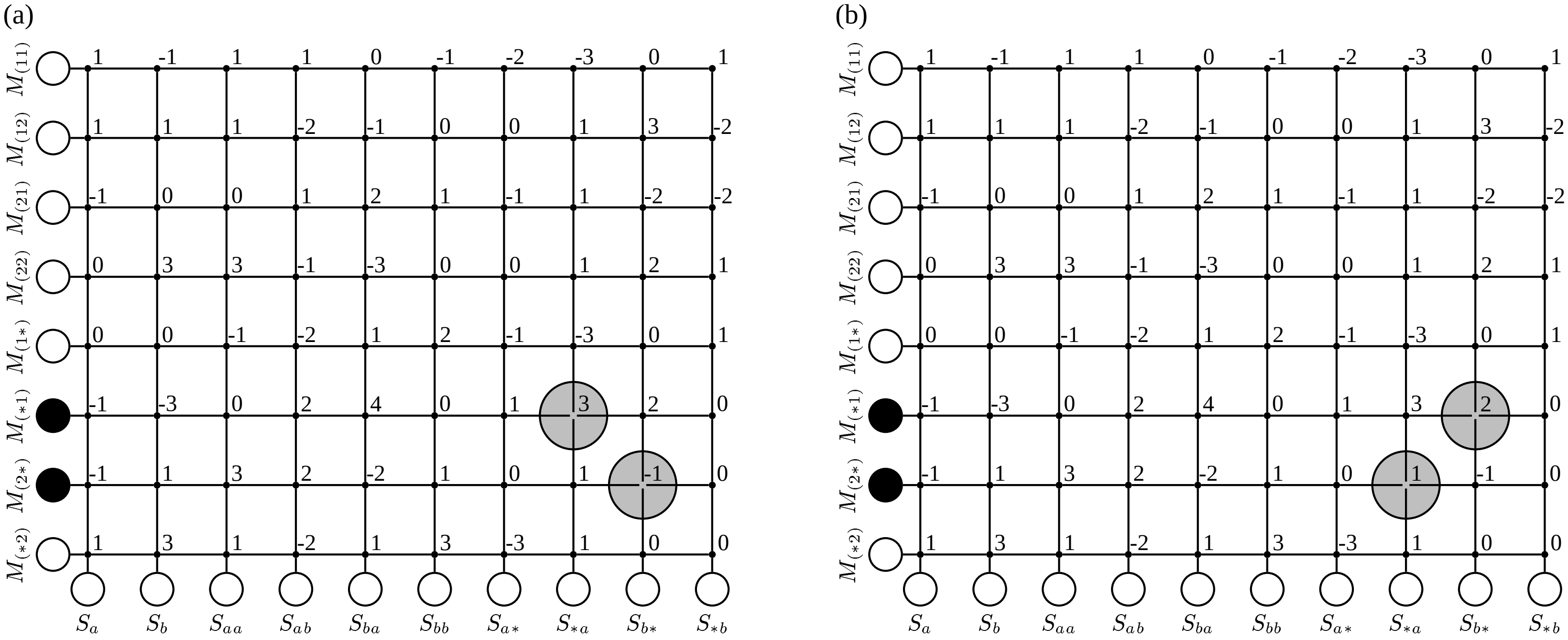

(a)
(b)
$M_{(11)}$
$M_{(12)}$
$M_{(21)}$
$M_{(22)}$
$M_{(1*)}$
$M_{(*1)}$
$M_{(2*)}$
$M_{(*2)}$
S_a S_b S_{aa} S_{ab} S_{ba} S_{bb} S_{a*} S_{*a} S_{b*} S_{*b}
0 0 0 0 0 0 0 0 0 0
d d d d c d d c c d
d d d d c d d c c d
b b b b a b b a a b
d d d d c d d c c d
d d d d c d d c c d
b b b b a b b a a b
b b b b a b b a a b
d d d d c d d c c d

Figure 5.2: Learning of the meaning-signal pair $\langle (2\ 1)\,, ba\rangle$ using the weight-update rule $W = (a\ b\ c\ d)$. In (a), the nodes in $\mathcal{N}_M$ and $\mathcal{N}_S$ have been set to the patterns of activation representing the components of $(2\ 1)$ and ba. All connections have weight 0. In (b) the result of the application of the learning process is shown — all connections now have weights of a, b, c or d, depending on the activations of the nodes they connect.



(a)

	S_a	S_b	S_{aa}	S_{ab}	S_{ba}	S_{bb}	S_{a*}	S_{*a}	S_{b*}	S_{*b}
$M_{(11)}$	1	-1	1	1	0	-1	-2	-3	0	1
$M_{(12)}$	1	1	1	-2	-1	0	0	1	3	-2
$M_{(21)}$	-1	0	0	1	2	1	-1	1	-2	-2
$M_{(22)}$	0	3	3	-1	-3	0	0	1	2	1
$M_{(1*)}$	0	0	-1	-2	1	2	-1	-3	0	1
$M_{(*1)}$	-1	-3	0	2	4	0	1	3	2	0
$M_{(2*)}$	-1	1	3	2	-2	1	0	1	-1	0
$M_{(*2)}$	1	3	1	-2	1	3	-3	1	0	0

(b)

	S_a	S_b	S_{aa}	S_{ab}	S_{ba}	S_{bb}	S_{a*}	S_{*a}	S_{b*}	S_{*b}
$M_{(11)}$	1	-1	1	1	0	-1	-2	-3	0	1
$M_{(12)}$	1	1	1	-2	-1	0	0	1	3	-2
$M_{(21)}$	-1	0	0	1	2	1	-1	1	-2	-2
$M_{(22)}$	0	3	3	-1	-3	0	0	1	2	1
$M_{(1*)}$	0	0	-1	-2	1	2	-1	-3	0	1
$M_{(*1)}$	-1	-3	0	2	4	0	1	3	2	0
$M_{(2*)}$	-1	1	3	2	-2	1	0	1	-1	0
$M_{(*2)}$	1	3	1	-2	1	3	-3	1	0	0

Figure 5.3: Evaluation of two meaning analysis-signal analysis pairs, encoded as patterns of activation over $\mathcal{N}_M$. In (a) the pair $\langle\{(2\,*)\,,(*\,1)\}\,,\{b*,*a\}\rangle$ is evaluated by taking the connection weights $w_{M_{(2*)},S_{b*}}$ and $w_{M_{(*1)},S_{*a}}$ (highlighted in grey) and calculating g. $g\left(\{(2*)\,,(*1)\}\,,\{b*,*a\}\right)=1$. In (b) this process is repeated for the pair $\langle\{(2\,*)\,,(*\,1)\}\,,\{*a,b*\}\rangle$, yielding $g\left(\{(2*)\,,(*1)\}\,,\{*a,b*\}\right)=1.5$.

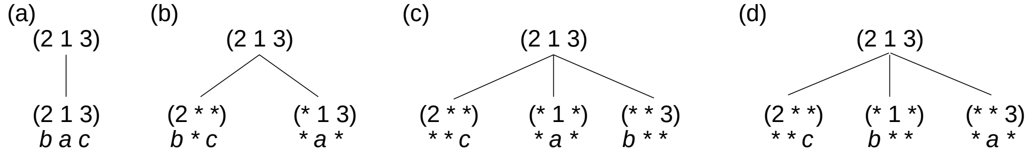


Figure 5.4: Parse trees corresponding to four of the possible 169 analyses pairs of the meaning-signal pair $\langle (2\ 1\ 3)\,, bac\rangle$. (a) gives the parse tree for the analysis pair $\langle \{(2\ 1\ 3)\}\,, \{bac\}\rangle$. There is a single node in the tree, which is labelled with the single component from both meaning and signal analysis. (b) gives the parse tree for the analysis pair $\langle \{(2\ *\ *)\,, (*\ 1\ 3)\}\,, \{b*c, *a*\}\rangle$. The left daughter node is labelled with the 1st component of both analyses, and the right daughter node is labelled with the 2nd component of both analyses. (c) and (d) give the parse trees for the analyses pairs $\langle \{(2\ *\ *)\,, (*\ 1\ *)\,, (*\ *\ 3)\}\,, \{**c, *a*, b**\}\rangle$ and $\langle \{(2\ *\ *)\,, (*\ 1\ *)\,, (*\ *\ 3)\}\,, \{**c, b**, *a*\}\rangle$ respectively. These differ in the order of the second and third components of the signal, which leads to different interpretations of the semantics of string-initial b and medial a.

How is this process to be interpreted? A meaning analysis-signal analysis pair can be interpreted as a parse tree where each terminal node of the tree is labelled with both a component of meaning and a component of signal. The ith node of the tree is labelled by the ith component of the meaning analysis and the ith component of the signal analysis. This yields the fairly natural interpretation that the ith component of the meaning analysis is conveyed by the ith component of the signal analysis. This is illustrated in Figure 5.4.

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

5.2.3 The Iterated Learning Model

In previous Chapters, I presented ILMs with fairly large population sizes (100). Here we will look at the simplest possible case, where the population consists of a single individual at any one time. That individual produces some observable behaviour and then is removed. This observable behaviour is then observed and learned from by a new individual, and the process iterates.

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

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

One key difference should be noted between this ILM and the ILMs outlined in Chapter 3. In previous ILMs each of the e exposures consisted of the observation of the *complete* set of meaning-signal pairs produced by a particular agent — if $e = 3$ then the learner observes three *complete* sets of meaning-signal pairs. In this ILM, each of the e exposures consists of the *single* observation of a *single* meaning-signal pair. There is therefore a potential bottleneck (as defined by Kirby) on cultural transmission — learners are not guaranteed to make observations of every meaning in the environment, and therefore may subsequently be required to produce a signal for a meaning which they themselves have not observed paired with a signal. We can calculate the expected *coverage* for a given $\mathcal{E}$ and e, $c\left(\mathcal{E}, e\right)$ (Equation taken from Brighton (2002)):

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

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

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

$m \in \mathcal{E}$ is selected in turn, therefore the learner observes the full set of observable behaviour produced by the preceding agent.

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

5.2.4 Environments

As discussed in Section 5.2.1, a distinction has been made between the meaning space $\mathcal{M}$ and the environment $\mathcal{E}$, the set of meanings for which agents are required to produce signals. In Brighton's model, described in Section 5.1, meanings in the environment are selected at random from the meaning space — $\mathcal{E}$ is random subset of $\mathcal{M}$. However, it is not necessarily the case that meanings in the environment should be selected at random from the space of possible meanings. I will introduce a notion of *environment structure*, in contrast to Brighton's notion of meaning space structure. In an *unstructured* environment, meanings in $\mathcal{E}$ are selected at random from $\mathcal{M}$. In a *structured* environment, meanings in the environment are drawn from a hypercube subset of the space of possible meanings.

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

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

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

Figure 5.5 shows three unstructured environments, of various densities. Figure 5.6 shows three structured environments, of various densities[1].

[1] Recall that values within a feature are unorganised. Therefore, structured environments are not simply *smaller* than unstructured environments — each structured environment could be rearranged so as to appear to more fully fill the meaning space. It is the degree of sharing of feature values which defines environment structure, not apparent closeness.

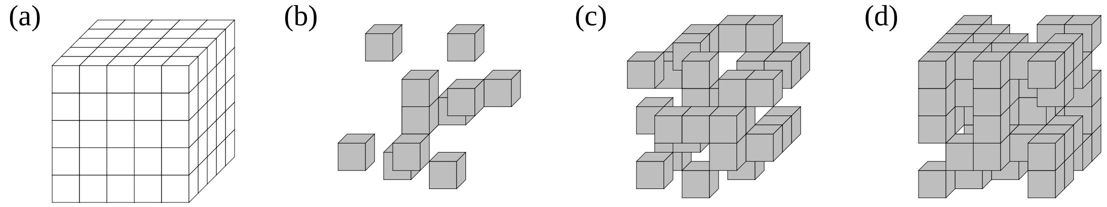


Figure 5.5: Unstructured environments. F and V define a meaning space $\mathcal{M}$. For $F = 3$ and $V = 5$ the meaning space can be visualised as a cube, as in (a). The three dimensions of the cube each correspond to the three feature values. The five subdivisions on each dimension correspond to the five values for each feature. Each point in this cube corresponds to a particular meaning. (b) is a sparse, unstructured environment ($p = 0.096$), where the meanings in $\mathcal{E}$ are highlighted in grey. (c) is a medium density, unstructured environment ($p = 0.248$). (d) is a dense, unstructured environment ($p = 0.504$).

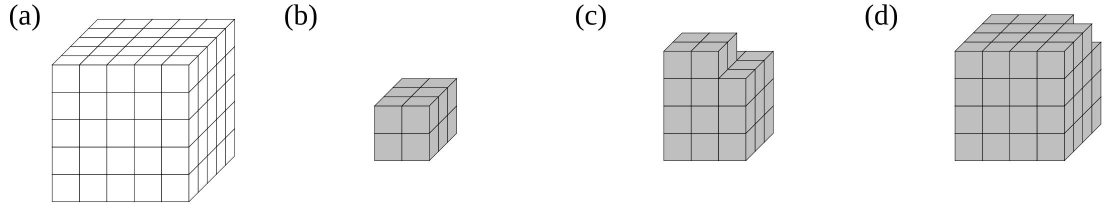


Figure 5.6: Structured environments. (a) is the meaning space. (b) is a sparse, structured environment ($p = 0.096$). (c) is a medium density, structured environment ($p = 0.248$). (d) is a dense, structured environment ($p = 0.504$).

5.2.5 *Measuring compositionality*

In Chapter 1 compositionality, one of the fundamental design features of language, was defined in fairly loose, impressionistic terms — in a compositional language the meaning of an expression is a function of the meanings of its parts and the way in which they are combined. In contrast, in a non-compositional (or *holistic*) system, the meaning of an expression is not dependent on the meaning of its parts. In Chapter 2 this impressionistic definition was applied to the computational models of Kirby and Batali, who demonstrate that compositional languages can emerge through cultural processes — in the final, stable emergent systems in these models, the meaning of expressions depends on the meanings of subparts of those expressions (syntactic subtrees in Kirby's model, sub-exemplars in Batali's model) and the way in which those subparts are combined. I will define two measures of compositionality here, both drawing on slightly different interpretations of the informal definition given above.

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

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

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

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

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

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

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

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

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

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

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

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

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

5.3 The impact of transmission bottleneck and environment structure

I will begin by presenting results for an ILM where every agent uses the weight-update rule $W = (1\ \ -1\ \ -1\ 0)$. The agents in the initial generation of each ILM have connection weights of 0, and therefore use the maximum entropy system where every meaning analysis-signal analysis pair occurs with equal probability — the initial language L is random. The meaning space ($F = 3$ and $V = 5$) and six environments illustrated in Figures 5.5 and 5.6 were used. The signal space is given by $l_{max} = 3$, $\sigma = \{a, b, c, d, e, f, g, h, i, j\}$.

5.3.1 Linguistic evolution in the absence of a bottleneck

Runs of the ILM were carried out, using the no-bottleneck variant of step 4 — each individual observes the full set of observable behaviour produced by the preceding agent.

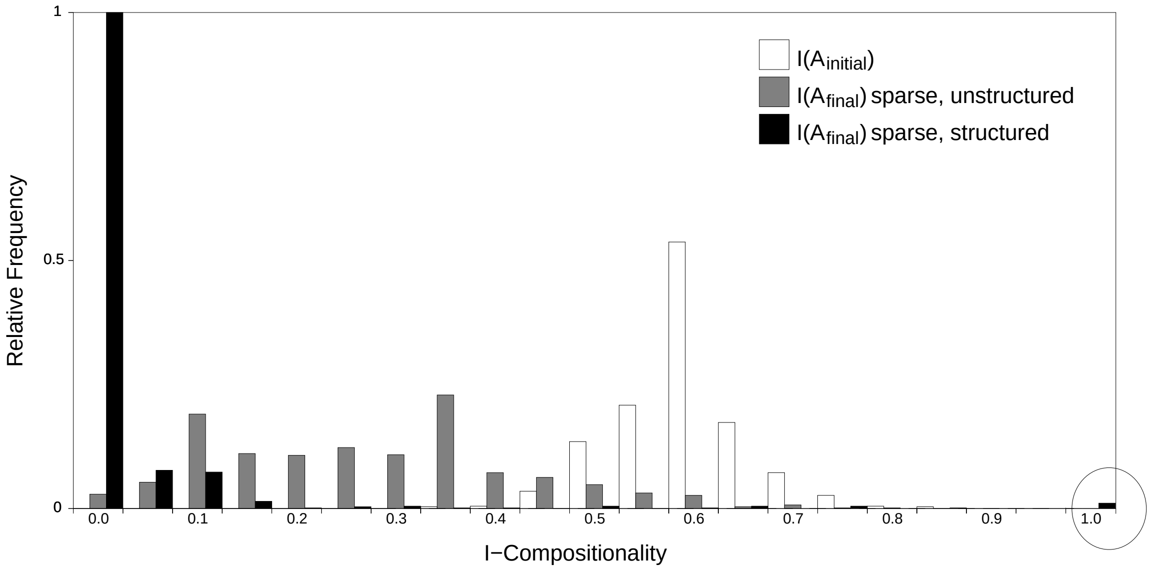


Figure 5.7: I-compositionality of initial and final, stable systems in sparse environments, when there is no bottleneck on cultural transmission. The initial systems are partially i-compositional. The final systems are less i-compositional, although highly i-compositional systems (circled) do occur with very low frequency when the environment is structured.

1000 runs were carried out for each of the six environments. Each run was allowed to proceed to a stable state, where parent and child produce identical observable behaviour. At this point, in the absence of a bottleneck on cultural transmission, further change is impossible. Figures 5.7 and 5.8 give the distributions of systems with respect to $I\left(\mathcal{A}\right)$ and $E\left(\mathcal{O}\right)$, for the 1000 runs of the ILM with the sparse unstructured and sparse structured environments[2].

In Figure 5.7, values of $I\left(\mathcal{A}_{initial}\right)$ are distributed around 0.6, while values for $I\left(\mathcal{A}_{final}\right)$ are typically lower. This is due to the random behaviour of the initial agents — each meaning analysis-signal analysis pair occurs with equal probability, and given that there are more multi-component analyses pairs than single-component analyses pairs, the initial random behaviour scores highly in terms of internal compositionality. The final stable systems tend to have lower internal compositionality. In structured environments, the i-compositionality of the final systems tends to be around 0. In unstructured environments,

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

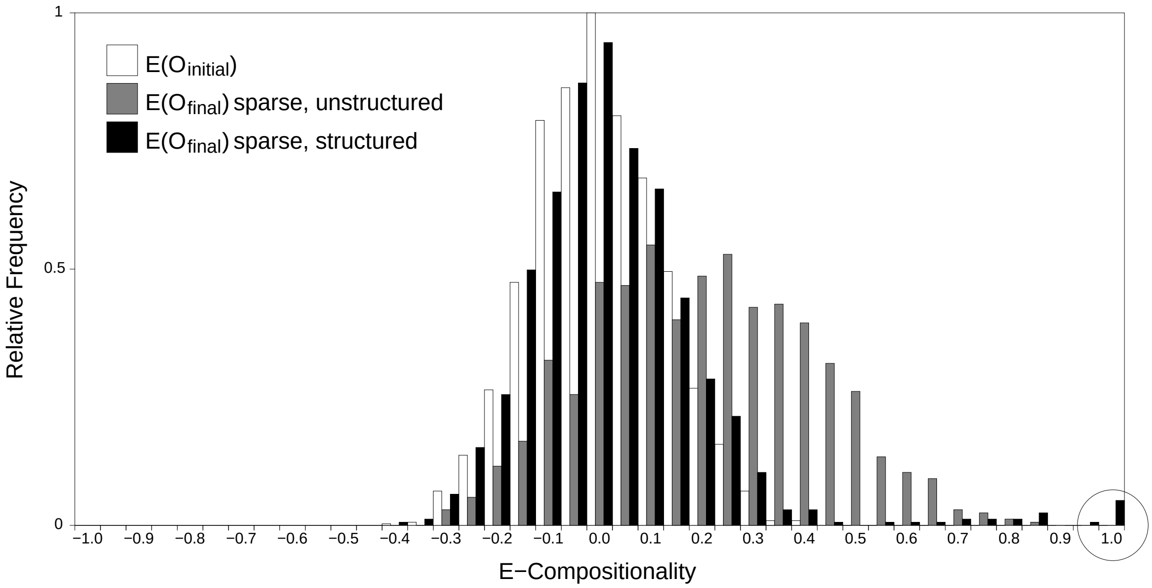


Figure 5.8: E-compositionality of initial and final, stable systems in sparse environments, when there is no bottleneck on transmission. The initial systems have low e-compositionality. The final systems are also of low or medium e-compositionality. Highly e-compositional systems (circled) occur infrequently, and only when the environment is structured.

final i-compositionality tends to be somewhat higher. Highly i-compositional systems occur very infrequently and only where the environment is structured.

In Figure 5.8, values of $E\left(\mathcal{O}_{initial}\right)$ are distributed around 0, indicating that the initial, random systems are not highly e-compositional. As with the internal compositionality measure, the final stable systems tend not to be highly compositional according to the external measure, with unstructured environments leading to a slightly higher level of compositionality. Highly e-compositional systems occurring infrequently and only when the environment is structured.

The i-compositionality measure has the somewhat undesirable property of treating the random initial behaviour as partially compositional. However, for the stable states the internal and external measures are equivalent — for the data in Figures 5.7 and 5.8, there is a high degree of correlation between $I\left(\mathcal{A}_{final}\right)$ and $E\left(\mathcal{O}_{final}\right)$ ($r = 0.842$, $p < 0.001$). This reflects the fact that agents produce an e-compositional language in an i-compositional manner, and similarly produce an e-holistic system in an i-holistic manner — i-compositional internal representation leads to e-compositional language. For the remaining results I will focus on the external compositionality measure, and unless otherwise indicated, "compositional" will mean "e-compositional". The internal measure will be returned to below in Section 5.4.

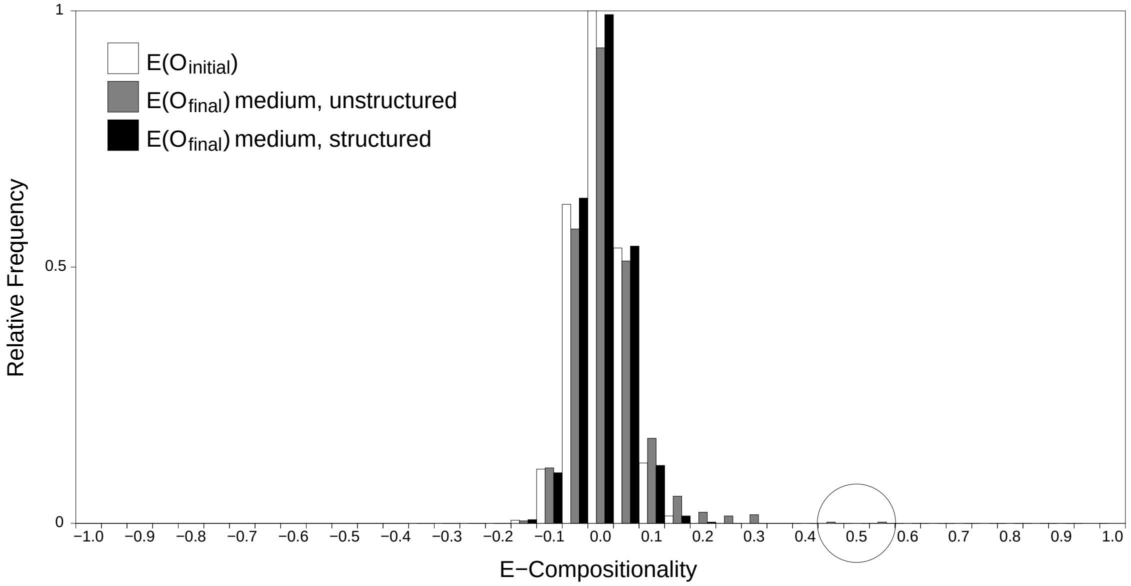


Figure 5.9: E-compositionality of initial and final, stable systems in medium density environments, when there is no bottleneck on transmission. The initial systems and the vast majority of the final systems have low e-compositionality. Partially e-compositional final systems (circled) occur with very low frequency, and only when the environment is unstructured.

Figures 5.9 and 5.10 give the distributions of systems with respect to $E(\mathcal{O})$, for 1000 runs of the ILM with the medium and dense environments.

Comparison of Figures 5.8, 5.9 and 5.10 shows that, as environment density increases the frequency of final non-compositional systems increases. In the sparse environments, partially compositional systems do occur, and are more frequent in the unstructured environment. Highly compositional systems occur with very low frequency in the sparse, structured environment. Partially e-compositional systems occur with very low frequency in the medium, unstructured environment. In the dense environments all systems are non-compositional, regardless of the degree of environment structure.

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

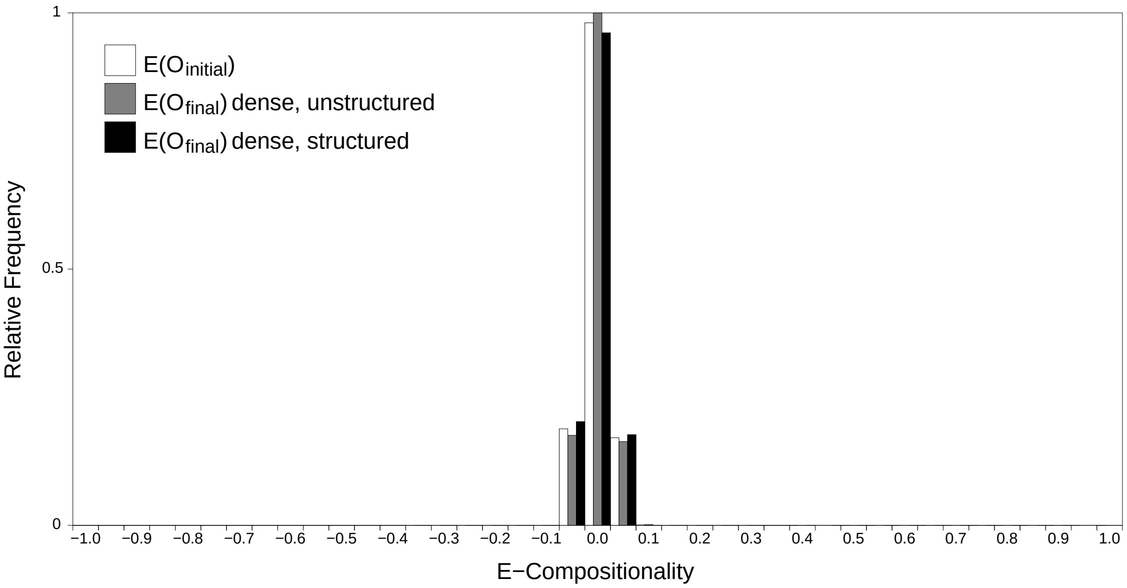


Figure 5.10: E-compositionality of initial and final, stable systems in dense environments, when there is no bottleneck on transmission. Both the initial and final systems have low e-compositionality.

Secondly, why does environment density impact on the compositionality of the emergent systems? Figure 5.11 plots the values of $E\left(O_{initial}\right)$ against $E\left(O_{final}\right)$ for the simulation runs in the sparse, structured environment. As can be seen from the Figure, the runs can be split into three groups:

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

All environments exhibit runs falling into groups (a) and (b). Only when the environment is sparse and structured do group (c) points occur, representing runs which converge on highly compositional languages. Is $E\left(O_{final}\right)$ related to $E\left(O_{initial}\right)$? Table 5.1 gives the mean and standard deviations of the initial values of $E\left(O_{initial}\right)$, categorised according to which group they fall into.

As can be seen from the first column of the Table, runs in all environments have a mean value of $E\left(\mathcal{O}_{initial}\right)$ of approximately 0. However, these initial values are much more tightly distributed around the mean in the more densely filled environments. The second column gives the mean $E\left(\mathcal{O}_{initial}\right)$ for simulation runs where $E\left(\mathcal{O}_{initial}\right) = E\left(\mathcal{O}_{final}\right)$.

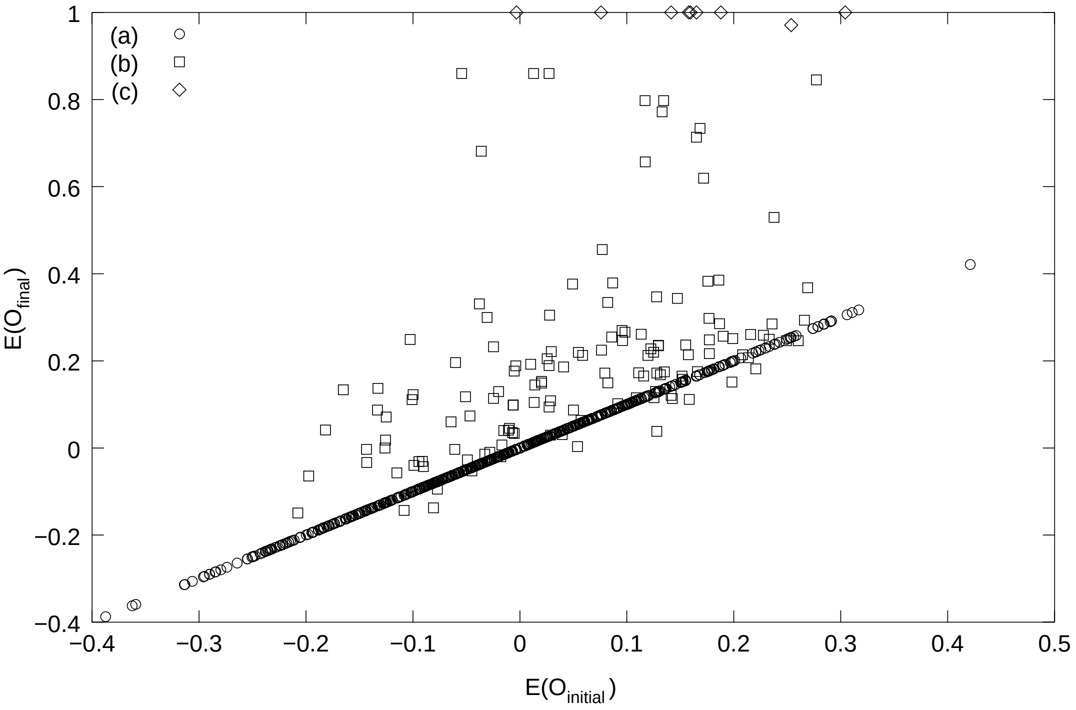


Figure 5.11: Initial e-compositionality against final e-compositionality for runs in the sparse structured environment. Each point represents a single run. The runs can be separated into three groups. Points labelled as (a) have the same initial and final e-compositionality. For points labelled as (b) the e-compositionality of the initial and final systems is different, but the final system is not highly e-compositional. Points labelled (c) represent runs where the final language is highly e-compositional.

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

Table 5.1: Sensitivity to initial conditions. The table gives the mean (μ) and standard deviation (σ) of the e-compositionality of the initial systems in the various environments (sp = sparse, m = medium, d = dense, us = unstructured, s = structured), broken down by the three groups identified in Figure 5.11. Standard deviation is given once only, as σ for each subgroup is approximately the same as that for all groups combined. The mean for group c points is higher than that for group b points, which is generally higher than that for group a points. As environment density increases, the initial values are clustered more tightly around the mean.

These are somewhat lower than the overall mean, and are lower than the mean $E\left(\mathcal{O}_{initial}\right)$ for simulation runs which move away from initial value (excluding the values for the dense environments, which buck the overall trend). Also, for the group b runs in sparse and medium environments, the mean value of $E\left(\mathcal{O}_{initial}\right)$ is lower for unstructured environments than for structured environments. Finally, the mean $E\left(\mathcal{O}_{initial}\right)$ for simulation runs which converge on highly compositional languages is higher still.

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

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

However, while shared feature values allow the possibility of the spread of compositionality, they also inhibit it — in a structured environment, any compositional tendency in the initial random mapping has to cover a large number of meanings which share feature values. If only some of these meanings share a character for a particular feature value, then the other meanings, which do not share the character, are likely to outweigh the slight compositional tendency. In contrast, in unstructured environments fewer meanings share feature values, therefore the initial random system has to be less 'lucky' in the assignment of characters to feature values. This is reflected in the fact that the mean $E\left(\mathcal{O}_{initial}\right)$ has to be higher in structured environments before $E\left(\mathcal{O}_{final}\right)$ moves away from $E\left(\mathcal{O}_{initial}\right)$, and also in the fact that the average $E\left(\mathcal{O}_{final}\right)$ in unstructured environments is higher (see Figure 5.8). In structured environments, the initial compositional tendency has to be strong to escape the attraction of the overall non-compositional mapping, but once this attraction has been escaped highly compositional systems can emerge. In contrast, in unstructured environments the attraction of the initial non-compositional mapping is weaker, due to the reduced degree of feature-value sharing, but the potential spread of compositionality is reduced.

5.3.2 Linguistic evolution in the presence of a bottleneck

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

To this end, runs of the ILM were carried out, with step 4 of the iteration algorithm being replaced by the bottleneck condition — each individual observes e meaning-signal pairs, randomly selected from the set of observable behaviour produced by the preceding agent. e will be experimentally varied. Selection of a value of e depends on the desired degree of coverage $c\left(\mathcal{E}, e\right)$, but also on the number of distinct feature values included in the meanings in $\mathcal{E}$ and the rate of seeing distinct feature values with respect to the rate of seeing distinct meanings. For example, in the sparse unstructured environment (see Figure 5.5 a) there are 12 distinct meanings, and 15 distinct feature values (values 1,2,3,4 and 5 for each feature). In the case where $e < 12$ in this environment it is therefore impossible for a learner to observe all possible feature values paired with a (sub)signal. This will have consequences for the stability of the communication systems through the bottleneck. Consequently, as a simplifying rule of thumb we will not consider the case where $e < 12$. This rules out simulation runs in the sparse environments where $c\left(\mathcal{E}, e\right) < 0.65$.

100 runs of the ILM were carried out for:

- the sparse unstructured and structured environments given in Figures 5.5 (b) and 5.6 (b) with $c = 0.8$ ($e = 19$).
- the medium unstructured and structured environments in Figures 5.5 (c) and 5.6 (c) with $c = 0.4$ ($e = 16$), $c = 0.6$ ($e = 28$) and $c = 0.8$ ($e = 49$).
- the dense unstructured and structured environments in Figures 5.5 (d) and 5.6 (d) with $c = 0.25$ ($e = 18$), $c = 0.4$ ($e = 32$), $c = 0.6$ ($e = 57$) and $c = 0.8$ ($e = 101$)

The distribution of the initial and final systems for these runs in terms of e-compositionality is given in Figures 5.12–5.19.

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

The main result apparent from these Figures is that, in the presence of a bottleneck, highly compositional languages emerge with high frequency, and emerge most frequently when the environment is structured. The nuances of this general result will be returned to below. First, however, the main result must be explained.

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

Why are compositional languages so strongly preferred when there is a bottleneck on transmission? Holistic languages cannot persist in the presence of a bottleneck. The meaning-signal pairs of a holistic language have to be observed to be reproduced. When

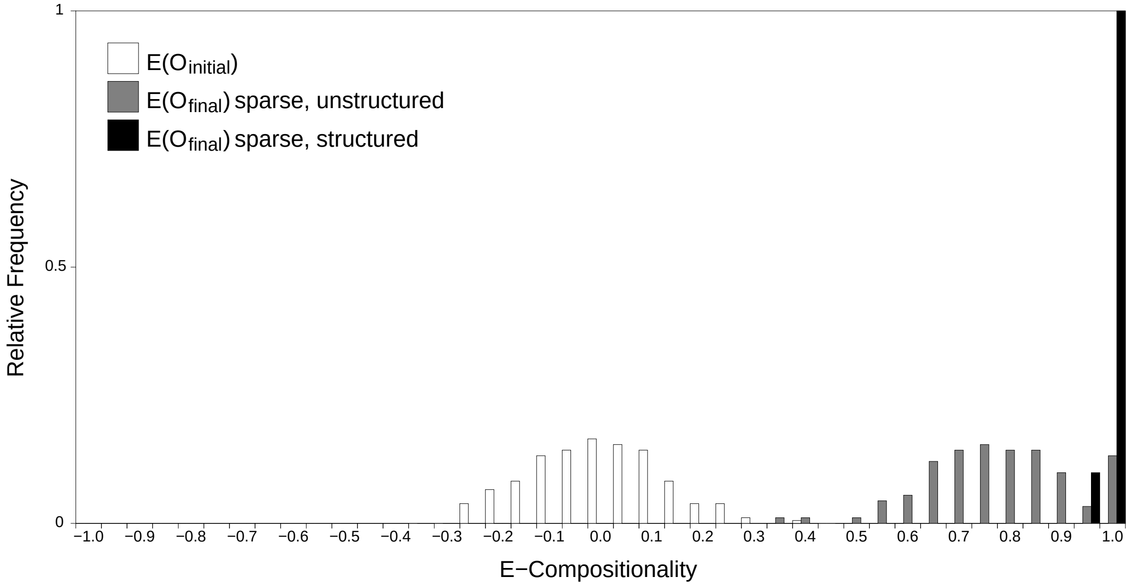


Figure 5.12: Compositionality of initial and final languages in sparse environments, in the presence of a bottleneck on cultural transmission ($c\,(\mathcal{E}, e) = 0.8$). Highly compositional languages are highly frequent, and are most frequent when the environment is structured.

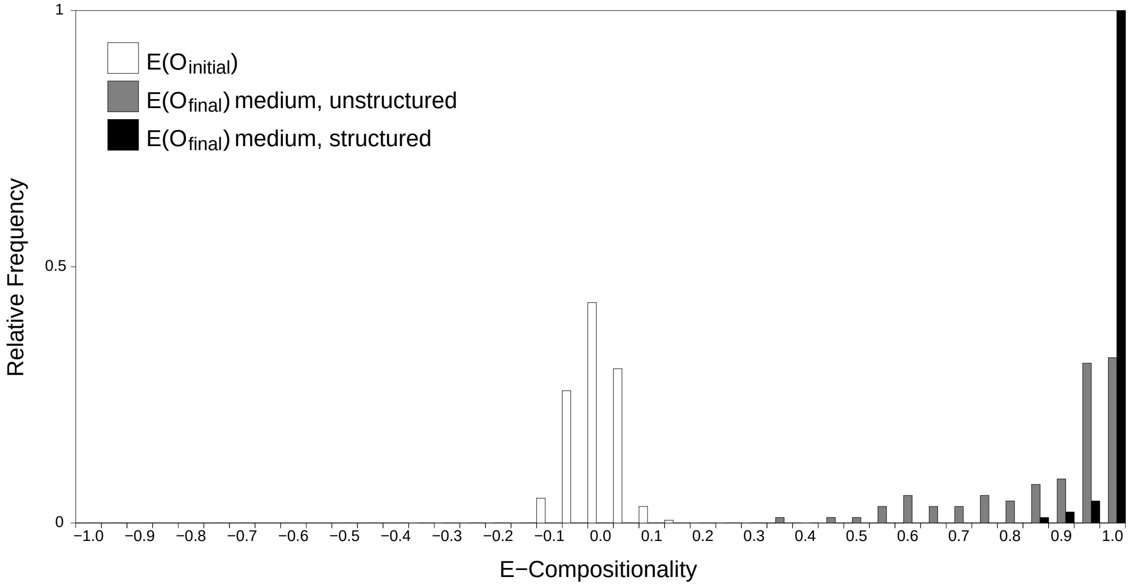


Figure 5.13: Compositionality of initial and final languages in medium density environments, in the presence of a bottleneck on cultural transmission ($c\,(\mathcal{E}, e) = 0.4$). As in Figure 5.12, highly compositional languages are highly frequent, and are most frequent when the environment is structured.

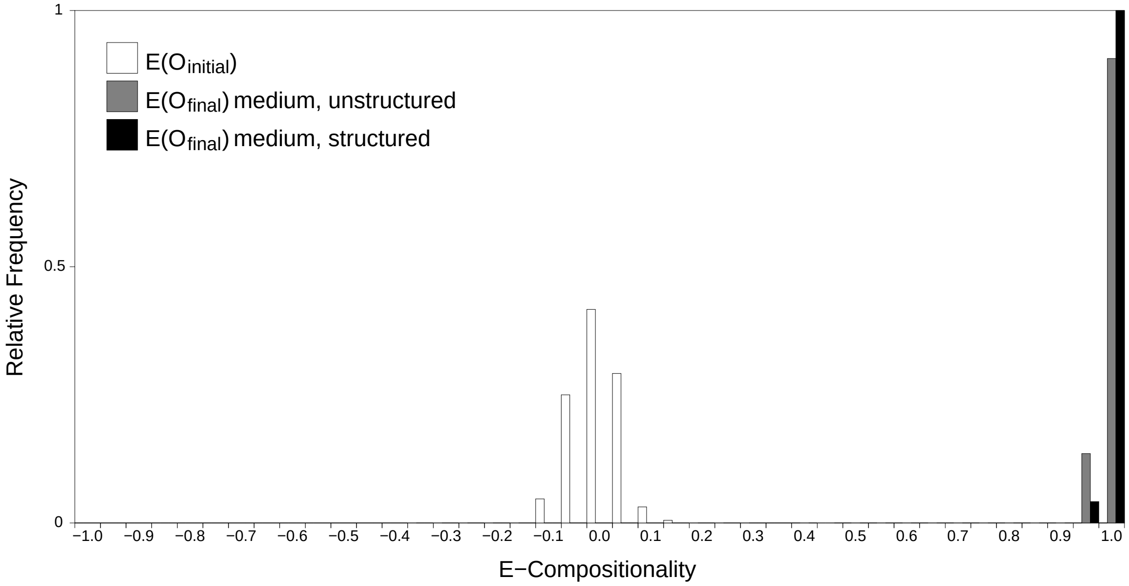


Figure 5.14: Compositionality of initial and final languages in medium density environments, in the presence of a bottleneck on cultural transmission ($c\left(\mathcal{E}, e\right) = 0.6$). There is less disparity between unstructured and structured environments.

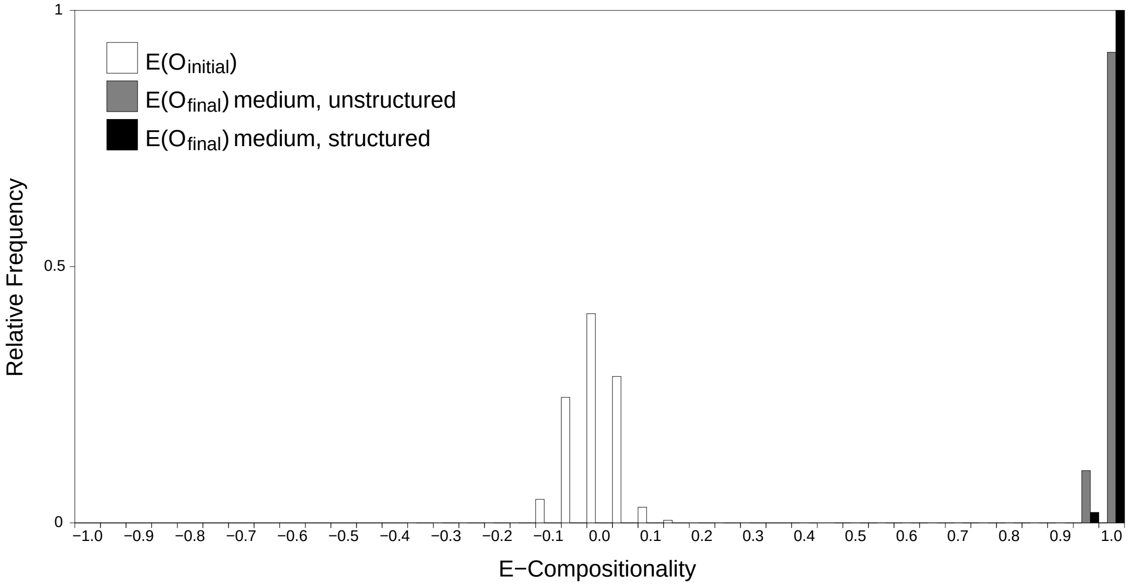


Figure 5.15: Compositionality of initial and final languages in medium density environments, in the presence of a bottleneck on cultural transmission ($c\left(\mathcal{E}, e\right) = 0.8$).

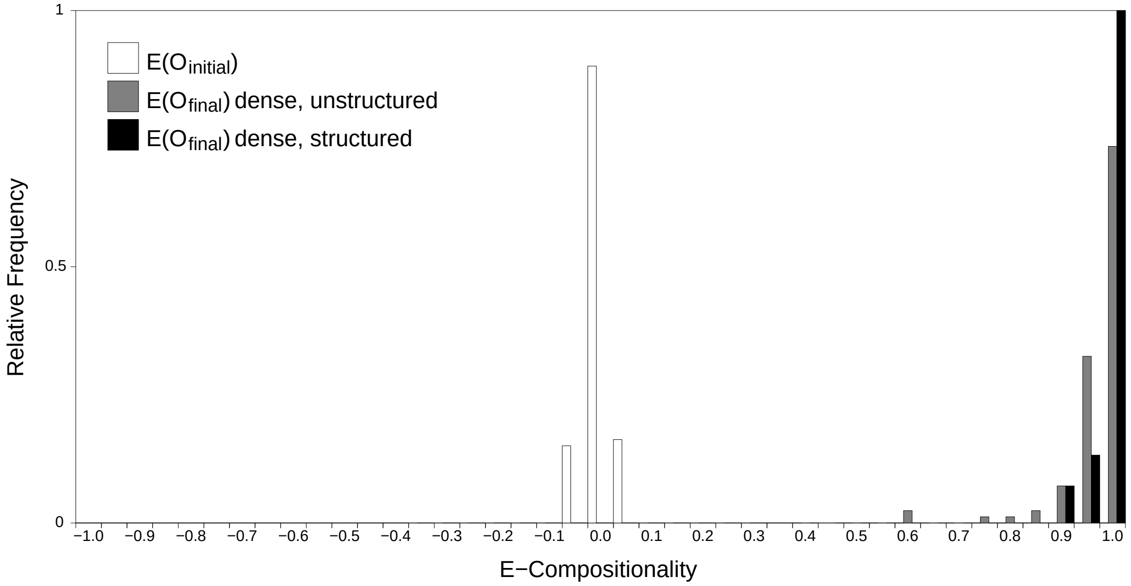


Figure 5.16: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c\left(\mathcal{E},e\right)=0.25$). While the majority of final languages are highly compositional, some partially compositional systems do exist when the environment is unstructured.

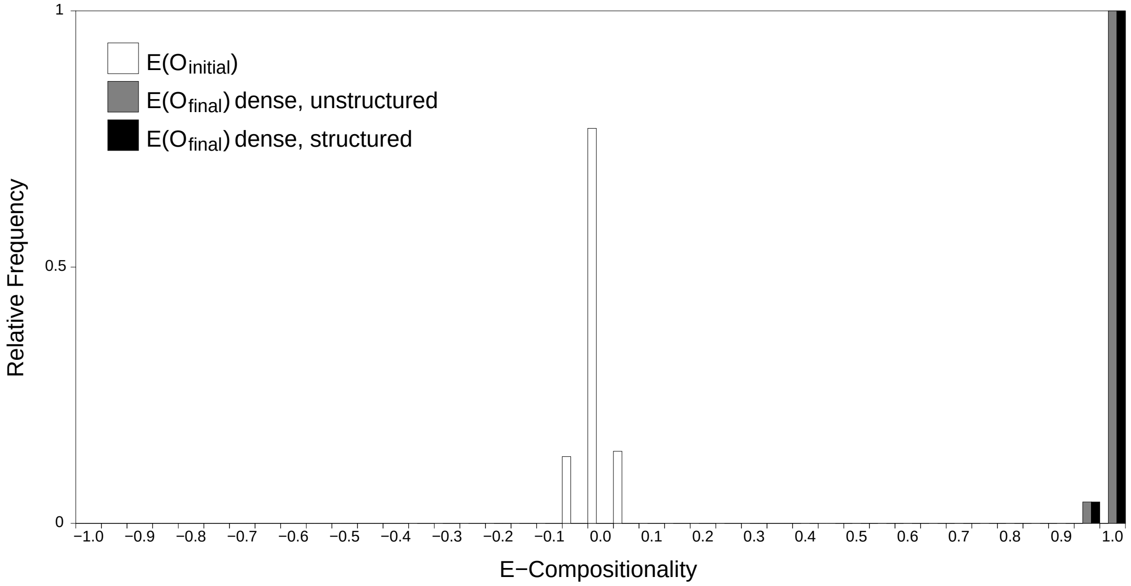


Figure 5.17: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c\left(\mathcal{E},e\right)=0.4$).

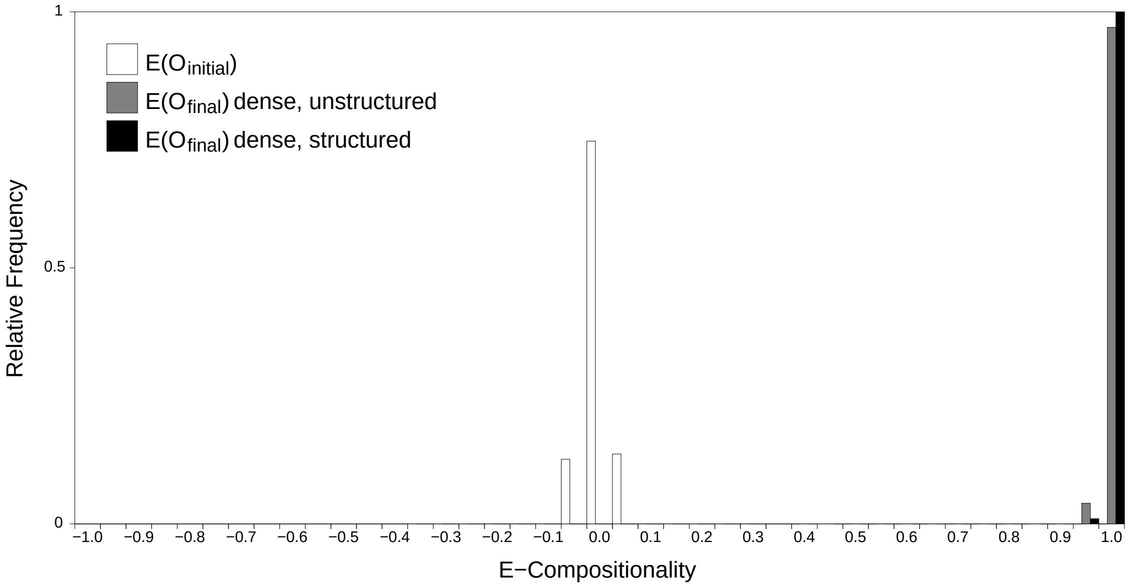


Figure 5.18: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.6$).

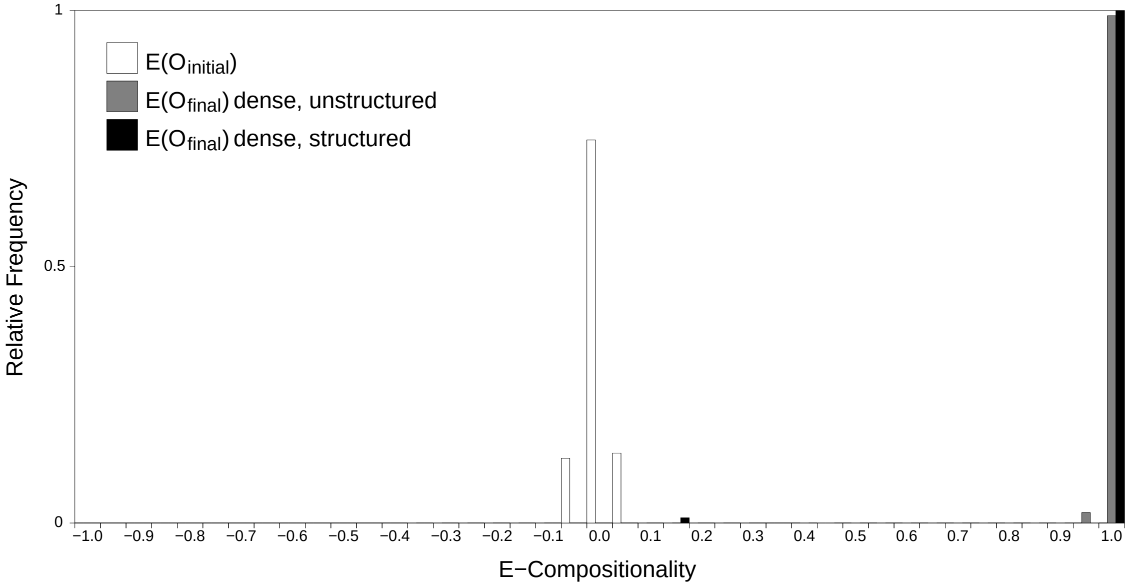


Figure 5.19: Compositionality of initial and final languages in dense environments, in the presence of a bottleneck on cultural transmission ($c(\mathcal{E}, e) = 0.8$).

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

In a structured environment the advantage of compositionality is at a maximum. In such environments, meanings share feature values with several other meanings. A language mapping these feature values to a signal substring is highly generalisable. When the environment is unstructured, meanings share feature values with few or no other meanings. In the most extreme case, a meaning may have a value for a particular feature which no other meaning has. The signal associated with that meaning cannot then be deduced from observations of the signals associated with other meanings, and has to be observed to be learned. Consequently, compositional language in an unstructured environment is less stable through the transmission bottleneck. The resultant languages are a compromise between the push towards compositionality introduced by the bottleneck and the pull back towards randomness resulting from the possibility of not observing a particular feature value paired with a subsignal.

The severity of transmission bottleneck, $c\left(\mathcal{E}, e\right)$, does appear to have an impact on the compositionality of the emergent languages — for example, comparison of Figure 5.13 with Figures 5.14 and 5.15 suggests that the difference between structured and unstructured environments is most pronounced when c is low. However, consideration of all the results taken together suggests that variation in c alone may not explain the observed patterns of behaviour. The results presented here can be split into two groups on gross qualitative grounds:

1. Situations where highly compositional systems are highly frequent for both structured and unstructured environments, with emergent languages in structured environments tending to be slightly more compositional. This occurs in the medium density environments for $c\left(\mathcal{E}, e\right) = 0.6$ or 0.8 (Figures 5.14 and 5.15), and in dense environments for $c\left(\mathcal{E}, e\right) = 0.4$, 0.6 or 0.8 (Figures 5.17–5.19)
2. Situations where highly compositional systems emerge with high frequency in structured environments, and the emergent systems in unstructured environments exhibit a range of compositionality, from partial to high. The sparse environment

runs where $c(\mathcal{E}, e) = 0.8$ (Figure 5.12) match this description, as do the results for the medium density environment where $c(\mathcal{E}, e) = 0.4$ (Figure 5.13) and the dense environment where $c(\mathcal{E}, e) = 0.25$ (Figure 5.16).

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

This explanation, based purely on $c(\mathcal{E}, e)$, breaks down when confronted with the results for sparse environments, with high c (Figure 5.12). While the theory predicts little difference between the compositionality of the languages in structured and unstructured environments, the results show a large difference. The importance of environment structure is *greater* than the theory predicts for that level of c. The theory also fails somewhat to account for the results for dense environments with $c(\mathcal{E}, e) = 0.25$. In this case the theory predicts a wider disparity between structured and unstructured environments than observed in the medium density environment where $c(\mathcal{E}, e) = 0.4$, given the lower value of c. However, the structure of the environment in fact has slightly *less* impact.

A more satisfying analysis can be gained by adapting Brighton's (2002) equations for calculating the probability of seeing a particular feature value given a particular number of exposures. After e observations a learner will have accumulated a set of observations of values for the particular ith feature f_i. Let us call this set of observations O_{f_i}. Brighton gives the probability of a particular value v being in this set of observations, $Pr\left(v \in O_{f_i}\right)$, as:

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

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

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

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

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

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

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

In other words, the probability of being able to express a given meaning compositionally is the product of the probabilities of having seen each feature value paired with a subsignal. We can then average $Pr\left(m|O_{f_1}, O_{f_2}, O_{f_3}\right)$ for all $m \in \mathcal{E}$, to give the average probability of being able to produce an utterance compositionally. Table 5.2 compares the values of $\overline{Pr}\left(m\right)$ (the value of $Pr\left(m|O_{f_1}, O_{f_2}, O_{f_3}\right)$, averaged over all meanings in $\mathcal{E}$) for structured and unstructured environments, for various values of e.

The difference between values of $\overline{Pr}\left(m\right)$ for structured and unstructured environments provide a useful measure of the relative stability advantage of compositional language in structured environments over compositional language in unstructured environments. This

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

Table 5.2: Comparison of $\overline{Pr}\,(m)$ for structured and unstructured environments of various densities, for various values of e. The Difference column gives $\overline{Pr}\,(m)$ for structured environments minus $\overline{Pr}\,(m)$ for unstructured environments, and is a measure of the relative stability advantage of compositional systems in the structured environment — the greater the difference, the greater the stability advantage of compositionality in structured environments.

size of this value corresponds fairly well to the differences between the final systems in structured and unstructured environments observable in Figures 5.12–5.19. For example, the observable difference is greatest in Figure 5.12, followed by Figure 5.13, and these two settings of environment density and e yield the largest and second-largest differences in the Table.

5.3.3 Summary

To summarise the results presented so far, it has been shown that environment density, environment structure, and bottleneck impact on the cultural evolution of compositionality. In the absence of a bottleneck, highly compositional language is unlikely to evolve. Highly compositional languages only evolve in structured environments, due to the increased potential for spread of compositionality arising from the large number of shared feature values between meanings. The emergence of such systems is highly sensitive to the initial, random assignment of signals to meanings.

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

5.4 Exploring the impact of learning bias

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

The investigations outlined in this Section will roughly follow the format of the experiments outlined in Chapter 3, Section 3.4 — the ability of networks with various weight-update rules to acquire, maintain and construct an optimal system will be explored in Sections 5.4.1–5.4.3. In Section 5.5 I will describe the key properties of the learning bias required to acquire, maintain and construct optimal compositional languages.

5.4.1 Acquisition of a compositional system

The first issue is to ascertain whether individual agents, in isolation, can acquire a perfectly compositional, unambiguous communication system. Such a language, $\mathcal{L}$, was constructed according to the feature value–character mapping given in Table 5.3. As in the previous section, $F = 3$, $V = 5$, $l_{max} = 3$ and $\Sigma = \{a, b, c, d, e, f, g, h, i, j\}$. The medium density, structured environment illustrated in Figure 5.6 (c) provided the set of meanings $\mathcal{E}$. With respect to this environment (or indeed any other), L is perfectly e-compositional — $E\left(\mathcal{L}\right) = 1$.

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

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

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

Table 5.3: A feature value lookup table for a compositional language. The signal characters are concatenated in the order of the feature values — for example, $(1\ 1\ 1)$ would be expressed as jeh.

- Production of the signal associated with m_i always[3] resulted in s_j being produced, i.e. $\langle m_i, s_j \rangle$ can be reproduced in production.
- **and** reception of s_j always resulted in the interpretation m_i, i.e. $\langle m_i, s_j \rangle$ can be reproduced in reception, meaning that the agent would communicate optimally with itself or another agent using the same weight-update rule exposed to $\mathcal{L}$.

18 of the 81 possible rules succeeded in the acquisition task, with the remaining 63 failing to reproduce the observed system. These 18 successful rules were classified as [$+$maintainer, ±constructor][4] in the simple associative network tests outlined in Chapter 3.

Note that there is both a degree of continuity with the previous classification — the same 18 rules are separated out in both classifications — and a discontinuity — 31 rules were able to acquire unambiguous systems in the previous classification, whereas in the new classification only 18 can. Rules which were [$+$learner, $-$maintainer] in the earlier categorisation[5] are incapable of acquiring $\mathcal{L}$. This is due to the fact that the learning procedure for the structured communication systems requires that learners be able to handle multiple active input nodes when learning — all components of a meaning and signal are presented simultaneously to the learner. As discussed in Chapter 3, [$+$learner, $-$maintainer] weight-update rules are biased in favour of many-to-one mappings, and this bias, given the multiple active units present in each exposure, leads to an inability to acquire $\mathcal{L}$.

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

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

[5] Bear in mind that the terms $+$learner, $+$maintainer and so on are based on the classification given in Chapter 3, and are shorthands for a set of restrictions on the relationship between the values of α, β, γ and δ in the weight-update rules. While the mnemonic of $+$learner was quite transparent in Chapter 3, it is less so here — certain weight-update rules which fit the pattern glossed as $+$learner cannot learn a compositional system. However, these terms will be preserved, for reasons which will become obvious.

Of those rules which were classified as capable of acquiring $\mathcal{L}$, a further evaluation was made as to whether they reproduced their acquired mapping in an i-compositional or non-i-compositional manner. Agents using the 18 [+maintainer] rules were trained on $\mathcal{L}$, as before. They were then called upon to produce for each $m \in \mathcal{E}$ to give a set of meaning-signal pairs, with an associated underlying set of winning meaning analysis–signal analysis pairs $\mathcal{A}_p$. Similarly, they were prompted with each s used in $\mathcal{L}$, to yield a set of meaning analysis-signal analysis pairs $\mathcal{A}_r$. $\mathcal{A}_p$ and $\mathcal{A}_r$ were evaluated according to the internal compositionality measure given in Section 5.2.5. Weight-update rules which yielded sets of analysis pairs where $I\left(\mathcal{A}_p\right) = I\left(\mathcal{A}_r\right) = 1$ were classified as [+ic-preserver], otherwise they were classified as [$-$ic-preserver].

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

5.4.2 Maintenance through a bottleneck

Next, maintenance tests, similar to those outlined for the associative network model in Chapter 3, were carried out to assess the ability of the various weight-update rules to maintain an optimal system. In Chapter 3 the maintenance tests assessed whether populations of agents using the weight-update rules were able to acquire an optimal system in the presence of noise, without a bottleneck. The maintenance test here measures the ability of populations of agents using these weight-update rules to preserve the optimal system through a bottleneck on cultural transmission.

Recall from the description of the ILM given in Section 5.2.3 above that the agents in the initial population use some predefined communication system L. For the experiments outlined in this section, the initial population's set of weights $\mathcal{W}$ were constructed such that the $p(m)$ of the initial L generates the unambiguous, perfectly e-compositional meaning-signal pairs encoded in $\mathcal{L}$, described above — in other words, the initial language in these simulations is predefined and perfectly compositional. ILMs were run with each of the 81 possible learning rules, with each learning receiving 28 exposures to the communication system of the previous generation ($e = 28$, $c\left(\mathcal{E}, e\right) = 0.6$). Populations were defined as having maintained a compositional system if $E\left(\mathcal{O}\right)$ and $I\left(\mathcal{A}\right)$ remained above 0.95 for every generation of ten 100 generation runs.

No [$-$maintainer] rules succeeded in this task. All [+maintainer,$-$ic-preserver] weight-update rules exhibited behaviour similar to run (a) in Figure 5.20, and failed to maintain

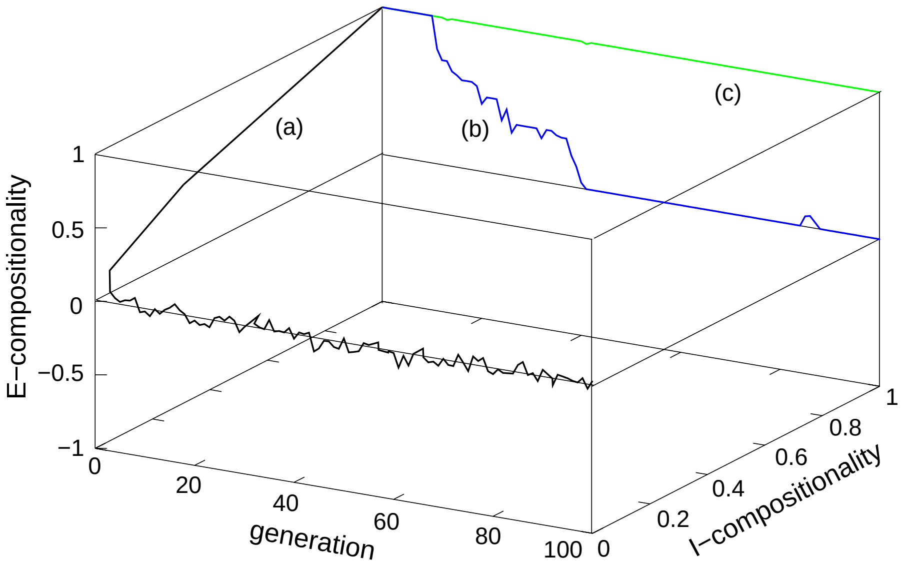


Figure 5.20: Three characteristic patterns of behaviour in populations using the 81 weight-update rules when attempting to maintain a perfectly compositional language. The population in run (a) rapidly collapses from using the initial system to an i- and e-holistic system. This behaviour characterises [+maintainer, −ic-preserver] rules. The population in run (b), which is using a weight-update rule classified as [+maintainer, −constructor, +ic-preserver], loses the initial language, although its final language is i-compositional. The population in run (c), which characterises [+constructor, +ic-preserver] weight-update rules, maintains the perfectly compositional initial language.

the perfectly compositional system. Of the seven [+maintainer, +ic-preserver] rules, five behaved in a similar fashion to run (b) in Figure 5.20. Only two [+maintainer, +ic-preserver] weight-update rules succeeded in maintaining a compositional system. Populations using these two weight-update rules behaved like run (c) in Figure 5.20.

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

5.4.3 Construction through a bottleneck

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

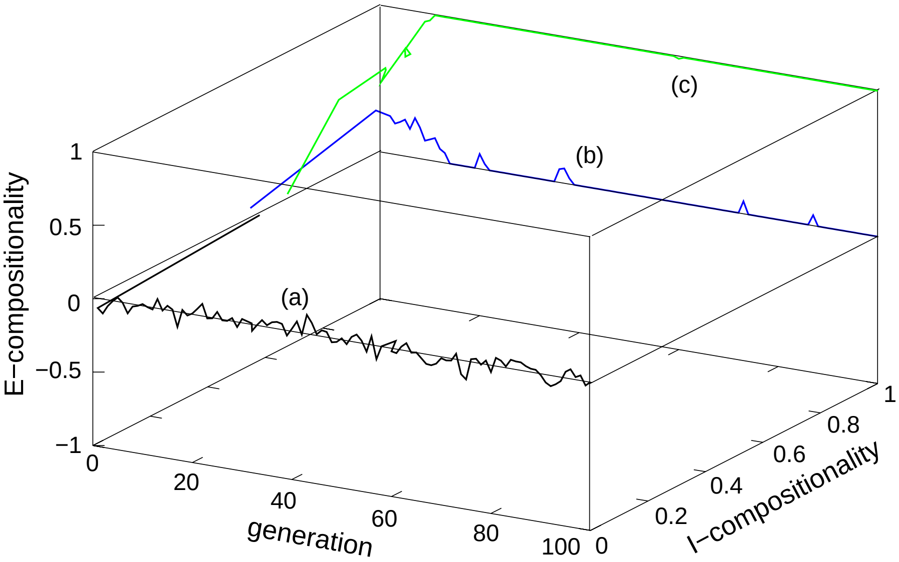


Figure 5.21: Three characteristic patterns of behaviour in populations using the 81 weight-update rules when attempting to construct a compositional language. The population in run (a) converges to an i- and e-holistic system. This behaviour characterises [+maintainer, −ic-preserver] rules. The population in run (b), which is using a weight-update rule classified as [+maintainer, −constructor, +ic-preserver] converges on a language which is i-compositional but e-holistic. The population in run (c), which characterises the two [+constructor, +ic-preserver] weight-update rules, constructs a language which is highly i- and e-compositional.

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

Perhaps unsurprisingly, only the two weight-update rules which were capable of maintaining a preexisting optimal system were capable of constructing a compositional system from random initial behaviour. As mentioned earlier, these rules have the [+constructor, +ic-preserver] classification. Populations using these weight-update rules behaved in a similar fashion to population (c) in Figure 5.21. The [+maintainer, −ic-preserver] and [+maintainer, −constructor, +ic-preserver] weight-update rules were incapable of constructing an optimal system, and behaved similarly to population (a) and (b) respectively in Figure 5.21.

5.4.4 The classification hierarchy

The numbers of weight-update rules with the various complete classifications are given in Table 5.4. Note that it is no longer possible to draw up a neat hierarchy of weight-update rules — while the hierarchy relating the [±learner], [±maintainer] and [±constructor] features persists, the [±ic-preserver] distinction cuts across this neat division, grouping [$+$maintainer, $-$constructor] rules together with [$+$constructor] rules.

5.5 The key bias

What pattern of assignment of values to α, β, γ and δ results in some weight-update rules being able to construct compositional languages from scratch, whereas other weight-update rules cannot maintain or learn such a system? In Chapter 3 this type of question was tackled by considering the connection weights in a small network before and after exposure to a single meaning-signal pair. A similar strategy is pursued here. I will consider the simple case where $F = V = 2$, $l_{max} = 2$, $\Sigma = \{p, q\}$. A network of appropriate dimensions was trained on two meaning-signal pairs, $\langle (1\ 1), pp\rangle$ and $\langle (1\ 2), pq\rangle$, using the weight-update rule $(a\ b\ c\ d)$. The connection weights in this network after observing these two meaning-signal pairs are given in Figure 5.22.

5.5.1 An overview of the learning biases

In this Section I will focus on the g values for the various possible analyses in the network as a whole, in fairly broad terms, returning in more detail to smaller parts of the network in Section 5.5.2.

The g for a particular meaning analysis–signal analysis pair depends on one or two connection weights from the network shown in Figure 5.22. Table 5.5 gives the g values for various meaning analysis-signal analysis pairs. To simplify matters, only one possible ordering of components in the signal is given in the Table. For example, there are three possible analyses of the signal pp — $\{pp\}$, $\{p*, *p\}$ and $\{*p, p*\}$, but only the first two are included in Table 5.5.

5.5.1.1 [$+$maintainer, $-$constructor, $\pm$ic-preserver] rules

We have already established that [$+$maintainer] rules are characterised by the restriction:

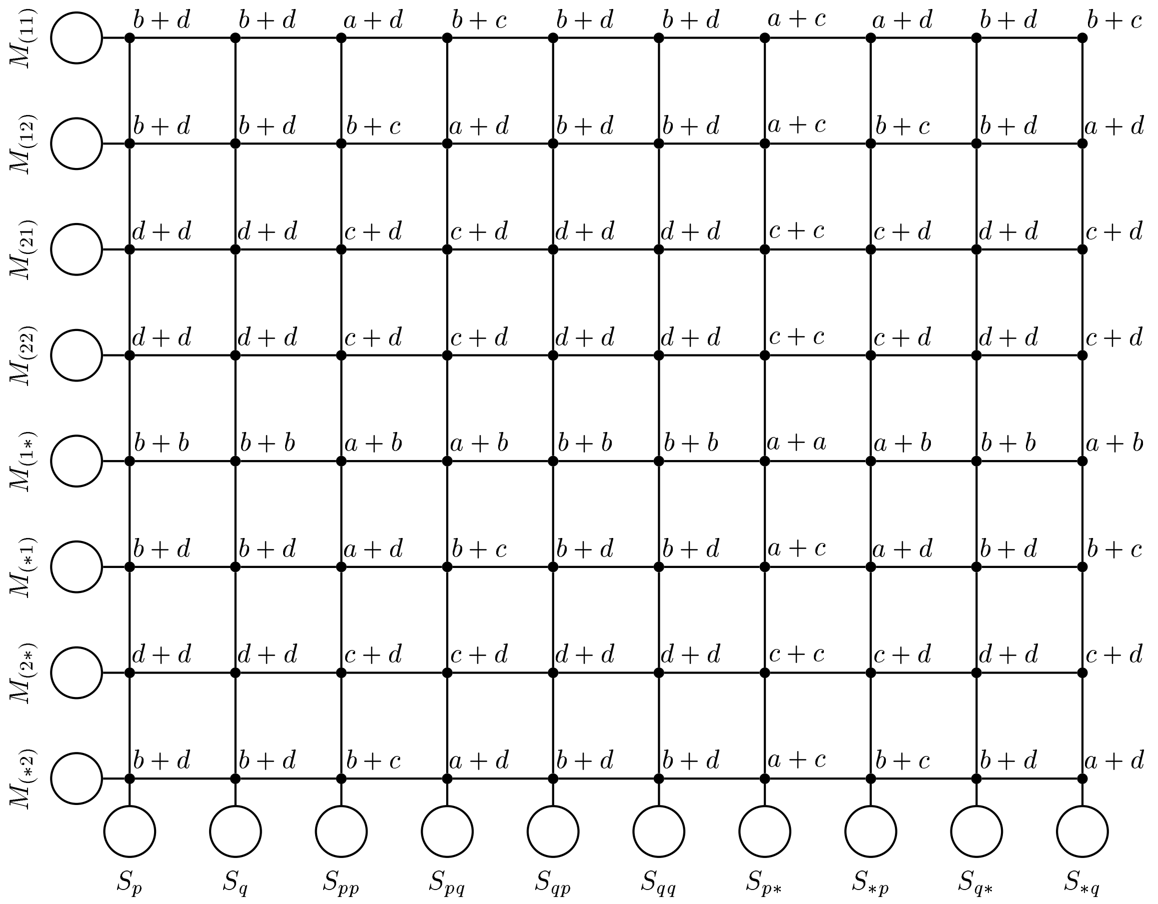
> A weight-update rule is [$+$maintainer] if $\alpha > \beta \wedge \delta \geq \gamma$

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

Table 5.4: The number of weight-update rules of each particular complete classification, from the sample of 81, and a summary of their properties with respect to compositional languages. “Acquire?” indicates whether agents using rules with this classification can acquire a perfectly compositional language. “Acquire i-compositionally?” indicates whether they can reproduce an acquired system in a perfectly i-compositional manner. “Maintain?” indicates whether agents using a weight-update rule with this classification can maintain a perfectly compositional language through a bottleneck, in the context of the ILM. “Construct?” indicates whether such agents can construct a highly compositional system from random initial behaviour in the presence of a bottleneck, in the context of the ILM.

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

Table 5.5: The g values of various meaning analysis -signal analysis pairs in the network after storing $\langle(1\ 1)\ , pp\rangle$ and $\langle(1\ 2)\ , pq\rangle$ using learning rule $(a\ b\ c\ d)$. Meanings which have not been observed paired with a signal are given in italics. Note that only one ordering of two-component signal analyses is given — there are in fact another 4 two-component signal analyses ($\{*p, p*\}$, $\{*q, p*\}$, $\{*p, q*\}$, $\{*q, q*\}$), but these other analyses can be safely ignored for the purposes of this analysis.



	S_p	S_q	S_{pp}	S_{pq}	S_{qp}	S_{qq}	S_{p*}	S_{*p}	S_{q*}	S_{*q}
$M_{(11)}$	$b+d$	$b+d$	$a+d$	$b+c$	$b+d$	$b+d$	$a+c$	$a+d$	$b+d$	$b+c$
$M_{(12)}$	$b+d$	$b+d$	$b+c$	$a+d$	$b+d$	$b+d$	$a+c$	$b+c$	$b+d$	$a+d$
$M_{(21)}$	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$c+c$	$c+d$	$d+d$	$c+d$
$M_{(22)}$	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$c+c$	$c+d$	$d+d$	$c+d$
$M_{(1*)}$	$b+b$	$b+b$	$a+b$	$a+b$	$b+b$	$b+b$	$a+a$	$a+b$	$b+b$	$a+b$
$M_{(*1)}$	$b+d$	$b+d$	$a+d$	$b+c$	$b+d$	$b+d$	$a+c$	$a+d$	$b+d$	$b+c$
$M_{(2*)}$	$d+d$	$d+d$	$c+d$	$c+d$	$d+d$	$d+d$	$c+c$	$c+d$	$d+d$	$c+d$
$M_{(*2)}$	$b+d$	$b+d$	$b+c$	$a+d$	$b+d$	$b+d$	$a+c$	$b+c$	$b+d$	$a+d$

Figure 5.22: Connection weights after observing and learning the meaning-signal pairs $\langle (1\ 1)\,, pp\rangle$ and $\langle (1\ 2)\,, pq\rangle$ using weight-update rule $(a\ b\ c\ d)$.

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

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

Why does this pattern of weight changes lead to the network exploiting the internal compositional representations? I will focus first on weight-update rules which are characterised as [+maintainer,−constructor]. For these rules, $\delta = \gamma$. Table 5.6 gives the g values for various analyses for [+maintainer, −constructor, ±ic-preserver] rules.

Tables 5.6 (a) and (b) highlight the relevant values of g for [+maintainer, $-$constructor, $-$ic-preserver] rules. For these rules, $\alpha < \delta$ (as is the case in Table 5.6 (a)) or $\alpha = \delta$ (as is the case in Table 5.6 (b)). In the former case, the networks are strongly biased against compositional systems — the compositional representational capacities of the network are not exploited due to the fact that δ dominates α. In the latter case, the networks are neutral with respect to compositionality — both one- and two-component analyses are possible, due to the fact that $\alpha = \delta$. Finally, Table 5.6 (c) highlights the relevant values of g for [+maintainer,$-$constructor,+ic-preserver] rules. For these rules, α dominates δ. Consequently, analyses involving multiple components are preferred to those involving single components, indicating a bias in favour of i-compositional systems.

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

5.5.1.2 [+constructor, ±ic-preserver] rules

[+constructor] rules are characterised by the restriction:

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

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

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

Why does this pattern of weight changes lead to the network exploiting the internal compositional representations? Tables 5.7 (a) and (b) highlight the relevant values of g for [+constructor,$-$ic-preserver] rules. For these rules, $\alpha < \delta$ (as is the case in Table 5.7 (a)) or $\alpha = \delta$ (as is the case in Table 5.7 (b)). The parallels with the [+maintainer, $-$constructor, $-$ic-preserver] rules are clear. In the case where $\alpha < \delta$, the networks are strongly biased against i-compositional systems — the compositional representational capacities of the network are not exploited due to the fact that δ dominates α. In the case where $\alpha = \delta$, the networks are neutral with respect to i-compositionality — both one- and two-component analyses are possible. Finally, Table 5.7 (c) highlights the relevant values of g for [+constructor,+ic-preserver] rules. For these rules, α dominates δ.

(a)

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

(b)

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

(c)

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

Table 5.6: The g values for meaning analysis–signal analysis pairs in the network after training on two meaning-signal pairs using the weight-update rule $(a\ b\ c\ d)$. This table focuses on [+maintainer, −constructor, ±ic-preserver] rules, for which $c = d = D$ and $a > b$. The highest g values in each row are highlighted in grey. Meanings which have not been observed paired with a signal are given in italics. (a) The [+maintainer, −constructor, −ic-preserver] rule where $D > a$. The observed meaning-signal mappings can be reproduced. Only the one-component analyses are used, indicating a bias against i-compositionality. Meanings which have not been observed map to any single-component analysis with equal probability, indicating no bias against many-to-one meaning-signal mappings. (b) The [+maintainer,−constructor,−ic-preserver] rule where $a = D$. The observed meaning-signal mappings can be reproduced, with both one-component and two-component analyses being equally probable. This indicates neutrality with respect to i-compositionality. The non-observed meanings map to both one- and two-component analyses, again with no bias against many-to-one mappings. (c) The [+maintainer,−constructor,+ic-preserver] rule where $a > D$. The observed meaning-signal pairs are reproduced using two-component analyses, indicating a bias in favour of i-compositionality. Signals for the unobserved meanings are also produced using two-component analyses, but once again many-to-one mappings are not avoided.

Consequently, analyses involving multiple components are preferred to those involving single components, indicating a bias in favour of i-compositional systems.

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

5.5.2 The two parts of the bias

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

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

The learning biases of the weight-update rules can be characterised along these two distinct dimensions — a bias in favour of (or against) exploiting internally compositional representations, and a bias in favour of (or against) a principled system of mappings from

(a)

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

(b)

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

(c)

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

Table 5.7: The g values for meaning analysis–signal analysis pairs in the network after training on two meaning-signal pairs using the weight-update rule $(a\ b\ c\ d)$. This table focuses on [+constructor,±ic-preserver] rules, for which $a > b$ and $d > c$. (a) The [+constructor,−ic-preserver] rule where $d > a$. The observed meaning-signal mappings can be reproduced. Only the one-component analyses are used, indicating a bias against i-compositionality. Meanings which have not been observed map to signals which have not been observed paired with any meaning, indicating a bias in favour of one-to-one mappings. (b) The [+constructor,−ic-preserver] rule where $d = a$. The observed meaning signal mappings can be reproduced, with both one-component and two-component analyses being equally probable. This indicates neutrality with respect to i-compositionality. As with (a), non-observed meanings map to non-observed signals, indicating a one-to-one bias. However, both one- and two-component analyses are possible. (c) The [+constructor,+ic-preserver] rule where $a > d$. The observed meaning-signal pairs are reproduced using two-component analyses, indicating a bias in favour of i-compositionality. As before, non-observed meanings map to non-observed signals, with only the two component analyses being used. The mapping is in fact perfectly one-to-one.

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

Table 5.8: Average internal compositionality of production and reception behaviour combined ($I(\mathcal{A})$), production behaviour alone ($I(\mathcal{A}_p)$) and reception behaviour alone ($I(\mathcal{A}_r)$). Results are for all 81 weight-update rules, organised according to the relationship between α and δ

feature values to signal characters. Only when the correct biases for both aspects of the problem are in place will compositional languages be maintained or constructed.

5.5.2.1 The internal compositionality bias

Comparison of Sections 5.5.1.1 and 5.5.1.2 reveals a common element to the rules which are [+maintainer, ±constructor, +ic-preserver]:

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

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

The acquisition tests outlined in Section 5.4.1 were repeated for all 81 weight-update rules. However, on this occasion the ability to reproduce the observed mapping was ignored, and the internal compositionality of their production ($I(\mathcal{A}_p)$) and reception ($I(\mathcal{A}_r)$) behaviour was measured. The results are summarised in Table 5.8.

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

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

5.5.2.2 The one-to-one bias

As discussed above in Sections 5.5.1.1 and 5.5.1.2, the biases of weight-update rules identified in Chapter 3 carry over into the more complex model — [+maintainer, $-$constructor, $\pm$ic-preserver] agents are neutral with respect to the one-to-one nature of the meaning-signal mapping, whereas [+constructor, $\pm$ic-preserver] agents are biased in favour of acquiring one-to-one mappings.

The relative values of α and δ determine whether an agent produces or receives in an internally-compositional manner. Parallel to this are the (α,β) and (δ,γ) relationships, which determines the bias with respect to the one-to-one quality of mappings. In other words, the (α,β) and (δ,γ) relationships determine whether internally-holistic and internally-compositional mappings are one-to-one or not, then the (α,δ) relationship determines which of the internally-holistic or internally-compositional analyses is actually used.

The learning biases with respect to holistic analyses and compositional analyses can therefore be looked at separately. In the previous Sections connection weights in a network exposed to two meaning-signal pairs were examined. In this Section it is sufficient to look at a network, identical in structure to the network given in Figure 5.22, which has been trained on the meaning-signal pair $\langle (1\ 1)\,, aa\rangle$. The g values of interest for the i-holistic analyses are summarised in Table 5.9 (a).

Table 5.9 (b) highlights the dominant connection weights for [+maintainer, $-$constructor, $\pm$ic-preserver] weight-update rules. The observed meaning-signal pair can be reproduced holistically. The other meanings map to every possible holistic analysis with equal probability, including the already-observed signal pp. This indicates neutrality with respect to the one-to-one nature of the meaning-signal mapping, and is an identical result to that reported in Chapter 3 for the simpler associative network model.

Table 5.9 (c) highlights the dominant connection weights in a network using a [+constructor, $\pm$ic-preserver] weight-update rule. The observed meaning-signal pair $\langle (1\ 1)\,, aa\rangle$ can be

(a)

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

(b)

Meaning	Seen?	Signal					
		p	q	pp	pq	qp	qq
$(1\ 1)$	yes	b	b	$\boxed{a}$	b	b	b
$(1\ 2)$	no	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$
$(2\ 1)$	no	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$
$(2\ 2)$	no	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$	$\boxed{D}$

(c)

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

Table 5.9: (a) The g values for i-holistic meaning analysis–signal analysis pairs after learning the meaning–signal pair $\langle(1\ 1)\,, pp\rangle$ using the weight-update rule $(a\ b\ c\ d)$. (b) The g values for a network using a [+maintainer, $-$constructor, ±ic-preserver] weight-update rule. In such rules $a > b$ and $c = d = D$. The highest g value in each row is highlighted in grey. The observed meaning–signal pair can be reproduced, but there is no bias against many-to-one mappings from meanings to signals. (c) The g values for a network using a [+constructor, ±ic-preserver] weight-update rule. In such rules $a > b$ and $d > c$. The observed meaning–signal pair can be reproduced, and there is a bias against many-to-one mappings from meanings to signals — signal pp is avoided for all meanings apart from $(1\ 1)$.

reproduced. The other meanings map to p, q, pq, qp or qq with equal probability — pp is avoided. The results for the associative network model outlined in Chapter 3 still hold in the more complex model — [+constructor] agents are biased in favour of acquiring one-to-one mappings.

[$-$maintainer] weight-update rules, as discussed in Chapter 3, either cannot acquire observed holistic meaning-signal mappings (in the case of [$-$learner, $-$maintainer] rules) or can acquire such systems but are biased in favour of many-to-one mappings between meanings and signals (in the case of [+learner, $-$maintainer] rules). In the context of the i-holistic analyses part of structured networks, this has the consequence that networks using such rules either:

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

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

Table 5.10 (a) gives the weights of the connections in the network between partially-specified components of meaning (organised according to the feature value which is specified) and partially-specified signal components (once again, different orderings are ignored — the subsignals given in the Table refer to signal components where the specified character is in the same position in the signal as the specified feature value — for example, row 1 column 1 of the Table gives the connection weight between $\{(1\ *)\}$ and $\{p*\}$).

Table 5.10 (b) highlights the dominant connection weights in the relevant portion of the network after learning using a [+maintainer, $-$constructor, ±ic-preserver] weight-update rule. The observed pairings of $(1\ *)$ with $p*$ and $(*\ 1)$ with $*p$ can be reproduced. Meaning components which have not been observed paired with any signal map to both possible signal substrings with equal probability — there is no bias against having a many-to-one mappings from feature values to signal substrings.

Table 5.10 (c) shows the dominant connection weights in a network trained on the meaning-signal pair $\langle (1\ 1)\ , aa \rangle$ using a [+constructor, ±ic-preserver] weight-update rule. As with the [+maintainer, $-$constructor, ±ic-preserver] rule outlined above, the observed feature value-subsignal pairs can be reproduced. Unlike the [+maintainer, $-$constructor, ±ic-preserver] rules, use of a [+constructor, ±ic-preserver] weight-update rule results in a one-to-one mapping between feature values and subsignals — the subsignal q is preferred

(a)

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

(b)

			Subsignal	
Feature	Value	Seen?	p	q
1	1	yes	$\boxed{a}$	b
1	2	no	$\boxed{D}$	$\boxed{D}$
2	1	yes	$\boxed{a}$	b
2	2	no	$\boxed{D}$	$\boxed{D}$

(c)

			Subsignal	
Feature	Value	Seen?	p	q
1	1	yes	$\boxed{a}$	b
1	2	no	c	$\boxed{d}$
2	1	yes	$\boxed{a}$	b
2	2	no	c	$\boxed{d}$

Table 5.10: (a) The g values for meaning component–signal component pairs after learning the meaning–signal pair $\langle (1\ 1), pp \rangle$ using the weight-update rule $(a\ b\ c\ d)$. As before, alternative orderings of signal components can be safely ignored — it is assumed that the value of the nth feature maps on to the nth character in the signal. (b) The g values for a network using a [+maintainer, $-$constructor, $\pm$ic-preserver] weight-update rule. In such rules $a > b$ and $c = d = D$. The observed feature value–signal character pairs can be reproduced, but there is no bias against many-to-one mappings from feature values to signal characters. (c) The g values for a network using a [+constructor, $\pm$ic-preserver] weight-update rule. In such rules $a > b$ and $d > c$. The observed feature value–signal character pairs can be reproduced, and there is a bias against many-to-one mappings from feature values to signal characters — signal character p is reserved for feature value 1.

to p for unseen feature values. This bias results in the preferential acquisition of perfectly compositional, perfectly one-to-one mappings for networks using [$+$constructor, $+$ic-preserver] weight-update rules.

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

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

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

5.5.3 The bias in other models

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

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

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

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

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

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

The bias against synonymy is less direct. Consider the case where there are two possible ways of expressing a given meaning, both with equal costs. During production of an utterance, the agent will be neutral with respect to these two variants, and will select one randomly. Another agent will learn from that production, and may reuse that exemplar when speaking to the agent who produced the utterance. The original agent may then learn based on that production (given the negotiation framework, agents can learn from individuals who they themselves have taught), in which case the exemplar they used in the first place will have its cost reduced. This exemplar will then be used more frequently than its synonymous alternative. The reduction in costs leads to the reinforcement of one way of expressing each meaning, leading to the elimination of synonymy.

As a final example from the symbolic models of grammar induction, the learners used in Hurford (2000) are biased in favour of using internally compositional representations. As discussed in Section 5.1, Hurford's learners are strongly biased towards acquiring compositional rules — they can acquire such rules on the basis of a single observation, and also invent in a compositional manner. Hurford also includes a direct bias against synonymy, operating at the production level — if an individual has several ways of expressing a meaning, it uses the expression it acquired first. This production bias will lead to the rapid elimination of synonymy in the population's language. The bias of Hurford's learners with respect to homonymy is less clear. There is no obvious bias against homonymy built into the learning model — unlike Kirby and Batali's models, there is no prohibition on acquiring homonymous utterances. However, homonymy is perhaps less of a problem in Hurford's model due to the extremely large character alphabet available to his agents. While Kirby and Batali use the 26 letters of the alphabet, Hurford's agents have access to 2000 distinct 'syllables', which are combined to form utterances. The probability of homonymous utterances occurring by chance is therefore very low. I would anticipate that, assuming there is no hidden bias against homonymy in Hurford's learning model, a smaller syllable inventory would lead to the more frequent emergence of homonyms in his model, with a concomitant loss of e-compositionality in the emergent systems.

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

Biases against homonymy and synonymy also determine the behaviour of ILMs involving neural network learners. Batali (1998) and Kirby & Hurford (2002) use feedforward networks with the obverter architecture. As discussed in Chapter 3, Section 3.5.2, this network architecture leads to a bias in favour of one-to-one mappings between whole meanings and whole signals. This bias also applies at the level of individual parts of meanings and parts of signals. Parts of meanings are represented by individual output nodes in these networks, and parts of signals are either represented by patterns of activation over the input nodes (in Batali's (1998) model) or as individual input nodes (in

Kirby & Hurford's (2002) model). In either case, the obverter network architecture leads to a pressure for one-to-one mappings between individual input nodes, or patterns over groups of such nodes, and individual output nodes — many-to-one mappings are unstable, whereas one-to-many mappings are unlearnable.

In contrast, Hare & Elman (1995) use an imitator feedforward network architecture. As discussed in Chapter 3, Section 3.5.3, this choice of network architecture leads to the loss, rather than emergence, of linguistic structure. This is a consequence of the many-to-one bias inherent in the imitator architecture, which applies at both the level of whole meanings and signals and at the level of subparts of meanings and signals.

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

5.5.4 Summary

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

5.6 One-to-one biases and the acquisition of linguistic structure

In Chapter 3 I discussed evidence from vocabulary acquisition research that suggests that children are biased in favour of one-to-one mappings between meanings and words. This type of bias, as demonstrated by the computational models outlined in Chapter 3, will result in the emergence of functional communication through purely cultural processes. In this Chapter we have seen that a one-to-one bias, in combination with a tendency to exploit internally-compositional representations and a bottleneck on transmission, leads to the cultural emergence of compositional language. In this Section I will argue that human language learners bring one-to-one biases to the acquisition of linguistic structure, as well as to the acquisition of unstructured lexical items. I will have less to say about the tendency of humans to exploit internally compositional representations — this capacity is fairly uncontroversially present in humans, and may form the basis for most learning capacities in most species. Most, if not all, models of learning assume that neural architecture is biased to exploit similarity structures in the environment (Rosenblatt 1958). The computational models outlined earlier in the Chapter suggest that the application of these two learning biases will lead to the cultural emergence of a communication system with some of the characteristic structure of human language.

There is in fact a large body of evidence that children, when acquiring morphological and syntactic systems, are biased in favour of acquiring one-to-one mappings. I will first present a general overview of this argument, which is based on work by Dan Slobin and John Haiman. I will then present more specific empirical and theoretical arguments that these one-to-one biases apply both on the sub-word level, in the acquisition of morphological systems, and on the super-word level, in the acquisition of syntactic structures.

5.6.1 One-to-one biases in general: clarity and isomorphism

On the general level, under the guise of the maxim “be clear”, Slobin (e.g. Slobin (1973), Slobin (1977), Slobin (1985)) suggests that children “strive to maintain a one-to-one mapping between underlying semantic structures and surface forms” (Slobin 1977:186). For Slobin, this bias on the part of learners is evident during language acquisition. He gives several examples, both from morphology and syntax. Italian children go through a phase of never omitting optional pronouns, thereby preserving the argument which appears in the underlying semantic representation. Similarly, English-speaking children avoid zero-morphemes in inflectional paradigms, and as a result over-generalise in irregular cases. Slavic-speaking children use an inflectional suffix for each grammatical case, even when some cases are supposed to remain unmarked in certain genders, thereby maintaining an overt realisation of case. English-speaking children go through a period of avoiding contracting auxiliaries, preferring the more analytic “I will not” to “I won’t”.

Slobin also claims that systems which conform to the one-to-one bias are easier to acquire. In Serbo-Croat relative clause formation, the surface form preserves the underlying semantic representation fairly clearly, the only changes necessary being insertion of a relative pronoun and the deletion of a repeated object. The same could be said to be true in English — “I lost the ball which Henry bought” corresponds reasonably closely with the (presumed) underlying semantic representation $\exists x\,(ball\,(x)) \wedge lost\,(me, x) \wedge bought\,(henry, x))$. In contrast, in Turkish the relationship between the relative clause and the underlying semantic representation is highly opaque. Slobin claims that the lack of a one-to-one, structure-preserving mapping between meaning and surface form in the Turkish system explains why Turkish children do not reliably acquire the system until the age of five, whereas Serbo-Croat speaking children master the one-to-one mapping by the age of two.

Slobin also gives evidence, again from Turkish and Serbo-Croat, that language-learners are biased against many-to-one, homonymous mappings when acquiring morphological systems and systems conforming to these biases are easier to learn. In Turkish, the agglutinating inflectional morphology system is very regular — each acoustically-salient suffix bears one element of the meaning (e.g. person, number, case) of the word. In contrast the Serbo-Croat inflectional system is "a classic Indo-European synthetic muddle . . . there are many irregularities, a great deal of homonymy, and scattered zero morphemes" (Slobin 1977:191). The Turkish morphological system conforms more closely to the one-to-one mapping principle than the Serbo-Croat system. Slobin suggests that this explains why the entire Turkish morphological system is mastered well before the age of two, whereas the Serbo-Croat system is not mastered until around age five — morphological systems which conform closely to learner biases are easier to acquire than systems exhibiting numerous many-to-one mappings.

John Haiman (e.g. Haiman (1980), Haiman (1985)) pursues a similar line of argument to Slobin, under the banner of *isomorphism*. According to Haiman, isomorphism, "whose existence is universally (though often implicitly) recognized in practice, is that of a one-to-one correspondence between the signans and the signatum, whether this be a single word or a grammatical construction . . . such a bi-unique correspondence must exist" (Haiman 1980:515). Whereas Slobin is keen to emphasise the importance of the one-to-one bias and its implications for learnability, Haiman is less concerned with why one-to-one mappings are preferred and more concerned with defending the idea and identifying when it can be overridden. The majority of Haiman (1980) is devoted to a fairly detailed account of specific conditions under which this bias can be overridden, based on an intricate analysis of the systematically homonymous medial verb morphology in Hua, a language of Papua New Guinea. However, Haiman does provide some more general defences against classic attacks on theories proposing a one-to-one bias in syntax acquisition:

> “In the realm of syntax, the Katz-Postal hypothesis (Katz & Postal 1964) that transformations do not change meaning has entailed a commitment to the belief that both neutralization (many deep structures, one surface structure) and diversification (many surface structures, one deep structure) must exist. Recently, this view has come under attack from two fronts . . . First, it is claimed that different surface structures invariably *do* correspond to different meanings, however fine-grained [there is no syntactic synonymy]. . . . A somewhat different approach has been taken in the study of neutralization [syntactic homonymy]. Syntactic targets, or structurally ambiguous forms, do exist, and in one sense are violations of isomorphic bi-unique correspondence. But there is another sense in which they are not violations, and it is in this sense that neutralization itself is iconic [reflecting underlying semantic similarity]” (Haiman 1980:517).

Some theoretical arguments against the syntactic equivalent of synonymy will be outlined below. Haiman’s second point is that distinct deep structures sharing a single surface structure is a reflection of underlying shared semantic structure. He gives several examples, two of which I will outline briefly here. Firstly, he highlights the case of the modal auxiliaries in English, which are all similar morphologically and syntactically, and are similar semantically, conveying the idea of futurity or potentiality. His second example is based on the semantic similarities between relative clauses and cleft clauses, such as “That’s the room that I found it in” (relative clause) and “It’s Max’s room that I found it in” (cleft). Haiman claims that relatives and clefts are similar cross-linguistically, reflecting a shared element of meaning.

5.6.2 One-to-one biases in morphology

Mańczak (1980) presents a set of “laws of analogical evolution” for morphological change, the first of which is that “[t]he number of morphemes having the same meaning more often diminishes than increases” (Mańczak 1980:284) — paradigms tend to lose synonymous morphemes. Mańczak bases his laws on a survey of historical grammars and etymological dictionaries, and his laws (including the first law) have had some statistical verification.

McMahon (1994) presents an account of the change of the inflectional system in English, which I have also mentioned briefly in connection with Hare & Elman (1995). In modern English, plurality is generally marked with the suffix -s. The system in Old English was rather more complex, and plurality was bound up in inflections which also marked gender and case. McMahon argues that the -s suffix was reinterpreted as the marker for the plural and possessive in certain paradigms. “Analogical extension next stepped in, gradually generalising the new /s/ plural marker to many nouns . . . which had never used /s/ to mark the plural in any case” (McMahon 1994:72) — multiple ways of expressing the plural were replaced by a single dominant strategy, and furthermore, as predicted by Slobin, this suffix singled plurality only, rather than a bundle of features.

McMahon’s account is one of historical change, and is only implicitly based on an acquisition bias in learners in favour of regular, one-to-one morphological paradigms. Vennemann (1978) presents several pieces of empirical evidence which suggest that, for children acquiring language, “[s]uppletion is undesirable, uniformity of linguistic symbolization is desirable: Both roots and grammatical markers should be unique and constant” (Vennemann 1978:259).

It has long been acknowledged that, as alluded to earlier in discussion of Slobin’s general argument, children learning English go through a phase of over-generalising the regular past tense form to words which should correctly be marked in an irregular fashion or not at all — for example, “goed” rather than “went” or “keeped” rather than “kept” (Brown 1973). This indicates a preference for a uniform system of markers of tense. This type of over-generalisation is of course not unique to English. Vennemann presents several examples of similar phenomena in other languages. I will summarise his examples from Russian, which shows a preference by learners for uniform roots, and Spanish, which shows a preference for uniform stems.

Russian marks nouns for one of three genders and one of six cases. The Russian paradigm for three cases of interest and the three genders can be summarised as:

	Masculine	Neuter	Feminine
Nominative	null	-o	-a
Accusative	null	-o	-u
Instrumental	-om	-om	-oy

There are additional complications in that masculine nouns are further subdivided into animate and inanimate classes, which follow slightly different inflectional paradigms. However, that need not concern us here. Vennemann reports that Russian children typically generalise from the feminine to mark the Accusative case in every noun class with the -u suffix, leading to a uniform marker of the Accusative and getting rid of the homonymous -o and null suffixes. It could be argued that this is due to a frequency effect — according to Vennemann, 70% of the forms a child sees will be of the feminine gender, partly because diminutives are feminine. However, children learning Russian also generalise the -om marker of instrumental case, originating outside the feminine, across all noun classes. This is later replaced by the -oy suffix from the feminine, again across all classes, before the final system is settled on. This acquisition behaviour indicates the presence of a learning bias in favour of a one-to-one mapping from aspects of morphological meaning (case) to surface realisations of those markers.

Vennemann's second example concerns a learner preference for uniform roots. Spanish underwent a process known as "velar softening", whereby [k] changed to [s] and [g] changed to [x] prior to [i] or [e]. This change should have led to the partial paradigm:

	'mark'	'pay'
Indicative	mar[k]amos	par[g]amos
Subjunctive	∗mar[s]emos	∗par[x]emos

However, the attested paradigm is in fact:

	'mark'	'pay'
Indicative	mar[k]amos	par[g]amos
Subjunctive	mar[k]emos	par[g]emos

Vennemann argues that velar softening did not take place in the subjunctive paradigm due to a learner preference for uniformity of roots — during the process of change, where both alternatives were presumably available, children preferentially acquired the second paradigm, which has a one-to-one mapping from meanings to roots.

These accounts of morphology acquisition are completely compatible with my model of one-to-one biases — learners prefer mappings where elements of meaning map, in a one-to-one fashion, onto morphemes, sub-parts of words.

5.6.3 One-to-one biases in syntax

There is also a body of evidence that one-to-one biases apply to the acquisition of grammatical systems, which is compatible with my account of one-to-one biases, in particular the work by Slobin and Wanner & Gleitman on mappings from meanings to function words. There is also a body of theoretical work which assumes a more global one-to-one bias, between complete deep structures and surface structures.

Slobin (1985) reports on two examples from French of children exhibiting a preference for explicit one-to-one mappings between grammatical functions and grammatical constructions. The first case is based on data from a single child. In the adult system in French, the preposition "de" is used for both partitive and possessive uses. When the preposition occurs with a masculine complement NP the preposition and the article from the NP are contracted. With feminine NPs the preposition and article are not contracted. This contraction proceeds regardless of the partitive/possessive distinction:

	Masculine	Feminine
Partitive	J'ai du pain	Y'a de la neige
	I have some bread	*There is some snow*
Possessive	le chapeau du monsieur	le soulier de la dame
	the man's hat	*the woman's shoe*

The child discussed by Slobin mistakes the contraction device for an explicit marker of the partitive use of the preposition and produces:

	Masculine	Feminine
Partitive	J'ai du pain	*Y'a da neige
Possessive	*le chapeau de le monsieur	le soulier de la dame

This appears to be a case of a child expecting a one-to-one correspondence between a semantic distinction (partitive versus possessive) and an aspect of the surface form, where in fact this correspondence does not exist in the target language.

Slobin's second, more general, example from French concerns the use of "que" and "qui". In the adult French system, "qui" functions as a relative pronoun, used for relative clauses where the pronoun is the subject of the verb in the relative clause (for example, "L'homme *qui* a tué le chien", glossed in English as "The man who killed the dog"). "que" is also used as a relative pronoun, for relative clauses where the relativizer is the object of the verb in the relative clause (for example, "L'homme *que* nous avons vu", glossed as "The man who(m) we saw"), and also as a complementizer (for example, "Je pense *que* l'homme est mort", glossed as "I think that the man is dead"). Slobin outlines two general stages in the typical acquisition of these function words in French. In the first stage, "que" is acquired and used for both complementizer and relativizer. In the second stage, the child acquires "qui" and uses this exclusively for relatives, with "que" used exclusively as a complementizer (resulting in production of the ungrammatical $*$"L'homme qui nous avons vu"). This deviation from the adult system on the part of the child appears to be a consequence of a preference for a clear, one-to-one mapping from grammatical functions to grammatical function words.

Wanner & Gleitman (1982) make a similar point to Slobin, arguing that in general children find it difficult to acquire systems which mark negation morphologically. They attribute this to a preference for explicit one-to-one mappings from grammatical functions to grammatical function words, rather than polyfunctional morphemes.

Turning to syntactic theory, it can be noted that formal studies of learnability tend to make an assumption, equivalent to the Contrast or ME principles, that for any underlying representation there is at most one surface structure — one-to-many mappings from structured semantic representations to structured signals are ruled out, or dispreferred, by learners. Wexler & Culicover (1980) make the strong form of this assumption — their Uniqueness principle states that every deep structure has a single possible surface structure. Pinker (1984) makes a weaker statement, that the bias against one-to-many mappings can be overridden given sufficient evidence. In both cases, the Uniqueness principle is given as a necessary method for children to solve the “no negative evidence problem”, an aspect of the poverty of the stimulus. Given this principle, as soon as a child hears a particular surface structure it can rule out all transformations that would have generated a different surface structure for the underlying semantic representation.

Theories based around isomorphism, the assumption of a preference for one-to-one mapping between meanings and words or grammatical structures, are typically seen (not least by the proponents of such theories) as antagonistic to the generative tradition in linguistics. Newmeyer (1998), however, rightly points out that most generative theories, starting with Chomsky’s (1965) standard theory, assume that syntactic structure and propositional content are intimately related at deep structure — deep structure determines both surface syntactic structure and semantic interpretation. In earlier versions of the classical generative paradigm, movement operations deriving surface structure from deep structure can presumably perturb the isomorphism of meaning and syntactic structure. Later theories (e.g. Chomsky (1995)) abandon the distinction between deep structure and surface structure — there is a single derivation to a level of Logical Form and at some point during that derivation (Spell Out) phonological form is stripped out to provide the input to the phonological system. While the semantic and phonological interpretation processes may also conceal some underlying isomorphism, the main point is that generative theories typically assume an intimate, roughly isomorphic relationship between meaning and signal.

The projection principle, one of the basic assumptions of generative syntax, also suggests that semantic and syntactic structures should be closely related. The projection principle states that lexical information is represented syntactically — the syntactic properties of lexical items project up through the syntax, and determine, in conjunction with structural restrictions, the syntactic structure of sentences. The syntactic properties of lexical items are frequently related to their semantic properties — for example, the argument structure and theta-role assignments of lexical items are dependent on the semantic properties of those items, and have consequences for the surface forms of sentences involving those items. There is therefore a general assumption in syntactic theory that semantics and surface forms are closely related and dependent on the same underlying information.

Langacker (1977), in a paper on grammatical change through reanalysis, bases much of his argument around a notion of "transparency":

> "the ideal or optimal linguistic code, other things being equal, will be one in which every surface unit . . . will have associated with it a clear, salient, and reasonably consistent meaning or function, and every semantic element in a sentence will be associated with a distinct and recognizable surface form. Languages are thus optimal along this parameter to the extent that they show a one-to-one correspondence between units of expression and units of form, and languages should therefore tend to change towards this situation rather than away from it" (Langacker 1977:110)

Langacker sees this pressure acting in direct opposition to two other tendencies, a preference for signal simplicity, or economy of production, and code simplicity, a preference for fewer fixed expressions which have to be memorised (for example, lexical items — see the Langacker quote in Chapter 3, Section 3.6.2). While Langacker clearly sees signal simplicity as a real-time preference by speakers, and code simplicity as a preference by learners, he is less explicit on the source of the transparency preference. However, this preference, as we have seen in this Chapter, can naturally be expressed as a preference by learners, in the form of a bias in favour of one-to-one mappings between elements of meaning and elements of signal.

My final example of an appeal to one-to-one biases will come from Bever & Langendoen (1971). Bever & Langendoen are concerned with how pressures acting on language use and language learning effect linguistic evolution. As such, their aims are highly compatible with this thesis and one quote from their paper would be equally at home in most recent papers on computational modelling approaches to linguistic evolution: "the linguistic future is highly constrained by the structural and behavioral systems implicit in the linguistic present. One consequence of this is that certain universals of language, which appear to be aspects of synchronic 'linguistic structure' have sources in the ways in which language is learned and used" (Bever & Langendoen 1971:451–452).

Bever & Langendoen's argument is based around an analysis of changes in the way in which relative clauses are formed between Old English and modern English. As the inflectional system of Old English changed, overt relativizers on relative clauses became obligatory in certain contexts. Bever & Langendoen argue that the overt relativizers became obligatory in order to solve ambiguity introduced by the loss of inflections — "the loss of inflections created certain perceptually ambiguous constructions which were ruled out of the language by the changes in relative clause formation . . . one cannot require of a language that it never generates a sentence which violates a perceptual generalization, only that the internal organization of actually uttered sentences be perceptually recoverable in general" (Bever & Langendoen 1971:444–445). Bever & Langendoen assume that the change in the OE inflectional system was a consequence of a combination of pressures for signal simplicity and a learner tendency to regularize inflectional paradigms (as discussed above). In contrast, they claim that the restructuring of relative clauses was a consequence of language *use*, based on concerns of perceptual ambiguity and processing difficulty. However, these two distinct pressures can be unified if we assume, as I have done here, that learners bring a one-to-one bias to the acquisition of all levels of linguistic structure — while simultaneously normalising morphological paradigms, a bias against structural ambiguity (many-to-one mappings from meanings to structures) lead learners to reinterpret, restructure, and restrict the structure of relative clauses.

There therefore appears to be a bias, applying at all levels of language acquisition, in favour of one-to-one mappings between meanings and signals. The simulation results outlined in this Chapter show that without such a bias compositional structure cannot emerge through cultural processes, or be maintained in the presence of a bottleneck on cultural transmission. However, given such a bias compositional structure can emerge, but only in the presence of a bottleneck on cultural transmission.

5.7 Summary of the Chapter

In this Chapter I have shown that compositional language reliably emerges when there is a bottleneck on cultural transmission, the environment is structured and learners are appropriately biased. The transmission bottleneck forces languages to be generalisable, and the generalizability of languages is maximised when the environment is structured.

The appropriate learning bias consists of two components — a bias to exploit internally compositional representations, and a bias to map each element of meaning to a single, unambiguous element of the signal. This first part of the bias may be a general property of learning — learning exploits regularities in the environment. I have presented evidence that the second element of the bias, a preference for one-to-one mappings between elements of semantic representations and elements of surface form, applies during children's acquisition of morphology and syntax — humans possess the learning bias which can lead, through cultural processes, to compositionality.

CHAPTER 6

The evolution of compositionality in populations

In the previous Chapter the cultural evolution of compositionality was investigated in the context of a single individual learning their communication system from a single cultural parent, and transmitting their system to another single individual. In such a context the notion of communication has no place, as solitary individuals have no-one to communicate with. It is also meaningless to study genetic transmission in such a context, as a population consisting of a single individual will be genetically homogeneous by definition.

In this Chapter I will investigate the transmission of structured communication systems in the context of *populations* of individuals. Do the findings outlined in the previous Chapter hold in the context of cultural transmission within populations? What consequences does this have for communication within such populations? And how does the dual transmission of communication systems by cultural transmission and weight-update rules by genetic transmission impact on the population's communicative behaviour?

A review of population-level ILMs and EILMs which deal with the evolution of structured communication is carried out in Section 6.1. This review suggests that the ILM and EILM approaches can be extended to a consideration of linguistic evolutions in populations. However, they also highlight the fact that studies of the evolution of learning bias have typically been restricted to parameter-setting models of learning, or variants thereof. In Section 6.2 I discuss methods of measuring communicative accuracy among populations of individuals using the type of languages discussed in the previous Chapter. In Section 6.3 I go on to describe an extension of the ILM from the previous Chapter to non-trivial population sizes. Finally, in Section 6.4, I expand this model to a full Evolutionary Iterated Learning simulation, and outline results pertaining to the evolution of learning biases which support communicatively optimal, compositional language.

6.1 Models of the evolution of linguistic structure in populations

Iterated Learning Models of the cultural evolution of structured communication in populations do exist, and are discussed in Section 6.1.1. The population approach also leads fairly naturally to the use of Evolutionary Iterated Learning Models, and examples of these models are discussed in Section 6.1.2.

6.1.1 Cultural evolution in populations

I have discussed three models which demonstrate that the repeated expression and induction of linguistic form within a population can lead to the emergence of structured systems of meaning-signal mappings. Two models by John Batali (Batali (1998), described in Section 2.3.3.4 of Chapter 2, and Batali (2002), discussed in Section 2.3.6.2 of Chapter 2) demonstrate that morphological and syntactic structure can emerge in populations. However, as discussed earlier, the Negotiation Model framework makes it difficult to isolate the relative importance of learning bias and transmission bottleneck in shaping linguistic behaviour in these populations. Hurford (2000), covered in Section 5.1 of Chapter 5, demonstrates that recursively compositional language can emerge in small populations (Hurford uses a population size of 5) in a gradual ILM population model. Similarly, Kirby (2000) demonstrates that compositional language can emerge in small populations through purely cultural processes. Kirby's (2000) model is an earlier version of the model described in Kirby (2002). A gradual turnover ILM is used, with a simple, non-embedding semantics and a stochastic grammar inducer.

Parameter-setting models of language acquisition have also been used in population-level ILMs. Niyogi & Berwick (1997), using a mathematical model, consider the spread of linguistic variants in populations through Iterated Learning. In Niyogi & Berwick's model, there are two competing linguistic variants, L_1 and L_2, which differ with respect to the setting of a particular parameter[1]. Learners sample the languages of the adult population and decide, using a parameter-setting procedure known as the Trigger Learning Algorithm (Gibson & Wexler 1994), on which variant to acquire. Niyogi & Berwick demonstrate that, if one language, say L_1, produces triggers which are consistent with both L_1 and L_2, while L_2 produces triggers which are only consistent with L_2, then L_2 will come to dominate the population.

[1] Niyogi & Berwick's model need not be interpreted as a parameter-setting model — we could take the two linguistic variants to represent linguistic systems which differ arbitrarily from one another. However, the single parameter interpretation is the most natural one, given their use of the Trigger Learning Algorithm.

Briscoe (2000a) (discussed in Section 2.3.5.2, Chapter 2) and Kirby (1999) (described in Section 2.3.3.2, Chapter 2) present models which are similar to that of Niyogi & Berwick, which demonstrate the convergence of populations on shared parameter settings. This convergence is driven by a frequency-dependent bias in Briscoe's model, whereas a direct bias in favour of parsability drives populations in Kirby's model to a parameter setting which is yields optimally parsable utterances.

6.1.2 Gene-culture coevolution in populations

Parameter-setting approaches have also been used to model the co-evolution of languages and parameter-setting language acquisition devices. Kirby & Hurford (1997) describe an extension to Turkel's (2002) model of the evolution of co-ordination. As in Turkel's model, each individual's genotype consists of a string of 1s and 0s (representing inviolable principles) and ?s (representing settable parameters). An individual's mature phenotype is a string of 1s and 0s, with all genotype ?s being set to either 1 or 0. Unlike in Turkel's model, Kirby & Hurford use a generational EILM, with the setting of parameters in an individual's mature phenotype being determined by cultural transmission. Immature individuals receive a number of trigger utterances from mature individuals in the previous generation. Triggers specify the setting of a single parameter (as either 1 or 0), and mature individuals produce triggers consistent with their own grammar. Learners then set the values of ?s in their phenotype according to observed triggers and a learning procedure based on the Trigger Learning Algorithm. Mature individuals breed according to communicative success. Communicative success between two individuals is dependent on the number of matching settings they share in their mature phenotypes, but also on the 'parsability' of the utterances they produce — an arbitrary subset of the set of possible mature phenotypes were considered to produce more parsable utterances.

Kirby & Hurford report that, under these conditions, maximally functional (parsable) grammars do *not* emerge — the simulated populations converge on mature phenotypes which do not produce maximally parsable utterances. Not only does the population converge culturally on suboptimal grammars, but they nativize those suboptimal grammars — the Baldwin effect leads to the emergence of disfunctional, inviolable principles which prevent learners in the population from acquiring an optimal system. Kirby & Hurford attribute this to the overriding pressure for learners to learn the language of their community, regardless of whether it is optimally parsable or not.

In a second set of experiments, Kirby & Hurford introduce selection for parsability on cultural transmission. With a certain small probability, learners preferentially retain parameter settings which yield higher parsability over settings which yield lower parsability. Learners are therefore directly biased in favour of acquiring optimal parameter settings. Under this revised setup, the simulated populations converge on optimal grammars, and nativize those optimal grammars via the Baldwin effect.

Briscoe (2000b) presents an extension of his Iterated Learning Model (Briscoe (2000a), discussed above), which also demonstrates the role of the Baldwin effect in the nativization of linguistic structure, with languages which minimise working memory load and therefore improve parsability being preferentially nativized. His model shows that Kirby & Hurford's (1997) second result still stands under more realistic assumptions about grammars, parsability and population dynamics — more parsable and learnable languages emerge culturally, and the LADs of learners evolve a default setting which matches the dominant language in the population.

A paper by Martin Nowak and colleagues is of particular relevance to this Chapter, and merits discussion in detail. Nowak *et al.* (2000) compare the fitnesses of individuals pursuing two possible learning strategies — holistic learners and compositional learners. In their model, events consist of an action and an object acted upon. Nowak *et al.* vary the number of possible objects and actions, and also the possible combinations of objects with actions — for example, events involving action 1 and object 1 may occur with a certain frequency, whereas events involving action 1 and object 2 may never occur. Nowak *et al.*'s model of events is therefore equivalent to (and formed the basis for) my model of meaning spaces and environments — their event space corresponds to a two-dimensional meaning space, with V for each dimension given by the number of possible objects and actions, and their *event rate matrix* (which specifies which action-object combinations may occur) is equivalent to an environment in my model.

Nowak *et al.* assume there are two possible types of learners — holistic learners, who attempt to learn a single word for each event, and compositional learners, who learn separate words for actions and objects and then combine those words to form descriptions of events. Nowak *et al.* calculate the equilibrium frequency of words in populations consisting solely of holistic or compositional learners, calculate the levels of communicative accuracy associated with these word frequencies and compare these values across populations.

In populations of holistic learners, the frequency of individuals who use a word W_{ij} to refer to an event E_{ij} is given by $x\left(W_{ij}\right)$. Nowak *et al.* assume a generational ILM where each individual receives b exposures to the linguistic behaviour of the previous generation. During each of these exposures, a holistic learner successfully learns a word with probability q. $\phi\left(E_{ij}\right)$ gives the frequency of occurrence of event E_{ij}. After a single generation, the frequency of word W_{ij} to refer to event E_{ij} is given by:

$$x'\left(W_{ij}\right) = R\left(W_{ij}\right) \cdot x\left(W_{ij}\right) \cdot \left(1 - x\left(W_{ij}\right)\right) - x\left(W_{ij}\right)$$

where $x'\left(W_{ij}\right)$ is the proportion of individuals who know the word after transmission, $x\left(W_{ij}\right)$ is the frequency of individuals who know the word prior to transmission and $R\left(W_{ij}\right)$ is the reproductive rate of the word, and is given by:

$$R\left(W_{ij}\right) = bq\phi\left(E_{ij}\right)$$

In other words, the reproductive rate of word W_{ij} is the product of the number of exposures each individual receives (b), the probability of learning that word after a single exposure (q) and the probability that you hear someone talking about the event E_{ij} ($\phi\left(E_{ij}\right)$). In the population equation, the change in the proportion of people who know the word depends on this reproductive rate and the cultural variance in the population ($x\left(W_{ij}\right) \cdot \left(1 - x\left(W_{ij}\right)\right)$). The term $\ldots - x\left(W_{ij}\right)$ simply keeps the population size constant.

Given this equation, Nowak *et al.* go on to calculate the equilibrium frequency of individuals who know word W_{ij} for event E_{ij} — the frequency which, possibly infinitely many, iterated learning events will lead to, assuming that the word has a reproductive rate of greater than 1. This quantity, $x^{*}\left(W_{ij}\right)$, is given by:

$$x^{*}\left(W_{ij}\right) = 1 - \frac{1}{R\left(W_{ij}\right)}$$

In other words, the equilibrium frequency of W_{ij} will depend on its reproductive rate. If the reproductive rate of a word is very high then virtually everyone in the population will know the word. As all events which actually occur are assumed to occur with equal frequency, the key factors in determining $R\left(W_{ij}\right)$ are number of exposures b and learnability q.

The communicative accuracy, and therefore fitness, of a population of such holistic learners at equilibrium is simply the probability that any two individuals will know the word for an event, summed over all possible events:

$$F_{holistic} = \sum_{i,j} \gamma_{ij} \cdot \left[x^* \left(W_{ij}\right)\right]^2$$

where γ_{ij} is 1 if event E_{ij} is allowed by the event rate matrix (environment) and 0 otherwise. Note two things about this measurement. Firstly, it is implicit in this measure that there is only ever one word for each event in the population — the measure is not summed over all possible words for each event. Secondly, it is also implicit that each event is associated with a unique word — words for one event are never confused with words for other events.

Nowak *et al.* go through a similar processes for calculating word frequencies in populations of compositional learners. Compositional learners learn separate words for objects O_i and actions A_j of events E_{ij}. The word associated with object O_i is N_i (a noun), while the word associated with action A_j is V_j (a verb). The assumption is made that learners spend half their time learning nouns and half their time learning verbs. The reproductive rates of N_i and V_j are given by:

$$R\left(N_i\right) = \left(b/2\right) q_c \phi\left(E_i\right)$$

$$R\left(V_j\right) = \left(b/2\right) q_c \phi\left(E_j\right)$$

where q_c is the probability of learning a noun or verb after a single exposure (taken to be lower than q), $\phi\left(E_i\right)$ gives the frequency of events involving object O_i and $\phi\left(E_j\right)$ gives the frequency of events involving action A_j. These equations are clearly variants of the holistic learner equations. The equilibrium frequency of individuals who know both N_i and V_j is given by:

$$x^* \left(N_i V_j\right) = \frac{\left(1 - 1/R\left(N_i\right)\right) \cdot \left(1 - 1/R\left(V_j\right)\right)}{1 - 1/\left(R\left(N_i\right) + R\left(V_j\right)\right)}$$

This leads to a communicative accuracy at equilibrium of:

$$F_{compositional} = \sum_{i,j} \gamma_{ij} \cdot \left[x^* \left(N_i V_j\right)\right]^2$$

In other words, as for the holistic learners, communicative accuracy depends on the probability of two individuals being able to make a signal for each event, multiplied by the probability of that event occurring.

Based on these equations, Nowak *et al.* go on to identify the conditions under which compositional learners will be preferred to holistic learners — the circumstances under which $F_{compositional} > F_{holistic}$. To do this they make several more assumptions. Firstly, they assume that there are n objects and n actions — there are as many objects as actions. This yields n^2 possible events. Suppose some fraction p of these events occur (for example, if $p = 0.5$ then only half of all possible combinations of objects and actions actually occur in the environment), and further assume that these events are distributed randomly through the space of possible events. Under these conditions, compositional learners will be preferred when:

$$n > \frac{3q}{pq_s}$$

In other words, there is a critical value for the number of objects and actions that must be exceeded before compositional learners are favoured. Nowak *et al.* give the following illustrative example. If nouns and verbs are twice as hard to memorise as holistic words ($q_c = q/2$) and if one third of all possible events actually occur in the environment ($p = 1/3$), then n must be greater than 18 before compositional learners are favoured.

This is an interesting result, and Nowak *et al.* successfully demonstrate that mathematical techniques can be fruitfully applied to the investigation of the evolution of learning apparatus underlying language. However, several criticisms of their model can be made. Firstly, as noted above, they assume that the relationship between events and words, or nouns and objects, or actions and verbs, is perfectly one-to-one. As we have seen throughout this thesis, arriving at this situation is far from straightforward. Secondly, they also rule out any gene-culture coevolution — they calculate the communicative accuracy of genetically homogeneous populations at equilibrium, then compare across populations. The situation is likely to be more complex in heterogeneous populations, and more complex still in heterogeneous populations undergoing selection for communicative success. Finally, they assume that events are randomly scattered in the event matrix — as we saw in Chapter 5, a non-random sampling from the space of possible meanings can have consequences for cultural evolution in populations, and therefore might have implications for gene-culture coevolution.

6.2 Languages, communication and communicative agents

The model of structured languages is identical to that used in the previous Chapter. A language L consists of a production function $p\left(m\right)$, mapping from meanings m to signals s, and a reception function $r\left(s\right)$, mapping from signals s to meanings m. Each $m \in \mathcal{M}$ is a vector drawn from an F-dimensional space, where each dimension has V possible values, and each signal $s \in \mathcal{S}$ is a string of characters of length 1 to l_{max}, where the characters are drawn from the alphabet Σ.

We can calculate the communicative accuracy of two individuals in exactly the same way as that outlined in Chapter 3, Section 3.2. If $p(m)$ is converted to a probabilistic function $p(s_j|m_i)$, which gives the probability of producing signal s_j given meaning m_i, and $r(s)$ is similarly viewed as a probabilistic function $r(m_i|s_j)$ then the communicative accuracy of a producer P with production function $p(s|m)$ signalling to a receiver R with reception function $r(m|s)$, averaged over all meanings, is:

$$ca\left(P,R\right) = \frac{\sum_{i=1}^{i=|\mathcal{M}|}\sum_{j=1}^{j=|\mathcal{S}|} p\left(s_j|m_i\right)\cdot r\left(m_i|s_j\right)}{|\mathcal{M}|}$$

In a population possessing an optimal communication system $ca(P,R) = 1$ for any choice of P and R.

Note that, given the distance function between meanings given in Chapter 5, based around Hamming distance, it would be possible to have a communicative accuracy measurement which awarded partial credit for getting a proportion of the meaning across to the receiver. This is the approach taken in Batali (1998) and Batali (2002). In this case $ca\left(P,R\right)$ would be defined as:

$$ca\left(P,R\right) = \frac{\sum_{i=1}^{i=|\mathcal{M}|}\sum_{j=1}^{j=|\mathcal{S}|}\sum_{k=1}^{k=|\mathcal{M}|} p\left(s_j|m_i\right)\cdot r\left(m_k|s_j\right)\cdot sim\left(m_i,m_k\right)}{|\mathcal{M}|}$$

where $sim\left(m_i,m_k\right)$ gives the degree of similarity between two meanings, and is defined as:

$$sim\left(m_i,m_k\right) = \frac{\left(F - HD\left(m_i,m_k\right)\right)}{F}$$

For the moment I will persevere with the all-or-nothing measurement. However, in Section 6.4, I will return to the partial payoff measure when considering the dual-transmission of compositional systems.

The model of a communicative agent is identical to that used in the preceding Chapter — each individual is modelled by an associative network capable of manipulating mappings between structured meanings and structured signals, and each individual acquires its system based on observation and the application of a weight-update rule W, specified by the 4-tuple $(\alpha\ \beta\ \gamma\ \delta)$.

6.3 Cultural evolution in populations

The simulations described in Chapter 5 demonstrate that, in populations consisting of a single individual, the cultural transmission of meaning-signal mappings leads to the emergence of compositional language. This is dependent on learners being biased in an appropriate fashion, the presence of a bottleneck on cultural transmission, and a degree of structure in the environment. Do these results scale up when we consider populations which consist of more than one individual at any one time?

This question can be addressed using a gradual population turnover ILM. The model of languages, communication and communicative agents is as given above in Section 6.2. The initialisation and iteration processes are given below.

Initialisation Create a population of N agents[2], each using the weight-update rule W and having an initial set of connection weights $\mathcal{W}$, where each $w \in \mathcal{W}$ has a weight of 0.

Iteration

1. Select an agent at random from the population and remove it.
2. For every remaining member of the population, generate a set of meaning-signal pairs by applying the network production process to every $m \in \mathcal{E}$.
3. Create a new agent with connection weights of 0 who uses weight-update rule W.

[2] $N = 20$ for all ILMs outlined in this Section. In previous Chapters $N = 100$ was typically used. However, the more complex models of communication and communicative agents increases the computational cost of each cohort. In the gradual population turnover model computational complexity is constant with respect to population size, as each cohort involves replacing a single individual. However, larger populations take longer to converge on a shared system. $N = 20$ reduces this factor, while still allowing meaningful population-level dynamics.

4. The new agent receives e exposures to the population's observable behaviour and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W. See below for more detail.
5. The new agent joins the population. Return to 1.

Each pass through the iteration process will be termed a *cohort*, and as with other ILMs there is no genetic diversity within the population and no selection based on communicative ability.

Step 4 of the iteration process offers some complications. In the ILM outlined in Chapter 3, each of the e exposures consists of exposure to the complete set of observable behaviour generated by a single, randomly selected individual. In the model outlined in Chapter 5, each of the e exposures consisted of an exposure to a single meaning-signal pair produced by the individual's single cultural parent, and the exposures were either selected exhaustively from the environment $\mathcal{E}$ (in the no-bottleneck condition) or randomly (in the bottleneck condition).

In this model, neither of these methods of transmission is entirely suitable. If each of the e exposures consisted of exposure to the complete set of observable behaviour generated by a single, randomly selected individual then we immediately rule out a bottleneck on cultural transmission. If each of the e exposures consists of an exposure to a single meaning-signal pair produced by a single cultural parent, then convergence within the population will occur only by chance — true, non-random convergence requires that individuals sample the behaviour of several individuals. It is therefore necessary to introduce a new parameter τ, which is the number of cultural parents an individual has. Two versions of step 4 of the iteration process will be defined, one for the no-bottleneck condition and one for the bottleneck condition.

4 (No-bottleneck) The new agent selects τ cultural parents[3] at random from the population. The new agent receives $e = |\mathcal{E}|$ exposures to the communicative behaviour produced by those τ parents. During each of these e exposures the new agent observes the meaning-signal pairs produced by each parent for a single meaning $m \in \mathcal{E}$ and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W. Each $m \in \mathcal{E}$ is selected in turn, therefore the learner observes the full set of observable behaviour produced by each of the τ parents.

[3] $\tau = 3$ for all simulation runs reported here. This means that each individual will observe the behaviour of three individuals, which was the case for the associative network ILM discussed in Chapter 3.

4 (Bottleneck) The new agent selects τ cultural parents at random from the population. The new agent receives e exposures to the communicative behaviour produced by those τ parents. During each of these e exposures the new agent observes the meaning-signal pairs produced by each parent for a single, randomly selected, meaning and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W. The agent will therefore observe approximately $|\mathcal{E}| \cdot c(\mathcal{E}, e)$ distinct meanings, paired with the corresponding signals produced by each of the τ parents.

The no-bottleneck version leads, as will be discussed in Section 6.3.1, to the emergence of shared stable communication systems. However, the bottleneck version as given above does not. This appears to be due to the high level of variability in the behaviour observed by learners during the early stages of a simulation run. The bottleneck version of step 4 is therefore revised as follows. Each agent selects θ individuals at random from the population, where θ is randomly selected from the range $[1, \tau]$. The agent then selects τ cultural parents at random with replacement from among these θ individuals — in other words, learners are exposed to the same size of data set regardless of the number of distinct cultural parents they have, but the data set can contain the behaviour of between 1 and τ individuals. Each individual therefore receives τ exposures to each meaning, as in the bottleneck version of step 4 given above. However, these exposures will be to the behaviour of at most τ distinct individuals. Alternatively, given that $\tau = 3$ for all runs reported here, they will have 2 distinct cultural parents and observe one of them twice, or have a single cultural parent and observe that individual's behaviour three times[4].

I will consider an ILM where every agent uses the weight-update rule $W = (1\ -1\ -1\ 0)$. As shown in Chapter 5, this is one of the two [+constructor, +ic-preserver] rules. For the results described in Sections 6.3.1 and 6.3.2, $F = 3$, $V = 5$, $l_{max} = 3$ and $\sigma = \{a, b, c, d, e, f, g, h, i, j\}$. The initial agents have connection weights of 0, and therefore use the maximum entropy system where every meaning analysis-signal analysis pair occurs with equal probability. This is the same experimental setup as for the ILM described in Section 5.3, the only difference being the scaling up to larger populations.

[4] As part of my current research project, I am working on an extension to Kirby's (2002) model of the evolution of recursive syntax. One part of the project involves scaling this model up from populations consisting of a single individual to larger populations. Interestingly, a similar problem is encountered with Kirby's model — the set of cultural parents for each individual must be fairly tightly constrained, otherwise stable systems of meaning-signal mappings never emerge.

6.3.1 Linguistic evolution in the absence of a bottleneck

Runs of the ILM described above were carried out, using the no-bottleneck variant of step 4 — each individual observes the complete set of behaviour of $\tau = 3$ members of the population. 100 runs[5] of the ILM were carried out for each of the sparse environments shown in Figures 5.5 and 5.6 in Chapter 5. 50 runs of the ILM were carried out for each of the medium density environments shown in Figures 5.5 and 5.6 in Chapter 5. As in Chapter 5, the e-compositionality of the emergent languages (averaged over all members of the population) is the key measure of linguistic structure. The population's communicative accuracy is also measured, to establish whether the emergent languages are functional and shared by all members of the population. Communicative accuracy is estimated by evaluating every individual's average communicative accuracy as both producer and receiver with two randomly selected partners according to the all-or-nothing measure $ca(P, R)$ given in Section 6.2, averaging over all individuals in the population. Runs were allowed to proceed to a stable state, where the population exhibits no linguistic diversity.

In all simulations runs in each environment the populations converge upon an optimal shared communication system which yields $ca(P, R) = 1$ for any choice of P and R. This is as expected, given the one-to-one learning bias associated with the weight-update rule used by learners in these populations. Figures 6.1 and 6.2 plot the compositionality of the initial and final, stable systems for the sparse and medium-density environments. The results for the medium-density environment are similar to those shown in Figure 5.9 in Chapter 5 (for the same environments with an ILM involving isolated individuals).

The results for the sparse environments are rather different from the results from the single-individual ILM. In the isolated individual ILM (see Figure 5.8), the majority of runs converged on non-compositional systems. Partially compositional systems did occur, with their frequency being greatest when the environment was unstructured. Highly compositional systems were very infrequent, and occurred only when the environment was structured.

The results for the population ILM shown in Figure 6.1 show a much stronger tendency towards compositionality. The majority of the final systems are not holistic. Partially compositional systems occur with comparatively high frequency in both unstructured

[5] 1000 runs were carried out for the no-bottleneck condition of the single-individual ILM. A smaller number of runs were carried out in the population ILM due to two factors: 1) the increased computational memory requirements introduced by having 20, rather than 1, associative network in the population and 2) the increased number of cohorts required for a population to reach a stable state.

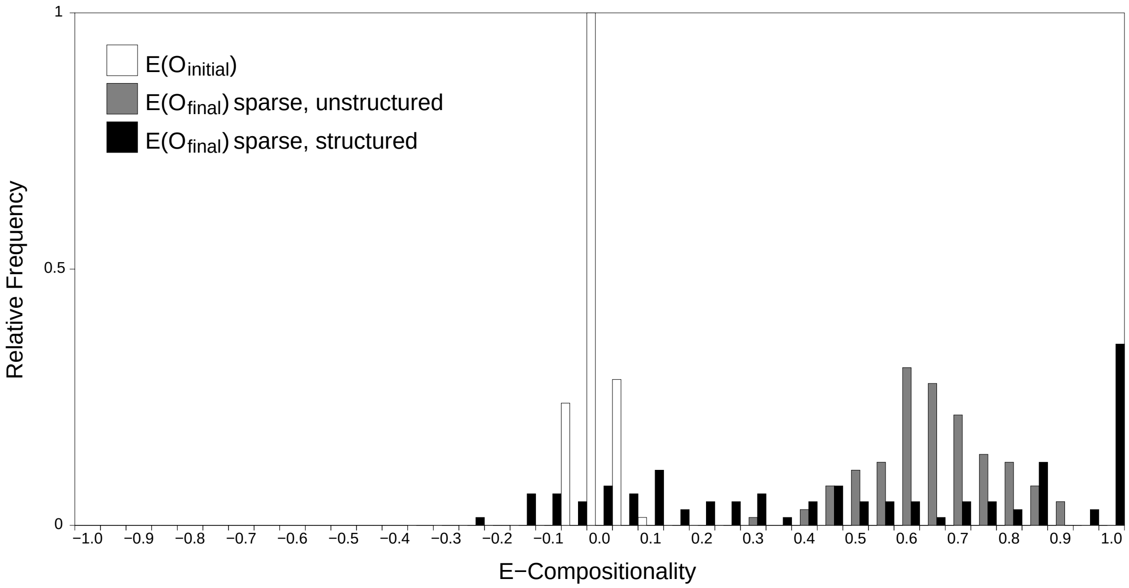


Figure 6.1: E-compositionality of initial and final, stable systems in sparse environments, when there is no bottleneck on transmission. The initial systems have low e-compositionality. The final systems are of partial or high e-compositionality. Highly e-compositional systems occur most frequently when the environment is structured.

and structured environments. Perfectly compositional systems emerge with fairly high frequency, but only when the environment is structured.

Why does the expansion to non-trivial population size lead to the more frequent emergence of compositionality, but only in the sparse environment? Recall from Chapter 5 that [+constructor, +ic-preserver] agents are biased in favour of acquiring i-compositional systems, and are further biased in favour of acquiring one-to-one mappings between feature values and signal substrings, which leads to a bias in favour of e-compositional language. In the single-agent population case, this bias can lead to the emergence of highly compositional language even in the absence of a bottleneck on cultural transmission, but only if the initial, random language already exhibits slight compositional tendencies. In the single-agent population case, learners are essentially stuck with the system of their single cultural parent. Their learning bias has a reduced impact, due to the absence of competing variants to select between (recall from B&R's model given in Chapter 2 that the rate of increase of the variant favoured by directly-biased transmission is dependent on the degree of cultural variation present in the population).

In the population ILM, each individual has several cultural parents and therefore biased acquisition potentially has a greater impact. An individual attempting to acquire two systems will be more influenced by the system which conforms more fully to their bias. In the population ILM, this means that highly compositional systems will be preferred to

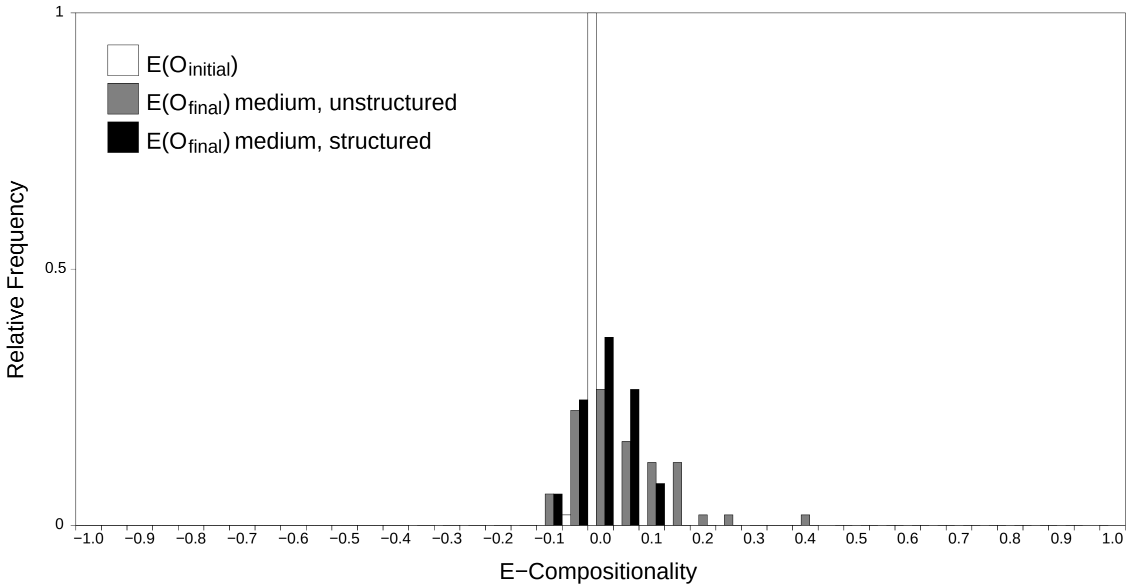


Figure 6.2: E-compositionality of initial and final, stable systems in medium density environments, when there is no bottleneck on transmission. The initial systems and the vast majority of the final systems have low e-compositionality. Partially e-compositional final systems occur with very low frequency, and only when the environment is unstructured.

less compositional systems. This results, in sparse environments, in the frequent emergence of highly compositional systems. The difference between unstructured and structured sparse environments is due, as discussed in Chapter 5, to the greater potential spread of compositional mappings in the structured environment, due to the number of feature values shared between meanings.

The fact that each learner has several cultural parents, and therefore several possible communication systems to choose from, increases the force of the learner's bias and results in the emergence of systems which conform to that bias. However, this only happens in sparse environments — in medium density environments, the final stable systems tend overwhelmingly to be holistic, with partially compositional systems occurring very infrequently and only when the environment is unstructured. Why?

Recall from Chapter 5 that the compositionality of the final system in the single-individual ILM is sensitive to the compositionality of the initial, random system. Where this initial mapping exhibits compositional tendencies, yielding $E\left(\mathcal{O}_{initial}\right)$ above the mean, there is an increased likelihood of the system moving, over iterated learning events, towards more compositional languages. The compositional tendencies of the initial system spread to other parts of the system over time, resulting in an increase in compositionality. For the more densely-filled environments, partially or highly compositional systems emerge infrequently due to the fact that the initial systems tend to be clustered more tightly around the non-compositional mean. When the environment contains few meanings the initial system may, by chance, exhibit some compositional tendencies. However, when the environment contains a large number of meanings such tendencies are likely to be drowned out by the majority non-compositional mapping.

In the population ILM, with medium density environments, the lack of compositional tendencies in the early, random mappings of the population prevents highly compositional systems from ever emerging. Even though each learner observes several individuals, their set of cultural parents is essentially homogeneous with respect to compositionality — each parent uses a non-compositional system (although not necessarily the same one). This lack of cultural variation effectively nullifies the learner preference for compositionality. In contrast, in sparse environments the initial random systems are more widely distributed, and more likely to exhibit some compositional tendencies which the learner bias can exploit.

Partially compositional systems do emerge with low frequency in medium density, unstructured environments. This is due to the fact that, in such environments, fewer meanings share feature values, therefore the initial random system is more likely to exhibit slight compositional tendencies — the initial systems in unstructured environments has to be less ‘lucky’ in the assignment of characters to feature values. This can provide some cultural variation among an individual’s cultural parents, allowing the learner bias to have some effect.

6.3.2 Linguistic evolution in the presence of a bottleneck

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

To this end, runs of the ILM described above were carried out, using the bottleneck variant of step 4 — each individual observes e meaning-signal pairs, randomly selected from the set of behaviour produced by $\theta \leq \tau = 3$ different members of the population. 10 runs of the ILM were carried out for each of the sparse environments shown in Figures 5.5 and 5.6 in Chapter 5, with a bottleneck of $c\left(\mathcal{E}, e\right) = 0.8$ (e = 19) and 10 runs were carried out for each of the medium density environments shown in Figures 5.5 and 5.6

(a)

Density	$c(\mathcal{E}, e)$	Proportion compositional	Average $E(O_{final})$	Average *ca* (final)
sparse	0.8	0	0.59	0.51
medium	0.4	0	0.28	0.05
medium	0.5	0.2	0.47	0.26
medium	0.6	1.0	0.99	1.0

(b)

Density	$c(\mathcal{E}, e)$	Proportion compositional	Average $E(O_{final})$	Average *ca* (final)
sparse	0.8	1	1.0	1.0
medium	0.4	1	0.99	0.98
medium	0.5	1	0.99	0.99
medium	0.6	1	0.99	1.0

Table 6.1: Summary of results for the population ILM. (a) gives the proportion of runs converging on a highly compositional system, the average e-compositionality of the final systems and the average communicative accuracy yielded by the final systems for *unstructured* environments. Highly compositional, communicatively-optimal languages only reliably emerge in the medium density when the bottleneck is wide. (b) gives the same measurements for runs of the ILM in *structured* environments. Highly compositional, communicatively-optimal languages always emerge when the environment is structured.

with bottlenecks of $c(\mathcal{E}, e) = 0.4$, 0.5 and 0.6 (e = 16, 21 and 28 respectively)[6]. Runs were allowed to proceed for 5000 cohorts.

Table 6.1 summarises the results of these simulation runs, in terms of the proportion of runs converging on a highly compositional system ($E(\mathcal{O}_{final}) > 0.95$), and the average final levels of e-compositionality and communicative accuracy. Figure 6.3 show compositionality and communicative accuracy against time in five representative runs of the ILM.

As shown in the Table, environment structure has a significant impact on the compositionality of the emergent systems. In the unstructured environments, highly compositional, communicatively-optimal systems only reliably emerge in the medium density environment with a relatively wide bottleneck ($c(\mathcal{E}, e) = 0.6$). In contrast, in the structured environments, highly compositional, optimal systems *always* emerge.

[6] This is obviously a significantly smaller number of runs than was carried out for the single-individual ILM, and is due to the increased computational cost of population-level ILMs, as discussed above. Each run of the population ILM with a medium density environment takes approximately 36 hours on a 2.5GHz Pentium 4 processor.

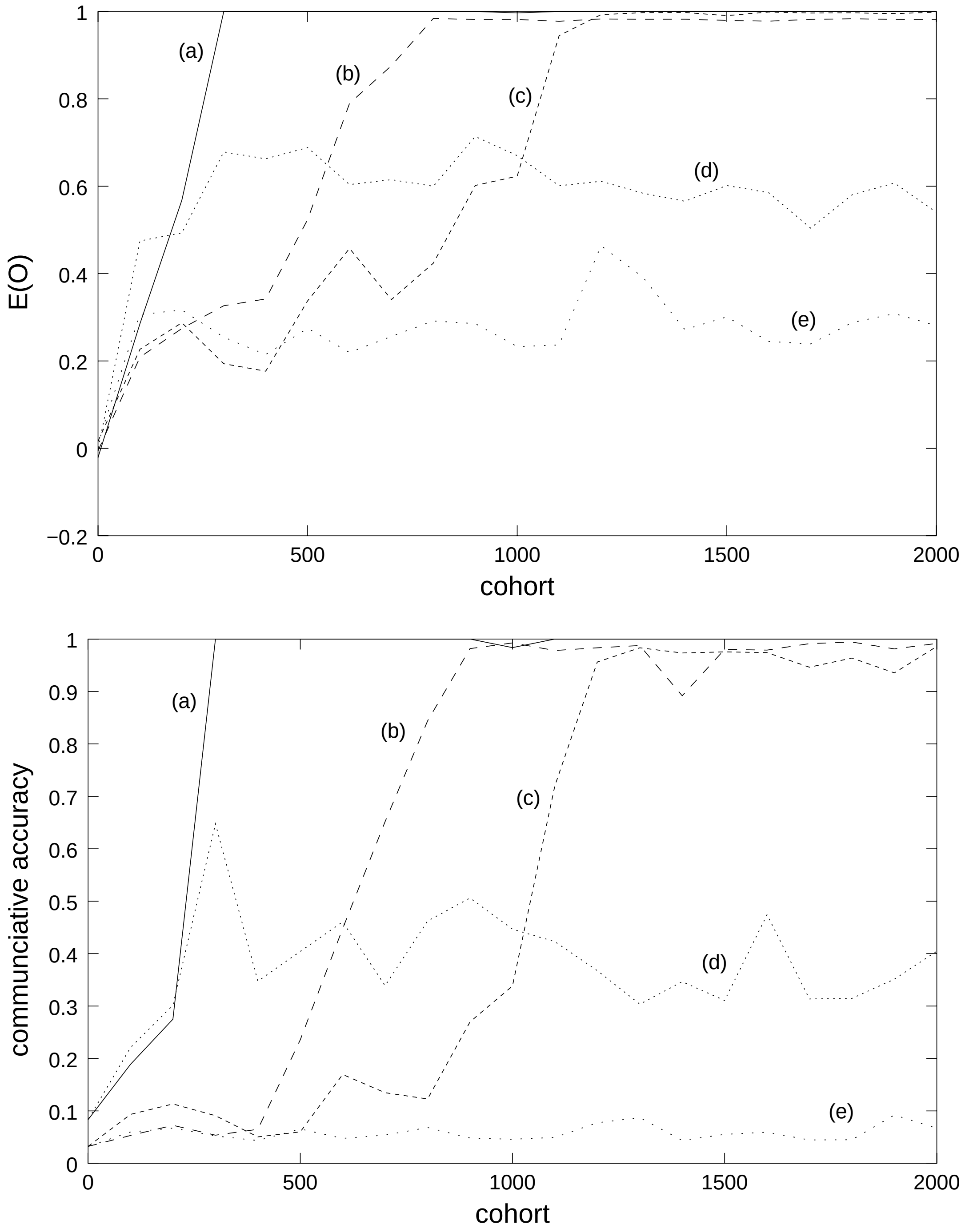


Figure 6.3: Plots of e-compositionality (top) and communicative accuracy (bottom) against time in runs of the population ILM. (a) shows the progress of a simulation run in the sparse, structured environment. (b) is an example of a convergent run in a medium density, structured environment. (c) is a convergent run from a medium density, unstructured environment where $c = 0.6$. (d) is a non-convergent run from the sparse, unstructured environment. (e) is a non-convergent run from a medium density, unstructured environment with a tight bottleneck on transmission ($c = 0.4$).

Learners in the population ILM observe and learn from the linguistic behaviour of between one and three cultural parents. If these cultural parents have strongly conflicting languages, or if their languages are non-compositional, then the learner will tend to arrive at a non-compositional or partially compositional system, depending on the number of meanings in the environment (as discussed above, the e-compositionality of random systems is sensitive to the number of meanings expressed in that system). If, on the other hand, there is broad agreement in the linguistic behaviour of a learner's cultural parents, then the learner will converge on a system which is similar to that of its cultural parents, and which exhibits a similar level of compositionality.

Stability and regularity at the individual level therefore form the basis for convergence in the population. In structured environments, this individual-level stability emerges fairly straightforwardly. Compositional languages have a strong advantage in such environments, due to their generalizability. When the environment is structured, individual members of the population will converge fairly quickly on systems which are at least partially compositional. These systems will then spread fairly rapidly through the population, until the population converges on a communicatively optimal, compositional language. This is reflected in the reliable emergence of compositional language in structured environments, regardless of the degree of environment density or the severity of transmission bottleneck. These results show that the ILM results from the previous Chapter can scale up to the case where the population at any one time consists of more than one individual.

However, in unstructured environments stability at the individual level is rather more difficult to achieve, due to the lack of shared feature values between meanings. This problem can be overridden by a relatively wide bottleneck. For example, when learners see 60% of the language of the previous generation, there is very little difference between structured and unstructured environments, as can be seen in the final distribution of languages in Figure 5.14 in Chapter 5. The wide bottleneck allows individuals to reach highly compositional, consistent systems. As shown in the Table, in the population ILM these systems spread through the population — when the bottleneck is wide ($c = 0.6$), highly compositional languages always emerge.

For the lower levels of bottleneck in the medium environments, and in the sparse environment, the lack of structure in the environment is more problematic. As can be seen from Figures 5.12 and 5.13 in Chapter 5, in the single individual ILMs, for tight bottlenecks, unstructured environments lead to languages which are partially compositional, and therefore only partially stable.

What consequences does this have in the multi-agent population model? Each learner observes and learns from the communicative behaviour of several other individuals. If these individuals are using a partially compositional system then there will be some randomness in their linguistic behaviour — while a learner's cultural parents may share some part of the meaning-signal mapping, some of their behaviour will be random and therefore unlikely to be shared. This means that the learner will receive contradictory learning input — each of their cultural parents will produce different signals for certain meanings. As discussed above, this means that the learner's bias and the number of meanings in the environment become important. In the sparse environment, partially compositional systems can still emerge — based on a set of conflicting observations, the learners tend to arrive at a partially compositional system. The system never becomes perfectly compositional due to the bottleneck, which reintroduces instability.

However, in the medium density environments, such compositional systems never get off the ground — as can be seen from the Table, highly compositional systems emerge infrequently in unstructured environments when the bottleneck is relatively tight ($c = 0.4$ or 0.5). In these circumstances, learners will be faced with a large set of contradictory input. As a consequence, they will tend to acquire a non-compositional system — as discussed above, given a large number of meanings, the majority non-compositional mapping tends to drown out any weak compositional tendencies. As a consequence, partially compositional systems do not emerge in unstructured environments — the system of meaning-signal mappings remains random, with each individual's system tending to be rather different from the systems of other members of the population.

6.3.3 Summary

Communicatively optimal, compositional languages can emerge in populations through purely cultural processes. When there is no bottleneck on cultural transmission, the fact that learners make observations of several cultural parents increases the impact of their learning bias, effectively allowing them to pick the system of meaning-signal mappings which most closely matches their bias. This leads to the emergence of compositional languages with reasonably high frequency, although only when the environment is sparse and structured — in the absence of a bottleneck, the results are sensitive both to the compositionality of the early systems and to the potential for spread of compositionality.

In the presence of a bottleneck, highly compositional systems emerge with high frequency when the environment is structured. However, when the environment is unstructured such systems only emerge when the bottleneck is relatively wide. Convergence on a compositional language first requires a degree of stability at the individual level. This is straightforwardly achieved when the environment is structured, due to the high potential for generalisation. In unstructured environments, this stability can be achieved when the bottleneck on transmission is not too tight. However, when the bottleneck is tight individual members of the population never arrive at stable systems, and as a consequence the population never converges on a shared language.

6.4 The evolution of learning biases for compositional language

We have established that compositional language can evolve through cultural processes in a population, provided that learners have the appropriate learning bias (of the [$+$constructor, $+$ic-preserver] classification). The final question is to investigate whether this learning bias can evolve through natural selection for communicative success. The simulation results described in Chapter 4 suggest that one-to-one biases for vocabulary acquisition are unlikely to evolve specifically for their communicative function, due to the time delay between the emergence of such a bias and a communicative payoff for individuals possessing it. This should make us skeptical as to whether such a bias can evolve for the acquisition of a (potentially) structured system of meaning-signal mappings.

The model of languages and communication is as described in Section 6.2. As discussed in that Section, there are two possible methods of evaluating communicative accuracy between two individuals — one which counts a communicative episode as a success only if speaker and hearer arrive at exactly the same meaning, and one which gives partial credit for speaker and hearer arriving at partially overlapping meanings. I will investigate both alternatives here.

6.4.1 Genotypes, phenotypes and reproduction

The model of a phenotype communicative agent is as described in Section 6.2 — an associative network capable of representing structured meaning-signal mappings, with an initial set of connection weights $\mathcal{W}$ and a weight-update rule W.

As in the EILM for the simple associative network outlined in Chapter 4, Section 4.5, I will assume that an individual's weight-update rule W is genetically-encoded. A genotype is specified by the 4-tuple $(a_\alpha\ a_\beta\ a_\gamma\ a_\delta)$ where a_x is an allele drawn from the set $\{-1, 0, 1\}$. The process of mapping from a genotype to a phenotype involves converting such a 4-locus chromosome into a $\langle \mathcal{W}, W \rangle$ phenotype. Each weight-update rule W is specified by a 4-tuple $(\alpha\ \beta\ \gamma\ \delta)$. During genotype-phenotype mapping α is set to the value of allele a_α, β is set to the value of allele a_β and so on. The genotype therefore specifies the phenotype's weight-update rule. All $w_{i,j} \in \mathcal{W}$ are set to 0 — every agent has all their initial connection weights set to 0.

To recap, there are 81 possible genotypes, which encode the 81 possible weight-update rules discussed in Chapter 5, Section 5.4. These 81 weight-update rules can be split into four classifications:

- 63 are classified as [$-$maintainer], and are therefore unable to acquire an e-compositional language.
- 11 are classified as [$+$maintainer, $\pm$constructor, $-$ic-preserver], and are able to acquire an e-compositional language, but represent it in an internally-holistic fashion.
- 5 are classified as [$+$maintainer, $-$constructor, $+$ic-preserver], and are able to acquire an e-compositional language, but unable to maintain such a language in the presence of a bottleneck.
- 2 are classified as [$+$constructor, $+$ic-preserver], and are able to acquire, maintain and construct an e-compositional language.

Individuals inherit their genes from their parents. As in earlier EILMs, organisms are haploid but sexual recombination (involving crossover, in an identical fashion to that outlined for the previous EILMs) is used. Newly-formed genotypes are also subject to mutation.[7]

6.4.2 The EILM

A gradual EILM is used — at each cohort, a single individual is selectively removed from the population, the remaining members of the population breed according to communicative success to produce a new individual, and that new individual acquires its communication system based on observations of the population's behaviour.

Initialisation Create a population of N agents[8]. Each initial agent has a random genotype, with the allele at each locus selected randomly from the range of possible alleles. Each initial individual's phenotype is determined by their genotype and the genotype-phenotype mapping.

[7] Point mutations occur on the newly-formed genotype with probability $p_{\ m}$ ($p_m = \frac{0.04}{l_g}$ for all simulations outlined in this section, where l_g is the length of the genome.) Mutation involves replacing the allele a_i at the mutated locus with another allele $a_{j \neq i}$, where a_j is selected from the set of possible alleles.

[8] $N = 50$ for all simulations outlined in this section.

Iteration

1. Select an individual from the population according to the death procedure outlined below and remove it.
2. For every remaining member of the population, generate a set of meaning-signal pairs by applying the network production process to every meaning m in the environment $\mathcal{E}$.
3. Create a new agent. The new agent inherits their genotype from their parents, who are selected from the population according to the reproduction procedure outlined below.
4. The new agent selects θ individuals at random from the population, where θ is randomly selected from the range $[1, \tau]$. The agent then selects τ cultural parents at random from among these θ individuals[9]. The new agent receives e exposures to the communicative behaviour produced by those cultural parents. During each of these e exposures[10] the new agent observes the meaning-signal pairs produced by each parent for a single, randomly selected meaning and updates their connection weights according to the observed meaning-signal pairs and their weight-update rule W. The agent will therefore observe approximately $|\mathcal{E}| \cdot c(\mathcal{E}, e)$ distinct meanings, paired with their corresponding signals produced by each of the τ parents.
5. The new agent joins the population. Return to 1.

As with the EILM outlined in Chapter 4, Section 4.5, tournament selection is used to determine reproduction and death. During each tournament T individuals [11] are selected from the population at random and evaluated. Each individual is scored according to their average communicative accuracy (according to one of the two measures) when acting as both producer and receiver with two randomly selected partners. During selection to decide death, the individual with the lowest communicative accuracy from among the T selected individuals 'wins' the tournament and is removed from the population. During selection to decide reproduction, the individual with the highest communicative accuracy wins the tournament and reproduces.

Note from the iteration procedure that each agent observes a subset of the language of its cultural parents — there is a bottleneck on cultural transmission.

[9] $\tau = 3$ for the runs outlined here.

[10] $e = 24$ for all EILMs outlined in this section, which yields a bottleneck of $c(\mathcal{E}, e) = 0.6$ with respect to the environment described below — each learner observes approximately 60% of the language of its cultural parents.

[11] As in the simple associative network EILM, $T = 3$.

6.4.3 The environment

As discussed above with reference to the population ILM, the associative network model is computationally expensive, both in terms of memory and CPU cycles, particularly when used in the context of a population. This is largely due to the large size of the associative network, and the large number of analysis pairs which have to be evaluated during production and reception.

These problems can be alleviated by reducing the number of feature values (V) and the size of the character inventory ($|\Sigma|$). To this end, for all EILMs outlined in this Section $F = 3$, $V = 3$, $l_{max} = 3$, $\Sigma = \{a, b, c, d, e, f\}$.

This selection of F and V means that the environments used in Chapters 5 and Section 6.3 of this Chapter can no longer be used, due to the changed space of possible meanings $\mathcal{M}$. Instead, an environment is used where $\mathcal{E} = \mathcal{M}$ — every possible meaning in the meaning space is present in the environment. This allows us to simplify away from the structured-unstructured distinction with respect to environments.

The change in environment, number of exposures ($e = 24$ is used in the EILM) and also the change in the population size ($N = 20$ in the population ILM, whereas $N = 50$ in the population EILM) compared to Section 6.3 makes it necessary to rerun the ILM with the new environment and population size. Ten runs of the ILM were carried out, using a [+constructor, +ic-preserver] weight-update rule. All runs converged on a communicatively optimal, highly compositional ($E(\mathcal{O}) \geq 0.95$) language, with the mean time to convergence being 3680 cohorts, although half of the runs converged on a stable system within 2000 cohorts. These runs are plotted in Figure 6.4. This demonstrates that, as before, given the appropriate learning bias, communicatively optimal, compositional language can emerge through cultural processes given this experimental setup.

6.4.4 A negative result

Ten runs of the EILM were carried out, using the all-or-nothing evaluation of communicative accuracy given in Section 6.2 — a communicative episode was only considered a success if speaker and hearer arrived at exactly the same meaning. Runs were allowed to proceed for 10000 cohorts. None of these runs converged on a communicatively-optimal or compositional communication system — all runs remained stuck with a random, e-holistic communication system, which yields chance levels of communicative accuracy. All populations became fixated on genotypes which encoded [$-$maintainer] weight-update rules.

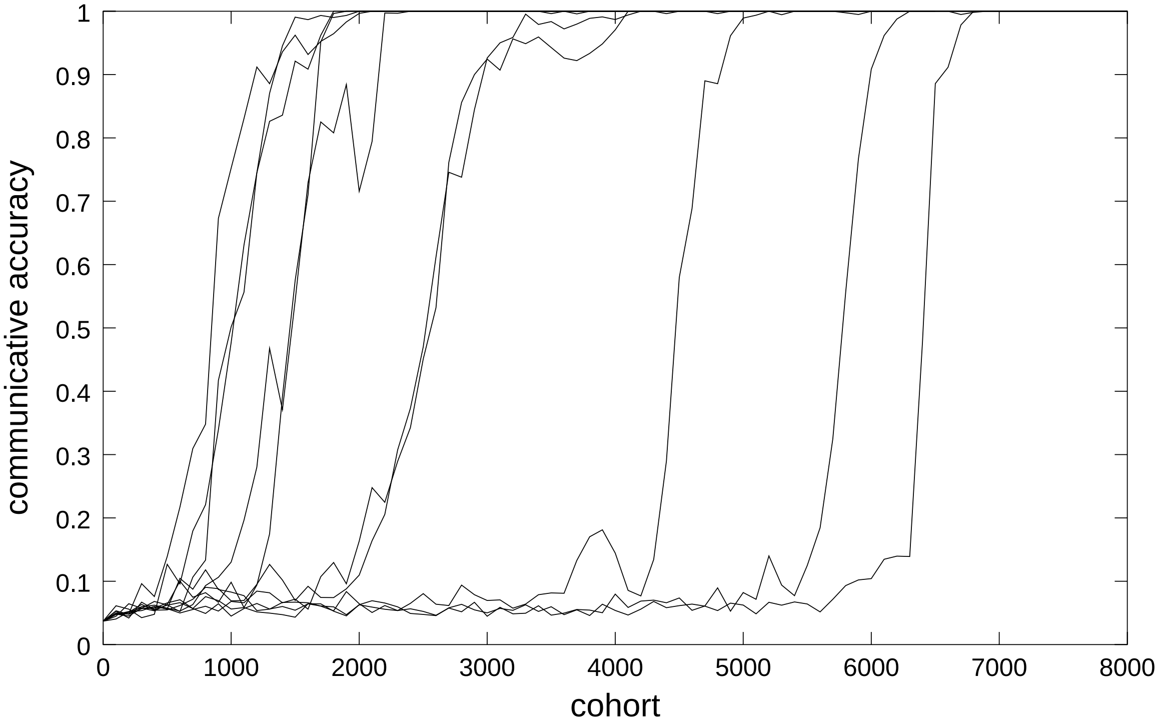


Figure 6.4: Communicative accuracy against time in the ILM with the parameter setting which will be used for the EILM. Communicative accuracy was evaluated according to the all-or-nothing measure. All runs converge on an optimal system, although time to convergence varies considerably from run to run.

These results are unsurprising. We saw in Chapter 4 that appropriate learning biases are unlikely to evolve, given the time-lag between the emergence of such biases and a payoff to individuals possessing them. The evolutionary task for the populations here is much harder — only two of the 81 genotypes are any use (as opposed to nine of 81 in the associative network EILM in Chapter 4) and cultural convergence, even given the correct learning bias, is potentially somewhat slow.

6.4.5 A positive result: the evolution of learning biases for compositional language

A further ten runs of the EILM were carried out, using the partial credit evaluation of communicative accuracy given in Section 6.2 — individuals receive a payoff from communication which is proportional to the similarity between the meaning the speaker was attempting to convey and the meaning the hearer arrives at.

Figures 6.5 and 6.6 show the progress of a simulation run where the population constructs a communicatively optimal, compositional language. This is a typical example of a successful run. Five of the ten EILM runs were successful in this respect — learning

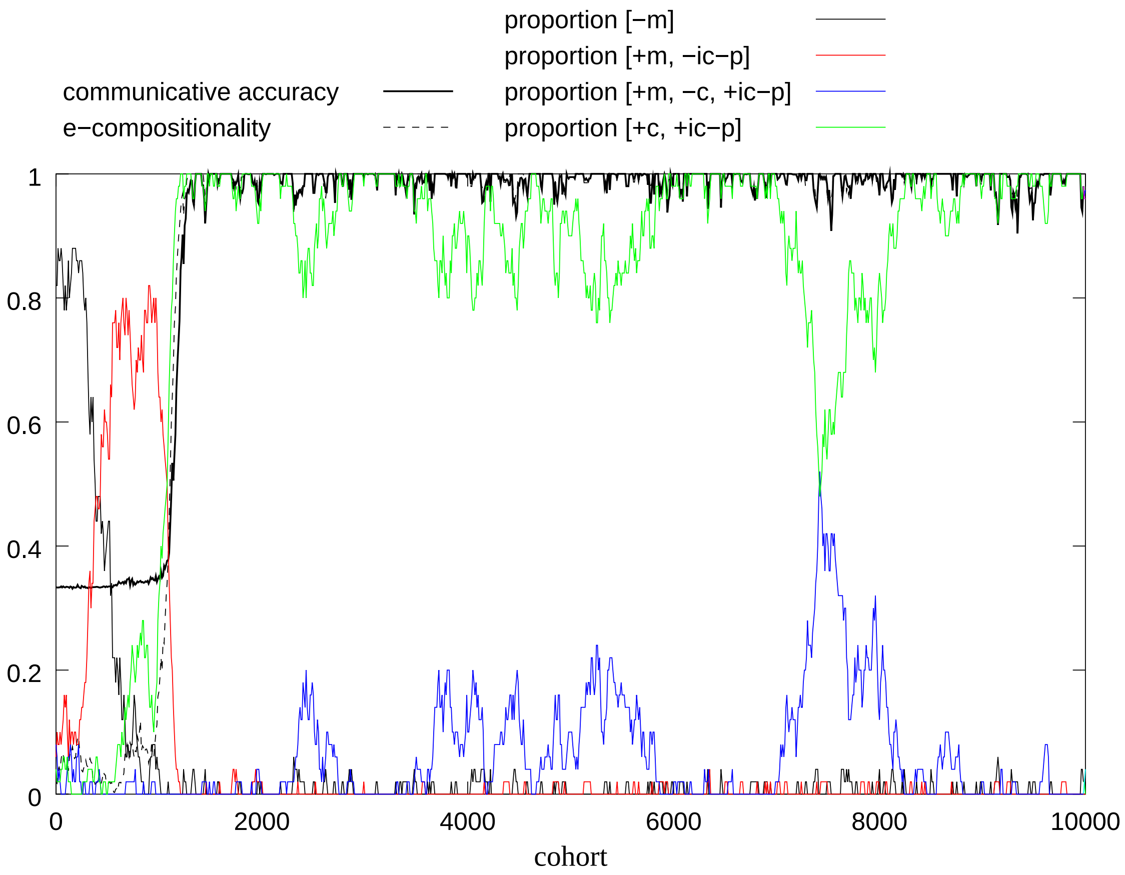


Figure 6.5: The evolution of learning bias leading to communicatively optimal, compositional language. Proportions of various groups of genotypes are given ([±m] stands for [±maintainer], [±c] stands for [±constructor], [±ic-p] stands for [±ic-preserver]). The population's communicative accuracy and compositionality reach maximal values with 2000 cohorts. The population comes to be dominated by [+constructor, +ic-preserver] weight-update rules, although [+maintainer, −constructor, +ic-preserver] weight-update rules do drift in and out after 2000 cohorts.

biases supporting the cultural evolution of a compositional language emerge 50% of the time. Figure 6.7 shows the relationships between the population's average communicative accuracy, e-compositionality and the proportion of individuals in the population with weight-update rules encoding [+constructor, +ic-preserver] weight-update rules. There is a clear relationship — as the number of [+constructor, +ic-preserver] individuals in the population increases, so too does compositionality and, latterly, communicative accuracy.

In Chapter 4, we saw that the evolution of one-to-one biases for vocabulary acquisition consisted of three stages — an initial stage of drift, a stage of selection for the appropriate learning biases, then a further stage of drift. This same three-stage process is evident

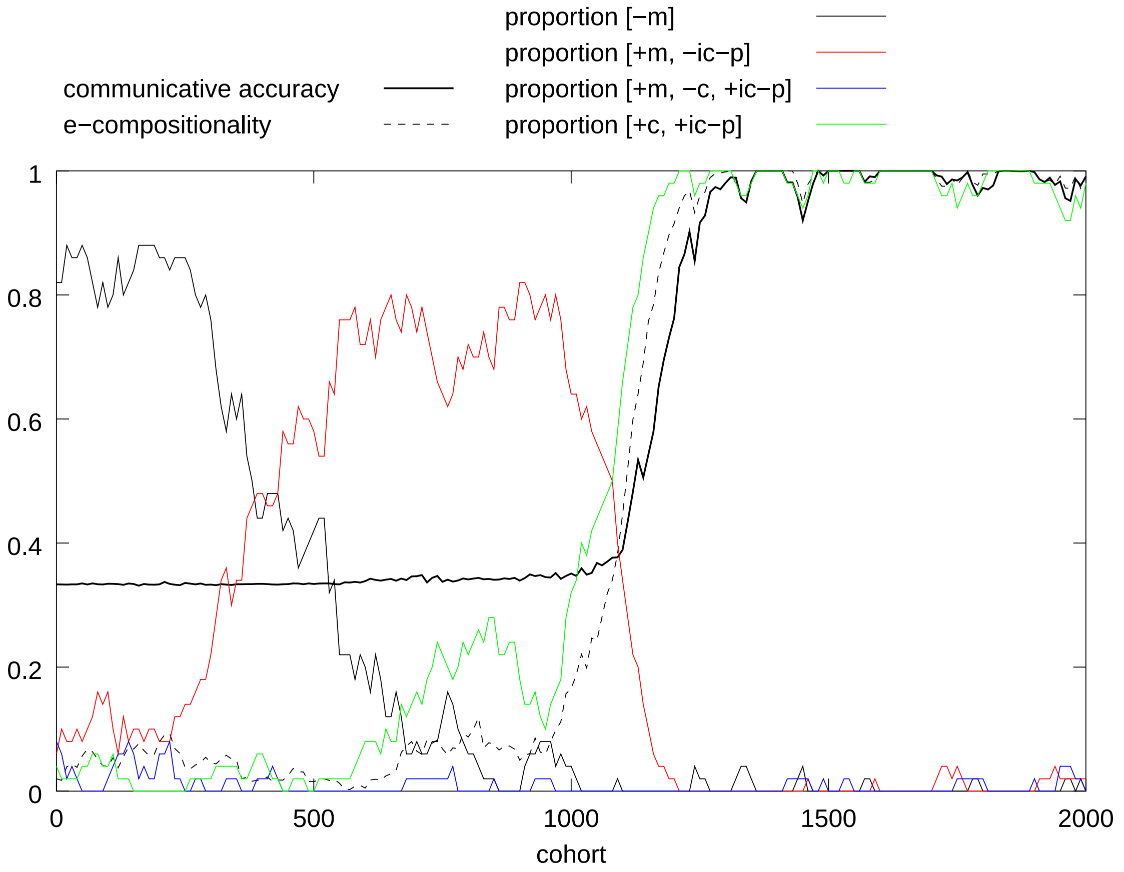


Figure 6.6: The first 2000 cohorts of the simulation run in Figure 6.5. At about 900 cohorts the population is dominated by [+maintainer, ±constructor, −ic-preserver] individuals. There are also a small number of [+constructor, +ic-preserver] individuals present. The numbers of the [+constructor, +ic-preserver] individuals increases sharply from 900 cohorts, and as a consequence the communicative accuracy and e-compositionality of the population's language increases to maximum values.

in runs where learning biases supporting communicatively optimal, compositional language emerge. Figures 6.8–6.11 plot the relative communicative accuracies (*rca*s) for four classes of genotypes in this run. As can be seen from these Figures, the first 2000 cohorts of the simulation run consists of a stage of drift, followed by a stage of selection. The initial drift stage lasts from 0 to 900 cohorts. During this time the *rca* of all four classes of genotypes remains around 1, indicating that no genotype is associated with above-average levels of communicative accuracy. Access to breeding in the population is random for this time. Genetic drift results in an increase in the numbers of [+maintainer, −ic-preserver] individuals (incapable of constructing an optimal system through a bottleneck), and later an increase in the number of [+constructor, +ic-preserver] individuals (capable of constructing such a system).

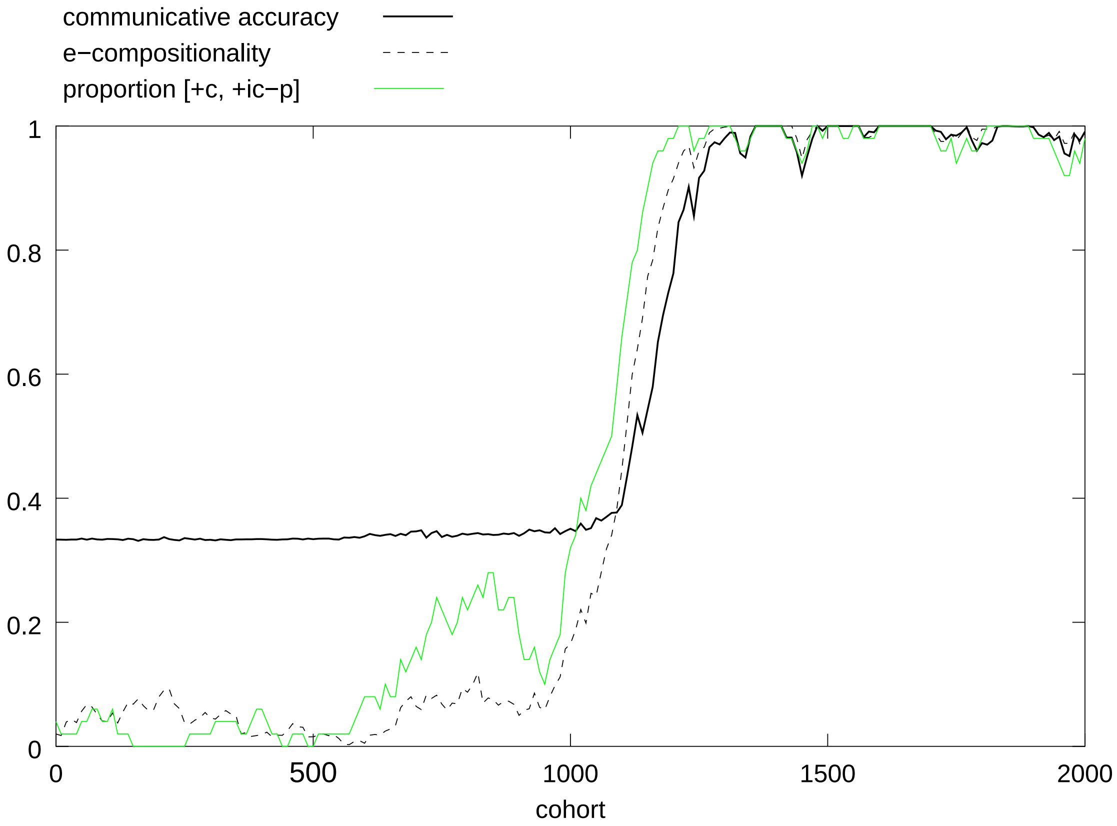


Figure 6.7: The relationship between communicative accuracy, e-compositionality and proportion of [+constructor, +ic-preserver] individuals in the successful run.

The mini peak of [+constructor, +ic-preserver] individuals results in the beginnings of a communicatively useful, partially compositional system of meaning-signal mappings. Consequently, individuals with such genotypes, who are capable of acquiring and contributing to the construction of such a system, receive a communicative payoff — *rca* for these genotypes rises above 1, they receive disproportionate access to breeding roles and their numbers increase sharply in the population. *rca* for other genotypes (particularly [+maintainer, −ic-preserver] genotypes, which form a significant proportion of the population up until 900 cohorts) drops below 1, and their numbers decrease sharply. Individuals with genotypes which make them capable of acquiring and constructing an optimal, compositional language are selected for, to the detriment of other genotypes.

This selection proceeds until the population consists entirely of [+constructor, +ic-preserver] individuals. Shortly after, compositionality and communicative accuracy reach maximum levels — the population converges on a communicatively optimal, compositional language. A second period of drift then ensues. During this period, as can be seen in Figure 6.5, [+maintainer, −constructor, +ic-preserver] individuals are introduced

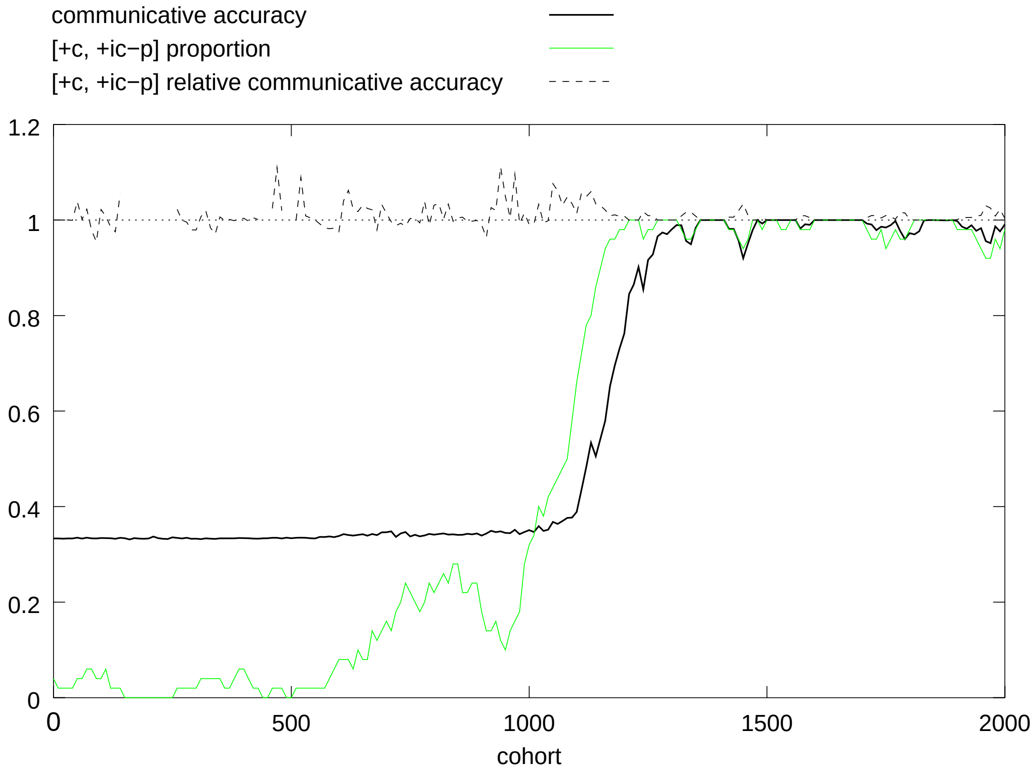


Figure 6.8: The relative communicative accuracy of individuals with [+constructor, +ic-preserver] weight-update rules. This value fluctuates around 1, and is clearly above one for the period from 900 to 1200 cohorts, at which point the numbers of such individuals increase sharply.

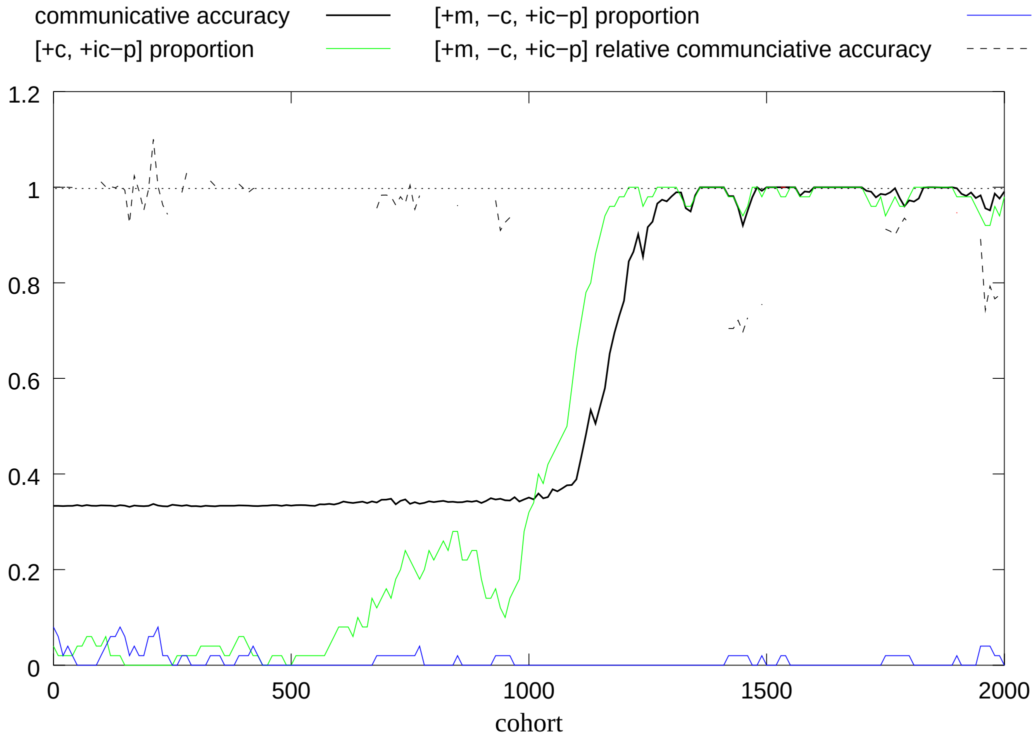


Figure 6.9: The relative communicative accuracy of individuals with [+maintainer, −constructor, +ic-preserver] weight-update rules. These individuals are not present in significant numbers in the early stages of the run.

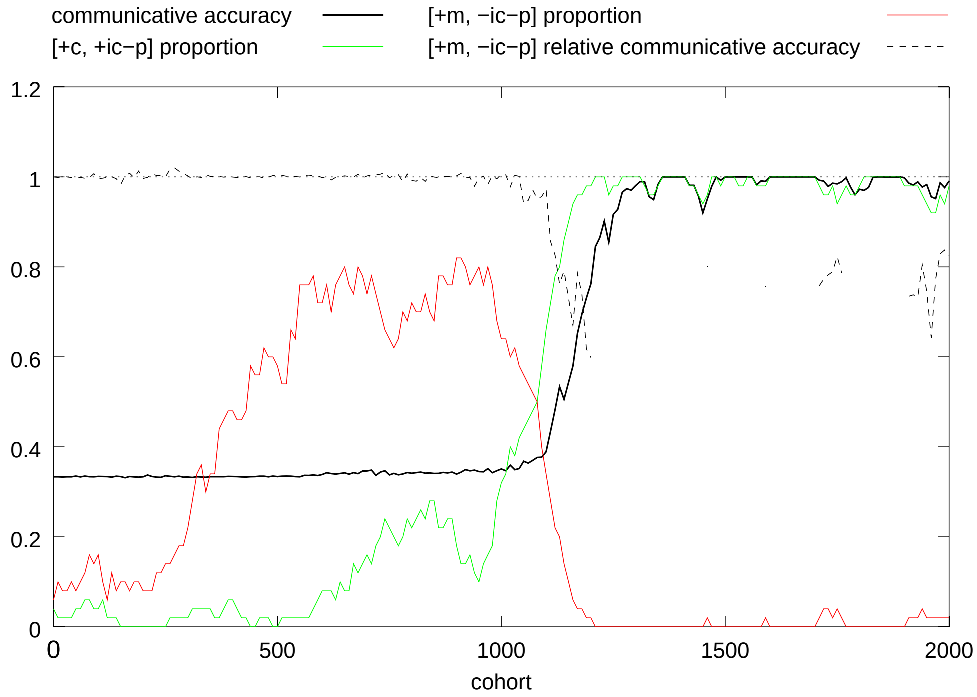


Figure 6.10: The relative communicative accuracy of individuals with [+maintainer, ±constructor, −ic-preserver] weight-update rules. The numbers of these individuals increases from 200 to 600 cohorts. However, the fact that their *rca* remains at 1 suggests that this increase is due to drift. Their numbers drop sharply from 900 cohorts, at which point the number of [+constructor, +ic-preserver] individuals increases sharply. The *rca* of [+maintainer, ±constructor, −ic-preserver] drops below 1 around this point.

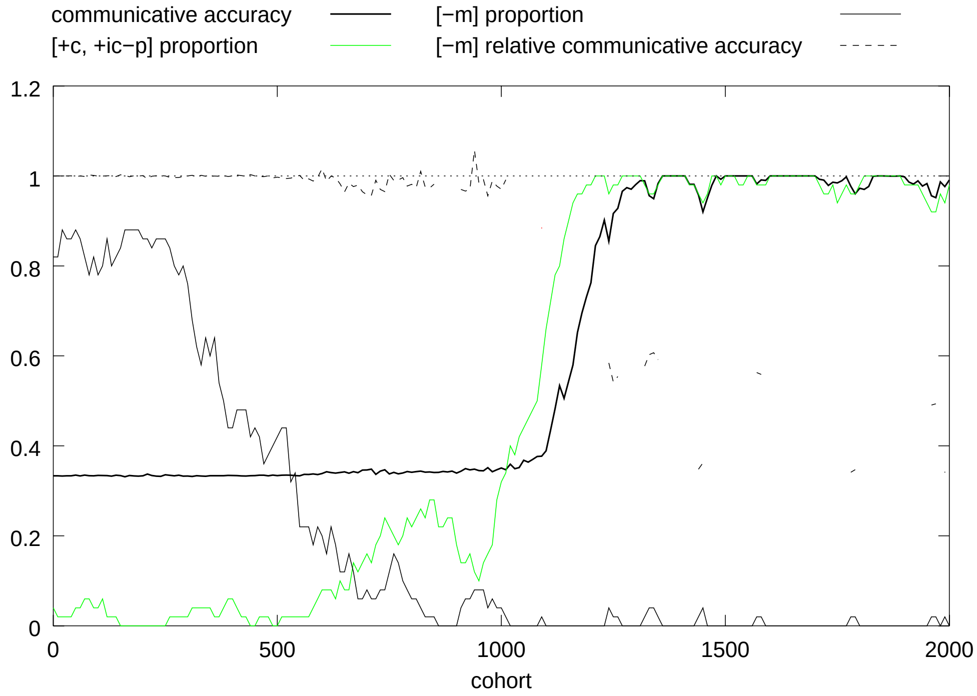


Figure 6.11: The relative communicative accuracy of individuals with [−maintainer] weight-update rules. Their numbers decline from 200 to 700 cohorts. The fact that their *rca* remains around 1 at this point suggests that this is due to drift.

into the population, and their numbers fluctuate randomly, although generally remaining fairly low. We saw in Chapter 5 that individuals using [+maintainer, −constructor, +ic-preserver] cannot maintain an optimal compositional system in a single-agent ILM. However, in a mixed population with [+constructor, +ic-preserver] individuals, given a reasonably wide bottleneck on transmission, such individuals can maintain the optimal system, provided their numbers do not get too great. Consequently, drift allows them to enter the population. At around 7500 cohorts they make up 50% of the population. At this point, they begin to lose the optimal system, the population's overall communicative accuracy drops somewhat, and the [+maintainer, −constructor, +ic-preserver] individuals are selected against until their numbers drop to lower levels.

The successful runs of the EILM therefore exhibit the same three-stage process of drift-selection-drift that we saw in Chapter 4, although the second period of drift is somewhat more constrained. The emergence of the optimal learning bias is therefore dependent on an initial period of drift. In the EILM in Chapter 4 this resulted in a very low number of runs converging on optimal system. In this EILM 50% of runs converge on optimal systems. Why?

The partial-credit measurement of communicative accuracy plays an important role — if the all-or-nothing measurement is used, optimal systems never emerge. The partial-credit measurement reduces the time it takes individuals with [+constructor, +ic-preserver] weight-update rules to get some communicative payoff, and therefore reduces the period of vulnerability to drift. Individuals who learn using [+constructor, +ic-preserver] weight-update rules will not necessarily arrive at the same overall system of meaning-signal mappings, but they may, based on commonalities in the observable behaviour they learn from, arrive at a shared system of mappings from one or two feature values to signal substrings. In the all-or-nothing measurement of communicative accuracy, there is no reward for this — the whole meaning must be correct. However, under the partial-credit scheme, these individuals will receive some small communicative payoff, and therefore be more likely to breed and add more [+constructor, +ic-preserver] individuals to the population. The partial-credit measurement of communicative accuracy smoothes the fitness landscape, making the evolution of appropriate learning biases more straightforward.

A second reason that appropriate learning biases evolve so rapidly is due to the speed of cultural convergence in the EILM — the population moves from random levels of communicative accuracy to optimal levels in around 300 cohorts. As discussed in Chapter 4, speed of cultural convergence plays an important role in the EILM, with a reduction in the time to cultural convergence leading to less sensitivity to drift and consequently more frequent genetic convergence.

The fact that the population in the EILM converges so quickly appears to be at odds with the results from the ILM discussed in Section 6.4.3 above — in a pure ILM, convergence takes on the order of thousands, rather than hundreds, of generations. There are two factors that can explain this disparity. Firstly, in the EILM there is (weak) natural selection of cultural variants, which weeds out systems of meaning-signal mappings which offer below-average communicative accuracy. Secondly, and more importantly, in the ILM the initial system is completely random. In the EILM, at the point where the construction of the optimal system takes off, the population's communication system is not entirely random — 1000 cohorts of 'preparatory' cultural evolution have occurred in the population. This preparatory stage ensures that new individuals entering the population will observe and learn from linguistic behaviour which has common elements. These common elements will be tend to be picked out by [+constructor, +ic-preserver] individuals, who will consequently share some subset of the system of meaning-signal mappings with other [+constructor, +ic-preserver] individuals, thereby reducing the time to cultural convergence.

6.4.6 Summary

Learning biases which lead to the evolution of communicatively optimal, compositional language can evolve through natural selection acting on genetic transmission. However, this only occurs when the partial-credit measurement of communicative accuracy is used. This measurement ensures that individuals with appropriate learning biases receive a fitness payoff fairly rapidly.

Even with the partial-credit measurement scheme, learning biases for compositionality only emerge 50% of the time. This is due to a dependence on genetic drift — as seen in Chapter 4, successful runs exhibit a drift-selection-drift pattern, where the initial period of drift is required to provide appropriate genotypes in sufficient numbers for cultural evolution to get underway. In the structured model of communication, there is less dependence on this initial period of drift, due to the partial-credit evaluation function and the rapid cultural convergence observed in the population. However, the fact remains that this initial period of drift is necessary — there is no immediate advantage in being biased to acquire a communicatively optimal, compositional language in a population which has no established communication system.

6.5 Discussion

What can the results from the Evolutionary Iterated Learning Model tell us about the evolution of language acquisition biases in humans? Much of the discussion relating to the evolution of vocabulary acquisition biases given in Chapter 4 remains pertinent. We can either draw a positive conclusion, and argue that these results show that human-like learning biases *can* evolve through natural selection, or we can take a negative position and argue that these results show that learning biases in humans must have arisen by (initially) non-adaptive mechanisms.

The negative conclusion remains the strongest one — despite the comparatively high level of success in the EILM using the partial-credit communicative accuracy evaluation, the point remains that the evolution of one-to-one biases requires an initial, fortuitous period of genetic drift. The natural conclusion to draw from this is that some mechanisms other than natural selection for communication must have provided appropriate genotypes in sufficient numbers to allow the cultural construction process to get underway. As discussed in Chapter 4, perhaps this one-to-one property was an incidental feature of some learning apparatus which evolved for some other purpose (i.e. the one-to-one bias is a spandrel). Alternatively, the appropriate learning bias may have evolved for some other function, then been pressed into service for communication.

These results do suggest an interesting alteration to the positive interpretation, however. Comparison of the EILM results in this chapter and those in Chapter 4 show that one-to-one learning biases are more likely to evolve when the communication system is potentially structured, given the partial credit fitness function. This suggests that an appropriate learning bias is more likely to evolve for the acquisition of a structured language than an unstructured language — the possibility of regularities in subparts of the meaning-signal mapping smoothes the fitness landscape and simplifies the evolutionary problem. In other words, if we accept the positive interpretation, the models in this Chapter and Chapter 4 show that evolution is more likely to go the whole hog and evolve a learning bias for language, rather than evolving a learning bias for unstructured vocabulary then later elaborating this bias.

6.6 Summary of the Chapter

In the first part of this Chapter I demonstrated that compositional language can emerge in populations of linguistic individuals through cultural processes, given a bottleneck on cultural transmission, a structured environment and the appropriate learning bias. I then went on to show that this learning bias can evolve under natural selection for communication. However, the evolution of this bias is dependent on an initial period of genetic drift — there is no immediate advantage to individuals possessing the appropriate bias in a population with no established communication system. The results and discussion from Chapter 4 therefore pertain to the evolution of learning biases for structured communication — such biases may best be explained as a spandrel or exapted trait.

CHAPTER 7

Conclusions

I began this thesis by highlighting three features which, in combination, make language unique among the communication systems of the natural world — language is the only communication system which is culturally transmitted and symbolic and compositional. Theories of the evolution of language must explain the origins of these three distinctive features of language.

A recurring theme of this thesis has been to explain the latter two hallmarks of language in terms of the first — to explain the conventionalised symbolic vocabulary and compositional structure of language as features which arise during the cultural transmission of language. Pressures acting on language during cultural transmission result in the emergence of conventionalised symbolic vocabulary and compositional linguistic structure, through the adaptation of the linguistic system to the medium of cultural transmission.

A feature of this thesis has been the extensive reference to, and use of, formal models — as discussed towards the end of Chapter 1, theories of language evolution typically appeal to the interactions between several complex adaptive systems, and our intuitions about the behaviour of these kinds of systems are notoriously poor. Formal modelling techniques, such as those I have outlined here, allow us to investigate the outcomes of complex adaptive processes, and, by experimentation, to uncover underlying regularities and determinants of the behaviour of such systems. This kind of experimentation has allowed me to identify what I consider to be the two key determinants of linguistic structure — a bottleneck on cultural transmission and a one-to-one bias in language learners.

Can other features of language be explained as a consequence of pressures acting on language during cultural transmission? The comparison of design features of language and other communication systems carried out in Chapter 1 suggests that duality of patterning, whereby a small number of meaningless elements (phonemes) are combined to form a large number of meaningful elements (words), may be a fourth unique characteristic of language. Duality of patterning seems to be a prime candidate for explanation by this kind of cultural adaptation account. We can envisage a scenario under which agents produce sequences of articulatory gestures, initially without any underlying system of phonemic coding, which are reinterpreted over cultural time as sequences of gestures representing underlying phonemic representations. Formal models have demonstrated that, for vowel systems at least, this kind of cultural self-organisation can lead to the emergence of phonemic coding systems, via reinforcement learning (e.g. de Boer (2000), de Boer (2001)) or a coupling of sensory and motor systems (Oudeyer 2002).

These models of the emergence of phoneme systems need to be extended to explain the combination of phonemes to form words. However, the initial results of de Boer and Oudeyer, in conjunction with the research on symbolicism, compositionality and recursion reviewed and described in this thesis, suggest that cultural processes may offer a powerful and general explanation for most of the hallmarks of human language.

In Chapter 2 I introduced Boyd & Richerson's (1985) general model of cultural transmission, and discussed more language-specific treatments of cultural processes. There are three main issues in treating language as a culturally transmitted system. The first is to identify the units of cultural replication. It may be the case that there is no true replication in a cultural system — the cycle from E-language to I-language means that there is no direct replication of either E-language or I-language. The second problem is identifying the cultural traits of interest. Given that the formation of I-language is dependent on observation of E-language, and the production of E-language is dependent on an underlying I-language, both (or neither) external and internal language may be considered as culturally transmitted traits.

I have followed Boyd & Richerson's assumption, implicit in Boyd & Richerson (1985) and explicit in Boyd & Richerson (2000), that we can ignore such issues, at least until we have established the usefulness of cultural approaches to explaining the human behaviour of interest. As Boyd & Richerson put it,

> "[i]f the application of Darwinian thinking to understanding cultural change depended on the existence of replicators, we would be in trouble. Fortunately, culture need not be closely analogous to genes. Ideas must be gene-like to the extent that they are somehow capable of carrying the cultural information necessary to give rise to cumulative evolution of complex cultural patterns that differentiate human groups . . . this can be accomplished by a

> most ungene like, replicatorless process of error-prone phenotypic imitation. All that is really required is that culture constitutes a system maintaining heritable variation" (Boyd & Richerson 2000:158).

In other words, we can ignore the precise mechanisms underlying cultural transmission, provided that we accept that cultural transmission, or cultural heritability, is possible in some sense. The Darwinian theory of evolution by natural selection was conceived, and broadly accepted, in spite of the lack of a precise understanding of the units underlying biological heritability, or the units of biological replication. Theories of cultural evolution should be allowed similar leeway, at least for the time being.

The third problem of modelling the cultural transmission of language concerns the nature of the primary linguistic data available to learners. I have assumed throughout this thesis that learners observe meaning-signal pairs produced by other members of their linguistic community. I have further assumed that the mental representations corresponding to meanings are shared within a population and, in the case of the models of structured language discussed in Chapters 5 and 6, that these meanings are structured. My position has been similar to that of Schoenemann in that I "take for granted that there are features of the real world which exist regardless of whether an organism perceives them . . . [d]ifferent organisms will divide up the world differently, in accordance with their unique evolved neural systems . . . [i]ncreasing semantic complexity [or meaning space structure] therefore refers to an increase in the number of divisions of reality which a particular organism is aware of" (Schoenemann 1999:318). Similarly, I take it that perception of structure in the environment is a consequence of focusing on regularities in the real world which exist regardless of whether organisms perceive them. This capacity for structured semantic representations may form a preadaptation for language, or as discussed below, may co-evolve with linguistic forms.

Under this set of assumptions, I have shown that cultural evolution can deliver up some of the characteristic structure of language. But how safe are these assumptions?

As discussed in Chapter 2, there is some evidence that human infants can deduce the meaning which should be associated with an utterance, through strategies such as intentional inference (Baldwin 1991; Baldwin 1993a; Baldwin 1993b; Tomasello & Barton 1994; Tomasello *et al.* 1996), and biases to attend to, for example, whole objects (Gentner 1982; Macnamara 1982). However, I am not suggesting that this process is error free. It has been argued that if this process *were* error free, and humans were effectively telepathic, then there would be no need to produce utterances at all. I think this is perhaps over-stating the case. We might accept that the process of identifying the referents of nouns like “book” or verbs like “jump” were error-free, while simultaneously accepting that the process of deducing the meaning associated with utterances such as “Last Tuesday, my brother told me he was in love with a widow” would be error-prone — the straightforward deduction of the meaning of simple terms might form a building block for the acquisition of some language, which then allows the acquisition of more opaque terms by syntactic context, more complex processes of deduction and so on. As a further defence, it has been demonstrated that the basic Expression/Induction framework can work even if we assume that the observation of meaning is error-prone (Hurford 1999) — if we weaken the assumption that learners receive error-free meaning-signal pairs, the whole E/I approach does not come crashing down.

Formal modelling approaches have been used to tackle the issue of language evolution in the absence of explicit meaning transfer (e.g. Steels (1997), Steels (1998), Smith (2001a) and Smith (forthcoming)). These demonstrate that stable communication systems can emerge if we do not assume that learners are presented with meaning-signal pairs. In these models, learners are instead presented with a signal and a set of objects in a simplified world, where one feature of one of the objects is the referent of the signal. These models have been broadly interpreted as showing that explicit meaning transfer is an unnecessary assumption, and therefore should not be made. I would draw the opposite conclusion: these models show that not too much rests on the assumption of explicit meaning transfer, therefore we should make this assumption in the interests of simplicity. However, these models are still extremely valuable in that they shed light on a second issue related to the observation of meaning: how do members of a population converge on a shared set of meanings, and how are these meanings structured?

Andrew Smith (Smith forthcoming) demonstrates that agents will only arrive at similar semantic representations through self-organisation if they 1) construct meanings in an intelligent way, so that they discriminate objects in the world from other objects in the world and 2) inhabit a world which is “clumpy”, such that objects cluster together in object space, rather than being randomly distributed throughout the space of possible objects. I have assumed that all members of a population have access to a set of shared semantic representations. Smith shows that such shared semantic representations are essential for successful communication within a population, if we assume an observational, rather than reinforcement, learning paradigm. It is intriguing to note that Smith finds the emergence of such shared semantic representations only where the world is clumpy, and I find that compositional languages emerge most frequently when the shared semantic representations are “clumpy” (i.e. occupying a structured subspace of the space of possible meanings). This suggests a scenario where a world exhibiting a certain limited and regular degree of variation in the objects it permits leads to the formation of shared semantic representations in communicative agents, which in turn leads to the emergence of a shared, compositional language.

The second half of Chapter 2 was spent outlining Boyd & Richerson's taxonomy of pressures operating on cultural transmission — natural selection of cultural variants, guided variation, and biased transmission (which can be further subdivided into directly-biased transmission, indirectly-biased transmission and frequency-dependent bias). As discussed in that Chapter, all of these processes have been implicated in the cultural evolution of language. In this thesis I have focussed almost exclusively on directly-biased transmission, specifically cultural evolution resulting from learner biases with respect to the one-to-one nature of meaning-signal mappings, although in Chapter 3 I briefly consider the possibility of the natural selection of cultural variants. This force turns out to be weak in comparison to the pressure arising from learner biases — learner biases tend to eliminate the cultural variation which natural selection feeds upon.

Chapters 3–6 constitute the bulk of the thesis, and outline original research into the cultural and biological evolution of language and language-learning biases. In Chapter 3 I show that a bias in favour of one-to-one mappings is required for the cultural evolution of a symbolic, communicatively functional vocabulary. I further argue that this bias is present in humans. In Chapter 5, I demonstrate that a one-to-one bias is necessary for the cultural evolution of compositionally-structured language, although such languages only emerge frequently when there is a bottleneck on cultural transmission. Once again, I show that children appear to bring precisely this kind of bias to the language acquisition task.

The theme of Chapters 3 and 5 is that one-to-one biases play a crucial role in the evolution of language, and that one-to-one biases apply during language acquisition by human infants. The clear implication of these Chapters is therefore that linguistic structure may be a reflex of human learning biases applied to the cultural transmission of language, rather than a consequence of an innate specification of linguistic structure. A strong nativist might argue that linguistic structure is prespecified, and in order to understand language we need only to understand this innate prespecification. My position is that language is only prespecified in the sense that language learners come to the language learning task with an innate learning bias. Understanding this learning bias is important, but not the whole story — we also have to understand how the learning bias interacts with the medium of cultural transmission. It is only through this interaction that linguistic structure emerges. For example, if there is no bottleneck on cultural transmission then the one-to-one learning bias does not lead to the reliable emergence of compositional structure. However, a bottleneck on transmission radically changes this picture — the bottleneck forces language to be generalisable, and this pressure, in combination with the biases of learners, leads to the emergence of compositionality. Compositionality is not specified entirely in the learner, but emerges through the interactions between learner biases and transmission of linguistic structure on a cultural substrate (Brighton *et al.* forthcoming).

There are several important implications, and potential elaborations, of the research described in Chapters 3 and 5. Firstly, it highlights the importance of one-to-one biases in explaining the cultural evolution of linguistic structure. Learning bias has typically taken a back seat in explanations of linguistic evolution, with the role of the transmission bottleneck and the resultant pressure to generalise receiving more attention. However, Chapter 5 demonstrates (both through my original modelling and a review of the learning biases used in Kirby (2002), Batali (2002) and others) that one-to-one biases are crucial for the evolution of compositionality. I suspect that we might find them to be important on almost every level of linguistic structure.

Generalisation alone leads to the loss of linguistic structure — the generalisation that any meaning can be expressed by any linguistic form covers all possible combinations of data, and subsumes all other generalisations. We therefore might expect, given a pressure to generalise, that the most stable (or only stable) system would be one which maps any meaning to any form. This is clearly not what we see in language — there must be something counteracting the pressure to generalise in an unconstrained way. When considering this possibility, Hurford concludes that natural selection (of cultural variants or of genetic variants) must be responsible for preventing overgeneralisation:

> “in hearing a particular syllable used to express a particular atomic meaning, an acquirer might in theory make the absurd overgeneralization than *any* syllable can be used to express *any* meaning . . . [a]ny tendency to make overgeneralizations of such an absurd kind would presumably be eliminated by natural selection based on success in correctly divining a speaker’s meaning and/or successfully signalling ones own meaning. Any mutant [cultural or biological] displaying any tendency to generalize from the primary linguistic data in ways which will lead to her being misunderstood, as she would be if she used *any* form to convey *any meaning*, will be at a disadvantage.” (Hurford 2000:348).

We can probably rule out natural selection acting on overgeneral cultural variants — as discussed above, such pressures are fairly weak in comparison to those arising from learner biases, and are also likely to be weak in comparison to the pressure to generalise arising from a bottleneck on transmission. This leaves us with Hurford's (intended) conclusion, that natural selection must weed out individuals who are capable of learning in an overgeneral way. This of course implies that the learning capacity with respect to generalisation is genetically encoded. I would further argue that this genetically encoded constraint on generalisation comes in the form of a bias in favour of one-to-one mappings between meanings and signals — such biases constrain generalisations in *precisely* the right way. They allow the bottleneck on transmission to determine the degree of generalizability of the linguistic system, while at the same time constraining the system so that communicative function is maximised.

Let me take a more concrete example. In the G&B framework of syntactic analysis, reference is commonly made to a general movement operator, move-α. Move-α states that any element in a syntactic structure can be moved to any other position in that syntactic structure. Of course, this generates syntactic structures which no speaker of a language would accept as grammatical. Move-α is therefore constrained in certain ways — for example, in terms of the possible landing sites of the moved element or in terms of "islands" from which elements may not move.

Move-α is clearly an overgeneralisation. This generalisation would have a high yield in cultural terms, and we might therefore expect it to emerge over cultural time, given a bottleneck on transmission. The fact that unconstrained movement does not occur suggests that there is some counteracting force at play. This could be a consequence of an arbitrary bias in learners, resulting from the structure of internal representations or from the wiring of the language centers of the brain. However, it could also be a consequence of a preference for one-to-one mappings between underlying semantic representations and surface forms. Unconstrained application of move-α would scramble such relationships, potentially beyond repair. A preference for one-to-one biases would constrain movement so that, by and large, the correspondence between underlying semantic forms and surface forms is preserved.

A second implication, if we accept that one-to-one biases have an important role to play in shaping linguistic evolution, is that more research should focus on identify these biases (or their absence) in the process of language acquisition. The Contrast/Mutual Exclusivity bias has been fairly well established, and I have suggested that similar experiments could be used to identify a bias against homonymy in lexical acquisition. One-to-one biases in the acquisition of structured linguistic form might be more difficult to isolate experimentally. However, language acquisition in naturalistic circumstances provides fairly good evidence on this, and the reanalysis of existing data in terms of one-to-one biases might shed further light on this issue.

Hauser *et al.* (2002) emphasise the importance of cross-species comparison in informing theories of language evolution. I would suggest that one-to-one biases provide an ideal candidate for this type of comparative approach. The tests for Contrast/Mutual Exclusivity, and the proposed test for a bias against homonymy, are fairly simple experiments which should be relatively straightforward to apply to non-human species, or at least the higher primates. However, to date only a single such experiment has been carried out (Lyn & Savage-Rumbaugh 2000), using two subjects, and the results are somewhat equivocal.

In the concluding sections of Chapters 3 and 5 I make the point that one-to-one biases in the learner are not the only possible biases at play in the cultural transmission of language — in particular, there are two further pressures, summarised by Langacker (1977) as pressures for signal simplicity (a least-effort preference on the part of individuals producing utterances) and code simplicity (a bias, in learners, against having to memorise large numbers of fixed expressions, such as words). I pointed out that these two pressures work *with* any one-to-one bias in learners in eliminating synonymy, but work *against* the one-to-one bias by favouring homonyms. This was advanced as one possible explanation for the fact that homonyms are fairly common in language, whereas synonyms are not. A worthwhile extension to the models in Chapter 3 and (particularly) Chapter 5 would be to introduce an explicit treatment of such pressures. This would show whether stability is a possibility given a tension between several learner/speaker biases and, if so, what structure the linguistic systems exhibits at stable states.

A second possible line of extension is to consider a fuller range of pressures operating at the level of the Arena of Use, the interface between the grammatical competence of one individual and the primary linguistic data of another. The transmission bottleneck is one aspect of the Arena of Use (an infinitely expressive competence is represented only by a finite number of utterances), as is the proposed bias in favour of signal simplicity. However, there are other possible pressures which could act at this level. Firstly, there are aspects of the poverty of the stimulus other than the transmission bottleneck. In Chapter 1 I summarised Pullum & Scholz's (2002) analysis of aspects of the poverty of the stimulus problem, which I repeat here.

1. Children are not specifically or directly rewarded for their advances in language learning.

2. Children's data-exposure histories are finite, but they acquire an ability to produce or understand an infinite number of sentences.
3. Children's data-exposure histories are highly diverse, yet language acquisition is universal.
4. Children's data-exposure histories are incomplete in that there are many sentences they never hear, yet can produce and understand.
5. Children's data-exposure histories are solely positive — they are never given details of what is ungrammatical.
6. Children's data exposure histories include numerous errors, such as slips of the tongue and false starts.

The transmission bottleneck corresponds to points 2 and 4 in this list, and possibly point 3 — diversity of input might be a consequence of taking a small sample from an infinite set. Diversity of input might also add a further pressure for generalisation. Consider the case where children receive learning input both from adults, and from other children. If each child receives different input from adults, this will tend to lead different children to converge on different grammars. However, if children also learn from one another, there will be a pressure for the children to converge on a grammar which is more general than that suggested by each individual's input from adults — learner-learner contact might force children's grammars to accept the union of the sets of sentences produced by the (adult) cultural parents of each child. Similarly, point 5 in the list above might introduce a bias towards generality — the lack of explicit negative evidence will tend to allow children to misconverge on superset grammars. The possible consequences of the first and final points in this list of aspects of the poverty of the stimulus are less obvious.

Finally, pressures arising from the passage of language through the Arena of Use might offer one possible explanation for *iconicity* in syntax. Iconicity has been used as a fairly general term which subsumes what I have called isomorphism, the one-to-one relationship between semantic representations and surface structures. Iconicity has also been invoked to cover cases where linguistic form appears to mirror non-linguistic reality. McMahon (1994) gives the scarcity of languages where objects precede subjects as a possible example of iconicity — this phenomenon "might be ascribed to the greater relevance or perceptual salience of the Subject in real-world situations; in linguistic representations of those situations, the Subject therefore comes first" (McMahon 1994:86). If this type of iconicity is to arise anywhere, it must be in the Arena of Use.

In Chapters 4 and 6 I describe models designed to test whether natural selection acting on genetic transmission can identify learning biases which lead, through cultural processes, to optimal (and compositional, in Chapter 6) communication. The main result of these Chapters is to show that this was not the case, or at least not reliably so. The cultural evolution of a functional communication system takes time, and consequently there is no immediate fitness payoff for individuals who are appropriately biased. This leads to random genetic transmission in the early stages of the construction process, which can stop cultural evolution before it gets started. These experiments support the general logical point that there is no advantage in being able to acquire an optimal communication system if there is no meaningful communication system to acquire.

These experiments indicate to me that the default assumption about the evolution of the human language acquisition biases should be non-adaptationist, rather than adaptationist. Even in the simple models in Chapters 4 and 6, with a fairly simple genetic search space and strong selection pressure acting in favour of successful communication, such biases are unlikely to evolve. It therefore seems safer to conclude either that the biases applied to the acquisition of language either a) originally evolved under selection pressure for some other function or b) did not evolve for any specific function, but are a spandrel, a coincident feature of some learning apparatus which was selected for. On the face of it, this is a rather frustrating conclusion — while I give a strong argument on the learning biases required for language, my account of the evolution of such biases appeals to extra-linguistic factors.

However, things are not as bleak as all that. Firstly, I am not ruling out the possibility of evolutionary reappropriation followed by adaptation specifically for language. For example, the human language learning biases might have evolved for some other purpose, then been pressed into service for language and subsequently specialised solely for language. I would also allow the possibility that the appropriate learning biases *arose* for some non-linguistic or general purpose, but were almost immediately identified as being good for communication and *spread* widely as a direct result of their utility for communication.

In part because of the rather unsatisfactory conclusions forced upon us by the results of Chapters 4 and 6, this area is ripe for further exploration. I have focussed on the assumption that learning biases are genetically encoded. An intriguing alternative possibility is that they are learned. Quinn (2001) shows that the capacity to communicate can evolve in populations which originally do not communicate, and are not pre-configured for communication. The cultural parallel to this would be to show that the capacity to learn from others was itself a learned capacity. More subtly, individuals might learn which type of learning is best for acquiring a communication system, in which case I anticipate they would learn to apply a one-to-one learning bias. Of course, the evolution of the meta-learning process, whereby individuals learn to learn from others, then has to be explained.

One possibility is that this meta-learning capacity is rooted in our understanding of others as intentional agents.

A second implication of these results relates to the role of the Baldwin effect in explaining language evolution. The Baldwin effect has received a great deal of attention of late (witness Jackendoff's comment that he "agree[s] with practically everyone that the 'Baldwin effect' had something to do with it [language evolution]" (Jackendoff 2002:237)), but it is not clear how the Baldwin effect relates to accounts of biological evolution of a learning bias which effects cultural evolution. The Baldwin effect predicts that behaviours which are initially learned tend to become innate — in other words, genes adapt to fit cultures. However, in my models the opposite is true — cultures adapt to fit genes, because genes encode learning biases which guide cultural evolution. What would Baldwinian evolution look like in this framework? We could envisage a scenario where the genes encode a range of strengths of bias. The emergence of a weak bias, say in favour of one-to-one mappings, results in the emergence of cultural traits conforming to this bias. Via the Baldwin effect, stronger and stronger forms of the one-to-one bias then evolve biologically, to allow individuals to reliably acquire the dominant cultural trait. Learned traits do not strictly speaking become innate, but learning becomes more restrictive and more biased. This seems to me to be a form of the Baldwin effect which is very appealing for explaining language evolution — we do not necessarily want to say that there is a tendency for language to become innate, but we might want to say that there has been evolution to more and more constraining forms of learning.

Finally, Chapters 4 and 6 indirectly highlight the importance of population dynamics in understanding the cultural evolution of language. In Chapter 4, the spatial organisation of populations was shown to have consequences for the speed of cultural evolution in those populations, which in turn impacted on the biological evolution of learning biases. In Chapter 6, in the population Iterated Learning Model, factors such as the number of cultural parents each individual has impacts on the possibilities for cultural evolution. It is reassuring that these models are sensitive to population structure and population dynamics — as discussed in Chapter 1, population dynamics play a role in creolization events, and the population dynamics have to be in some sense 'right' before creolization can take place (Ragir 2002). However, it is not clear what the right population dynamics are, or why precisely these matter. Computational modelling could be profitably applied to the investigation of such questions.

To summarise, the central theme of this thesis has been to explain the unique features of language in terms of the cultural adaptation, by language, to two pressures:

1. a pressure to be generalisable, arising from a bottleneck on cultural transmission
2. a pressure to conform to a learner preference for one-to-one mappings between meanings and signals, and parts of meanings and parts of signals.

Linguistic structure is not specified entirely in the learner, but emerges through the interactions between learner biases and transmission of linguistic structure on a cultural substrate.

APPENDIX A

Mathematical models of transmission

In this appendix I will present details of mathematical models of cultural and genetic transmission. In Section A.1 I outline B&R's treatment of cultural transmission and the factors influencing cultural evolution. In Section A.2 I outline a simple mathematical model of genetic transmission and biological evolution by natural selection. Finally, in Section A.3 I describe the mathematical details of B&R's model of the dual transmission of cultural traits and a genetically-encoded direct bias.

A.1 Models of cultural transmission

This Section covers B&R's basic models of unbiased cultural transmission (Section A.1.1), and their treatment of the various pressures acting on cultural transmission (Section A.1.2).

A.1.1 Basic cultural transmission models

A.1.1.1 Transmission of dichotomous traits

B&R provide a simple model of the cultural transmission of a dichotomous trait, where individuals are either characterised as having cultural trait c or d. p is the proportion of individuals in the population with cultural variant c, and $1 - p$ is the proportion of individuals with variant d. p' is the proportion of individuals in the population with cultural variant c after cultural transmission.

Each individual acquires their cultural variant based on their observations of the cultural variants of n cultural parents, or models. The probability that a naive individual acquires variant c based on the behaviour of n models is therefore:

$$Prob(individual = c|X_1, \ldots, X_n) = \sum_{i=1}^{n} A_i X_i$$

where $X_i = 1$ if the ith model possesses variant c and $X_i = 0$ if the ith model possesses variant d, A_i is the probability that the naive individual acquires the variant of the ith model and $\sum_i A_i = 1$. A_i therefore gives the importance of the ith cultural parent in the enculturation process.

Given this equation we can now calculate p', the proportion of individuals with variant c after cultural transmission. For this we require the probability that a given set of models $(X_1, \ldots, X_n)$ is formed, $Prob(X_1, \ldots, X_n)$. This leads to:

$$p' = \sum_{x_1=0}^{1} \ldots \sum_{x_n=0}^{1} Prob(c|x_1, \ldots, x_n) Prob(X_1 = x_1, \ldots, X_n = x_n)$$

In other words, the proportion of individuals with variant c is equal to the probability that an individual will acquire variant c based upon exposure to a specific set of n models, multiplied by the probability of the formation of that set of models and summed over all sets of models. p' therefore depends on the probability of forming sets of models. If we assume that the probability of any cultural parent possessing variant c is equal to p (i.e. cultural parents are drawn at random from the population) then it can be shown that cultural transmission leaves the frequency of cultural variants in the population unchanged i.e.

$$p' = p$$

Therefore, if the original population exhibits variation for some cultural trait, cultural transmission itself will not reduce that variation or alter the distribution of variants, assuming random selection of cultural parents — cultural transmission alone will not lead to cultural evolution or cultural adaptation.

A.1.1.2 Transmission of continuous traits

In the continuous trait model each individual is characterised by a single number, X, representing the value of their culturally-acquired character. In this case a population cannot be characterised by the proportion of one cultural variant, as in the dichotomous character model, but must rather be modelled as a distribution over values of X, $P(X)$.

Making the simplifying assumption that $P(X)$ can be approximated by a normal distribution allows a population to be characterised by the mean value of X in the population, $\overline{X}$ and the variance of X in the population, V.

B&R consider how a blending inheritance model would alter the mean and the variance of a cultural characteristic in a population. As in the dichotomous model, each naive individual is exposed to the behaviour of n models, with the cultural variant of the ith model being X_i. Based on these observations, the naive individual makes an estimate of the ith model's cultural rule, Z_i, where:

$$Z_i = X_i + e_i$$

where e_i is a random variable drawn from the normal distribution with mean 0 and variance E_i, representing errors in the naive individual's estimate of the model's cultural character. As the name suggests, in a blending inheritance model the enculturated individual's cultural variant, X_0, is simply the average of their estimates of their n models' variants:

$$X_0 = \sum_{i=1}^{n} A_i Z_i$$

where, as in the dichotomous model, A_i is the importance of the ith model.

By a similar method to that used for the dichotomous case, it is possible to calculate the mean value of X in the population after cultural transmission, $\overline{X}'$ and variance of X in the population after transmission, V'. B&R show that, assuming non-selective formation of sets of models:

$$\overline{X}' = \overline{X}$$

In other words, blending inheritance does not change the population mean of the cultural variant, as with the dichotomous model. However, the variance of the population does not necessarily remain unchanged. Assuming non-selective formation of sets of models, equal weighting for all models ($A_i = 1/n$ for all i) and no correlation between errors in a given set of models ($Cov(e_i, e_j) = 0$ for all i and j):

$$V' = (1/n)(V + \overline{E})$$

where $\overline{E}$ is the average value of E_i (recall that E_i gives the variance of a normal distribution, and the errors made by the learner when estimating the trait of the ith cultural parent come from this distribution) for the set of n models. In other words, there are two forces acting on the population. Assuming no errors in transmission ($\overline{E} = 0$), blending transmission tends to reduce the variance in the population, with variance being reduced faster for larger numbers of cultural parents. The counteracting force, dependent on the average error introduced during transmission ($\overline{E}$), tends to increase the variance in the population, with errors of larger variance increasing the variance in the population more.

A.1.2 Pressures acting on cultural transmission

B&R provide mathematical accounts of how three pressure acting on transmission can result in cultural change and cultural evolution. These are:

1. Natural selection of cultural variants, resulting from selective removal of enculturated individuals.
2. Guided variation, resulting from individual learning by enculturated individuals.
3. Biased transmission, resulting from the strategy of learners during cultural transmission. The forces of biased transmission can be further subdivided into three forms:
   (a) Directly biased transmission, resulting from a preference for learners to acquire one cultural variant over another.
   (b) Indirectly biased transmission, resulting from a preference for learners to acquire cultural traits which are associated with other cultural traits.
   (c) Frequency-dependent transmission, resulting from a disproportionate preference for learners to acquire the most (or least) frequent cultural trait in the population.

In Sections A.1.2.1 to A.1.2.5 B&R's models for these pressures are reviewed. In the interests of clarity, a separate section is devoted to each of the three subtypes of biased transmission.

A.1.2.1 Natural selection of cultural variants

B&R model the natural selection of cultural variants by assuming that there are a set of n distinct social roles (e.g. mother, father, uncle, priest, teacher). Each naive individual acquires their cultural characteristic based on observation of a subset of these roles τ_j.

As before, the weight of social role k is A_k. The weight of social role k with respect to a subset of social roles τ_j, A_{kj} is:

$$A_{kj} = \begin{cases} A_k & \text{if } k \text{ belongs to } \tau_j \\ 0 & \text{otherwise} \end{cases}$$

Working within the dichotomous traits model, the probability that a naive individual acquires variant c based on the behaviour of the set of individuals with phenotypic values $X_1, \ldots, X_n$ and the set of cultural parents with roles τ_j is therefore:

$$Prob(individual = c|\tau_j, X_1, \ldots, X_n) = \frac{\sum_{k=1}^{k=n} A_{kj} X_k}{\sum_{k=1}^{k=n} A_{kj}}$$

This equation normalises the weight of the cultural parent with role k by the weights of all roles present in the set of roles τ_j and is clearly related to the earlier equation for the cultural transmission of dichotomous traits.

In order to model natural selection we must assume that the probability that an individual attains a particular social role k depends on the cultural variant that that individual possesses. Let Ω_{ck} be the probability that an individual with cultural variant c attains social role k and, similarly, Ω_{dk} be the probability that an individual with cultural variant d attains social role k. As before, we will assume that the frequency of variant c in the population is p. The frequency of individuals with variant c attaining social role k, p'_k is therefore:

$$p'_k = \frac{\Omega_{ck} p}{\Omega_{ck} p + \Omega_{dk}(1-p)}$$

Working under the assumption that sets of roles τ_j are formed at random, and following a similar procedure to that outlined for the basic cultural transmission rule, B&R show that the frequency of cultural variant c in the population after differential attainment of social roles and cultural transmission by the linear rule, p'', is:

$$p'' = \sum_{k=1}^{n} \overline{A}_k p'_k$$

where $\overline{A}_k$ gives the importance of parents in the kth social role averaged over all possible sets of cultural parents τ_j according to the frequency with which those sets occur:

$$\overline{A}_k = \sum_j Prob(\tau_j) \left(\frac{A_{kj}}{\sum_l A_{lj}} \right)$$

This equation can be combine with the equation for p'_k given above. If σ_k is the selective advantage of variant c with respect to role k ($\sigma_k = (\Omega_{ck}/\Omega_{dk}) - 1$) and assuming that selection is weak, the equation becomes:

$$p'' = p + p(1-p) \left(\sum_{k=1}^{n} \overline{A}_k \sigma_k \right)$$

where the sum is the selection advantage of variant c in role k averaged over all social roles and weighted by the importance of each role ($\overline{A}_k$). Roles which offer a high selective advantage (σ_k) will have a strong influence, even if that social role is not weighted particularly highly in contribution to cultural transmission (i.e. $\overline{A}_k$ is not particularly high relative to $\overline{A}_{l \neq k}$). Variant c will increase if this quantity is positive and decrease if it is negative — if variant c offers a selective advantage when averaged over social roles then it will increase in frequency in the population. In other words, if possessing variant c makes an individual more likely to occupy a role which allows them to enculturate others and transmit that variant, then c will increase in frequency in the population.

A.1.2.2 Guided variation

A model of guided variation requires a model of individual learning. B&R assume that an individual can be characterised by a number X, the initial value of their phenotype prior to individual learning, and a number Y, the value of their phenotype after individual learning. This is therefore a continuous trait model of cultural characteristics. The goal of learning is determined by the environment, which is characterised by a number H. The aim of learning is essentially to move Y towards H. L is a parameter determining the reliance of an individual on individual learning, with high L indicating a high reliance on individual learning. Errors made during the learning process are represented by a normally distributed random variable ϵ with mean 0 and variance V_e. It can be shown that:

$$Y = aX + (1-a)(H + \epsilon)$$

where $a = V_e/(V_e + L)$ is a parameter that gives the importance of individual learning — $a \approx 1$ ($L \ll V_e$) corresponding to a tendency to rely on the initial value of the phenotype X and $a \approx 0$ ($L \gg V_e$) corresponding to a tendency to ignore the initial value of the phenotype and move towards the value preferred by the environment, H.

How does this type of individual learning change the mean value and the variance of a population's cultural characteristic? Prior to individual learning the mean value of the trait in the population is given by $\overline{X}$ and the variance is given by V. The mean value after individual learning, $\overline{Y}$ is:

$$\overline{Y} = a\overline{X} + (1 - a)H$$

where a is as before. As in the individual case, in the population case the mean value for the trait will tend towards the value favoured by the environment if $a < 0.5$. The variance of the population after individual learning, U, is:

$$U = a^2V + (1 - a)^2V_e$$

Individual learning both decreases the variance of the trait in the population through movement towards the environmentally-determined goal ($U = a^2V \ldots$) and increases it due to errors introduced by individual learning ($U = \ldots (1 - a)^2V_e$).

If we assume that the culturally-acquired value for the phenotype forms the initial value of the phenotype which can subsequently be altered by individual learning, this model of individual learning can be simply added to the blending model of cultural transmission. Assuming non-selective formation of sets of models, equal weighting for all models ($A_i = A_j = 1/n$) and no correlation between errors in a given set of models ($Cov(e_i, e_j) = 0$) the mean value of X in the next generation, $\overline{X}'$, is:

$$\overline{X}' = a\overline{X} + (1 - a)H$$

i.e. when individual learning is powerful ($a \approx 0$) the population moves towards the value of the phenotype favoured by the environment, due to the transmission of cultural traits favoured by individual learning, and when individual learning is weak ($a \approx 1$) the mean value of the population's cultural trait remains unchanged by individual learning. The variance after transmission, V', is:

$$V' = (1/n)(a^2V + (1-a)^2V_e + \overline{E})$$

i.e. blending both reduces ($(1/n)$) and increases ($\ldots + \overline{E}$) variance, and individual learning both reduces (a^2V) and increases ($(1-a)^2V_e$) variance.

A.1.2.3 Directly-biased transmission

Direct bias can be simply modelled using the model of the transmission of dichotomous characters given earlier. As before, the probability that an individual acquires cultural variant c given the set of cultural parents $X_1, \ldots, X_n$ is:

$$Prob(c|X_1, \ldots, X_n) = \sum_{i=1}^{n} A_i X_i$$

In the unbiased case, the value of a particular A_i is independent of X_i — the cultural variant used by a model does not affect the importance of that model to the naive individual. However, in the biased case, A_i depends on the intrinsic importance of the ith model, given by α_i, and the biasing function, $\beta(X_i)$:

$$A_i = \frac{\alpha_i(1 + \beta(X_i))}{\sum_{j=1}^{n} \alpha_j(1 + \beta(X_j))}$$

where the biasing function is:

$$\beta(X_i) = \begin{cases} B & \text{if } X_i = 1 \\ -B & \text{if } X_i = 0 \end{cases}$$

B gives the strength of the bias in favour of cultural variant c. Assuming $B > 0$ (variant c is favoured over variant d), if the ith model has cultural variant c then the intrinsic weight of that model will be increased by a factor $1 + B$, whereas if the ith model has variant d then the intrinsic weight of that model will be decreased by $1 - B$. Note that this model of a biasing function is in principle arbitrary with respect to the functionality of the cultural trait, with an arbitrary preference in favour of one variant over the other determined by B.

However, B could be linked to the expected fitness payoff of the variants, in which case the bias would be non-arbitrary and in favour of the cultural variant which is expected to yield the greatest fitness payoff.

B&R consider the case where each individual is exposed to two models, with intrinsic weights α_1 and α_2. This can be interpreted as either the case where each individual has two cultural parents, or the case where each individual has multiple cultural parents but is enculturated in a serial fashion, observing X_2 for each parent in turn, comparing it to their own value, (X_1) and deciding on which of the two possibilities (X_1 or X_2) to adopt. If p is the frequency of variant c in the population prior to such an episode of cultural transmission, its frequency after cultural transmission, p', will be:

$$p' = p + p(1-p)\left(\frac{4B\alpha_1\alpha_2}{1 - B^2\left(\alpha_1 - \alpha_2\right)^2}\right)$$

Assuming that both cultural parents have equal weight ($\alpha_1 = \alpha_2 = 0.5$), this reduces to:

$$p' = p + p(1-p)B$$

In other words, directly biased transmission will increase the frequency of the favoured variant in the population. The rate of increase depends on the strength of the bias (B) and the variance in the population ($p(1-p)$).

A.1.2.4 Indirectly-biased transmission

A model of indirect bias requires a model of the transmission of multiple cultural traits. B&R develop a model of the blending transmission of two quantitative cultural traits which is based on the basic transmission model for single continuous traits (outlined in Section A.1.1.2). The jth individual is characterised by a two-place vector $X_j = (X_{1j}, X_{2j})$. As before, a naive individual observing individual j forms an estimate of that individual's cultural variants, $Z_j = (Z_{1j}, Z_{2j})$ such that:

$$Z_{1j} = X_{1j} + e_{1j}$$

$$Z_{2j} = X_{2j} + e_{2j}$$

As before, e_{1j} and e_{2j} are random variables drawn from normal distributions with mean 0, variances E_{1j} and E_{2j} respectively and covariance E_{12j}.

As for the earlier definition of the blending rule, naive individuals observe and estimate the cultural character of n models and then form their own cultural character, $X_0 = (X_{10}, X_{20})$, by averaging over observed models:

$$X_{10} = \sum_{j=1}^{n} A_{1j} Z_{1j}$$

$$X_{20} = \sum_{j=1}^{n} A_{2j} Z_{2j}$$

where A_{ij} is the importance of the jth model in transmitting cultural characteristic i ($i =$ 1 or 2). As with the single-trait blending model, we need to know how this type of transmission will affect the mean value of trait i in the population, $\overline{X}_i$, and the variance of trait i, V_i. By similar methods, it can be shown that the mean and variance after transmission ($\overline{X}'_i$ and V'_i) are given by:

$$\overline{X}'_i = \overline{X}_i$$

$$V'_i = \sum_{j=1}^{n} A_{ij}^2 \left(V_i + \overline{E}_i\right)$$

where $\overline{E}_i$ is the weighted average of errors introduced during transmission:

$$\overline{E}_i = \frac{\sum_{j=1}^{n} A_{ij}^2 E_{ij}}{\sum_{j=1}^{n} A_{ij}^2}$$

If we assume that all cultural parents for a given trait have equal weight (i.e. $A_{ij} = A_{ik} = 1/n$) then this reduces to:

$$V'_i = (1/n)\left(V_i + \overline{E}_i\right)$$

i.e. as before, blending inheritance leaves the mean in the population unchanged and both decreases and increases variance, depending on the number of cultural parents and variance of the errors introduced.

Cultural transmission will also affect the covariance between the values of traits 1 and 2, C_{12}. For the simplified case where each cultural parent has equals weight ($A_{ij} = A_{ik} = 1/n$, which implies $A_{1j} = A_{2j}$ i.e the models are equally important in transmitting both traits, rather than some models being important in the transmission of one trait and other models being important in the transmission of other traits), the covariance of the values after transmission, C'_{12}, is give by:

$$C'_{12} = (1/n)\left(C_{12} + \frac{\sum_{j=1}^{n} E_{12j}}{n}\right)$$

i.e. as with variance, co-variance is reduced by the factor $(1/n)$ and increased by correlated errors, measured by the degree of correlation between errors averaged over all models ($\sum_{j=1}^{n} E_{12j}/n$). For more complex cases where the different traits are influenced by different sets of models (i.e. $A_{1j} \neq A_{2j}$) the covariance between the traits tends to decrease.

Given this blending model of the transmission of two quantitative characters, it is possible to model indirect bias. We will consider trait 1 to be the indicator trait and trait 2 to be the indirectly biased trait, so that an individual can be characterised by a two-place vector $X = (X_I, X_D)$ (I for Indicator trait, D for derived trait). As described above, individuals acquire their trait based on the weighted average of their estimate of the variants of their cultural parents i.e. $X_{i0} = \sum_{j=1}^{n} A_{ij} Z_{ij}$ where $i = I$ or D.

As discussed above, the indicator trait is a directly-biased trait — some values for the indicator trait are intrinsically preferred. A_{Ij} is therefore a function of the intrinsic influence of parent j with respect to trait I, α_{Ij}, and the estimated value of model j's trait I, Z_{Ij}:

$$A_{Ij} = \frac{\alpha_{Ij}\left(1 + \beta\left(Z_{Ij}\right)\right)}{\sum_{k=1}^{n} \alpha_{Ik}\left(1 + \beta\left(Z_{Ik}\right)\right)}$$

where $\beta(x)$ is a direct bias function. This equation should be familiar due to its similarity to the equation from the direct bias section.

The importance of the jth cultural parent with respect to the indirectly biased trait, A_{Dj}, will be a function of that parent's intrinsic importance, α_{Dj}, and the estimate of the jth model's value for the *indicator* trait, Z_{Ij} (rather than the estimate of the jth model's value for the indirectly biased trait Z_{Dj}):

$$A_{Dj} = \frac{\alpha_{Dj}\left(1+\theta\left(Z_{Ij}\right)\right)}{\sum_{k=1}^{n}\alpha_{Dk}\left(1+\theta\left(Z_{Ik}\right)\right)}$$

where $\theta\left(x\right)$ is the indirect bias function, of a similar form to the direct bias function.

Assuming the non-selective formation of sets of models, weak biasing functions and equal intrinsic weightings to all cultural parents ($\alpha_{ij} = \alpha_{ik} = 1/n$), the mean values of the traits after transmission, $\overline{X}'_i$ can be calculated given the mean values of the traits prior to transmission, $\overline{X}_i$:

$$\overline{X}'_I = \overline{X}_I + (1/n)\,Cov\left(Z_I, \beta\left(Z_I\right)\right)$$

$$\overline{X}'_D = \overline{X}_D + (1/n)\,Cov\left(Z_D, \theta\left(Z_I\right)\right)$$

$Cov\left(Z_i, f\left(Z_j\right)\right)$ is the covariance of the trait Z_i and the bias function f applied to some trait Z_j. If increases in Z_i tend to result in increases in $f\left(Z_j\right)$ then $Cov\left(Z_i, f\left(Z_j\right)\right) > 0$. On the other hand, if increases in Z_i tend to result in *decreases* in $f\left(Z_j\right)$ then $Cov\left(Z_i, f\left(Z_j\right)\right) < 0$. $Cov\left(Z_I, \beta\left(Z_I\right)\right)$ therefore gives the strength and direction of the direct bias — if $Cov\left(Z_I, \beta\left(Z_I\right)\right) < 0$ then the mean value of X_I must be above the value favoured by the direct bias and the mean value will decrease through transmission by an amount proportional to the magnitude of $Cov\left(Z_I, \beta\left(Z_I\right)\right)$. Similarly, if $Cov\left(Z_I, \beta\left(Z_I\right)\right) > 0$ then the mean value of X_I will increase.

$Cov\left(Z_D, \theta\left(Z_I\right)\right)$ gives the strength and direction of the indirect bias, and depends on whether values of Z_D and Z_I are correlated. Consider the case where Z_D and Z_I are positively correlated. Higher values of Z_D will be associated with higher values of Z_I. If the current value of Z_I associated with the current value of Z_D is below the optimum value given by $\theta\left(Z_I\right)$ then increases in Z_D will result in increases in $\theta\left(Z_I\right)$ and therefore $Cov\left(Z_D, \theta\left(Z_I\right)\right) > 0$. Similarly, if the current value of Z_I associated with the current value of Z_D is above the optimal value then increases in Z_D will result in decreases in $\theta\left(Z_I\right)$ and therefore $Cov\left(Z_D, \theta\left(Z_I\right)\right) < 0$. In either case, the mean of the population's value for X_D will tend towards the value associated with the value of X_I which maximises the indirect bias function θ — "variants of the indirectly biased trait that are positively correlated with the admired variants of the indicator trait will increase in frequency" (B&R p254). Similarly, variants of the indirectly biased trait which are negatively correlated with the admired variants of the indicator trait will decrease in frequency.

A.1.2.5 Frequency-dependent bias

In Section A.1.1.1 a model was described which gave the probability of acquiring cultural variant c on the basis of n models for the unbiased dichotomous case. The frequency-dependent bias case is very similar:

$$Prob(individual = c|X_1, \dots, X_n) = \sum_{i=1}^{n} A_i X_i + D\left(\sum_{i=1}^{n} A_i X_i\right)$$

Assuming that each model has equal importance this becomes

$$Prob(individual = c|j) = j/n + D\left(j\right)$$

where j is the number of parents with cultural variant c:

$$j = \sum_{i=1}^{n} X_i$$

and $D\left(j\right)$ is the frequency-dependent bias function. When $D\left(j\right) = 0$ for all j there is no frequency-dependent bias and the model reduces to the unbiased case. If $D\left(j\right) > 0$ for $j > n/2$ and $D\left(j\right) < 0$ for $j < n/2$ then transmission is biased in favour of conformity — the probability of acquiring the majority trait ($j > n/2$ indicates that the trait is possessed by more than half the set of n models) is increased by a factor $D\left(j\right)$, and the probability of acquiring the minority trait ($j < n/2$) is decreased by the factor $D\left(j\right)$. Conversely, if $D\left(j\right) < 0$ for $j > n/2$ and $D\left(j\right) > 0$ for $j < n/2$ then transmission is biased in favour of non-conformity — the probability of acquiring the majority variant is decreased and the probability of acquiring the minority variant is increased.

Assuming non-selective formation of sets of parents and some value k such that $k > n/2$ and k is minimised (i.e. the lowest value of k such that k represents more than half the number of models n), it can be shown that the proportion of individuals with variant c, p', after cultural transmission is:

$$p' = p + \sum_{j=k}^{n} D\left(j\right) \binom{n}{j} \left[p^{j}\left(1-p\right)^{n-j} - p^{n-j}\left(1-p\right)^{j}\right]$$

This rather complex equation deserves some explanation. There are $\binom{n}{j}$ ways to pick j individuals from a population of size n. The probability that one of these will exhibit exactly j individuals with variant c and $n-j$ individuals with variant d is $p^{j}\left(1-p\right)^{n-j}$ and the probability that one of these will exhibit exactly j individuals with variant d and $n-j$ individuals with variant c is $p^{n-j}\left(1-p\right)^{j}$. There are therefore $\sum_{j=k}^{n}\binom{n}{j}\left(p^{j}\left(1-p\right)^{n-j}\right)$ ways to pick sets of models from the population such that more than half of the models have variant c, and $\sum_{j=k}^{n}\binom{n}{j}\left(p^{n-j}\left(1-p\right)^{j}\right)$ ways to pick sets of models such that more than half have variant d. The proportion of c in the population therefore increases according to the frequency-dependent bias function $D\left(j\right)$ applied to the difference between the probability of picking sets of models such that the majority are of type c and the probability of picking sets of models such that the majority are of type d.

In the case of conformist transmission, $D\left(j\right) > 0$ for $j > k$ and $D\left(j\right) < 0$ for $j < k$. Therefore, the frequency of variant c, p, will increase whenever $p > 0.5$ (if c is the more frequent variant in the population then it will increase in frequency) and decrease when $p < 0.5$ (if c is the less frequent variant in the population then it will decrease in frequency). The rate of change of p is at its lowest as p approaches 1 or 0 (the two saturation points) or 0.5 (the point where the population is perfectly split between the two variants). Conformist transmission results in the spread of the most common cultural variant.

A.2 Genetic transmission and natural selection

The simplest models of natural selection acting on genetic transmission[1] deal with the changes in frequency of alleles of a single gene in asexually-reproducing haploid populations — each individual has a single gene drawn from a set of n alleles and each individual inherits the allele of their single parent. In sexually-reproducing diploid populations the equations are complicated by the fact that each individual has two alleles for each gene and receives one allele from each of their two parents.

[1] The mathematical model given here is based on the models given in B&R, Hartl & Clark (1997) and Futuyma (1998).

Ontogeny is typically treated in a very simplistic manner in mathematical models of population genetics. In the haploid organism, single gene case there are n distinct alleles and therefore n distinct genotypes $G_1 \ldots G_n$. It is typically assumed that there are n distinct phenotypes $F_1 \ldots F_n$ and ontogeny maps genotype G_i onto phenotype F_i. Selection then acts on the phenotype, but since there is a one-to-one correspondence between genotypes and phenotypes we can talk of selection acting on genotypes and effectively ignore ontogeny.

Suppose that each individual with phenotype F_i survives with probability s_i. If N_{G_i} is the number of individuals with genotype G_i (and therefore phenotype F_i) prior to selection then the number of individuals with genotype G_i after selection, N'_{G_i} is:

$$N'_{G_i} = s_i N_{G_i}$$

If we assume that every surviving individual with genotype G_i leaves, on average, o_i offspring then the number of individuals with genotype G_i the next generation, N''_{G_i}, is given by:

$$N''_{G_i} = o_i N'_{G_i} = s_i o_i N_{G_i} = f_i N_{G_i}$$

where f_i gives the overall fitness of genotype G_i, the probability that individuals with that genotype will survive to reproductive age multiplied by the average number of offspring produced.

Now consider a population with two genotypes G_a and G_b with fitness f_a and f_b respectively. Evolution by natural selection takes place in such a population where the two genotypes do not reproduce at equal rates — $f_a \neq f_b$. Typically we are not interested in the absolute numbers of the two genotypes, but the proportion of the population with genotype G_a and the proportion of the population with genotype G_b. We will define these as $P_{G_a} = \frac{N_{G_a}}{N}$ and $P_{G_b} = \frac{N_{G_b}}{N}$, where N is the overall population size ($N = N_{G_a} + N_{G_b}$). We can then calculate the proportion of genotype G_a at the next generation, which I will denote by P'_{G_a}:

$$P'_{G_a} = \frac{f_a N_{G_a}}{f_a N_{G_a} + f_b N_{G_b}} = \frac{f_a P_{G_a} N}{f_a P_{G_a} N + f_b P_{G_b} N} = \frac{f_a P_{G_a}}{f_a P_{G_a} + f_b P_{G_b}}$$

The proportion of genotype G_a at the next generation therefore depends on the proportions of genotypes G_a and G_b and their fitness. We can calculate how P_{G_a} will change over time:

$$\Delta P_{G_a} = \frac{f_a P_{G_a}}{f_a P_{G_a} + f_b P_{G_b}} - P_{G_a} = \frac{P_{G_a} P_{G_b} (f_a - f_b)}{f_a P_{G_a} + f_b P_{G_b}}$$

If $f_a > f_b$ then genotype G_a will increase in frequency, and if $f_a < f_b$ then it will decrease in frequency. The rate of change is at a maximum when $P_{G_a} P_{G_b}$ is at a maximum, which occurs when $P_{G_a} = P_{G_b} = 0.5$ — in other words, natural selection depends on genetic diversity, and the rate of evolution is higher when the population exhibits more diversity.

A.3 Dual transmission and direct bias

Within the dual transmission model, B&R consider the circumstances under which a biological capacity for individual learning and biased and unbiased cultural transmission will be favoured by natural selection. For the purpose of this thesis it is sufficient to review their model of the genetic evolution of direct bias. Recall from Section A.1.2.3 above that direct bias on cultural transmission will increase the frequency of the favoured variant in a population, with the rate of increase depending on the strength of the direct bias, given by the biasing function β, and the cultural variance in the population. B&R expand this model, following their general technique outlined above, to consider the case where β is determined genetically — an individual's genotype determines their preference for cultural variants.

B&R assume that there are two cultural variants, c and d, and two genetic variants in the population, e and f. Genotypes e and f define biasing functions β_e and β_f such that:

$$\begin{aligned} \beta_e(X_i) &= 0 \\ \beta_f(X_i) &= \begin{cases} B & \text{if } X_i = 1 \\ -B & \text{if } X_i = 0 \end{cases} \end{aligned}$$

Recall that $X_i = 1$ if the learner's ith cultural parent possesses variant c, and $X_i = 0$ otherwise. e is therefore the unbiased allele and f is the biased allele, where the bias is in favour of cultural variant c if $B > 0$. We can now calculate the probability that an individual with genotype G acquires cultural variant c given that it is exposed to cultural parents with the cultural variants $X_1, \ldots X_n$. This is given by:

$$Prob(c|X_1, \ldots, X_n, G) = \frac{\sum_{i=1}^{n} \alpha_i X_i \left(1 + \beta_G \left(X_i\right)\right)}{\sum_{i=1}^{n} \alpha_i \left(1 + \beta_G \left(X_i\right)\right)}$$

As before, α_i gives the intrinsic importance of the ith cultural parent. This is essentially identical to the equation for directly biased transmission given in Section A.1.2.3, with the addition of a specified genotype G which gives the particular biasing function β_G to be used.

B&R first assume that genetic parents are selected at random from the pool of possible parents, where the frequency of genotype G in that pool is q_G. The frequency of genotype G among offspring, q'_G, therefore remains unchanged — there is no natural selection acting on genetic transmission. p gives the frequency of cultural variant c in the parent population. Assuming that individuals have just two, equally-weighted, cultural parents, the frequency of individuals with genotype G and cultural variant X after cultural transmission, F'_{GX}, is therefore given by:

$$\begin{aligned}
F'_{ec} &= \quad q_e p \\
F'_{fc} &= \quad q_f \left(p + p(1-p)B\right) \\
F'_{ed} &= \quad q_e \left(1 - p\right) \\
F'_{fd} &= \quad q_f \left(1 - p - p(1-p)B\right)
\end{aligned}$$

As we would expect, individuals with the unbiased allele e have the same frequency of the two cultural variants as was present in the parent population — individuals with allele e and variant c occur with frequency given by the product of the frequency of genotype e and cultural variant c (p), and individuals with allele e and cultural variant d occur with frequency given by the product of the frequency of genotype e and cultural variant d ($1 - p$). Among individuals with the biased allele f cultural variant c increases in frequency according to the strength of the bias and the cultural variance in the parent population, and variant d decreases by a similar factor.

B&R then go on to add natural selection to the model. Natural selection weeds individuals out after cultural transmission and prior to breeding, with the probability that an individual with genotype G and cultural variant X survives to breeding age being given by W_{GX}. W_{GX} depends on the selective advantage of cultural variant c, s, and the cost of biased transmission, z:

$$\begin{aligned} W_{ec} &= 1+s \\ W_{fc} &= 1+s-z \\ W_{ed} &= 1 \\ W_{fd} &= 1-z \end{aligned}$$

Individuals with cultural variant c gain the fitness payoff s. Individuals with the biased genotype f suffer the cost of that bias, z. We can now calculate the expected frequency of individuals with genotype G and cultural variant X after selection, F''_{GX}, according to the equations given above for dealing with natural selection:

$$F''_{GX} = \frac{W_{GX}F'_{GX}}{W_{ec}F'_{ec} + W_{fc}F'_{fc} + W_{ed}F'_{ed} + W_{fd}F'_{fd}}$$

B&R then go on to make several simplifying assumptions. They assume that cultural variant c is always favoured by selection ($s > 0$) and that bias has no cost or a positive cost ($z \geq 0$) but that these factors are weak ($z, s \ll 1$). Given these assumptions, B&R work through a rather complex set of equations, keeping track of q_f, the frequency of the biased genotype (henceforth q) and p, the frequency of cultural variant c. q'' and p'' give the frequencies of these two characters in the next generation.

Assume for a moment that p, the proportion of individuals with cultural variant c, is fixed at some arbitrary value. What happens to q, the frequency of individuals with the biased genotype?

$$q'' = q + vq(1-q)$$

where v gives the “selection differential” of the biased allele and is given by:

$$v = B\left(sp\left(1-p\right)\right) - z$$

If v is positive the biased allele will increase in frequency. First consider the case where $z = 0$ — the biased genotype has no associated cost. If $p = 0$ or $p = 1$ then $v = 0$ and the biased allele does not change frequency — if the population exhibits no cultural variation then the biased allele has no fitness advantage over the unbiased allele and does not change in frequency. If the population exhibits cultural variation then $v > 0$ and the biased allele will increase in frequency. Now consider the case where $z > 0$ — the biased genotype has a cost. If the population exhibits cultural variance ($0 < p < 1$) then the sign of v will depend on the relative values of B, s, p and z. If the population exhibits no cultural variation ($p = 0$ or $p = 1$) then v will be negative and the biased genotype will decrease in frequency – the biased allele will suffer a fitness penalty due to its cost and no fitness benefit over the unbiased allele due to the lack of cultural variation. To summarise, in a population which is completely converged culturally (on either variant) the frequency of the biased variant should either remain constant (if biased learning is costless relative to unbiased learning), or decrease (if biased learning has a cost).

What can we predict about the frequency of cultural variant c, given by p? Variant c is always favoured by selection, and by biased transmission when $q > 0$. Therefore variant c will increase in frequency until the population reaches equilibrium at $p = 1$. As discussed in the previous paragraph, at this equilibrium state the biased genotype either has no advantage over the unbiased genotype or is at a disadvantage (where $z > 0$). Therefore, at equilibrium we should expect selection to either be neutral with respect to bias, or to see only the unbiased allele — directly biased transmission pushes the population to converge on the favoured cultural variant, at which point selection pressure on the population's genotypes either stops, or acts to reduces the frequency of the biased allele which drove cultural convergence in the first place.

Appendix B

Published papers

This appendix contains two journal articles accepted for publication prior to the completion of this thesis (Smith (2002) and Smith (in press)), and one further journal article which is currently under review (Smith *et al.* submitted). I also include two papers which have or will appear in edited collections (Smith (2001b) and Smith *et al.* (forthcoming)).

SMITH, K. 2002. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.

SMITH, K., in press. Natural selection and cultural selection in the evolution of communication. To appear in *Adaptive Behavior*.

Natural selection and cultural selection in the evolution of communication

Kenny Smith
Language Evolution and Computation Research Unit
Department of Theoretical and Applied Linguistics
University of Edinburgh
Adam Ferguson Building
40 George Square
Edinburgh, UK

kenny@ling.ed.ac.uk
Phone: +44 131 650 6658
Fax: +44 131 650 3962

To appear in *Adaptive Behavior*

Abstract

It has been postulated that aspects of human language are both genetically and culturally transmitted. How might these processes interact to determine the structure of language? An agent-based model designed to study gene-culture interactions in the evolution of communication is introduced. This model shows that cultural selection resulting from learner biases can be crucial in determining the structure of communication systems transmitted through both genetic and cultural processes. Furthermore, the learning bias which leads to the emergence of optimal communication in the model resembles the learning bias brought to the task of communication by human infants. This suggests that the iterated application of such human learning biases may explain much of the structure of human language.

Key words: communication, language, evolution, culture, learning

1 Introduction

Language is transmitted from generation to generation within a speech community. The precise nature of the intergenerational transmission remains a contentious issue. The transmission of language down the generations involves at least some cultural transmission – under normal circumstances children learn the language of their speech community through exposure to the linguistic behavior of that community. The most influential linguistic theories of modern times assume genetic transmission of the language faculty between generations in addition to this cultural transmission – language learners come to the language acquisition task equipped with some genetically-encoded Language Acquisition Device (Chomsky, 1987). The research outlined in this paper represents an attempt to understand the types of interactions which may occur between cultural transmission and genetic transmission of communication systems within a communicating population. Specifically, in this paper I argue that cultural selection resulting from learning biases is key in determining the structure of communication systems, such as language, which are both genetically and culturally transmitted.

In Section 2 the literature on the evolution of communication and language is reviewed. This review reveals a paucity of models on gene-culture interactions which are simple enough to be easily understood yet detailed enough to test hypotheses regarding the evolution of communication or language in the real world. In Section 3 such a model is proposed. Sections 4 and 5 present results generated by this model. These results suggest that cultural selection resulting from the biases inherent in the model of the learner is crucial in determining the structure of the emergent communication systems, and that natural selection is unable to override these biases. In Section 6 the learning biases of the model are examined in detail and related both to other agent-based models of the evolution of communication and to the communication-specific learning biases observed in humans. Finally, in the concluding section it is suggested that much of the structure of language may be best explained in terms of cultural evolution resulting from a pre-adapted learning mechanism.

2 The literature

The literature on the evolution of communication and language can be roughly divided into two main areas – that which addresses the question "when should we expect to see communication or language?" and that which addresses the question “what structure should we expect communication or language to exhibit?”.

The first question has typically been addressed by theoretical biologists but has recently been tackled by researchers using agent-based modeling techniques. Researchers in this area are concerned with the interlocking issues of signal costs and honesty (for example, from biology Zahavi (1975), Zahavi (1977), Krebs and Dawkins (1984) and Grafen (1990), from agent-based modeling Wheeler and de Bourcier (1995), Bullock (1997), Noble (1998)) and altruism (from agent-based modeling Ackley and Littman (1994), Oliphant (1996) and Reggia, Schulz, Wilkinson, and Uriagereka (2001)). These important issues will not concern us in this paper, other than to say that the presence of honest communication has been identified as a possibility in at least some circumstances by some researchers.

The second question can be viewed as consisting of three subquestions: “what structure would we expect communication or language to exhibit if it were shaped by purely biological processes?”; “what structure would we expect communication or language to exhibit if it were shaped by purely cultural processes?”; “what structure would we expect communication or language to exhibit if it were shaped by both biological and cultural processes?”.

There has been some research on the first question, both by biologists (see Hauser (1996) for a summary) and agent-based modelers. For example, Werner and Dyer (1992), Levin (1995) and Di Paolo (1997) show that agents can coordinate their actions or internal states optimally or near-optimally using innate communication systems given selection pressure for that coordination. Werner and Todd (1997) show that the reverse can also be true – agents can violate the innate expectations of receivers given innate signalling behavior and selection pressure for such violation. Cangelosi and Parisi (1998) show that an efficient biologically-transmitted communication system can emerge even without direct selection pressure, effectively due to evolution of internal representations and genetic drift of a communication system on top of this evolved substrate. However, human language, our ultimate object of study, consists of at least some learned component and these models are therefore of limited utility in understanding it.

The second question, concerning the structure of communication given purely cultural transmission processes, has received a considerable amount of attention in recent years. It appears that such processes may only be relevant to the study of communication in humans, given that Hauser states that “although call structure [in non-human primates] changes ontogenetically, no study has provided convincing evidence that acoustic experience is causally related to such changes" (Hauser, 1996, p315). Consequently, much of the research into this area has been carried out by linguists or cognitive scientists using agent-based modeling techniques to explain the cultural evolution of features of human language such as syntax (e.g. Batali (1998), Batali (in press), Hurford (2000), Kirby (2000), Kirby (in press), Brighton and Kirby (2001), see Hurford (2002) and Kirby and Hurford (2002) for an overview), regularity and irregularity (e.g. Kirby (2001), Worden (in press)) and other language universals (e.g. Christiansen and Devlin (1997), Kirby (1998)). Less work has been carried out on the more basic issue of the cultural evolution of non-syntactic communication systems. However, agent-based models directed at this issue can be found in Hutchins and Hazelhurst (1995), Oliphant and Batali (1997), Oliphant (1999) and Steels (1999).

This substantial body of literature presents a persuasive argument that the features of communication and language can be explained in terms of cultural processes. However, this work does have its weaknesses. Typically each paper considers a single model of learning. This lack of comparison between learning mechanisms makes it difficult to identify the biases of the chosen model of learning. Secondly, these models assume a degree of pre-existing mental apparatus, including a learning mechanism. This mental apparatus presumably evolved, although not necessarily for its later role in language-processing. But how might the evolution of learning mechanisms interact with the resulting cultural processes? Such models are not designed to address this question.

Finally, there is a small body of literature investigating the question of the structure of communication systems emerging through a mixture of genetic and cultural processes. Pinker and Bloom (1990) and Dor and Jablonka (2000) introduce hypothetical scenarios under which positive interactions between natural selection and cultural transmission lead to language. However, human intuitions regarding the behavior of such complex adaptive processes are notoriously poor and a formal model is desirable.

In an early paper MacLennan and Burghardt (1994) consider how reinforcement learning might interact with natural selection in the evolution of vocabulary-like systems. However, given that "a series of studies beginning with Brown and Hanlon (1970) have demonstrated that there is little reliable correlation between the grammaticality of children's utterances and the sorts of responses to these that their parents give" (Bloom and Gleitman, 2001), the relevance of models involving reinforcement learning to our understanding of language, our ultimate object of study, is doubtful. Batali (1994) considers interactions between selection and learning in populations of neural networks. The languages Batali's networks attempt to learn are externally determined, rather than emerging from the populations of agents themselves, and are therefore not truly culturally transmitted. This makes the relevance of this model to the field of human language less clear. Cangelosi (1999) uses neural networks to investigate gene-culture interactions in the evolution of symbolic communication systems. However, as in an earlier model (Cangelosi and Parisi, 1998), the structure of the communication system is determined by genetic factors with learning playing little role. Kirby and Hurford (1997), Turkel (in press) and Yamauchi (2001) consider possible interactions between natural selection and learning in the evolution of an innate language acquisition device and a language. However, their representations of language and learning are so abstract as to make any claims about the structure of human language difficult. Finally, in a recent paper Kvasnička and Pospíchal (1999) model interactions between natural selection and learning of culturally-emergent communication systems in a population of neural networks. This model is a step in the right direction, detailed enough to allow hypothesis to be formed about the structure of communication systems in the real world, abstract enough to be analyzable. However, the model suffers from two defects. Firstly, only one learning mechanism is considered. Secondly, only one level of selection pressure is considered. This means that the relationship between the learning bias of the learning mechanism and the forces of natural selection remain unclear.

This paper presents a model of the interactions between processes of biological transmission and cultural transmission in the evolution of simple communication. The model avoids defects of earlier models in investigating, in detail, the relationship between different selection pressures, different learning biases and different strengths of learning bias. This allows us to address the hypothesis, suggested by previous models, that natural selection and learning will interact positively to create optimal communication systems. The ability to manipulate the various pressures proves to be essential in understanding the key determinants of the behavior of the system. The model is simple enough to be analyzed but detailed enough to provide a starting point in understanding how these issues might apply to the evolution of communication and, ultimately, language, in the real world.

3 The model

The model consists of a simple model of communication (Section 3.1), a model of a communicative agent (Section 3.2) and a model of genetic and cultural transmission (Section 3.3).

3.1 The communication system

For the purposes of this model communication systems are mappings between a set of unstructured meanings m and a set of unstructured signals s. A communication system consists of a production function, $p(m)$, mapping from m to s, and a reception function, $r(s)$, mapping from s to m.

3.1.1 Communicative Accuracy

A measure of communicative accuracy can be defined for such communication systems. Given a signaler, P, producing signals using the function $p(m)$ and a receiver, R, interpreting signals using the function $r(s)$, the accuracy of communicating the meaning $m_i \in m$ between the two individuals, $ca(P, R, m_i)$, is:

$$ca(P, R, m_i) = \left\{ \begin{array}{ll} 1 & \text{if } r(p(m_i)) = m_i \\ 0 & \text{otherwise} \end{array} \right\}$$

When $ca(P, R, m_i) = 1$ the communication is successful[1] . A population's communicative accuracy can be estimated by taking the average $ca(P, R, m_i)$ for a random sample of P, R and m_i. In a population possessing an *optimal communication system* $ca(P, R, m_i) = 1$ for any choice of P, R and m_i.

3.1.2 Ambiguity

Such communication systems can be classified in terms of the degree of ambiguity they exhibit in the mapping from meanings to signals. Ambiguity arises when signals which are perceptually indistinguishable are associated with distinct meanings. A communication system of the type outlined above will be termed:

- *Unambiguous* if every meaning is associated with a distinct signal or signals.

[1] Note that this assumes that meanings are functionally distinct. For example, if two meanings m_i and m_j result in the same behavior on the part of the receiver and $r(p(m_i)) = m_j$ then the communication would be measured as a failure but could, at the behavioral level, be considered a success.

- *Partially ambiguous* if some, but not all, meanings are associated with identical signals.
- *Fully ambiguous* if all meanings are associated with identical signals.

These terms are more formally defined below.

3.2 The communicative agent

Feedforward neural networks are used to model communicative agents. The structure of the network used is shown in Figure 1. Networks with this configuration will be referred to as *imitator* networks. The input to the imitator network is considered to be the meaning to be communicated and the imitator's output is considered to be the signal used by that agent to communicate the input meaning, with the precise nature of the meaning-signal mapping determined by the connection weights in the network.

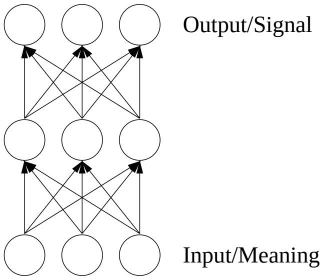


Figure 1: The structure of the neural network.

Communication systems therefore map from 3D meaning vectors to 3D signal vectors. Binary vectors are used, giving 2^3 possible meanings and 2^3 possible signals. A subset of the set of possible meaning vectors are considered to be communicatively relevant situations, where "communicatively relevant" means that agents receive a fitness payoff for communicating about those situations. For all simulations outlined in this paper, the set of communicatively relevant situations, m, consists of the unit vectors (1 0 0), (0 1 0) and (0 0 1). The set of available signals, s, consists of all 2^3 possible binary signal vectors.

Neural networks were chosen to model communicative agents for several reasons. Firstly, there is some tradition of using neural networks in research on the evolution of communication – neural networks of some form are used by Batali (1994), Hutchins and Hazelhurst (1995), Batali (1998), Cangelosi and Parisi (1998), Cangelosi (1999), Livingstone and Fyfe (1999) and Kirby and Hurford (2002). Continuing this tradition provides several benefits. In particular, using a similar model allows the results of this research to be more easily related to previous research and the generality of the results of earlier simulations to be tested.

Secondly, well-established mechanisms exist for training neural networks to learn input-output mappings (i.e. backpropagation). Using an established learning mechanism reduces the amount of novel elements contained in the model, as well as allowing our understanding of that mechanism to be expanded.

Finally, using neural networks allows both genetically-transmitted and culturally-transmitted information to influence, in principle, the eventual behavior of agents in the model. Some of the assumptions arising from the choice of neural networks are somewhat dubious – for example, the assumptions that there is a one-to-one correspondence between genotype and phenotype and that genetic information merely provides a starting point for unconstrained learning in the phenotype. Two points can be raised in defense of these assumptions. Firstly, these assumptions are not unprecedented in the computational modelling literature – see for example the models described in Belew, McInerney, and Schraudolph (1992), Nolfi, Elman, and Parisi (1994), Batali (1994) and Rolls and Stringer (2000). Secondly, using this approach avoids some even more arbitrary assumptions that would be required to model combined influences of genes and culture in a more abstract model. These assumptions will, however, be returned to in the concluding section.

The disadvantage of using feedforward networks is that the slightly contorted reversal process outlined below is required in order to allow bidirectionality. Why is bidirectionality desirable when modeling communication? It is a fundamental assumption of modern linguistics (originating with Chomsky (1965)) that production and reception depend upon a common underlying knowledge of language - an individual's linguistic *competence*. Competence can be distinguished from *performance*, which determines how the structures underlying competence are accessed during reception and production. This competence-performance distinction is maintained here, with an agent's competence being encoded in the set of connection weights in their neural network and their performance being determined by the production and reception processes used to access this competence.

3.2.1 Production and reception

Producing the signal associated with a given meaning $m_i \in m$ in such imitator agents is straightforward — the given meaning is used as the input to the network and activations are propagated forward through the network to give a real-valued output pattern of activation, which is thresholded at 0.5 to give the binary signal associated with the given meaning.

The deterministic nature of the feedforward network during production means that the definition of ambiguity for communication systems can be formally stated. Communication systems used by neural networks will be termed:

- *Unambiguous* if $p(m)$ is a one-to-one, or an injective, function.
- *Partially ambiguous* if $p(m)$ is a many-to-one function, but the range of $p(m)$ is not a singleton set.
- *Fully ambiguous* if the range of $p(m)$ is a singleton set.

Reception is slightly more complex, given that the networks are not bidirectional. All $m_i \in m$ are propagated through a given agent's network to produce a real-numbered output pattern of activation for each meaning. Each output pattern is given a confidence rating, corresponding to how closely that pattern matches the received signal, $s_r \in s$. The meaning which produces the signal closest to s_r, according to the confidence measure, is chosen as the interpretation of s_r. This method is based on the method used by Batali (1998) and Kirby and Hurford (2002) for producing outputs for similar networks.

The confidence measure that a given real-numbered output vector, o, of length n matches a target binary vector t of length n is given by $C(t|o)$. $C(t|o)$ is simply the product of the confidence scores for each individual node $1...n$ in the output vector i.e.

$$C(t[1 \ldots n]|o[1 \ldots n]) = \prod_{i=1}^{n} C(t[i]|o[i])$$

where the confidence measure for node i is

$$C(t[i]|o[i]) = \begin{cases} o[i] & \text{if } t[i] = 1, \\ (1 - o[i]) & \text{if } t[i] = 0. \end{cases}$$

(Equations adapted from Kirby and Hurford (2002))

3.3 The transmission of communication systems

In this section a model of the genetic transmission of network connection weights and cultural transmission of communication systems is outlined. A genetic algorithm (Holland, 1975) is used to model the process of genetic transmission and is combined with an iterated learning model (Brighton and Kirby, 2001) which is used to model cultural transmission.

3.3.1 The genetic algorithm

The genetic algorithm has four key components:

1. A model of population turnover.
2. A model of genotypes, phenotypes and the mapping from genotype to phenotype.
3. Breeding based on an evaluation of communicative ability.
4. A method of recombination and mutation of genes during breeding.

These four components are described below.

Population turnover A generational population model is used — at every time step of the simulation the entire population of size p is replaced by a new population of size p generated by breeding interactions between the members of the old population. $p = 100$ for all simulations outlined in this paper.

Genotypes and phenotypes The phenotype communicative agent used is as outlined in Section 3.2 — a three-layer, feedforward neural network mapping from input meanings to output communicative signals. Each individual's initial connection weights are specified by their genotype — each agent's genotype consists of a string of real numbers, with each locus in the genotype mapping to a particular connection in the phenotype network. The real-numbered allele at a particular locus in the genotype determines the initial weight of the associated connection in the phenotype network. This mapping from genotype to phenotype is illustrated in Figure 2. The agents in the initial population have random alleles in the range $[-1, 1]$. There is no restriction on the range of real-numbered alleles beyond the initial population.

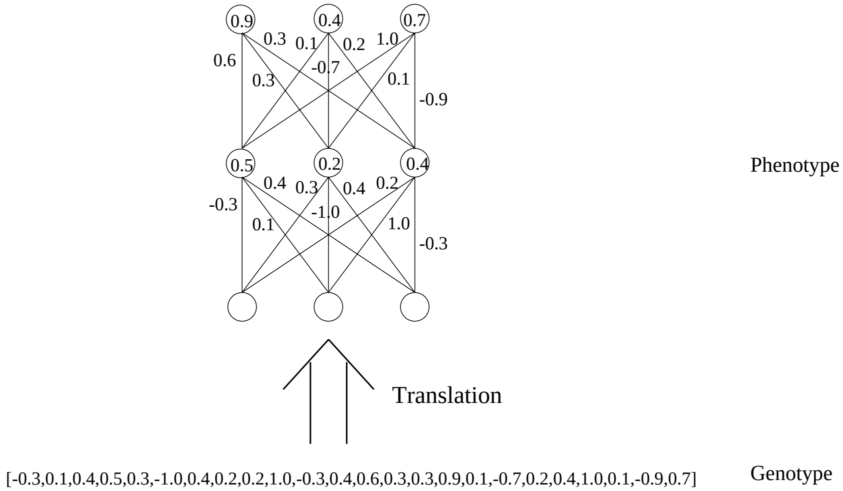


Figure 2: The mapping from genotype (a string of real numbers) to phenotype (a neural network). Bias node connection weights are shown in the associated node.

Selective breeding The probability of an agent breeding is determined by its success at communicating with other members of its generation of the population. The method of evaluating communicative success is given below:

```
• For each agent $A$ in the population:
    1 Remove $A$ from the population.
    2 Pick an agent $B$ at random from the population.
      3.1 Pick a meaning, $m_s$, at random from the set of communicatively relevant situations.
      3.2 Call $A$ the signaler and $B$ the receiver.
        3.3.1 Generate the signal $s_s$, that the signaler associates with $m_s$ (via the production mechanism outlined in Section 3.2.1).
        3.3.2 Identify the meaning, $m_r$, that the receiver associates with $s_s$ (via the reception process outlined in Section 3.2.1).
        3.3.3 Compare $m_r$ with $m_s$ and score the success of the communication. If $m_r$ is identical to $m_s$ score the communication as a success and increment $A$'s fitness. Otherwise, the communication is a failure.
    3.4 Call $A$ the receiver and $B$ the signaler and repeat step 3.3.
    3.5 Return $B$ to the population.
4 Repeat steps 2 and 3 $f$ times[2].
5 Return $A$ to the population
```

This fitness assessment algorithm corresponds to the measure of communicative accuracy outlined in Section 3.1.1. Agents receive a reward both for understanding and being understood and the rewards for both are equally weighted. The fittest b individuals in the population breed with equal probability to produce the next generation of agents, where $0 < b \leq p$.

Recombination of genes Breeding involves recombination of the genes of two parents, via crossover, and mutation. Single-point crossover occurs with probability P_{cross} [3]. Point mutations occur on the newly formed genotype with probability $P_{mutation}$[4]. Mutation results in the value at the mutated locus being increased by a random real number in the range [-1,1][5].

3.3.2 Cultural transmission

Iterated learning models have been used to examine the cultural evolution of communication (Oliphant, 1999) and compositional language (Kirby and Hurford, 1997; Batali, 1998; Batali, in press; Kirby, 1999; Kirby, 2000; Kirby, 2001; Brighton and Kirby, 2001; Kirby and Hurford, 2002). In the iterated learning model "each generation of language user acquires its linguistic competence by observing the behavior of the previous generation" (Brighton and Kirby, 2001, p592). This acquired linguistic competence then governs the behavior which is observed by the subsequent generation. The iterated learning model resembles the cultural equivalent of a genetic algorithm, although typically there is no notion of fitness.

In terms of the current model, individuals at generation $N+1$ observe and learn from the communicative behavior of generation N individuals. Each individual at generation $N+1$ receives e exposures to the communication systems of the population at generation N. These exposures are

[2] $f = 6$ for all simulations outlined in this paper.

[3] For all simulations outlined in this paper $P_{cross} = 0.95$.

[4] For all simulations outlined in this paper $P_{mutation} = \frac{0.1}{L_g} = 0.0042$, where L_g is the length of the genome.

[5] This mutation operator, in conjunction with the unrestricted range of alleles, allows the possibility of the emergence of extremely large-valued alleles. However, in practice such alleles do not occur. In the simulations outlined in this paper all alleles remain within the range [-5.41,5.29].

randomly distributed among the fittest t members of generation N. During each exposure, the set of meaning-signal pairs of the N generation agent is used to train the generation $N+1$ agent. The backpropagation algorithm was used to implement this learning process[6], with the starting point for learning being the connection weights specified in the learning agent's genotype. The learning agent's communication system will therefore be determined, at least to some extent, by the interactions between the processes of genetic transmission via breeding and cultural transmission via learning.

4 Results for imitators

In this section results generated by the model under two main parameter settings are presented:

1. Imitation learning and natural selection ($p = 100$, $20 \leq b \leq 100$, $t = 100$, $0 \leq e \leq 200$) (Section 4.1).
2. Selective imitation learning and natural selection ($p = 100$, $20 \leq b \leq 100$, $20 \leq t \leq 100$, $0 \leq e \leq 200$) (Section 4.2).

Under both general configurations the questions is asked: do the populations converge on communication systems resulting in high levels of communicative accuracy within the population? This was assessed by running 10 simulations under each parameter setting for a fixed number of generations (1000) and measuring the average communicative accuracy of the population, as defined in Section 3.1.1, for the last 10 generations of each run.

4.1 Imitation learning and natural selection

In this section we investigate the accuracy of the emergent communication systems in populations of size $p = 100$ for various numbers of learning episodes e and various amounts of selective pressure b. e ranges from 0 (no learning) to 200, and b ranges from 100 (no selection pressure on breeding) to 20 (very strong selection pressure on breeding). For all simulations outlined in this section, $t = 100$ – immature agents potentially sample and learn from the communicative behavior of the entire preceding generation, regardless of fitness.

[6] A learning rate of 0.5 and momentum of 0 were used.

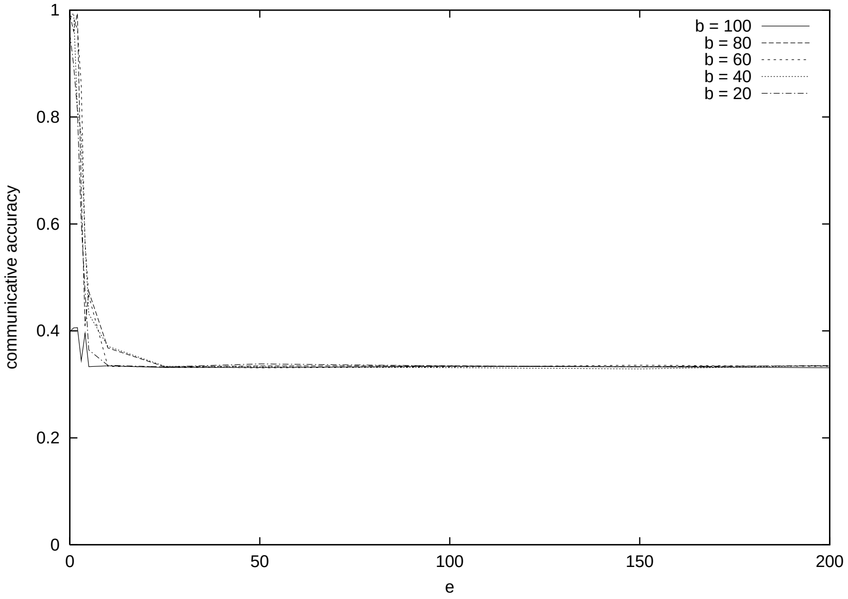


Figure 3: Average communicative accuracy in populations of imitator agents after 1000 generations, for various values of e and b. Each point represents the average of 10 simulations.

The average communicative accuracy of the populations for the full range of values of e are shown in Figure 3. Figure 4 shows the communicative accuracy of populations where $0 \leq e \leq 25$. For extremely low values of e optimal communication systems can emerge given selection pressure on breeding ($b < 100$). Average communicative accuracy rapidly tails off as e increases. For $e > 5$ there is little difference between simulations where breeding is random ($b = 100$) and breeding is non-random ($b < 100$) and for $e > 25$ there is no difference, with all populations converging on communication systems resulting in average communicative accuracy of 0.33. This corresponds to the chance level of performance of a population attempting to communicate three meanings with a fully ambiguous communication system. How can the emergence of these suboptimal communication systems be explained?

Table 1 shows the success of imitator networks at acquiring systems of differing levels of ambiguity for a given number of exposures, e. As can be seen from the Table, systems exhibiting a higher degree of ambiguity are easier to acquire than systems exhibiting a lower degree of ambiguity for all

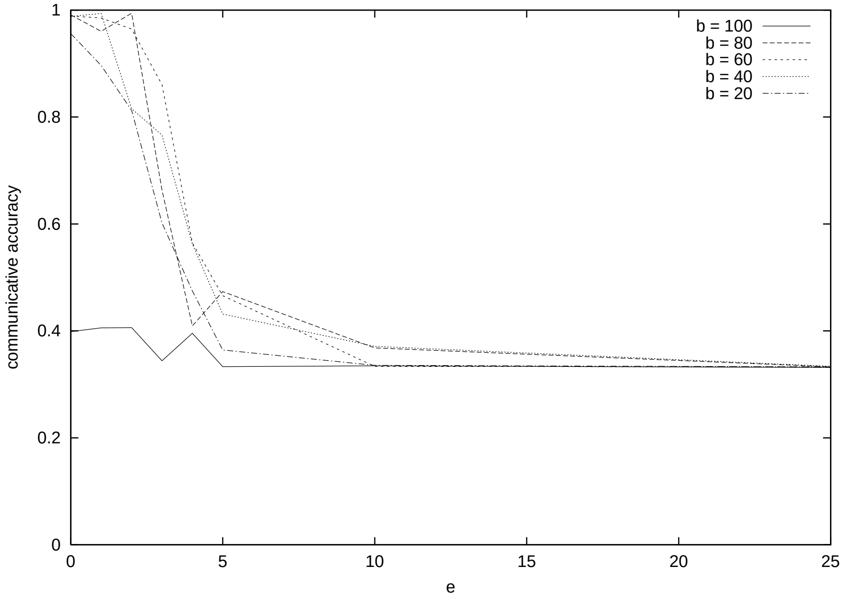


Figure 4: Average communicative accuracy in imitator populations, for small values of e.

values of e. The learning bias of imitator agents results from the imitator network architecture, as is discussed in Section 6.1. It should be noted that systems exhibiting a higher degree of ambiguity have an additional advantage, in that every exposure to an ambiguous system contains multiple exposures to the ambiguous signal. However, this is not the key factor, as can be seen by comparing success rates for fully ambiguous systems with low values of e to success rates for unambiguous systems with higher values of e – for example, five exposures to a fully ambiguous system (15 exposures to the ambiguous signal) gives a rate of success which can only be matched by 150 exposures to an unambiguous system, which gives 150 exposures to each of the three unambiguous signals.

As can be seen from Table 1, for extremely low e ($e < 5$) even fully ambiguous systems can not be reliably acquired. Figure 4 shows that levels of communicative accuracy significantly above the random level are only observed given $e < 5$ and selection pressure on breeding. In these circumstances learning is effectively disabled and natural selection is free to evolve the population's communication systems, resulting in levels of communicative accuracy above the chance level, with optimal communication for $e \leq 2$.

e	System Type		
	Fully Ambiguous	Partially Ambiguous	Unambiguous
1	24.9	0.4	0
2	50.1	0.6	0
3	76.0	1.6	0
4	90.9	1.8	0
5	97.9	1.4	0
10	100	0.4	0
25	100	1.5	0
50	100	32.9	13.3
100	100	93.1	82.8
150	100	99.5	97.9
200	100	100	99.8

Table 1: Percentage success at acquiring various types of system with various values of e. These results were empirically derived by generating 100 random communication systems of each type and training 100 networks with small random initial weights in the range $[-1, 1]$ on each system.

These optimal communication systems disappear given $e > 5$. Why is natural selection not developing optimal communication systems under these circumstances? It appears that the process of cultural transmission is overriding the process of natural selection. Fully ambiguous systems are always the easiest class of system to learn, and are therefore more likely to pass intact through the learning process. The repeated cultural transmission of communication results in the elimination of communication systems which do not conform to the learning biases of the agents – there is cultural selection in favor of systems which conform to the learner biases. In the case of imitator agents the learning bias happens to be in favor of communication systems which are extremely poor in terms of communicative accuracy. As we will see in Section 5, a different agent model leads to a different learning bias.

The learning bias of the agents in favor of ambiguous systems is a property of the imitator architecture (see Section 6.1), rather than the learning rate, which merely determines the strength of the bias for a particular value of e. Therefore the precise value of e at which cultural transmission overrides natural selection is dependent on the particular learning rate used – for example, if a lower learning rate was used then we would observe cultural transmission disabling natural selection only at larger values of e.

Importantly, however, natural selection would still be overridden by cultural selection at some point.

Why does natural selection not counteract the cultural adaptation of the communication systems to the learner biases and weed out poor communicators? Learning in the phenotype masks an individual's genetic makeup — with $e > 5$, no matter how good an agent's genes are, their effects are likely to be overtaken by learning, which almost fully determines an agent's communicative behavior. Shielding (Ackley and Littman, 1992) prevents natural selection from identifying good gene combinations and weeding out bad gene combinations. Only when $e < 5$ is natural selection not disabled by shielding of the genotype.

There are certain combinations of genes which make learning a particular communication system impossible — an agent's genes constitute the starting point for learning, and the backpropagation algorithm is sensitive to initial weights to a certain degree. Genetic drift does occasionally result in small numbers of agents being born whose genes are so good they cannot learn fully ambiguous communication systems. However, these agents must still communicate with their neighbors, and if those neighbors use a fully ambiguous system then using a better system to communicate with them yields no benefit[7]. The good gene combinations do not survive for long due to inter-breeding with agents whose genes allow them to acquire fully ambiguous systems. Cultural transmission leads to cultural stagnation in the simulated populations — the biases of the learners favor fully ambiguous communication systems and natural selection is powerless to counteract this.

4.1.1 Imitation learning and collapse

Cultural transmission not only prevents the development of an optimal communication system in the simulated populations — it prevents the maintenance of such a system. Figure 5 shows the average communicative accuracy of populations of imitator agents who start out with a shared, optimal, innate communication system — all the agents in the initial population have a hand-selected set of genes which encode an unambiguous communication system. $e = 200$ for all simulations in Figure 5. Various amounts of selection pressure (b) are used. $t = 100$, as in the simulations in the previous section. As can be seen from Figure 5, all populations collapse from using an unambiguous communication system to using a fully ambiguous communication

[7] Preferred interaction with genetically-related individuals might alleviate this problem somewhat, but was not investigated here.

system within 15000 generations.

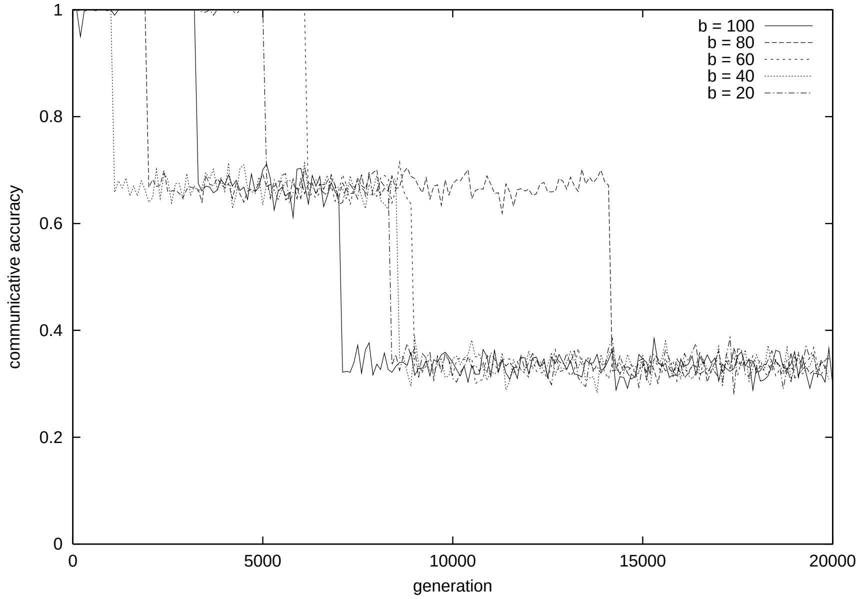


Figure 5: Average communicative accuracy of imitator populations over time, where $t = p = 100$.

As discussed above, learning in the phenotype almost completely masks an agent's genes but there are certain combinations of genes which make learning a particular communication system impossible. In each simulation shown in Figure 5 an agent will eventually be born whose genes are so bad that they cannot learn the unambiguous communication system in use by the rest of the population. This individual will learn a partially ambiguous or fully ambiguous communication system instead. Such agents will be unlikely to breed, given that their fitness will usually be lower than other agents in the population. Suboptimal communicators do have a negative effect on the fitness of optimally-communicating agents, given that those optimally-communicating agents suffer a penalty for not understanding or being understood by suboptimal communicators, although this will not usually depress the population's fitness enough to allow a suboptimal communicator to breed. However, while such individuals are unlikely to breed, their communication systems *will* be observed and learned from by agents in the next generation.

Table 2 shows the percentages of agents with random connection weights

System Type	% population
Unambiguous	2
Partially Ambiguous	25
Fully Ambiguous	73

Table 2: Percentage of initial population using systems of each degree of ambiguity.

in the range $[-1, 1]$ using communication systems of the three levels of ambiguity. Agents with random connection weights clearly tend to have a fully ambiguous communication system. This approximates the response of imitator agents to training on conflicting communication systems – training on conflicting data effectively randomizes the connection weights in the agents' network.

As discussed above, agents with bad genes will occasionally occur in the population due to genetic drift. The communication system of such agents will be observed by other agents in the subsequent generation. These individuals run the risk of acquiring a suboptimal communication system due to the randomizing effect of conflicting training data. If they do acquire a suboptimal system they will be unlikely to breed. Regardless of whether the suboptimal communicators breed or not, their communication systems will be observed by agents in the next generation. As increasing levels of ambiguity result in more successful cultural transmission, suboptimal communication systems spread through the population like a virus due to the processes of cultural transmission, until the whole population converges on a fully ambiguous communication system. Once again, natural selection is powerless to stop this process.

Note that $e = 200$ represents the best-case scenario for learning agents. $e = 200$ results in the highest level of learnability for unambiguous communication systems and also ensures that, at the early stages of collapse, suboptimal communication systems will constitute only a small part of an agent's observations. In populations where $e \leq 2$ the collapse phenomenon does not occur, but as discussed above the behavior of these populations is entirely determined by natural selection – they cannot truly be called learning populations.

4.2 Selective imitation learning and natural selection

The phenomenon observed in the simulations outlined in the previous section is purely a result of the bias of the learners towards acquiring fully ambiguous communication systems. In this section an additional learning bias, a preference of learners to learn from successful communicators, is added to the model.

Natural selection is implemented in this model by only allowing the top b members of the population to transmit their genetic information to the next generation via breeding. Similarly, only the fittest t members of the population transmit their communication systems culturally to the next generation, through the process of being observed and learned from. In the simulations in the previous section $t = p$ – all members of the population participate in cultural transmission, regardless of fitness. However, in this section simulations are described where $t < p$ – an agent's participation in cultural transmission depends to some extent on their fitness.

In these populations of discriminating learners there are therefore 3 potential selection pressures operating on the evolving populations and communication systems:

1. *Natural selection* (when $b < p$), operating on genetic transmission, favoring genes whose phenotype realizations are successful communicators.

2. *Cultural selection for learnability*, operating on cultural transmission, favoring communication systems which conform to the learning bias for fully ambiguous systems.

3. *Cultural selection for communicative success* (when $t < p$), operating on cultural transmission, favoring communication systems which result in successful communication.

Selection pressures 1 and 3 are clearly related, although operating on different modalities of transmission. Selection pressures 2 and 3 operate in the same modality of transmission, but are in direct competition.

Figures 6, 7, 8, 9 and 10 show the communicative accuracy of the emergent communication systems in populations of size $p = 100$, for various values of b, t and e.

The addition of the cultural selection pressure for communicative success has clearly failed to have a significant impact on the emergent communication systems – for $e > 10$ the populations' communication systems tend to be fully ambiguous. For very low values of t and high values of e partially ambiguous communication systems do occasionally emerge. However, the behavior of the population is still dominated by the intrinsic learning bias of the agents, which favors fully ambiguous systems.

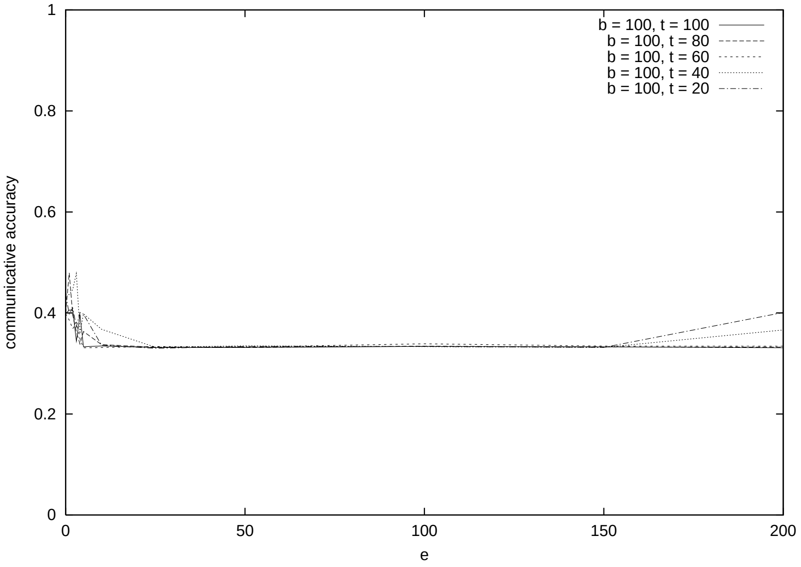


Figure 6: Average communicative accuracy in populations of imitator agents after 1000 generations, for various values of e and t ($b = 100$). Each point represents the average of 10 simulations.

4.2.1 Selective imitation and collapse

Figure 11 shows the communicative accuracy over time of imitator populations, for various values of b and t. $p = 100$ and $e = 200$ in all these simulations. As in the simulations outlined in Section 4.1.1, initially the populations are genetically converged on an optimal communication system.

As can be seen from Figure 11, the addition of selective imitation in the runs where $t < 100$ has failed to prevent the populations from moving away from the initial optimal communication system – in 3 runs the population has converged on a fully ambiguous system yielding chance levels of communicative accuracy, whereas in 2 runs ($b = t = 60$ and $b = t = 40$) the population has converged on a partially ambiguous communication system.

The populations have failed to maintain the original optimal system for the same reason as the populations discussed in Section 4.1.1 – shielding allows mutations to accumulate in the population, those mutations eventually prevent some agents acquiring the optimal communication system and the observation of that suboptimal behavior disturbs more individuals in subsequent generations. This results in the rapid spread of the communication

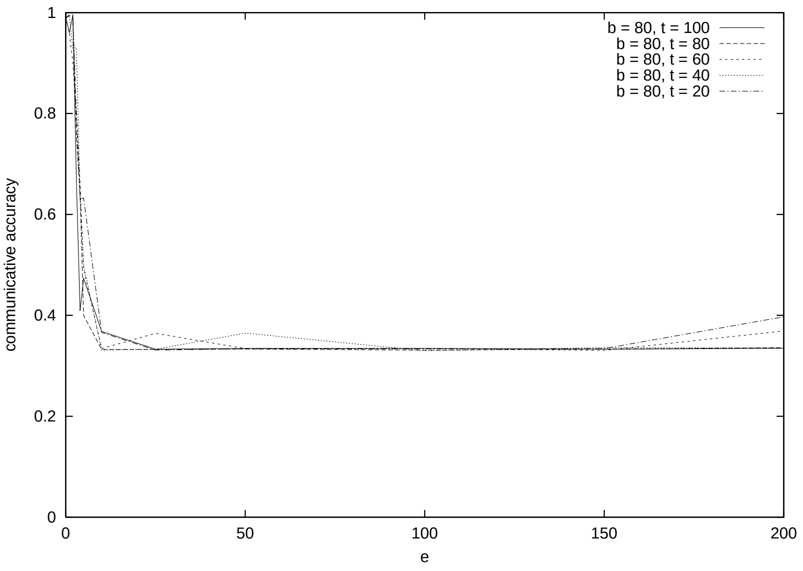


Figure 7: Average communicative accuracy in populations of imitator agents for various values of e and t ($b = 80$).

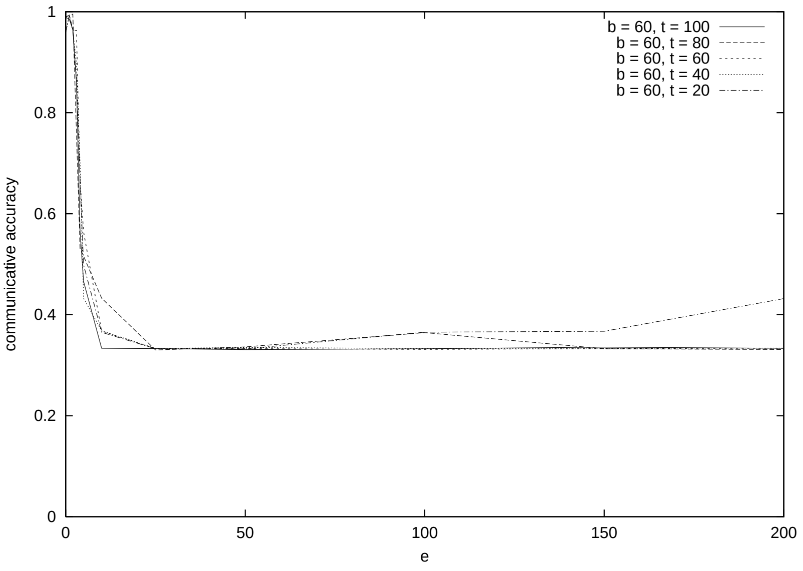


Figure 8: Average communicative accuracy in populations of imitator agents for various values of e and t ($b = 60$).

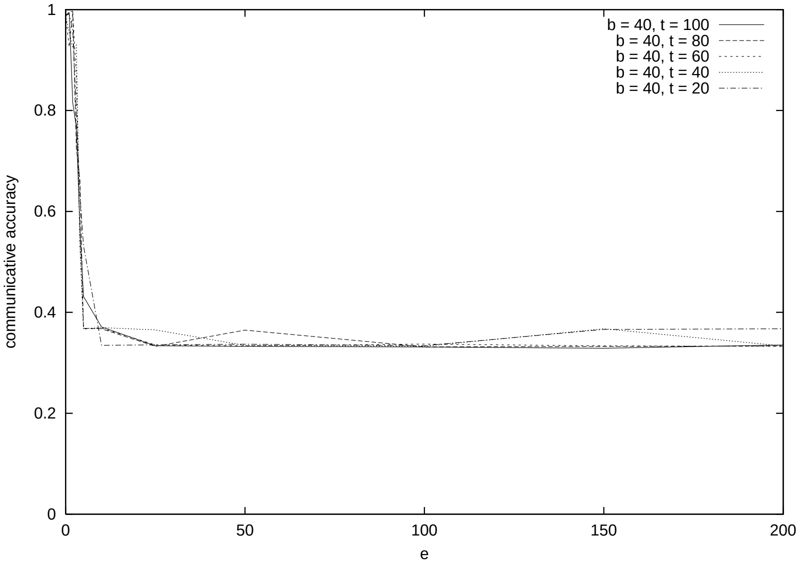


Figure 9: Average communicative accuracy in populations of imitator agents for various values of e and t ($b = 40$).

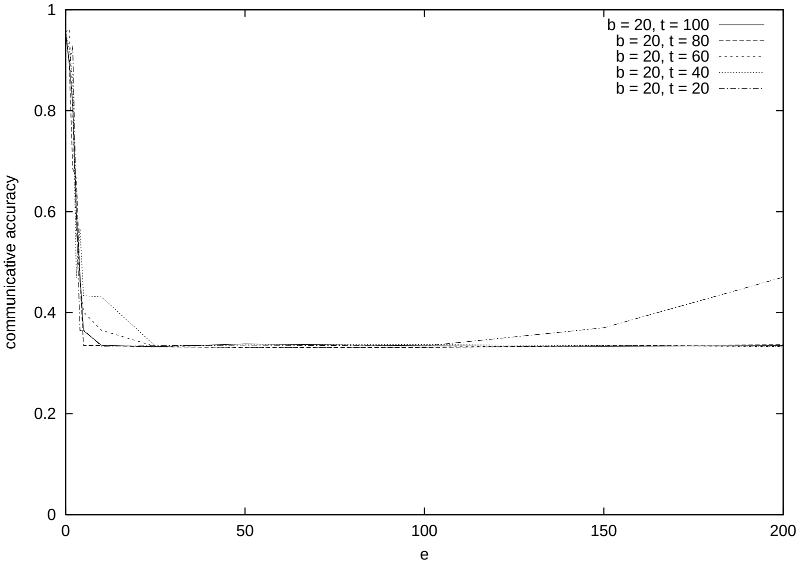


Figure 10: Average communicative accuracy in populations of imitator agents for various values of e and t ($b = 20$).

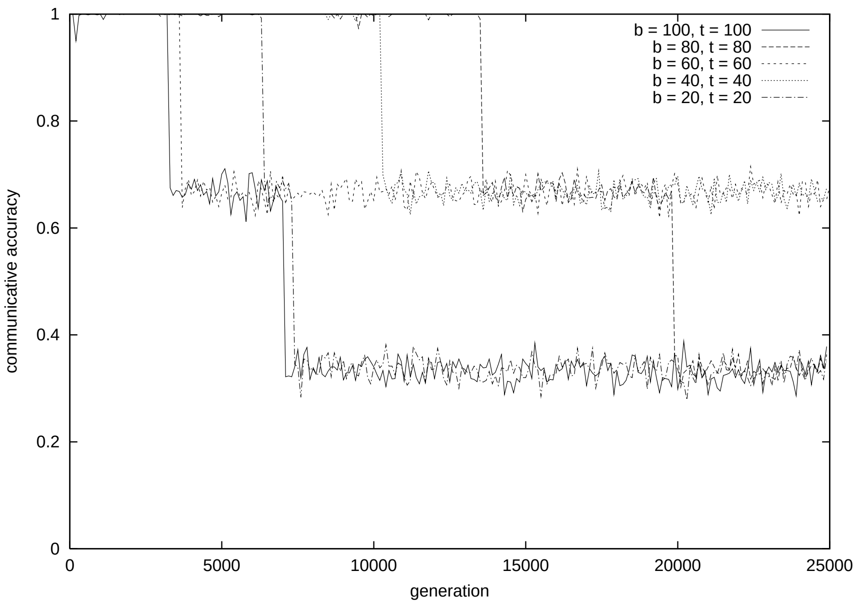


Figure 11: Average communicative accuracy of imitator populations over time.

systems which are most easy to acquire, which happen to be suboptimal in terms of fitness. Lower values of t makes it less likely that individuals with suboptimal communication systems will transmit those systems, therefore making it less likely that the population will collapse. However, as Figure 11 shows, collapses can occur given a sufficient number of mutation events at a single generation.

5 Tailoring the learning bias

In the simulations outlined in the previous section there were two learning biases — the intrinsic bias of the learners, which favors increased ambiguity, and the bias in favor of learning from successful communicators, which depended on t. In this section the model of a communicative agent is revised to build in a learning bias towards optimal, unambiguous communication systems. This bias results in the rapid and reliable emergence of such systems.

5.1 The new communicative agent

As outlined in Section 3.2, the communicative agents in all previous simulations were feedforward neural networks mapping from input meanings to output signals. Signal production for these imitator agents was merely a matter of propagating an input meaning pattern of activation through the network to produce an output signal. Reception was achieved by presenting all communicatively relevant meanings and selecting the meaning which maximizes confidence in the received signal. These networks are strongly biased in favor of acquiring fully ambiguous communication systems.

The new model of a communicative agent has exactly the same basic form as the imitator model, being a three-layer feedforward neural network. However, the crucial difference is that the new networks, which will be referred to as *obverter* (Oliphant and Batali, 1997)[8] networks, map from input signals to output meanings — the direction of the mapping has been reversed. Production and reception in these obverter networks operate as follows:

Production: Each of the set of possible signals is propagated through the network, producing a real-numbered output pattern of activation for each signal. The signal which produces the meaning closest to the meaning to be communicated, as determined by the confidence measure outlined in Section 3.2.1, is used to communicate the given meaning (as for imitator reception).

Reception: The received signal pattern is propagated forward through the network and the output pattern of activation is thresholded to produce a binary pattern of activation corresponding to that agent's interpretation of the received signal (as for imitator production).

The learning biases of these agents are shown in Tables 3 and 4. As can be seen from Table 3 these agents are strongly biased against learning fully ambiguous and partially ambiguous communication systems. Somewhat surprisingly, learnability never reaches 100%, even for unambiguous communication systems. It appears that certain unambiguous systems are unlearnable by obverter agents, while certain unambiguous systems are 100% learnable. The pattern to this learnable-unlearnable distinction is not important in this paper – the key point is that certain unambiguous systems are highly learnable whereas partially ambiguous and fully ambiguous systems are less learnable.

[8] Obverter networks are the equivalent of what Hurford (1989) termed Saussurean learners.

Table 4 shows that obverter networks with random weight settings in the range [-1,1] are strongly biased towards unambiguous communication systems — as discussed in Section 4.1.1, these random weight biases approximate the response of networks to exposure to conflicting communication systems.

e	System Type		
	Fully Ambiguous	Partially Ambiguous	Unambiguous
1	0.1	0.2	0.3
2	0.2	0.2	0.4
3	0.1	0.4	0.4
4	0	0.3	0.6
5	0	0.5	0.6
10	0	1.0	1.6
25	0	4.6	7.8
50	0	8.9	31.3
100	0	11.1	53.9
150	0	11.0	51.7
200	0	18.0	55.0

Table 3: Percentage success of obverter agents at acquiring various types of system with various values of e.

System Type	% population
Unambiguous	65
Partially Ambiguous	33
Fully Ambiguous	2

Table 4: The percentage of a population of obverter agents with small, random weights using communication systems of the given type

5.2 Obverter learning results in optimal communication

In the simulation runs plotted in Figure 12, the new obverter learner is substituted for the imitator learner used in previous sections. Excluding the change in the agent model, all other simulation details are identical to the simulation runs described in Section 4.1 — specifically, $p = 100$ and $t = 100$.

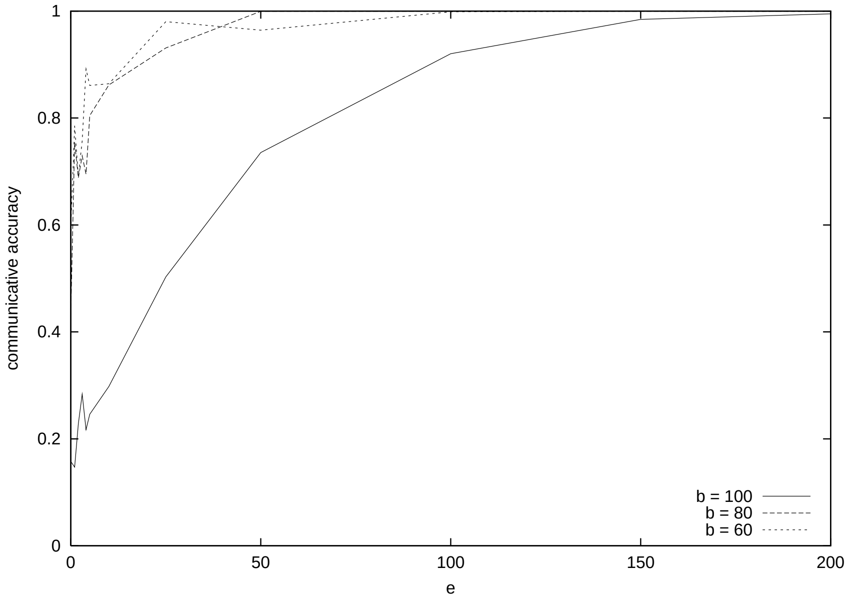


Figure 12: Average communicative accuracy in populations of obverter agents after 1000 generations, for various values of e and b ($t = 100$). Each point represents the average of 10 simulations.

Figure 12 shows a clear difference between simulation runs with no selection pressure on breeding ($b = 100$) and those with selection pressure on breeding ($b < 100$). For the runs with no natural selection, communicative accuracy is low with low e and increases as e increases. For $e \geq 100$ the populations reliably converge on optimal, unambiguous communication systems. As shown in Table 3, (a subset of the set of) unambiguous systems are highly learnable for these values of e and have a significant learnability advantage over ambiguous systems. As a result, ambiguous systems are selected against during cultural transmission until the populations converge on unambiguous systems.

In the runs with selection pressure on breeding the populations have higher communicative accuracy with $e < 100$ due to the development of innate communication systems through natural selection. As e increases the communicative accuracy of the population increases, indicating a positive interaction between natural selection and learning – natural selection favors genotypes which improve the learnability of unambiguous communication systems in use in the populations (i.e. the Baldwin Effect (Baldwin, 1896)).

As e increases this interaction decreases, and when $e \geq 100$ the interaction all but disappears.

Furthermore, populations of obverter agents are capable of maintaining such optimal communication systems indefinitely – unlike the populations shown in Sections 4.1.1 and 4.2.1 they do not suffer from the collapsing problem, even in the absence of selection pressure on breeding.

6 The key learning bias

Imitator agents cannot create or maintain optimal communication systems, even given a helping hand from natural selection, and obverter agents can construct and maintain optimal communication systems, given sufficient exposure, without help from natural selection. This is due to the inherent learning biases of the two types of agents (summarized in Tables 1, 2, 3 and 4). The relationship between these biases and the structure of the networks is explored in detail in Section 6.1. The key biases identified in Section 6.1 are discussed in terms of other models in Section 6.2.

6.1 The learning bias explored

In the terms of this paper, optimal communication systems are unambiguous mappings from meanings to signals – one-to-one (or injective) functions. Suboptimal systems are many-to-one or all-to-one functions. In terms of production and reception functions $p(m)$ and $r(s)$, in an optimal communication system:

1. $p(m)$ should be an injective function.

2. $r(s)$ should be a superset of the inverse of $p(m)$.

These two restrictions guarantee that every meaning is expressed using a distinct signal and that the reception process maps signals back on to the meanings they were originally intended to convey.

Feedforward neural networks learn many-to-one functions. Due to the deterministic nature of the feedforward propagation of activation values they cannot learn one-to-many mappings. The easiest function for a network to acquire is therefore an all-to-one mapping from inputs to outputs, the hardest learnable function is an injective (one-to-one) function and one-to-many mappings are unlearnable. The reversal process used to model reception behavior for imitators and production behavior for obverters is similarly biased – it generates a function, which may be injective or many-to-one, based on the function the feedforward network has acquired. In general, if the network has acquired a function $f(x)$ which has a range y, then the reversal process ensures that element $y_i \in y$ will map onto a single element $x_i \in x$ such that $f(x_i) = y_i$ – in simple terms, the reversal process deterministically reverses the function acquired by the network.

In imitator agents the feedforward network learns functions from meanings to signals – it learns $p(m)$. Since it is a feedforward network it will be biased towards acquiring a many-to-one or all-to-one $p(m)$. As illustrated in Figure 13 and discussed in the caption, the maximally stable $p(m)$ for imitator agents is therefore an all-to-one fully ambiguous function. Imitators therefore do not have a bias in favor of the 1st feature (above) of an optimal system. Reception in imitators will be based on their acquired $p(m)$ – as shown in Figure 13, in the case of an all-to-one $p(m)$, in $r(s)$ the signal s_i that constitutes the range of $p(m)$ will map onto a single element from m. Therefore a population of imitators agents will tend to produce the same signal for every meaning and interpret the ambiguous signal as communicating one arbitrary meaning. This situation results in performance equivalent to random guessing.

In obverter agents the feedforward network learns functions from signals to meanings – it learns $r(s)$. As illustrated in Figures 14 and 15 the only culturally stable system has an injective $p(m)$ (point 1 above) and an $r(s)$ which includes at least the inverse of $p(m)$ (point 2 above). Obverter agents are therefore strongly biased in favor of acquiring systems with the properties of optimal communication systems.

6.2 Other models

As mentioned in Section 2, there are several other models where cultural processes result in the emergence of optimal communication. Do the learning mechanisms used in these models include biases in favor of the two properties outlined above? This kind of analysis often requires a great deal of familiarity with the model involved. However, this bias can be identified in certain other models.

Beginning with the models involving cultural transmission and no natural selection, the two key biases can be observed in the auto-associator networks of Hutchins and Hazelhurst (1995), the "obverter" learner of Oliphant and Batali (1997), which is capable of constructing an optimal system of communication from random behavior, (but not the "imitator", which is not) and in the "constructor" agents of Smith (2002), capable of construct-

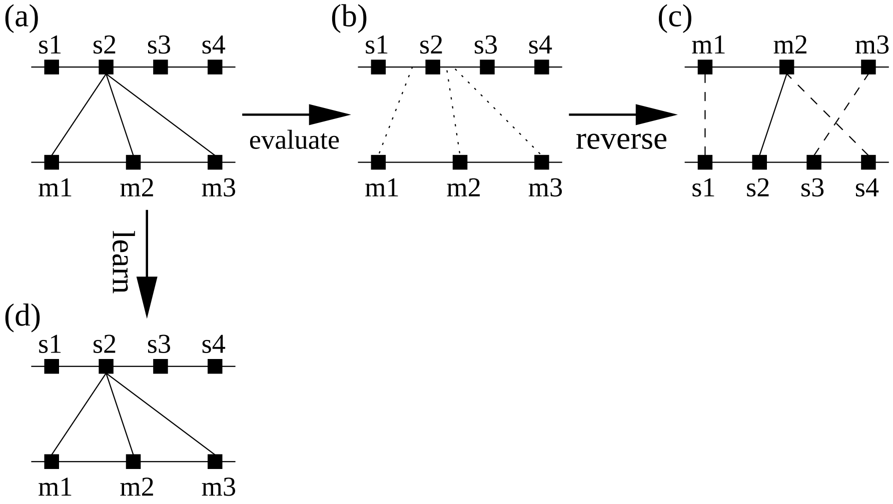


Figure 13: (a) is a representation of an imitator agent's feedforward network encoding an all-to-one $p(m)$ mapping three meanings onto a single signal, $s2$. The function from a domain of real numbers (input unit activations) to a codomain of real numbers (output unit activations) is represented by two lines, the lower line representing the domain, the upper representing the codomain. Squares represent particular points on the line corresponding to binary meanings or signals. Associations are shown with solid lines between elements in the domain and elements in the codomain. (b) represents the confidence-measuring step of the reversal process for the network underlying (a). In order to decide $r(s2)$, the real-number values of $p(m1)$, $p(m2)$ and $p(m3)$ are calculated. These real-numbered mappings are represented by dotted lines in (b). (c) represents the $r(s)$ derived from applying the reversal process to (a). $r(s2) = m2$ because $m2$ mapped closer to $s2$ than any other m in (b). The other associations are effectively random. The random nature of these mappings is represented by dashed lines. (d) represents the function acquired by an imitator network exposed to behavior generated by (a) – as it is an all-to-one function between meanings and signals it is easily learned by imitator agents. This is in fact the only stable function for imitators.

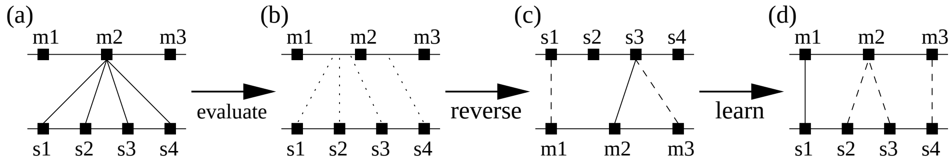


Figure 14: (a) represents an all-to-one $r(s)$ encoded in an obverter agent's feedforward network. As obverters map from signals to meanings this is the most learnable $r(s)$. (b) represents the confidence-measuring step of reversing this $r(s)$ to generate a $p(m)$ – as before, real-number mappings are shown as dotted lines. (c) shows the $p(m)$ derived from (a). $p(m2) = s3$ as $s3$ mapped closest to $m2$ in (b). The other associations are essentially random. The $p(m)$ in (c) produces the meaning-signal pairs $\{(m1, s1), (m2, s3), (m3, s3), \}$. The order of the meanings and signals in these pairs are reversed to train the next generation of obverter networks. (d) shows the $r(s)$ resulting from training an obverter network on the signal-meaning pairs $\{(s1, m1), (s3, m2), (s3, m3), \}$. $r(s1) = m1$, as expected. However, feedforward networks cannot learn one-to-many mappings so $r(s3)$ is effectively randomly assigned to a signal, in this case $m2$. As $s2$ and $s4$ are not represented in the training set they are effectively randomly assigned mappings. Notice that the mapping in (a) has been destroyed in (d) – (a) is not a culturally stable mapping.

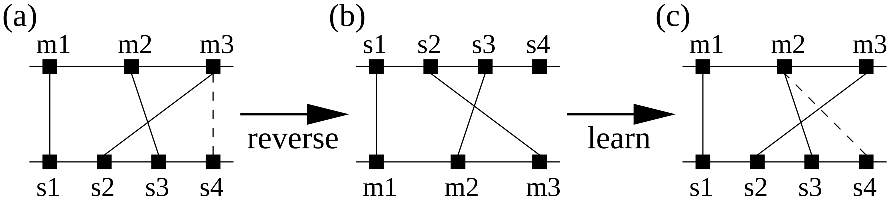


Figure 15: Only an unambiguous $p(m)$ is stable for obverter agents. (a) represents an obverter agent's $r(s)$. (b) is the $p(m)$ derived from reversal of (a) – it is an injective function. (c) illustrates that the $r(s)$ resulting from training the next generation of agents on data produced by (b) is effectively similar to (a) and will therefore lead to (b) once again – (a) and (b) are culturally stable. The only unstable aspect is the floating synonym $s4$. This synonym is highly unlikely to interfere with the mapping in (b) and the floating synonym phenomenon can be observed in the other obverter models outlined in the main text.

ing an optimal system, but not in any non-constructor agents. The class of constructor agents in these papers includes the Hebbian learner of Oliphant (1999), also capable of constructing an optimal system. The neural networks in Batali (1998) are essentially obverter agents, although the importance of their inherent bias is not identified.

The key bias can also observed in the model outlined in Kvasnička and Pospíchal (1999), which involves cultural transmission and natural selection. Given that the neural networks used by Kvasnicka and Pospichal are practically identical to the obverter network outlined in this paper, the behavior of their populations can probably be explained purely in terms of cultural processes. The absence of a contrastive learning bias or variance in natural selection pressure in their model obscures this fact. The models of MacLennan and Burghardt (1994) and Kirby and Hurford (1997) do have learning biases which are specifically directed towards optimizing acquired systems, but these models either model communication at a different level (Kirby and Hurford, 1997) or can be criticized for their use of reinforcement learning (MacLennan and Burghardt, 1994). Finally, it is worth pointing out that the theoretical models proposed in Pinker and Bloom (1990) and Dor and Jablonka (2000) do not take into account the role of learning biases in cultural evolution. The computational model outlined in this paper suggests that the consequences of such biases, which may be non-obvious, need to be taken into consideration.

The two key biases appear to be common, or at least recurring, in learning mechanisms capable of constructing optimal communication systems in the modeling literature. Oliphant (1999) claims that this kind of bias is in fact widespread in the natural world. But is there evidence that any species is actually biased in favor of learning one-to-one mappings in the domain of communication? As mentioned in Section 2, it is doubtful whether experience plays a role in determining the structure of communication in any non-human primate. However, the sole species in which experience definitely plays some role (humans) does appear to exhibit this bias. It has been suggested that word acquisition in humans is guided by the Contrast Principle (Clark, 1988), a bias in favor of one-to-one mappings between meanings and words. While Bloom (2000) suggests this principle is part of the human theory of mind, it can be conceived of as a communication-specific learning bias. The model suggests that some of the nature of the human communication system may be explicable in terms of cultural processes resulting from the iterated application of human learning biases. The role of such a learning bias in the evolution of syntax is a possible subject for future research - is such a bias sufficient for the cultural evolution of syntax, as well as simple communication? If not, what other components are required?

7 Conclusions

This paper outlines a computational model of the emergence of communication in a population of communicative agents. As previous work suggests, natural selection alone is capable of evolving optimal, innate communication systems in such populations. However, the addition of cultural transmission of communication systems does not necessarily assist the emergence of optimal communication systems. The biases of the learners involved in the cultural transmission process result in cultural selection — communication systems which conform to the biases of the learners are more likely to be successfully transmitted than communication systems which do not.

The results of simulations in which cultural selection is in direct conflict with natural selection are outlined in Section 4.1. In these circumstances, cultural selection resulting from the intrinsic biases of the agents proves to be the determining factor in the emergent behavior of the simulated populations. This is a clear case of what Durham (1991) would term gene-culture *opposition*. In Section 4.2, a second cultural selection pressure was introduced which was in direct conflict with the intrinsic learning biases of the simulated agents. This secondary pressure failed to override the intrinsic biases of the learners. In Section 5 the model of the learning agent was modified to build in a bias towards optimal communication systems. In populations of such agents optimal communication systems rapidly and reliably emerged, due to the cultural selection pressures arising from the learners' biases. As discussed in Section 6, these learning biases can be explained in terms of the networks' structure.

This model has several limitations. The model of communication used, with three unstructured meanings and eight unstructured signals, is very simple, although there is no reason to expect these results not to hold for more complex models of communication. The absence of any environment outwith the agents means that communicative accuracy must be measured at the agent-internal representations. While this is not unreasonable for referential communication systems, a behavior-based measure of communicative success would be more appropriate for modeling non-referential communication. Finally, natural selection has a fairly limited role to play in this model, a matter which is discussed further below.

These results have implications for both computational modeling of gene-culture interactions and research into the origins and evolution of language.

For computational modelers, the clear implication is that a particular choice of agent model or learning model can have a fundamental impact on the behavior of the systems as a whole. This suggests that modelers should be aware of how specific their results are to their model of learning and be prepared to justify their model of learning and its associated biases in terms of the real-world system that is being modeled. The bias in favor of one-to-one mappings associated with the the obverter agent corresponds to a learning bias observed in humans (the Contrast Principle, as discussed in Section 6.2), suggesting that, in terms of learning bias, the obverter model is preferable to the imitator model as a model of human language learning.

More generally, the simulations outlined in this paper suggest that research into the origins and evolution of language should not underestimate the role of cultural selection in this process. These simulations give an illustration of the fact that the learning biases of individual learners can have profound and far-reaching effects when placed in the context of iterated cultural transmission, and that in certain circumstances these cultural processes can effectively nullify the influence of natural selection during genetic transmission.

This is not to say that natural selection can have no role in the explanation of the evolution of language. In the simulations outlined in this paper, natural selection is restricted to tinkering with the starting point for the learning process. In a more realistic model, all aspects of the learning apparatus would be genetically transmitted. It would therefore be possible for natural selection to develop learning algorithms, and therefore modify learner biases and determine the precise nature of cultural selection occurring during cultural transmission. Under these circumstances, can natural selection identify learning algorithms which result in cultural selection for optimal communication? Preliminary modeling work in this area (Smith, 2001; Smith, in prep) suggests that natural selection may be unable to reliably identify such biases due to the significant delay between the appearance of the beneficial genes and the establishment of widespread beneficial culture. The story of the evolution of language may therefore be best told in two parts, with the development of the necessary pre-adaptation of an appropriate learning mechanism occurring in a geological time scale and the development of language parasitic on this learning mechanism occurring in a historical time scale.

Acknowledgements

Thanks to Prof Jim Hurford, Dr Simon Kirby, Henry Brighton, Andrew Smith and the anonymous reviewers for their helpful comments on this work.

References

Ackley, D. and M. Littman (1992). Interactions between learning and evolution. In C. Langton, C. Taylor, J. Farmer, and S. Rasmussen (Eds.), *Artificial Life 2*, pp. 487–509. Redwood City, CA: Addison-Wesley.

Ackley, D. and M. Littman (1994). Altruism in the evolution of communication. In R. Brooks and P. Maes (Eds.), *Artificial Life 4: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pp. 40–48. Redwood City, CA: Addison-Wesley.

Baldwin, J. M. (1896). A new factor in evolution. *American Naturalist 30*, 441–451.

Batali, J. (1994). Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In R. Brooks and P. Maes (Eds.), *Artificial Life 4: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*, pp. 160–171. Redwood City, CA: Addison-Wesley.

Batali, J. (1998). Computational simulations of the emergence of grammar. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight (Eds.), *Approaches to the Evolution of Language: social and cognitive bases*, pp. 405–426. Cambridge: Cambridge University Press.

Batali, J. (in press). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Belew, R., J. McInerney, and N. N. Schraudolph (1992). Evolving networks: Using the genetic algorithm with connectionist learning. In C. Langton, C. Taylor, J. Farmer, and S. Rasmussen (Eds.), *Artificial Life 2*, pp. 511–547. Redwood City, CA: Addison-Wesley.

Bloom, P. (2000). *How Children Learn the Meanings of Words*. MIT Press.

Bloom, P. and L. Gleitman (2001). Language acquisition. In R. A. Wilson and F. Keil (Eds.), *The MIT Encyclopaedia of the Cognitive Sciences*. Cambridge, MA: MIT Press.

Brighton, H. and S. Kirby (2001). The survival of the smallest: Stability conditions for the cultural evolution of compositional language. In J. Kelemen and P. Sosik (Eds.), *Advances in Artificial Life (Proceedings of the 6th European Conference on Artificial Life)*. Heidelberg: Springer-Verlag.

Brown, R. and C. Hanlon (1970). Derivational complexity and order of acquisition in child speech. In J. R. Hayes (Ed.), *Cognition and the Development of Language*. New York: Wiley.

Bullock, S. (1997). An exploration of signalling behaviour by both analytic and simulation means for both discrete and continuous models. In P. Husbands and I. Harvey (Eds.), *Fourth European Conference on Artificial Life*, pp. 454–463. Cambridge, MA: MIT Press.

Cangelosi, A. (1999). Modelling the evolution of communication: From stimulus associations to grounded symbolic associations. In D. Floreano, J. D. Nicoud, and F. Mondada (Eds.), *Advances in Artificial Life*, Number 1674 in Lecture notes in computer science. Springer.

Cangelosi, A. and D. Parisi (1998). The emergence of a 'language' in an evolving population of neural networks. *Connection Science 10*(2), 83–97.

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

Chomsky, N. (1987). *Knowledge of Language: Its Nature, Origin and Use*. Dordrecht: Foris.

Christiansen, M. and J. Devlin (1997). Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In M. Shafto and P. Langley (Eds.), *Proceedings of the 19th Annual Cognitive Science Society Conference*, pp. 113–118. Lawrence Erlbaum Associates.

Clark, E. (1988). On the logic of contrast. *Journal of Child Language 15*, 317–335.

Di Paolo, E. (1997). An investigation into the evolution of communication. *Adaptive Behaviour 6*, 285–324.

Dor, D. and E. Jablonka (2000). From cultural selection to genetic selection: a framework for the evolution of language. *Selection 1(1-3)*, 33–55.

Durham, W. H. (1991). *Coevolution: Genes, Culture and Human Diversity*. Stanford, CA: Stanford University Press.

Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology 144*, 517–546.

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

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Cambridge, MA: MIT Press.

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

Hurford, J. R. (2000). Social transmission favours linguistic generalization. In C. Knight, M. Studdert-Kennedy, and J. Hurford (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 324–352. Cambridge: Cambridge University Press.

Hurford, J. R. (2002). Expression/induction models of language evolution: dimensions and issues. In E. Briscoe (Ed.), *Linguistic Evolution Through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Hutchins, E. and B. Hazelhurst (1995). How to invent a lexicon: the development of shared symbols in interaction. In N. Gilbert and R. Conte (Eds.), *Artificial societies: the computer simulation of social life*. london: UCL Press.

Kirby, S. (1998). Fitness and the selective adaptation of language. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight (Eds.), *Approaches to the Evolution of Language: Social and Cognitive Bases*, pp. 359–383. Cambridge: Cambridge University Press.

Kirby, S. (1999). Learning, bottlenecks and infinity: a working model of the evolution of syntactic communication. In K. Dautenhahn and C. Nehaniv (Eds.), *Proceedings of the aisb'99 symposium on imitation in animals and artifacts*. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

Kirby, S. (2000). Syntax without natural selection: how compositionality emerges from vocabulary in a population of learners. In C. Knight, M. Studdert-Kennedy, and J. R. Hurford (Eds.), *The Evolutionary Emergence of Language: Social Function and the Origins of Linguistic Form*, pp. 303–323. Cambridge: Cambridge University Press.

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

Kirby, S. (in press). Learning, bottlenecks and the evolution of recursive syntax. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Kirby, S. and J. R. Hurford (1997). Learning, culture and evolution in the origin of linguistic constraints. In P. Husbands and I. Harvey (Eds.), *Fourth European Conference on Artificial Life*, pp. 493–502. Cambridge, MA: MIT Press.

Kirby, S. and J. R. Hurford (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi (Eds.), *Simulating the Evolution of Language*. Springer Verlag.

Krebs, J. R. and R. Dawkins (1984). Animal signals: mind-reading and manipulation. In J. R. Krebs and N. B. Davies (Eds.), *Behavioural ecology: an evolutionary approach*. Oxford: Blackwell Scientific Publications.

Kvasnička, V. and J. Pospíchal (1999). An emergence of coordinated communication in populations of agents. *Artificial Life 5*(4), 319–342.

Levin, M. (1995). The evolution of understanding: a genetic algorithm model of the evolution of communication. *Biosystems 36*, 167–178.

Livingstone, D. and C. Fyfe (1999). Modelling the evolution of linguistic diversity. In D. Floreano, J. Nicoud, and F. Mondada (Eds.), *Advances in artificial life: fifth european conference on artificial life*, pp. 704–708. Berlin: Springer.

MacLennan, B. and G. Burghardt (1994). Synthetic ethology and the evolution of cooperative communication. *Adaptive Behaviour 2*, 161–187.

Noble, J. (1998). Evolved signals: Expensive hype vs. conspiratorial whispers. In C. Adami, R. Belew, H. Kitano, and C. Taylor (Eds.), *Artificial Life 6: Proceedings of the Sixth International Conference on Artificial Life*. Cambridge, MA: MIT Press.

Nolfi, S., J. L. Elman, and D. Parisi (1994). Learning and evolution in neural networks. *Adaptive Behavior 3*(1), 5–28.

Oliphant, M. (1996). The dilemma of saussurean communication. *BioSystems 37*, 31–38.

Oliphant, M. (1999). The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior 7*(3/4), 371–384.

Oliphant, M. and J. Batali (1997). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter 11*(1).

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

Reggia, J., R. Schulz, G. Wilkinson, and J. Uriagereka (2001). Conditions enabling the evolution of inter-agent signaling in an artificial world. *Artificial Life 7*(1), 3–32.

Rolls, E. T. and S. M. Stringer (2000). On the design of neural networks in the brain by genetic evolution. *Progress in Neurobiology 61*, 557–579.

Smith, K. (2001). The importance of rapid cultural convergence in the evolution of learned symbolic communication. In J. Kelemen and P. Sosik (Eds.), *Advances in Artificial Life (Proceedings of the 6th European Conference on Artificial Life)*, pp. 381–390. Heidelberg: Springer-Verlag.

Smith, K. (2002). The cultural evolution of communication in a population of neural networks. To appear in Connection Science.

Steels, L. (1999). *The Talking Heads Experiment*, Volume I. Words and Meanings. Antwerpen: Laboratorium. Special pre-edition.

Turkel, W. J. (in press). The learning guided evolution of natural language. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Werner, G. and M. Dyer (1992). Evolution of communication in artificial organisms. In C. Langton, C. Taylor, J. Farmer, and S. Rasmussen (Eds.), *Artificial Life 2*, pp. 659–687. Redwood City, CA: Addison-Wesley.

Werner, G. and P. Todd (1997). Too many love songs: Sexual selection and the evolution of communication. In P. Husbands and I. Harvey (Eds.), *Fourth European Conference on Artificial Life*, pp. 434–443. Cambridge, MA: MIT Press.

Wheeler, M. and P. de Bourcier (1995). How not to murder your neighbour: Using synthetic behavioral ecology to study aggressive signaling. *Adaptive Behavior 3*(3), 273–309.

Worden, R. (in press). Words, memes and language evolution. In E. Briscoe (Ed.), *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Yamauchi, H. (2001). The difficulty of the baldwinian account of linguistic innateness. In J. Kelemen and P. Sosik (Eds.), *Advances in Artificial Life (Proceedings of the 6th European Conference on Artificial Life)*. Heidelberg: Springer-Verlag.

Zahavi, A. (1975). Mate selection - a selection for a handicap. *Journal of Theoretical Biology 53*, 205–214.

Zahavi, A. (1977). The cost of honesty (further remarks on the handicap principle. *Journal of Theoretical Biology 67*, 603–605.

SMITH, K., H. BRIGHTON, & S. KIRBY, submitted. Language evolution in a multi-agent model: the cultural emergence of compositional structure. Submitted to *Discrete Dynamics in Nature and Society*.

Language Evolution in a Multi-agent Model: the cultural emergence of compositional structure

K. Smith*, H. Brighton, S. Kirby
Language Evolution and Computation Research Unit
Theoretical and Applied Linguistics
University of Edinburgh
Adam Ferguson Building
40 George Square
Edinburgh, UK

{kenny,henryb,simon}@ling.ed.ac.uk

Abstract

Language arises from the interaction of three complex adaptive systems — biological evolution, learning, and culture. We focus here on cultural evolution, and present a multi-agent Iterated Learning Model of the emergence of compositionality, a fundamental structural property of language. Our key results is to show that the poverty of the stimulus available to language learners leads to a pressure for linguistic structure. When there is a bottleneck on cultural transmission, only a language which is generalisable from sparse input data is stable. Language itself evolves on a cultural time-scale, and compositionality is language's adaptation to stimulus poverty.

Key words: language, cultural evolution, compositionality
Running title: The Evolution of Compositionality

1 INTRODUCTION

Human language is at the nexus of several complex adaptive systems (Gell-Mann 1992). But what are these systems, and how did they interact to deliver up language, unique among the communication systems of the natural world? In what we will call the *standard adaptationist model*, language is seen primarily as a biological trait. Language can then be explained in terms of the interaction between biological evolution of the human "language instinct" (Pinker 1994) and individual learning of language.

The standard adaptationist model is based on the Chomskyan paradigm from linguistics, which focuses on the innate linguistic knowledge of the speaker. However, we argue that this de-emphasis of learning and cultural transmission obscures an important dynamic in language evolution. Language itself functions as a complex adaptive system, and the historical evolution of language interacts with individual learning and biological evolution of the language faculty.

We believe that an understanding of language evolution will require a thorough understanding of each of these three complex adaptive systems (biological evolution, learning and culture), but also, crucially, an understanding of how they interact. In this paper we will focus on modelling the cultural evolution of compositionality, one of the fundamental structural characteristics of

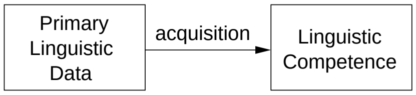


Figure 1: The Chomskyan Paradigm. The focus is on an individual's internal linguistic competence, and how this competence is acquired based on the available data. Acquisition is constrained and guided by an innate LAD and UG.

language. In our concluding remarks we will broaden our focus to discuss how cultural evolution might interact with biological evolution and learning.

We have modelled the cultural evolution of compositionality mathematically (Brighton 2002). Mathematical and computational models suggest different idealisations. The idealisations made by Brighton restrict us to examining the behaviour of perfectly compositional or completely non-compositional language. In this paper we present a multi-agent computational model of the dynamics arising from the cultural transmission of linguistic structure. We show that compositional language can emerge from an initially non-compositional system by cultural processes. The poverty of the stimulus available to language learners drives the evolution of linguistic structure — language itself evolves to be learnable, and compositionality is language's adaptation to the poverty of the stimulus problem.

2 COMPLEX SYSTEMS AND LANGUAGE EVOLUTION

2.1 The Standard Adaptationist Model

The standard adaptationist model places the Chomskyan approach to language within an evolutionary framework. In the Chomskyan paradigm (formulated and developed by Noam Chomsky, for example Chomsky (1965) and Chomsky (1995)), which has been highly influential in modern linguistics, language is viewed as an aspect of individual psychology. The object of interest is the internal linguistic competence of the individual, and how this linguistic competence is derived from the data the individual is exposed to. External linguistic behaviour is considered to be epiphenomenal, the uninteresting consequence of the application of this linguistic competence to a set of contingent communicative situations. From this standpoint, much of the structure of language is puzzling — how do children, apparently effortlessly and with virtually universal success, arrive at a sophisticated knowledge of language from exposure to sparse and noisy data? In order to explain language acquisition in the face of this poverty of the linguistic stimulus, the Chomskyan program postulates a sophisticated, genetically-encoded language organ of the mind, consisting of a Universal Grammar (UG), which delimits the space of possible languages, and a Language Acquisition Device (LAD), which guides the formation of linguistic competence based on the observed data. This scheme is illustrated in Figure 1.

Chomsky has been notoriously reluctant to offer an account of the evolution of UG and the LAD, preferring instead to appeal to architectural and developmental constraints:

> "We know very little about what happens when 10^{10} neurons are crammed into something the size of a basketball, with further conditions imposed by the specific manner in which this system developed over time. It would be a serious error to suppose that all properties, or the interesting properties of the structures that evolved, can be 'explained' in terms of natural selection." (Chomsky 1975:59)

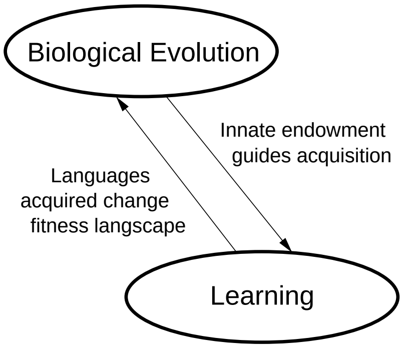

Biological Evolution
Innate endowment guides acquisition
Languages acquired change fitness langscape
Learning

Figure 2: The standard adaptationist model. Language is a consequence of the interaction between biological evolution and learning. The innate language capacity guides language acquisition. The functionality of the acquired language then has consequences for the biological evolution of the language capacity.

However, others have been less reticent in attempting to integrate the Chomskyan paradigm with evolutionary theory. Pinker & Bloom (1990) present the classic adaptationist account of language evolution, suggesting that "the ability to use a natural language belongs more to the study of human biology than human culture: it is a topic like echolocation in bats"(Pinker & Bloom 1990:707). They argue that language is adapted for the communication of propositional structures (in the internal representational "language of thought") over a serial channel. UG and the LAD have therefore evolved to facilitate the acquisition of language which performs this function.

Pinker & Bloom's (1990) account calls upon the interaction between two complex adaptive systems to explain the language capacity and linguistic structure. The process of language acquisition, constrained by UG and guided by the LAD, determines an individual's linguistic competence. This competence then contributes to an individual's reproductive fitness, resulting in selection in favour of an innate endowment which 1) facilitates language acquisition and 2) constrains the learner to learning languages which are communicatively useful. Biological adaptation of UG and the LAD then feeds back into the language acquisition process. This interaction is illustrated in Figure 2.

2.2 Culture: A Third Complex System

Those working within the Chomskyan paradigm typically play down the role of learning and the cultural transmission of language. For Chomsky, learning of a language 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), while Piattelli-Palmarini has suggested that "we would gain in clarity if the *scientific* use of the term [learning] were simply discontinued"(Piattelli-Palmarini 1989:2). This devaluation of learning, and consequently cultural transmission, arises from concerns based on poverty of the stimulus arguments. If the linguistic stimulus available to the child is too impoverished to allow language acquisition, then much of the structure of language must be prespecified, and learning effectively plays no role. However, focusing on the nature of this innate knowledge of language, to the detriment of the study of the cultural transmission of language, means that we overlook an important dynamic which can help explain some of the fundamental structural properties of language.

Following ideas developed by Hurford (1990), we place an understanding of cultural evolution

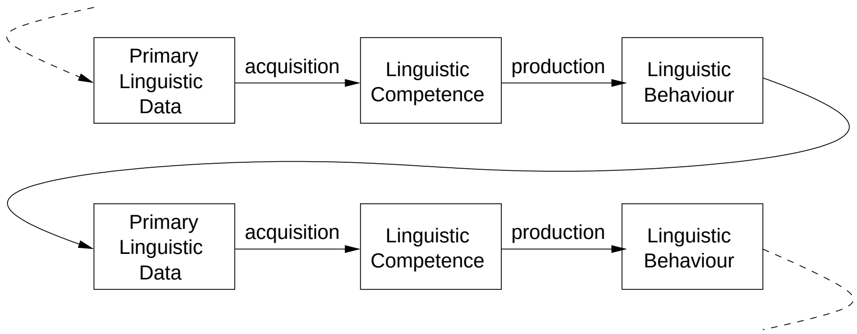


Figure 3: Language as a cultural phenomenon. As in the Chomskyan paradigm, illustrated in Figure 1, acquisition based on available data leads to linguistic competence. However, this competence in turn leads to linguistic behaviour, which becomes the linguistic data for the next generation of language learners.

at the heart of an explanatory approach. An individual's linguistic competence is derived from data which is itself a consequence of the linguistic competence of another individual. This view of language is illustrated in Figure 3.

What consequences does this view of language have for evolutionary explanations of language and the language faculty? The introduction of cultural transmission results in a third complex adaptive system, that of cultural evolution, operating on what has been dubbed a *glossogenetic* (Hurford 1990) time-scale, intermediate between the phylogenetic and ontogenetic time-scales. As in the standard adaptationist model, language acquisition is guided by an individual's innate endowment. The learner attempts to acquire the language of their cultural parents. Differences between the language of the parent and the child results in the cultural evolution of language itself. This cultural evolution further restricts the set of possible languages available to language learners at subsequent generations. The particular language acquired by a learner from the set of languages made available by this cultural evolution then contributes to the reproductive fitness of that individual, resulting in selection in favour of an innate endowment which 1) facilitates acquisition of those languages present in the culture and 2) constrains the learner to learning languages which are communicatively useful. Biological adaptation of UG and the LAD then feeds back into the language acquisition process. This interaction is illustrated in Figure 4.

The structure of language is dependent on the interaction between these three complex systems, and a full understanding of language evolution will require a treatment of all three adaptive processes. However, it is prudent to develop an understanding of each process in isolation before attempting to formulate a complete, unified model of the evolution of language. In this paper we will develop an account of the dynamics arising from the cultural transmission of language, then draw inferences as to how this dynamic might interact with the complex systems of individual learning and biological evolution.

3 THE ITERATED LEARNING MODEL

The Iterated Learning Model (ILM), as introduced in Kirby (2001) and Brighton (2002), provides a framework for studying the cultural evolution of language on a glossogenetic time-scale. The ILM in its simplest form is illustrated in Figure 5. In this model H_i corresponds to the linguistic competence of individual i, whereas U_i corresponds to the linguistic behaviour of individual i and

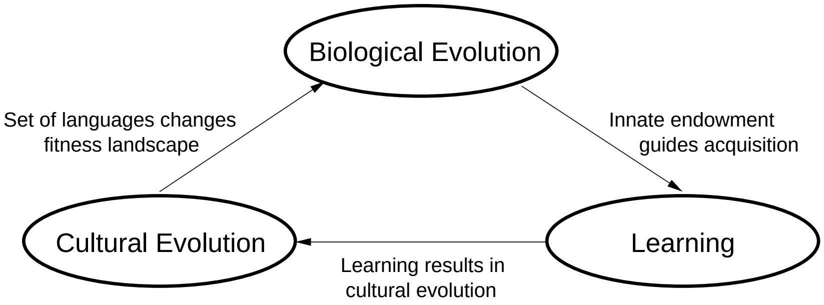

Biological Evolution
Set of languages changes fitness landscape
Innate endowment guides acquisition
Cultural Evolution
Learning
Learning results in cultural evolution

Figure 4: Adding cultural evolution. Language is now a consequence of the interaction between biological evolution, learning and cultural evolution. The innate language capacity guides language acquisition. The cultural transmission of language leads to cultural evolution, which then has consequences for the biological evolution of the language capacity.

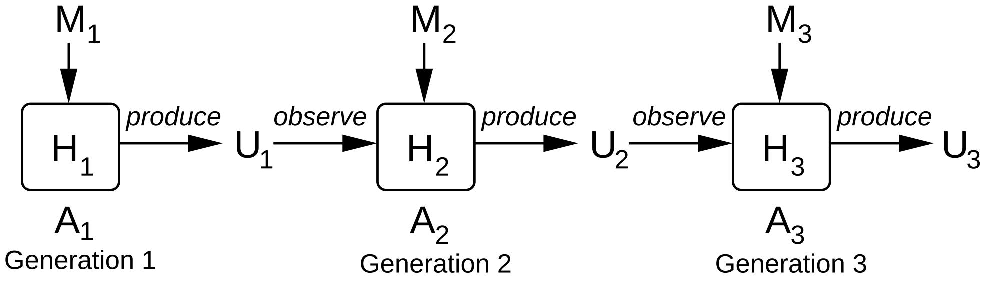


Figure 5: The ILM. In the simplest case, the ith generation of the population consists of a single agent A_i who has hypothesis H_i. Agent A_i is prompted with a set of meanings M_i. For each of these meanings the agent produces an utterance using H_i. This yields a set of utterances U_i. Agent A_{i+1} observes U_i and forms a hypothesis H_{i+1} to explain the set of observed utterances, and the cycle repeats.

the primary linguistic data for individual $i + 1$.

We focus here on the cultural evolution of language in the absence of any functional pressure for effective communication. While it has been suggested that functional considerations have an impact on language acquisition and production (for example, Hawkins (1990) suggests a preference by speaker/hearers for sentences which are easy to parse), ignoring such pressures allows us to make several simplifying assumptions. We can treat the population at any given generation as consisting of a single agent. This means that we can focus fully on vertical cultural transmission, and ignore for the moment horizontal, within-generation transmission. We can also ignore inter-generational communication. However, the ILM does not rule out a focus on the communicative function of language within or between generations in a population (see, for example, Smith (2002)) or the role of horizontal transmission (see Batali (2002) for an ILM where transmission is purely horizontal).

The ILM provides a powerful framework for investigating the cultural evolution of language. We have previously used the ILM to examine the emergence of word-order universals (Kirby 1999), the regularity–irregularity distinction (Kirby 2001) and recursive syntax (Kirby 2002). Here we will focus on the cultural evolution of compositionality, one of the characteristic structural properties of language.

4 MODELLING THE EVOLUTION OF COMPOSITIONALITY

4.1 Compositionality

In a compositional communication system the meaning of a signal is a function of the meaning of its parts (Krifka 2001). The morphosyntax of human language is highly compositional. For example, the relationship between the string *John walked* and its meaning is not completely arbitrary. It is made up of two components: a noun (*John*) and a verb (*walked*). The verb is also made up of two components: a stem and a past-tense ending. The meaning of *John walked* is thus a function of the meaning of its parts. Compositionality, in combination with recursive syntax, allows language users to produce and comprehend an infinite range of sentences.

Compositional language can be contrasted with non-compositional, or *holistic* communication, where a signal stands for the meaning as a whole, with no subpart of the signal conveying any part of the meaning in and of itself. Animal communication is typically viewed as holistic — no subpart of an alarm call or a mating display stands for part of the meaning "there's a predator about" or "come and mate with me".

How can we explain the compositionality of language? In the standard adaptationist model, compositionality must be viewed as a consequence of a biological adaptation of the unique human language organ — natural selection has favoured an innate endowment which restricts language learners to learning compositional systems. However, we demonstrate that compositionality can arise through purely cultural processes, as a result of the adaptation of language in the face of pressure to be learnable. This lifts some of the burden of explanation from the postulated language organ — cultural processes acting on a (possibly domain-general) biological substrate results in compositional language.

4.2 Necessary Elements of a Model

We treat language as a mapping between meanings and signals. A compositional language is a mapping which preserves neighbourhood relationships — neighbouring meanings will share structure, and that shared structure in meaning space will map to shared structure in signal space. A holistic language is one which does not preserve such relationships — as the structure of signals does not reflect the structure of the underlying meaning, shared structure in meaning space will not necessarily result in shared signal structure.

In order to model such systems we need representations of meanings and signals. Meanings are represented as points in an F-dimensional space where each dimension has V discrete values, and signals are represented as strings of characters of length 1 to l_{max}, where the characters are drawn from some alphabet Σ. More formally, the meaning space $\mathcal{M}$ and signal-space $\mathcal{S}$ are given by:

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

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

The world, which provides communicatively relevant situations for agents in our model, consists of a set of objects, where each object is labelled with a meaning drawn from $\mathcal{M}$. We will refer to such a set of labelled objects as an *environment*.

Utterances, the units of observable behaviour that individuals acquire their competence from, are considered to be meaning-signal pairs $\langle m, s\rangle$, where $m \in \mathcal{M}$ and $s \in \mathcal{S}$. We therefore assume that learners are able to deduce the communicative intentions of others during language acquisition. This is obviously an oversimplification — if the meaning of every signal was self-evident then the signal itself would serve little purpose. However, we can make several points in defence of this idealisation. Firstly, children do seem to have various strategies for deducing the meaning underlying an observed signal. Central to these abilities is the capacity to establish joint attention and perform intentional inference (Baldwin 1991; Bloom 2000). Secondly, computational simulations outlined in Hurford (1999) show that linguistic structure can be preserved if this idealisation is weakened, so that learners observe only partially-specified meanings in conjunction with signals. Finally, a strand of research parallel and complementary to our own abandons this idealisation completely (for example, Steels (1997), Steels (1998), Smith (2001), and Steels *et al.* (2002)). This work shows that shared linguistic structure can still emerge in the absence of explicit meaning transfer during learning.

We make a fundamental distinction between meanings and signals. An alternative approach is presented in Hashimoto (1998), who also treats language as a dynamical system but makes no distinction between the meanings of words and the words themselves. From this standpoint, the meaning of a signal is defined in terms of the relationship between that signal and other signals — the mesh of word-word associations constrains and guides the interpretation of signals. Language lies somewhere between these two extremes.

4.3 An Analysis of Stable States

Brighton (2002) presents a mathematical analysis of the relative stability of compositional and holistic language over cultural time. Brighton considers only perfectly compositional and completely holistic language. Addressing the question of relative stability allows us to predict when we should observe linguistic structure — when compositional and holistic language are equally stable we should expect them to emerge with equal frequency over cultural time, whereas when one type of language is more stable than the other we should expect that language to emerge more frequently and persist for longer.

One of Brighton's observations is that, in an iterated learning scenario, stability of a language over cultural time relates to the expressivity of learners exposed to that language. Consider the problem faced by a learner attempting to learn a holistic language. Given the lack of structure in the holistic language, the best strategy for the learner is simply to memorise meaning-signal pairs. The learner, when called upon to produce an utterance, will only be able to faithfully reproduce meaning-signal pairs that it itself has observed. Parts of the language which have not been observed cannot be expressed and will therefore be lost or will change — holistic language is only stable when the learner observes, and is therefore able to express, the complete language of the previous generation.

In contrast, the structure of a compositional language means that learners can acquire and express the complete language based on observation of a subset of that language. Consider a learner presented with a perfectly compositional language. In such a language each feature value will map onto a particular subsignal. The best strategy here is to memorise feature value–subsignal pairs. When called upon to produce an utterance for a given meaning, an individual is not restricted to reproducing meaning-signal pairs it itself has observed. A meaning will be expressible if every feature value of that meaning has been observed paired with a subsignal. A learner of a compositional language can therefore generalise from observed examples to express parts of the language that it has not actually observed, therefore the language will remain stable.

Brighton's key result is to show that the stability advantage of compositional language over holistic language is at a maximum when there is a *bottleneck* on cultural transmission. The transmission bottleneck occurs when learners only observe a subset of the language of the previous generation. This is one aspect of the poverty of the stimulus problem — the set of utterances of any human language is arbitrarily large, but a child must acquire their linguistic competence based on a finite number of sentences. The severity of the transmission bottleneck is given by the proportion of the language of the previous generation that a learner will observe.

Holistic languages cannot persist over time when the bottleneck on cultural transmission is tight — learners can only faithfully reproduce parts of the language which they have observed, and if they observe only a small subset of the language then the language will be unstable. In contrast, compositional languages are generalisable, due to their structure, and remain relatively stable even when a learner only observes a small subset of the language of the previous generation. Brighton shows that the poverty of the stimulus "problem" is actually a requirement for linguistic structure — were there no poverty of the stimulus, compositional language would have no advantage over unstructured holistic language.

4.4 A Model of Language Dynamics

Brighton's result is a fundamental one. However, by considering only perfectly compositional or completely holistic languages, Brighton is restricted to examining the Lyapounov stable states, places in language space that, if we start near, we stay near (Glendinning 1994). The model cannot explain the dynamics that occur when we move away from the extremes of compositionality, although insights taken from the model prove relevant to understanding the behaviour of dynamic models.

What happens to languages of intermediate compositionality during cultural transmission? Can compositional language emerge from initially holistic language, through a process of cultural evolution? We can investigate these question using techniques from artificial life, by developing a multi-agent computational implementation of the ILM. A neural network model of a linguistic agent, based on a simple model described in Smith (2002), is outlined below. This agent is inserted into the ILM, along with a model of environments, allowing us to model the dynamics arising from the cultural transmission of language.

4.4.1 A network model of a linguistic agent

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

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

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

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

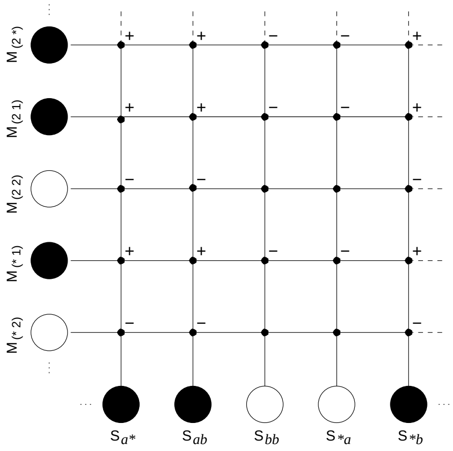


Figure 6: Storage of the meaning-signal pair $\langle (2\ 1), ab\rangle$. Nodes with an activation of 1 are represented by large filled circles, and are labelled with the component they represent. For example, $M_{(2\ *)}$ is the node which represents the meaning component $(2\ *)$. Small filled circles represent weighted connections. During the learning process, nodes representing components of $(2\ 1)$ and ab have their activations set to 1. Connection weights are then either incremented $(+)$, decremented $(-)$ or left unchanged.

Each node in $\mathcal{N}_M$ represents a component of a meaning, and there is a single node in $\mathcal{N}_M$ for each component of every possible meaning. Similarly, each node in $\mathcal{N}_S$ represents a component of a signal and there is a single node in $\mathcal{N}_S$ for each component of every possible signal.

Learning During a learning event, a learner observes a meaning-signal pair $\langle m, s\rangle$. The activations of the nodes corresponding to all possible components of m and all possible components of s are set to 1. The activations of all other nodes are set to 0. The weights of the connections in $\mathcal{W}$ are adjusted according to the weight-update rule:

$$\Delta W_{xy} = \begin{cases} +1 & \text{iff } a_x = a_y = 1 \\ -1 & \text{iff } a_x \neq a_y \\ \ \ 0 & \text{otherwise} \end{cases}$$

where W_{xy} gives the weight of the connection between nodes x and y and a_x gives the activation of node x. The learning procedure is illustrated in Figure 6.

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

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

The first condition states that an analysis may not consist of a set of components which all leave a particular feature unspecified — an analysis fully specifies a meaning. The second states that an analysis may not consist of a set of components where more than one component specifies

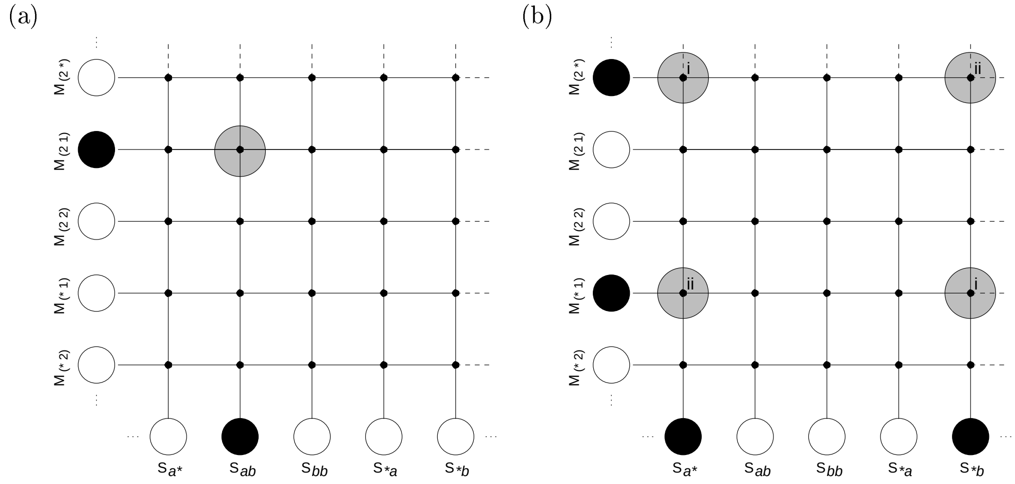


Figure 7: Retrieval of three possible analyses of $\langle (2\ 1), ab\rangle$. The relevant connection weights are highlighted in grey. (a) g for the one-component analysis $\langle \{(2\ 1)\}, \{ab\}\rangle$ depends on the weight of the connection between the nodes representing the components (2 1) and ab. (b) g for the two-component analysis $\langle \{(2\ *), (*\ 1)\}, \{a*, *b\}\rangle$ depends on the weighted sum of two connections, marked as i. The g for the alternative two-component analysis $\langle \{(2\ *), (*\ 1)\}, \{*b, a*\}\rangle$ is given by the weighted sum of the two connections marked ii.

the value of a particular feature — analyses do not contain redundant components. Valid analyses of signals are similarly defined.

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

$$g\left(\left\{c_m^1, c_m^2 \ldots c_m^N\right\}, \left\{c_s^1, c_s^2 \ldots c_s^N\right\}\right) = \sum_{i=1}^{N} \omega\left(c_m^i\right) \cdot W_{c_m^i, c_s^i}$$

where N is the number of components in the analysis of meaning and signal, $w_{c_m^i, c_s^i}$ gives the weight of the connection between the nodes representing the ith component of the meaning analysis and the ith component of the signal analysis and $\omega(x)$ is a weighting function which gives the non-wildcard proportion of x. The production process is illustrated in Figure 7.

4.4.2 Environment structure

An environment consists of a set of objects labelled with meanings drawn from $\mathcal{M}$. The number of objects in the environment gives the *density* of that environment — environments with few objects will be termed low-density, whereas environments with a large number of objects will be termed

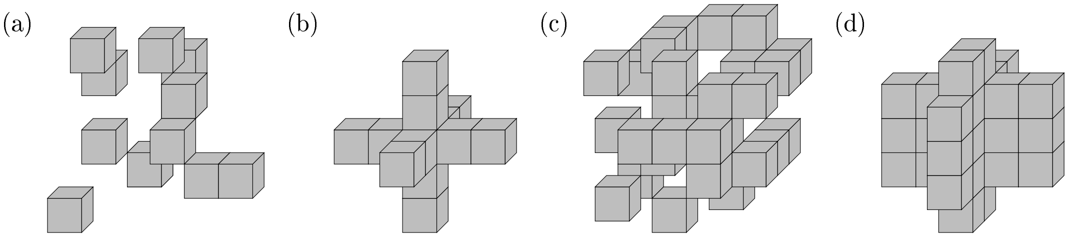


Figure 8: We will present results for the case where $F = 3$ and $V = 5$. This defines a three-dimensional meaning space. We highlight the meanings selected from that space with grey. (a) is a low-density, unstructured environment. (b) is a low-density, structured environment. (c) and (d) are unstructured and structured high-density environments.

high-density. When meanings are assigned to objects at random we will say the environment is *unstructured*. When meanings are assigned to objects in such a way as to minimise the average inter-meaning Hamming distance we will say the environment is *structured*. Sample low- and high-density environments are shown in Figure 8.

4.4.3 Measuring compositionality

As discussed above, a compositional mapping preserves neighbourhood relations when mapping between meanings and signals, whereas a holistic mapping does not, unless by chance. Our measure of compositionality, c, captures this and is calculated based on the set of meaning-signal pairs in an agent's language. c is the Pearson's Product-Moment correlation coefficient of the pairwise distances between pairs of meanings and the distance between the corresponding pairs of signals. We use the Hamming and Levenstein (string edit) distance measures to quantify inter-meaning and inter-signal distances respectively. c ranges between -1 and 1. A perfectly compositional language will have a c of 1, whereas $c \approx 0$ for holistic languages.

5 RESULTS

The network model of a linguistic agent outlined above is plugged into the ILM framework described in Section 3. We will vary two key parameters — the presence or absence of a bottleneck on cultural transmission, and the density and structure of the environment.

5.1 No Bottleneck on Cultural Transmission

First, runs of the ILM were carried out in the absence of a bottleneck on cultural transmission — each learner is presented with the complete language of the agent at the previous generation. The initial agents had all their connections weights set to 0, and therefore produced every meaning-signal pair with equal probability. Subsequent agents had connection weights of 0 prior to learning. Runs were allowed to progress until a stable state was reached, where agent A_i and A_{i+1} produced identical languages. At this point, in the absence of a bottleneck, further language change is impossible.

Figures 9 and 10 plot compositionality by frequency for the initial and final, stable languages for low-density and high-density environments. These results are based on 1000 independent runs of the ILM for each environment.

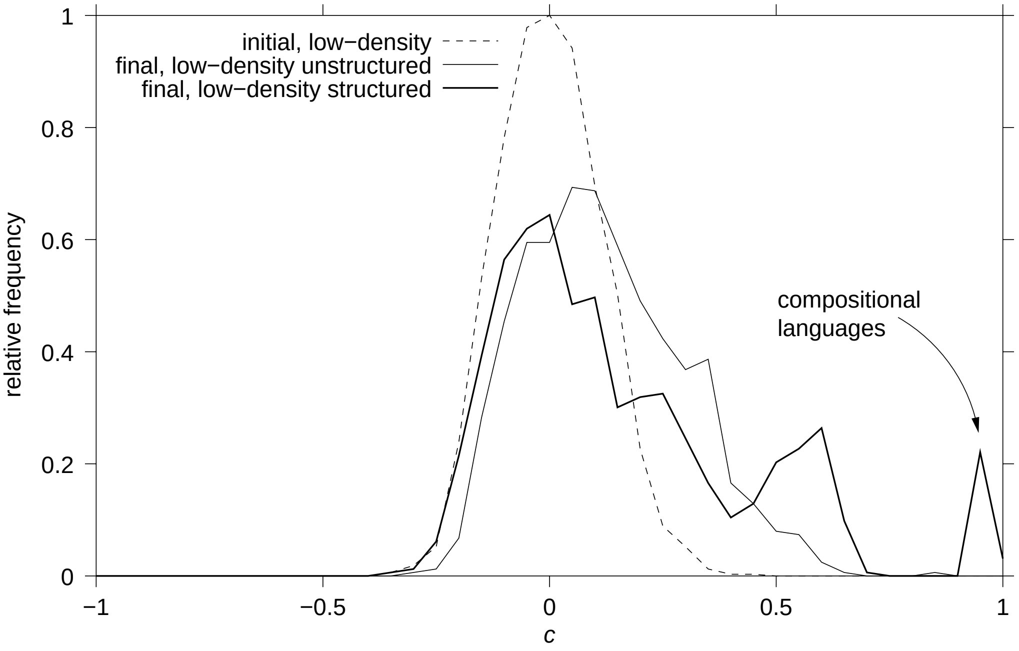


Figure 9: The relative frequency of initial and final systems of varying degrees of compositionality, where there is no bottleneck on cultural transmission. The results shown here are for the low-density environments given in Figure 8. The initial languages are largely holistic. Some final languages exhibit increased levels of compositionality. Highly compositional languages are infrequent.

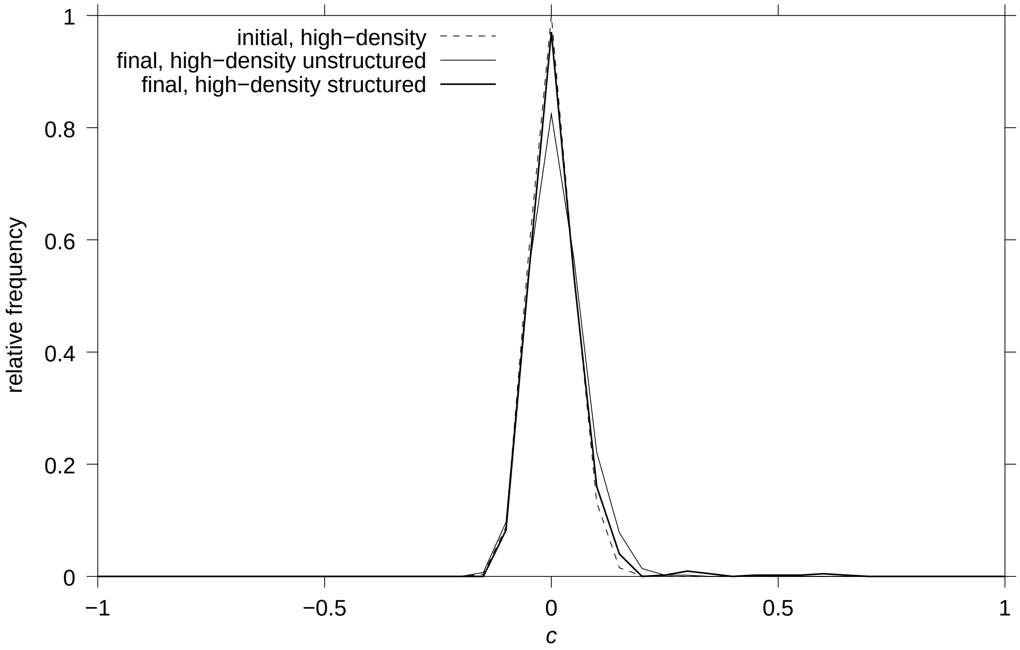


Figure 10: The relative frequency of initial and final systems of varying degrees of compositionality, where there is no bottleneck on cultural transmission. The results shown here are for the high-density environments given in Figure 8. Both the initial and final languages are holistic.

Environment	Initial Compositionality			
	all	$c_{initial} = c_{final}$	$c_{initial} \neq c_{final}$, $c_{final} < 0.8$	$c_{initial} \neq c_{final}$, $c_{final} \geq 0.8$
ld, u	$\mu = 0.0011$, $\sigma = 0.1135$	$\mu = -0.1484$	$\mu = 0.0076$	NA
ld, s	$\mu = -0.0002$, $\sigma = 0.1136$	$\mu = -0.0457$	$\mu = 0.0371$	$\mu = 0.0688$
hd, u	$\mu = 0.0004$, $\sigma = 0.0470$	$\mu = -0.0062$	$\mu = 0.0108$	NA
hd, s	$\mu = 0.0027$, $\sigma = 0.0454$	$\mu = -0.0002$	$\mu = 0.0101$	NA

Table I: Sensitivity to initial conditions. Environments are specified by a density — low-density (ld) or high-density (hd), and a degree of structure — unstructured (u) or structured (s). Simulation runs can be split into three groups, according to the values of $c_{initial}$ and c_{final}. The table gives the means (μ) of the initial language of the simulation runs, broken down by group. The runs in the low-density environments are sensitive to the compositionality of the initial system. The standard deviations (σ) is given once. σ for each subgroup is approximately the same as σ for the runs in that environment as a whole. The initial languages in the high-density environment are clustered more tightly around the mean.

Three results are apparent from these Figures:

1. Highly compositional systems are infrequent.
2. Compositional systems only occur when the environment is low-density.
3. Highly compositional systems only occur when the environment is structured.

The initial languages tend to be holistic. Brighton's (2002) results suggest that, with no bottleneck on cultural transmission, such systems will be highly stable. This seems to be the case for the majority of runs reported here, particularly in the high-density environment. The emergence of some partially or highly compositional systems in the low-density environments then seems somewhat surprising.

Individual simulation runs can be split into three groups – those where the final languages have the same level of compositionality as the initial languages ($c_{initial} = c_{final}$), those where the final compositionality is different from the compositionality of the initial language but not high ($c_{initial} \neq c_{final}$, $c_{final} < 0.8$), and those where the final systems are highly compositional ($c_{initial} \neq c_{final}$, $c_{final} \geq 0.8$). Table I gives the mean and standard deviations of the compositionality of the initial languages, organised into these three groups.

As can be seen from the first column of the Table, runs in all environments have a mean value of $c_{initial}$ of approximately 0. However, these initial values are much more tightly distributed around the mean in the high-density environments. The second column gives the mean $c_{initial}$ for simulation runs where $c_{initial} = c_{final}$. These are somewhat lower than the overall mean, and are lower than the mean $c_{initial}$ for simulation runs where $c_{initial} \neq c_{final}$. Finally, the mean $c_{initial}$ for simulation runs which converge on highly compositional systems is higher still.

These results suggest that, in the absence of a bottleneck on cultural transmission, there is a degree of sensitivity to the compositionality of the initial, random language. Where this initial system exhibits compositional tendencies, yielding $c_{initial}$ above the mean, there is an increased likelihood of the language moving, over iterated learning events, towards more compositional regions of language space. The compositional tendencies of the initial language spread to other parts of the language over time, resulting in an increase in compositionality. However, this progression is not guaranteed - not all simulation runs where $c_{initial}$ is above the mean eventually converge on more compositional systems. For the high-density environments partially or highly compositional systems do not emerge due to the fact that the initial systems tend to be clustered more tightly around the non-compositional mean. When the environment contains few meanings the initial language may, by chance, exhibit some compositional tendencies. However, when the environment contains a large number of meanings such tendencies are likely to be drowned out by the majority holistic mapping.

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

5.2 Bottleneck on Cultural Transmission

Next, runs of the ILM were carried out with a bottleneck on cultural transmission — each learner is presented with a subset of the language of the agent at the previous generation. The number of utterances produced by agents was set so that language learners observed utterances for approximately 40% of the language of the previous agent.

While in the absence of a bottleneck runs were allowed to proceed until a stable state was reached, in the bottleneck condition runs were terminated after a fixed number of generations (100). The random selection of objects from the environment for which to produce utterances means that, as with any stochastic system, a highly skewed distribution of objects could lead to the loss of structure. However, this is extremely unlikely. The results reported here accurately reflect the behaviour of the system — allowing the runs to proceed for several hundred more generations gives a similar distribution of languages.

Figure 11 plots the compositionality by frequency of the initial and final languages for unstructured and structured high-density environments. These results are based on 1000 independent runs of the ILM for each environment. Results are not shown for the low-density environments. At this severity of bottleneck, it is impossible for any learner to see every distinct feature value in the low-density environment, therefore no stable states exist.

Two results are apparent from Figure 11:

1. Highly compositional systems are frequent.
2. Highly compositional systems are most frequent when the environment exhibits structure.

Brighton's (2002) mathematical model predicts that, in the presence of a bottleneck on cultural transmission, compositional language will be more stable than holistic language. The results from the computational model bear this out, but also show that it is possible to move from an initially holistic system to a highly compositional system over time. Figure 12 illustrates the dynamics of the transition from holistic to compositional language. In the case where the environment is unstructured (Figure 12 (a)), there are two attractors — one where $c = 0.45$ and one where $c = 0.95$. This gives the bimodal distribution of systems for such environments seen in Figure 11. However, systems at these points are not highly stable — it is possible to move from one attractor

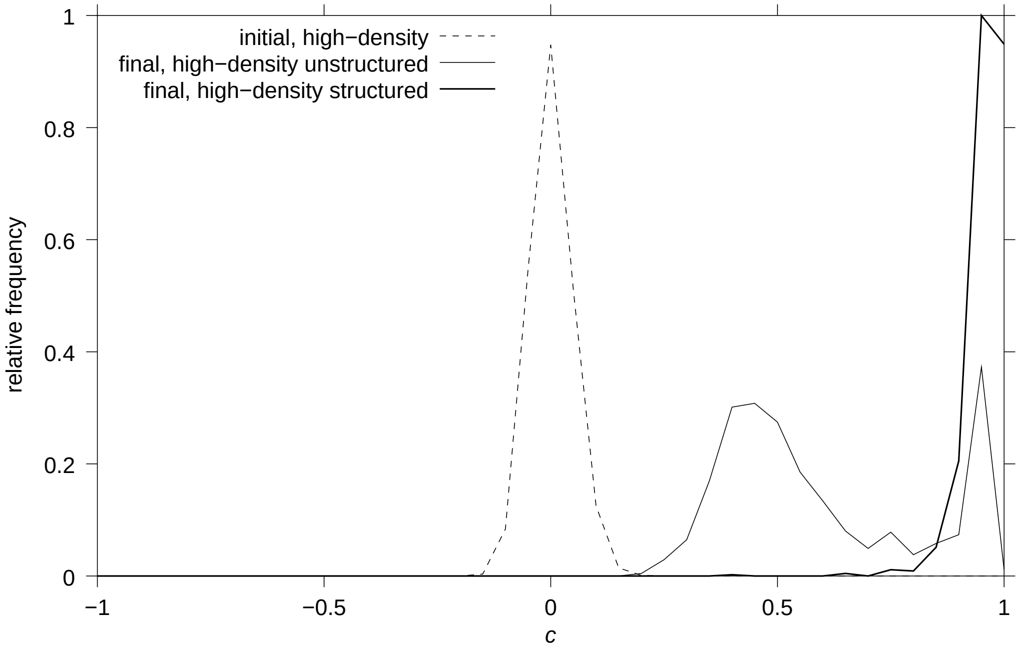


Figure 11: The relative frequency of initial and final systems of varying degrees of compositionality, where there is a bottleneck on cultural transmission. The results shown here are for the high-density environments. The initial languages are largely holistic. Highly compositional languages emerge with high frequency, and are most frequent when the environment is structured.

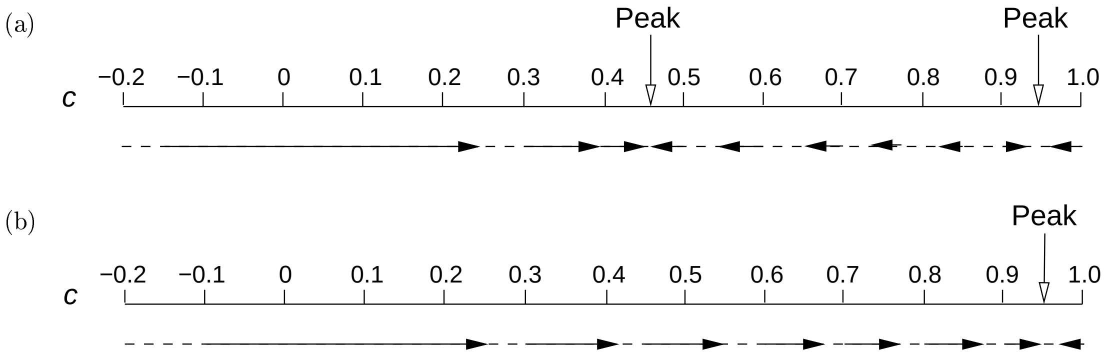


Figure 12: The dynamics of language change. Arrows represent the direction and magnitude of change of languages of a given level of compositionality, c. The origin of the arrow gives the compositionality of the language at time t. The direction and length of the arrow corresponds to the mean directionality and magnitude of change in compositionality for those systems at time $t + 1$. (a) Dynamic in the unstructured environment. There are two attractors, at $c = 0.45$ and $c = 0.95$, corresponding to the two peaks of the bimodal distribution given in Figure 11. Magnitude of change decreases as these attractors are approached. (b) Dynamics in the structured environment. There is a single attractor, at $c = 0.95$, corresponding to the peak in Figure 11. Again, magnitude of change decreases as this attractor is approached.

to another. In contrast, in the structured environment, as shown in Figure 12 (b), there is a single attractor at $c = 0.95$. Systems reaching this point are highly stable, and any perturbation away from the attractor is quickly reversed.

Why are compositional languages so strongly preferred when there is a bottleneck on transmission? Holistic languages cannot persist in the presence of a bottleneck. The meaning-signal pairs of a holistic language have to be observed to be reproduced. When a learner only observes a subset of the holistic language of the previous generation then certain meaning-signal pairs will not be preserved — the learner, when called upon to produce, will produce some other signal for that meaning, resulting in a change in the language. In contrast, compositional languages are generalisable, due to their structure, and remain relatively stable even when the learner observes a small subset of the language of the previous generation. Over time, language adapts to the pressure to be generalisable. Eventually, particularly when the environment is structured, the language becomes highly compositional, highly generalisable and consequently highly stable.

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

6 CONCLUSIONS

We have presented a multi-agent simulation model which demonstrates that compositional language can emerge from initially non-compositional language through purely cultural processes. Compositional language emerges when there is a bottleneck on cultural transmission — compositionality is an adaptation by language which allows it to slip through the transmission bottleneck. The advantage of compositionality is at a maximum when language learners perceive their world as structured — when the objects in the environment relate to one another in structured ways then a generalisable, compositional language is highly adaptive.

We are not, however, arguing that compositionality can be understood purely in terms of cultural evolution. The complex adaptive systems of learning and biological evolution still have a role to play. In the models described here, after exposure to a small set of utterances a learner's knowledge remains fixed. In the real case, however, an individual's knowledge and use of language is constantly changing and adapting, and this may impact on cultural evolution.

Similarly, we have offered no account of the biological evolution of the linguistic capacities of our simulated agents. Whereas the standard adaptationist model would hypothesise a complex, language-specific component of the brain designed to deal with compositionality, we make much weaker assumptions — a simple associative learning mechanism, in combination with a capacity to infer the communicative intentions of others, is sufficient to allow the cultural evolution of compositional language. It is not clear that either of these capacities are language specific, and an evolutionary account of their origins and development might be domain-general or exaptationist in flavour. The interaction between the evolution of this mental capacity and the ongoing cultural evolution of language is an exciting topic for future research, and computational modelling techniques will continue to be an invaluable tool in such research.

References

BALDWIN, D. A. 1991. Infants' contribution to the achievement of joint reference. *Child Development* 62.875–890.

BATALI, J. 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. Cambridge: Cambridge University Press.

BLOOM, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

BRIGHTON, H. 2002. Compositional syntax from cultural transmission. *Artifical Life* 8.25–54.

CHOMSKY, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

—— 1975. *Reflections on Language*. New York, NY: Pantheon.

—— 1980. *Rules and Representations*. London: Basil Blackwell.

—— 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

GELL-MANN, M. 1992. Complexity and complex adaptive systems. In *The Evolution of Human Languages*, ed. by J. A. Hawkins & M. Gell-Mann, 3–18. Reading, MA: Addison-Wesley.

GLENDINNING, P. 1994. *Stability, instability and chaos: an introduction to the theory of nonlinear differential equations*. Cambridge.

HASHIMOTO, T. 1998. Dynamics of internal and global structure through linguistic interactions. In *Multi-Agent Systems and Agent-Based Simulation*, ed. by J. S. Sichman, R. Conte, & N. Gilbert, 124–139. Berlin: Springer-Verlag.

HAWKINS, J. A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21.223–261.

HURFORD, J. R. 1990. Nativist and functional explanations in language acquisition. In *Logical Issues in Language Acquisition*, ed. by I. M. Roca, 85–136. Dordrecht: Foris.

—— 1999. Language learning from fragmentary input. In *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*, ed. by K. Dautenhahn & C. Nehaniv, 121–129. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

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

—— 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation* 5.102–110.

—— 2002. 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.

KRIFKA, M. 2001. Compositionality. In *The MIT Encyclopaedia of the Cognitive Sciences*, ed. by R. A. Wilson & F. Keil. Cambridge, MA: MIT Press.

PIATTELLI-PALMARINI, M. 1989. Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language. *Cognition* 31.1–44.

PINKER, S. 1994. *The Language Instinct*. Penguin.

——, & P. BLOOM. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13.707–784.

SMITH, A. D. M. 2001. Establishing communication systems without explicit meaning transmission. In *Advances in Artificial Life: Proceedings of the* 6^{th} *European Conference on Artifical Life*, ed. by J. Kelemen & P. Sosik. Berlin: Springer-Verlag.

SMITH, K. 2002. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.

STEELS, L. 1997. Constructing and sharing perceptual distinctions. In *Proceedings of the European Conference on Machine Learning, ECML '97*, ed. by M. van Someren & G. Widmer, 4–13. Berlin: Springer-Verlag.

—— 1998. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103.133–156.

——, F. KAPLAN, A. MCINTYRE, & J. VAN LOOVEREN. 2002. Crucial factors in the origins of word-meaning. In *The Transition to Language*, ed. by A. Wray, 252–271. Oxford: Oxford University Press.

SMITH, K. 2001b. The importance of rapid cultural convergence in the evolution of learned symbolic communication. In Kelemen & Sosík (2001), 637–640.

SMITH, K., 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.

Iterated Learning: a framework for the emergence of language

Kenny Smith[1], Simon Kirby[1], Henry Brighton[1]

[1] Language Evolution and Computation Research Unit, Theoretical and Applied Linguistics, University of Edinburgh, Adam Ferguson Building, 40 George Square, Edinburgh, UK. {kenny,simon,henryb}@ling.ed.ac.uk

To appear in "Self-organization and Evolution of Social Behaviour", edited by C. Hemelrijk.

Language is culturally transmitted. Iterated Learning, the process by which the output of one individual's learning becomes the input to other individuals' learning, provides a framework for investigating the cultural evolution of linguistic structure. We present two models, based upon the Iterated Learning framework, which show that the poverty of the stimulus available to language learners leads to the emergence of linguistic structure. Compositionality is language's adaptation to stimulus poverty.

Linguists traditionally view language as the consequence of an innate "language instinct" (Pinker 1994). It has been suggested that this language instinct evolved, via natural selection, for some social function — perhaps to aid the communication of socially relevant information such as possession, beliefs and desires (Pinker & Bloom 1990), or to facilitate group cohesion (Dunbar 1996). However, the view of language as primarily a biological trait arises from the treatment of language learners as isolated individuals. We argue that language should be more properly treated as a culturally transmitted system. Pressures acting on language during its cultural transmission can explain much of linguistic structure. Aspects of language which appear baffling when viewed from the standpoint of individual acquisition emerge straightforwardly if we take the cultural context of language acquisition into account. While we are sympathetic to attempts to explain the biological evolution of the language faculty, we agree with Jackendoff that "[i]f some aspects of linguistic behaviour can be predicted from more general considerations of the dynamics of communication [or cultural transmission] in a community, rather than from the linguistic capacities of individual speakers, then they should be"(Jackendoff 2002:101).

We present the Iterated Learning Model as a tool for investigating the cultural evolution of language. Iterated Learning is the process by which one individual's competence is acquired on the basis of observations of another individual's behaviour, which is determined by that individual's competence. This model of cultural transmission has proved particularly useful in studying the evolution of language. The primary goal of this paper is to introduce the notion of Iterated Learning and demonstrate that it provides a new adaptive mechanism for language evolution. Language itself can adapt on a cultural timescale, and the process of language adaptation leads to the characteristic structure of language. To this end, we present two models. Both attempt to explain the emergence of *compositionality*, a fundamental structural property of language. In doing so they demonstrate the utility of the Iterated Learning approach to the investigation of language origins and evolution.

In a compositional system the meaning of a signal is a function of the meaning of its parts and they way they are put together (Krifka 2001). The morphosyntax of language exhibits a high degree of compositionality. For example, the relationship between the string *John walked* and its meaning is not completely arbitrary. It is made up of two components: a noun (*John*) and a verb (*walked*). The verb is also made up of two components: a stem and a past-tense ending. The meaning of *John walked* is thus a function of the meaning of its parts.

The syntax of language is recursive — expressions of a particular syntactic category can be embedded within larger expressions of the same syntactic category. For example, sentences can be embedded within sentences — the sentence *John walked* can be embedded within the larger sentence *Mary said John walked*, which can in turn be embedded within the sentence *Harry claimed that Mary said John walked* and so on. Recursive syntax allows the creation of an infinite number of utterances from a small number of rules. Compositionality makes the interpretation of previously-unencountered utterances possible — knowing the meaning of the basic elements and the effects associated with combining them enables a user of a compositional system to deduce the meaning of an infinite set of complex utterances.

Compositional language can be contrasted with non-compositional, or *holistic* communication, where a signal stands for the meaning as a whole, with no subpart of the signal conveying any part of the meaning in and of itself. Animal communication is typically viewed as holistic — no subpart of an alarm call or a mating display stands for part of the meaning "there's a predator about" or "come and mate with me". Wray (1998) suggests that the protolanguage of early hominids was also holistic. We argue that Iterated Learning provides a mechanism for the transition from holistic protolanguage to compositional language.

In the first model presented in this paper, insights gained from the Iterated Learning framework suggest a mathematical analysis. This model predicts when compositional language will be more stable than non-compositional language. In the second model, techniques adopted from artificial life are used to investigate the transition, through purely cultural processes, from non-compositional to compositional language. These models reveal two key determinants of linguistic structure:

Stimulus poverty: *The poverty of the stimulus available to language learners during cultural transmission drives the evolution of structured language — without this stimulus poverty, compositional language will not emerge.*

Structured semantic representations: *Compositional language is most likely to evolve when linguistic agents perceive the world as structured — structured pre-linguistic repre-*

sentation facilitates the cultural evolution of structured language.

Two views of language

In the dominant paradigm in linguistics (formulated and developed by Noam Chomsky, for example Chomsky (1965) and Chomsky (1995)), language is viewed as an aspect of individual psychology. The object of interest is the internal linguistic competence of the individual, and how this linguistic competence is derived from the noisy fragments and deviant expressions of speech children observe. External linguistic behaviour (the set of sounds an individual actually produces during their lifetime) is considered to be epiphenomenal, the uninteresting consequence of the application of this linguistic competence to a set of contingent communicative situations. This framework is sketched in Figure 1 (a). From this standpoint, much of the structure of language is puzzling — how do children, apparently effortlessly and with virtually universal success, arrive at a sophisticated knowledge of language from exposure to sparse and noisy data? In order to explain language acquisition in the face of this poverty of the linguistic stimulus, the Chomskyan program postulates a sophisticated, genetically-encoded language organ of the mind, consisting of a Universal Grammar, which delimits the space of possible languages, and a Language Acquisition Device, which guides the "growth of cognitive structures [linguistic competence] along an internally directed course under the triggering and partially shaping effect of the environment" (Chomsky 1980:34). Universal Grammar and the Language Acquisition Device impose structure on language, and linguistic structure is explained as a consequence of some innate endowment.

Following ideas developed in Hurford (1990), we view language as an essentially cultural phenomenon. An individual's linguistic competence is derived from data which is itself a consequence of the linguistic competence of another individual. This framework is sketched in Fig. 1 (b). Under this view, the burden of explanation is lifted from the postulated innate language organ — much of the structure of language can be explained as a result of pressures acting on language during the repeated production of linguistic forms and induction of linguistic competence on the basis of these forms. In this paper we will show how the poverty of the stimulus available to language learners is the cause of linguistic structure, rather than a problem for it.

The Iterated Learning Model

The Iterated Learning Model, as introduced in Kirby (2001) and Brighton (2002), provides a framework for studying the cultural evolution of language. The Iterated Learning Model in its simplest form is illustrated in Fig. 2. In this model the hypothesis H_i corresponds to the linguistic competence of individual i, whereas the set of utterances U_i corresponds to the linguistic behaviour of individual i and the primary linguistic data for individual $i+1$.

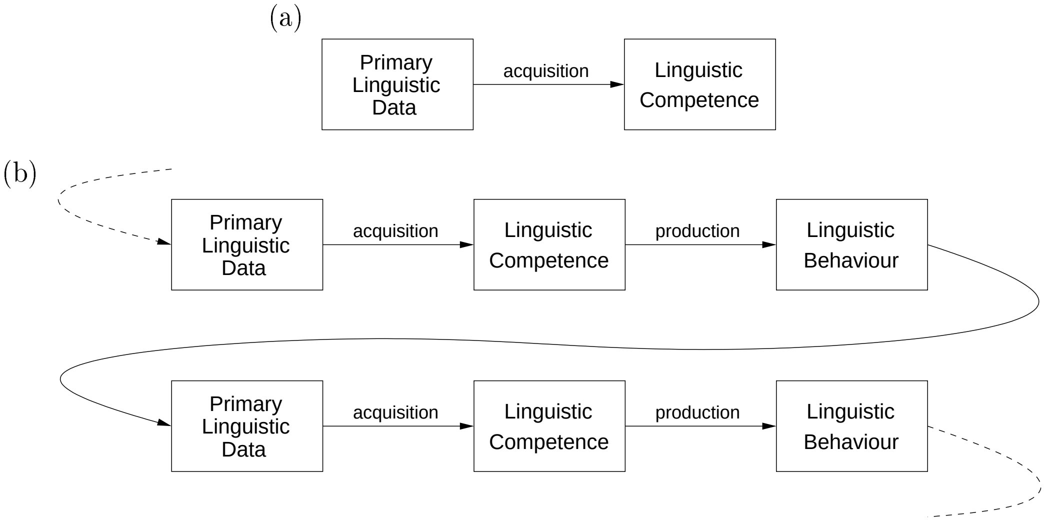


Figure 1: (a) The Chomskyan paradigm. Acquisition procedures, constrained by Universal Grammar and the Language Acquisition Device, derive linguistic competence from linguistic data. Linguistic behaviour is considered to be epiphenomenal. (b) Language as a cultural phenomenon. As in the Chomskyan paradigm, acquisition based on linguistic data leads to linguistic competence. However, we now close the loop — competence leads to behaviour, which contributes to the linguistic data for the next generation.

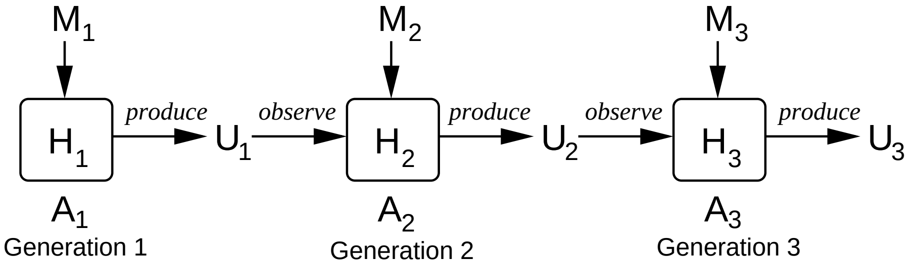


Figure 2: The Iterated Learning Model. The ith generation of the population consists of a single agent A_i who has hypothesis H_i. Agent A_i is prompted with a set of meanings M_i. For each of these meanings the agent produces an utterance using H_i. This yields a set of utterances U_i. Agent A_{i+1} observes U_i and forms a hypothesis H_{i+1} to explain the set of observed utterances. This process of observation and hypothesis formation constitutes learning.

We make the simplifying idealisation that cultural transmission is purely vertical — there is no horizontal, intra-generational cultural transmission. This simplification has several consequences. Firstly, we can treat the population at any given generation as consisting of a single individual. Secondly, we can ignore the intra-generational communicative function of language. However, the Iterated Learning framework does not rule out either intra-generational cultural transmission (see Livingstone & Fyfe (1999) for an Iterated Learning Model with both vertical and horizontal transmission, or Batali (2002) for an Iterated Learning Model where transmission is purely horizontal) or a focus on communicative function (see Smith (2002b) for an Iterated Learning Model focusing on the evolution of optimal communication within a population).

In most implementations of the Iterated Learning Model, utterances are treated as meaning-signal pairs. This is obviously an oversimplification of the task facing language learners — if the meaning of every signal were self-evident then the signal itself would be rather pointless. However, empirical evidence suggests that language learners have a variety of strategies for deducing the communicative intentions of others during language acquisition (see Bloom (2000) for review). We will assume for the moment that these strategies are error-free, while noting that the consequences of weakening this assumption is a current and interesting area of research (see, for example, Steels (1998), Smith (2001) and Steels *et al.* (2002)).

This simple model proves to be a powerful tool for investigating the cultural evolution of language. While we have previously used the Iterated Learning Model to explain the emergence of particular word-order universals (Kirby 1999), the regularity-irregularity distinction (Kirby 2001), and recursive syntax (Kirby 2002), here we will focus on the evolution of compositionality. The evolution of compositionality provides a test-case to evaluate the suitability of techniques from mathematics and artificial life in general, and the ILM in particular, to tackling problems from linguistics.

The cultural evolution of compositionality

We view language as a mapping between meanings and signals. A compositional language is a mapping which preserves neighbourhood relationships — neighbouring meanings will share structure, and that shared structure in meaning space will map to shared structure in the signal space. For example, the sentences *John walked* and *Mary walked* have parts of an underlying semantic representation in common (the notion of someone having carried out the act of walking at some point in the past) and will be near one another in semantic representational space. This shared semantic structure leads to shared signal structure (the inflected verb *walked*) — the relationship between the two sentences in semantic and signal space is preserved by the compositional mapping from meanings to signals. A holistic language is one which does not preserve such relationships — as the structure of signals does not reflect the structure of the underlying meaning, shared structure in meaning space will not necessarily result in shared signal structure.

In order to model such systems we need representations of meanings and signals. For both models outlined in this paper meanings are represented as points in an F-dimensional space where each dimension has V discrete values, and signals are represented as strings of characters of length 1 to l_{max}, where the characters are drawn from some alphabet Σ. More formally, the meaning space $\mathcal{M}$ and signal-space $\mathcal{S}$ are given by:

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

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

The world, which provides communicatively relevant situations for agents in our models, consists of a set of N objects, where each object is labelled with a meaning drawn from the meaning space $\mathcal{M}$. We will refer to such a set of labelled objects as an *environment*.

In the following sections two Iterated Learning Models will be presented. In the first model a mathematical analysis shows that compositional language is more stable than holistic language, and therefore more likely to emerge and persist over cultural time, in the presence of stimulus poverty and structured semantic representations. In the second model, computational simulation demonstrates that compositional language can emerge from an initially holistic system. Compositional language is most likely to evolve given stimulus poverty and a structured environment.

A mathematical model

We will begin by considering, using a mathematical model[1], how the compositionality of a language relates to its stability over cultural time. For the sake of simplicity, we will restrict ourselves to looking at the two extremes on the scale of compositionality, comparing the stability of perfectly compositional language and completely holistic language.

Learning holistic and compositional languages

We can construct a holistic language L_h by simply assigning a random signal to each meaning. More formally, each meaning $m \in \mathcal{M}$ is assigned a signal of random length l ($1 \leq l \leq l_{max}$) where each character is selected at random from Σ. The meaning-signal mapping encoded in this assignment of meanings to signals will not preserve neighbourhood relations, unless by chance.

Consider the task facing a learner attempting to learn the holistic language L_h. There is no structure underlying the assignment of signals to meanings. The best strategy here is simply to memorise meaning-signal associations. We can calculate the expected number of meaning-signal pairs our learner will observe and memorise. After R observations of randomly-selected objects paired with signals a learner will have a set of O observed meanings. We can calculate the probability that any arbitrary meaning $m \in \mathcal{M}$ will be included in O, $Pr\,(m \in O)$, with:

[1] This model is described in greater detail in Brighton (2002).

$$Pr\left(m \in O\right) = \sum_{x=1}^{N} \left(\begin{array}{l} \text{Probability that} \\ m \text{ is used to label} \\ x \text{ objects} \end{array} \right) \times \left(\begin{array}{l} \text{Probability of observing an} \\ \text{utterance being produced} \\ \text{for at least one of those } x \\ \text{objects after } R \text{ observations} \end{array} \right)$$

When called upon to produce utterances, such a learner will only be able to reproduce meaning-signal pairs they themselves observed. Given the lack of structure in the meaning-signal mapping, there is no way to predict the appropriate signal for a meaning unless that meaning-signal pair has been observed. We can therefore calculate E_h, the expected number of meanings an individual will be able to express after observing some subset of a holistic language, which is simply the probability of observing any particular meaning multiplied by the number of possible meanings:

$$E_h = Pr\left(m \in O\right) \cdot V^F$$

We can perform similar calculations for a learner attempting to acquire a perfectly compositional language. As discussed above, a perfectly compositional language preserves neighbourhood relations in the meaning-signal mapping. We can construct such a language L_c for a given set of meanings $\mathcal{M}$ using a lookup table of subsignals (strings of characters which form part of a signal), where each subsignal is associated with a particular feature value. For each $m \in \mathcal{M}$ a signal is constructed by concatenating the appropriate subsignal for each feature value in m.

How can a learner best acquire such a language? The optimal strategy is to memorise feature value-signal substring pairs. After observing R randomly selected objects paired with signals, our learner will have acquired a set of observations of feature values for the ith feature, O_{f_i}. The probability that an arbitrary feature value v in included in O_{f_i} is given by $Pr\left(v \in O_{f_i}\right)$:

$$Pr\left(v \in O_{f_i}\right) = \sum_{x=1}^{N} \left(\begin{array}{l} \text{Probability that } v \\ \text{is used to label } x \\ \text{objects} \end{array} \right) \times \left(\begin{array}{l} \text{Probability of observing an} \\ \text{utterance being produced} \\ \text{for at least one of those } x \\ \text{objects after } R \text{ observations} \end{array} \right)$$

We will assume the strongest possible generalisation capacity. Our learner will be able to express a meaning if it has viewed all the feature values that make up that meaning, paired with signal substrings. The probability of our learner being able to express an arbitrary meaning made up of F feature values is then given by the combined probability of having observed each of those feature values:

$$Pr\left(v_1 \in O_{f_1} \wedge \ldots \wedge v_F \in O_{f_F}\right) = Pr\left(v \in O_{f_i}\right)^F$$

We can now calculate E_c, the number of meanings our learner will be able to express after viewing some subset of a compositional language, which is simply the probability of being able to express an arbitrary meaning multiplied by N_{used}, the number of meanings used when labelling the N objects:

$$E_c = Pr\left(v \in O_{f_i}\right)^F \cdot N_{used}$$

We therefore have a method for calculating the expected expressivity of a learner presented with L_h or L_c. This in itself is not terribly useful. However, within the Iterated Learning framework we can relate expressivity to *stability*. We are interested in the dynamics arising from the iterated learning of languages. The stability of a language determines how likely it is to persist over iterated learning events.

If an individual is called upon to express a meaning they have not observed being expressed, they have two options. Firstly, they could simply not express. Alternatively, they could produce some random signal. In either case, any association between meaning and signal that was present in the previous individual's hypothesis will be lost — part of the meaning-signal mapping will change. A shortfall in expressivity therefore results in instability over cultural time. We can relate the expressivity of a language to the stability of that language over time by $S_h \propto E_h/N$ and $S_c \propto E_c/N$. Stability is simply the proportion of meaning-signal mappings encoded in an individual's hypothesis which are also encoded in the hypotheses of subsequent individuals.

We will be concerned with the *relative stability* of compositional languages with respect to holistic languages, S, which is given by:

$$S = \frac{S_c}{S_c + S_h}$$

When $S = 0.5$ compositional languages and holistic languages are equally stable and we therefore expect them to emerge with equal frequency over cultural time. When $S > 0.5$ compositional languages are more stable than holistic languages, and we expect them to emerge more frequently, and persist for longer, than holistic languages. $S < 0.5$ corresponds to the situation where holistic languages are more stable than compositional languages.

The impact of meaning space structure and the bottleneck

Relative stability S depends on the number of dimensions in the meaning space (F), the number of possible values for each feature (V), the number of objects in the environment (N) and the number of observations each learner makes (R). Unless each learner makes a large number of observations (R is very large), or there are few objects in the environment (N is very small), there is a chance that an agent will be called upon to express a meaning they themselves have never observed paired with a signal. This is one aspect of the poverty of the stimuli facing language learners — the set of utterances of any human language is arbitrarily large, but a child must acquire their linguistic competence based on a finite number of sentences. We will refer to this aspect of the poverty of stimulus as the *transmission bottleneck*. The severity of the transmission bottleneck depends on the number of observations each learner makes (R) and the number of objects in the environment (N). It is convenient to refer instead to the degree of object coverage (b), which is simply the proportion of all N objects observed after R observations — b gives the severity of the transmission bottleneck.

Together F and V specify the degree of structure in the meaning space. This in turn reflects the sophistication of the semantic representation capacities of agents — we follow Schoenemann in that we "take for granted that there are features of the real world which exist regardless of whether an organism perceives them ... [d]ifferent organisms will divide up the world differently, in accordance with their unique evolved neural systems ... [i]ncreasing semantic complexity therefore refers to an increase in the number of divisions of reality which a particular organism is aware of"(Schoenemann 1999:318). Schoenemann argues that high semantic complexity can lead to the emergence of syntax. The Iterated Learning model can be used to test this hypothesis. We will vary the degree of structure in the meaning space, together with the transmission bottleneck b, while holding the number of objects in the environment (N) constant. The results of these manipulations are shown in Fig. 3.

There are two key results to draw from these figures:

1. Relative stability S is at a maximum for small bottleneck sizes. Holistic languages will not persist over time when the bottleneck on cultural transmission is tight. In contrast, compositional languages are generalisable, due to their structure, and remain relatively stable even when a learner only observes a small subset of the language of the previous generation. The poverty of the stimulus "problem" is in fact required for linguistic structure to emerge.

2. A large stability advantage for compositional language (high S) only occurs when the meaning space exhibits a certain degree of structure, suggesting that structure in the conceptual space of language learners is a requirement for the evolution of compositionality. In such meaning spaces, distinct meanings tend to share feature values. A compositional system in such a meaning space will be highly generalisable — the signal associated with a meaning can be deduced from observation of other meanings paired with signals, due to the shared feature values. However, if the meaning space is too highly structured S is low, as few distinct meanings will share feature values and the advantage of generalisation is lost.

The first result outlined above is to some extent obvious, although it is interesting to note that the apparent poverty of the stimulus problem motivated the strongly innatist Chomskyan paradigm. The advantage of the Iterated Learning approach is that it allows us to quantify the degree of advantage afforded by compositional language, and investigate how other factors, such as meaning space structure, impact on the advantage afforded by compositionality.

A computational model

The mathematical model outlined above, made possible by insights gained from viewing language as a culturally-transmitted system, predicts that compositional language will

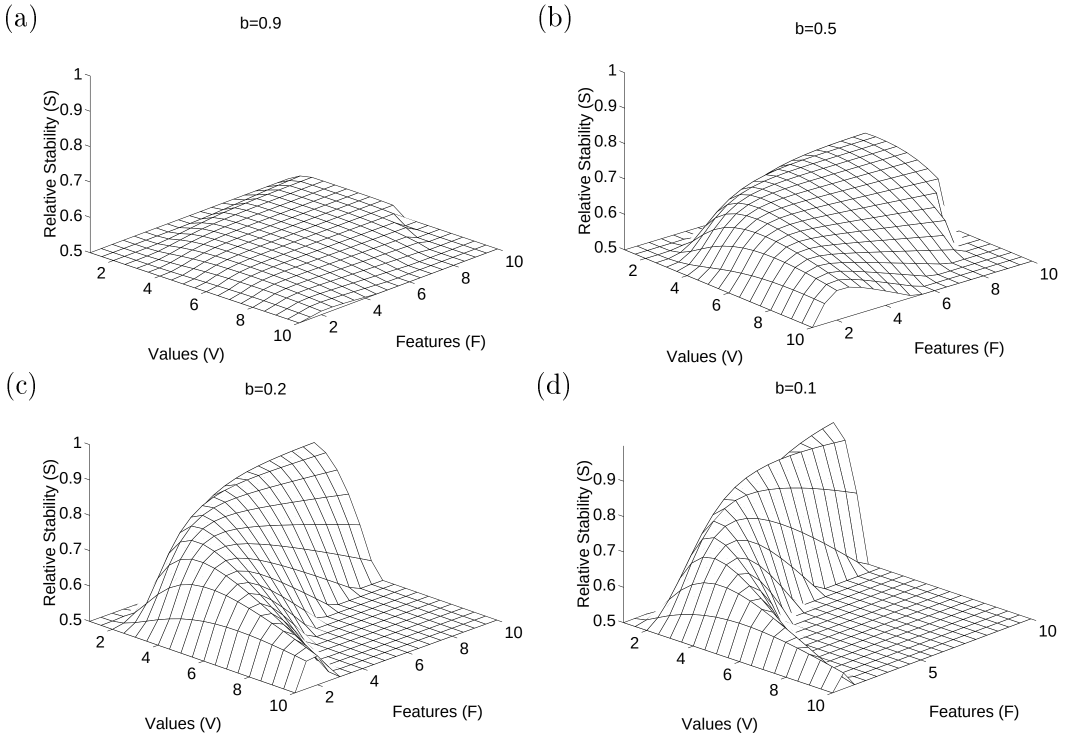


Figure 3: Severity of bottleneck and meaning space structure impact on the relative stability of compositional language. The relative stability advantage of compositional language increases as the bottleneck tightens, but only when the meaning space exhibits certain kinds of structure (in other words, for particular numbers of features and values). b gives the severity of transmission bottleneck, with low b corresponding to a tight bottleneck.

be more stable than holistic language when 1) there is a bottleneck on cultural transmission and 2) linguistic agents have structured representations of objects. However, the simplifications necessary to the mathematical analysis preclude a more detailed study of the dynamics arising from Iterated Learning. What happens to languages of intermediate compositionality during cultural transmission? Can compositional language emerge from initially holistic language, through a process of cultural evolution? We can investigate these question using techniques from artificial life, by developing a multi-agent computational implementation of the Iterated Learning Model.

A neural network model of a linguistic agent

Smith (2002b) presents a neural-network model of the evolution of holistic communication. We extend this model to allow the study of the cultural evolution of compositionality[2]. As in the mathematical model, meanings are represented as points in F-dimensional space where each dimensions has V distinct values, and signals are represented as strings of characters of length 1 to l_{max}, where the characters are drawn from the alphabet Σ.

Agents are modelled using networks consisting of two sets of nodes. One set represents meanings and partially-specified components of meanings ($\mathcal{N}_M$), and the other represents signals and partially-specified components of signals ($\mathcal{N}_S$). These nodes are linked by a set of bidirectional connections $\mathcal{W}$ connecting every node in $\mathcal{N}_M$ with every node in $\mathcal{N}_S$.

As with the mathematical model, meanings are sets of feature values, and signals are strings of characters. Components of a meaning specify one or more feature values of that meaning, with unspecified values being marked as a wildcard $*$. For example, the meaning (2 1) has three possible components, the fully-specified (2 1) and the partially specified (2 $*$) and ($*$ 1). These components can be grouped together into ordered sets, which constitute an analysis of a meaning. For example, there are three possible analyses of the meaning (2 1) — the one-component analysis $\{(2\ 1)\}$, and two two component analyses which differ in order, $\{(2\ *)\,,(*\ 1)\}$ and $\{(*\ 1)\,,(2\ *)\}$. Similarly, components of signals can be grouped together to form an analysis of a signal. This representational scheme allows the networks to exploit the structure of meanings and signals. However, they are not forced to do so.

Learners observe meaning-signal pairs. During a single learning episode a learner will store a $\langle m, s\rangle$ pair in its network. The nodes in $\mathcal{N}_M$ corresponding to all possible components of the meaning m have their activations set to 1, while all other nodes in $\mathcal{N}_M$ have their activations set to 0. Similarly, the nodes in $\mathcal{N}_S$ corresponding to the possible components of s have their activations set to 1. Connection weights in $\mathcal{W}$ are then adjusted according to the rule:

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

[2] We refer the reader to Smith (2002a) for a more thorough description of this model.

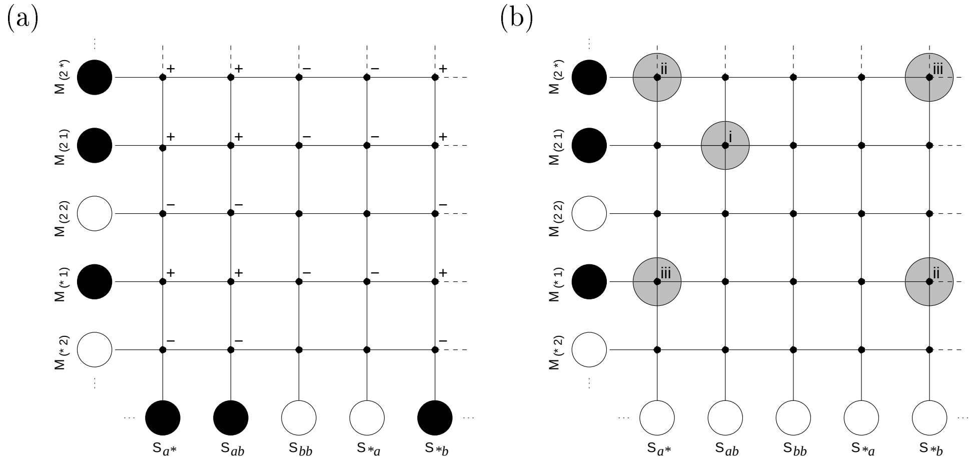


Figure 4: Nodes with an activation of 1 are represented by large filled circles. Small filled circles represent weighted connections. (a) Storage of the meaning-signal pair $\langle (2\ 1), ab \rangle$. Nodes representing components of (2 1) and ab have their activations set to 1. Connection weights are then either incremented (+), decremented (−) or left unchanged. (b) Retrieval of three possible analyses of $\langle (2\ 1), ab \rangle$. The relevant connection weights are highlighted in grey. The strength, g, of the one-component analysis $\langle \{(2\ 1)\}, \{ab\} \rangle$ depends of the weight of connection i. g for the two-component analysis $\langle \{(2\ *), (*\ 1)\}, \{a*, *b\} \rangle$ depends on the weighted sum of two connections, marked as ii. The g for the alternative two-component analysis $\langle \{(2\ *), (*\ 1)\}, \{*b, a*\} \rangle$ is given by the weighted sum of the two connections marked iii.

where W_{xy} gives the weight of the connection between nodes x and y and a_x gives the activation of node x. The learning procedure is illustrated in Fig. 4 (a).

In order to produce an utterance, agents are prompted with a meaning m and required to produce a signal s. All possible analyses of m are considered in turn with all possible analyses of every $s \in \mathcal{S}$. Each meaning analysis-signal analysis pair is evaluated according to:

$$g(\langle m, s \rangle) = \sum_{i=1}^{C} \omega(c_{mi}) \cdot W_{c_{mi} c_{si}}$$

where the sum is over the C components of the analysis, c_{mi} is the ith component of m and and $\omega(x)$ is a weighting function which gives the non-wildcard proportion of x. This process is illustrated in Figure 4 (b). The meaning analysis-signal analysis pair with the highest g is returned as the network's utterance.

Environment structure

In the mathematical model outlined above, the environment consisted of a set of objects labelled with meanings drawn at random from the space of possible meanings. In

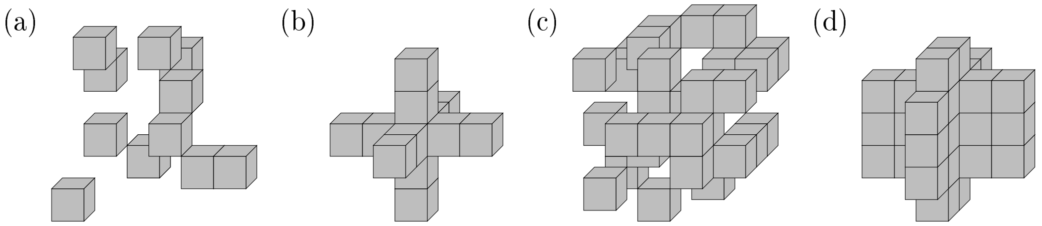


Figure 5: We will present results for the case where $F = 3$ and $V = 5$. This defines a three-dimensional meaning space. We highlight the meanings selected from that space with grey. (a) is a low-density, unstructured environment. (b) is a low-density, structured environment. (c) and (d) are unstructured and structured high-density environments.

the computational model we can relax this assumption, and investigate how non-random assignment of meanings to objects impacts on linguistic evolution. As before, an environment consists of a set of objects labelled with meanings drawn from the meaning space $\mathcal{M}$. The number of objects in the environment gives the *density* of that environment — environments with few objects will be termed low-density, whereas environments with a large number of objects will be termed high-density. When meanings are assigned to objects at random we will say the environment is *unstructured*. When meanings are assigned to objects in such a way as to minimise the average inter-meaning hamming distance we will say the environment is *structured*. Sample low- and high-density environments are shown in Fig. 5. Note the new usage of the term "structured" — while in the mathematical model we were concerned with structure in the meaning space, given by F and V, we are now concerned with the degree of structure in the environment. Different levels of environment structure are possible within a meaning space of a particular structure.

The impact of environment structure and the bottleneck

The network model of a language learner/producer is plugged into the Iterated Learning framework. We will manipulate three factors — the presence or absence of a bottleneck, the density of the environment and the degree of structure in the environment.

Our measure of compositionality is simply the degree of correlation between the distance between pairs of meanings and the distance between the corresponding pairs of signals. In order to measure the compositionality of an agent's language we first take all possible pairs of meanings from the environment, $\langle m_i, m_{j \neq i} \rangle$. We then find the signals these meanings map to in the agent's language, $\langle s_i, s_j \rangle$. This yields a set of meaning-meaning pairs, each of which is matched with a signal-signal pair. For each of these pairs, the distance between the meanings m_i and m_j is calculated using Hamming distance, and the distance between the signals s_i and s_j is calculated using Levenstein (string edit) distance.[3] This gives a set of distance pairs, reflecting the distance between all possible pairs

[3] Levenstein distance is a measure of string similarity. It is defined as the size of the smallest set of

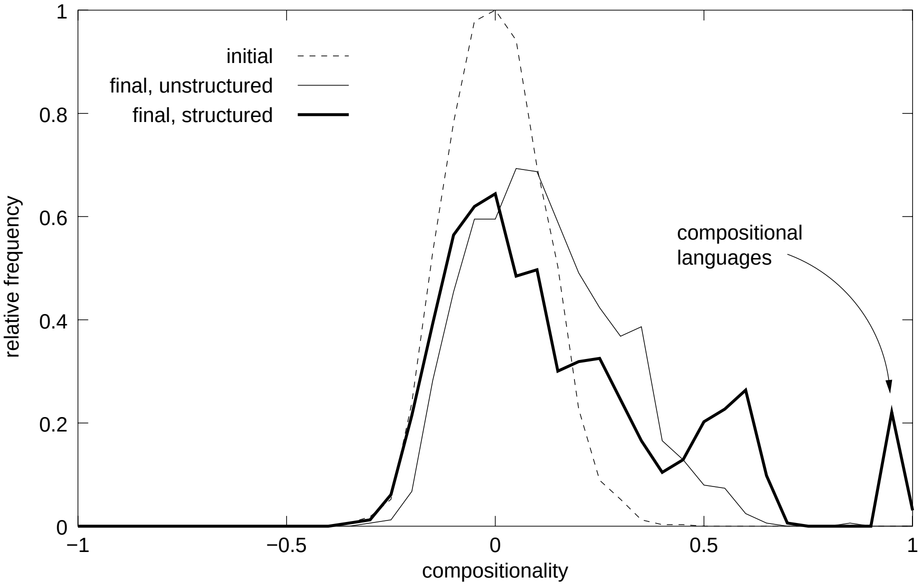


Figure 6: The relative frequency of initial and final systems of varying degrees of compositionality, when there is no bottleneck on cultural transmission. The results shown here are for the low-density environments given in Fig. 5. The initial languages are generally holistic. Some final languages exhibit increased levels of compositionality. Highly compositional languages are infrequent.

of meanings and the distance between the corresponding pairs of signals. A Pearson's Product-Moment correlation is then run on this set, giving the correlation between the meaning-meaning distances and the associated signal-signal distances. This correlation is our measure of compositionality. Perfectly compositional languages have a compositionality of 1, reflecting the fact that compositional languages preserve distance relationships when mapping between meanings and signals. Holistic languages have a compositionality of approximately 0 — holistic mappings are random, and therefore fail to preserve distance relationships when mapping between meaning space and signal space.

Fig. 6 plots the frequency by compositionality of initial and final systems in 1000 runs of the Iterated Learning Model, in the case where there is no bottleneck on cultural transmission. The initial agent has the maximum-entropy hypothesis — all meaning-signal pairs are equally probable. The learner at each generation is exposed to the complete language of the previous generation — the adult is required to produce utterances for every object in the environment. Each run was allowed to proceed to a stable state.

Two main results are apparent from Fig. 6.

1. The majority of the final, stable systems are holistic.

edits (replacements, deletions, or insertions) that could transform one string to the other.

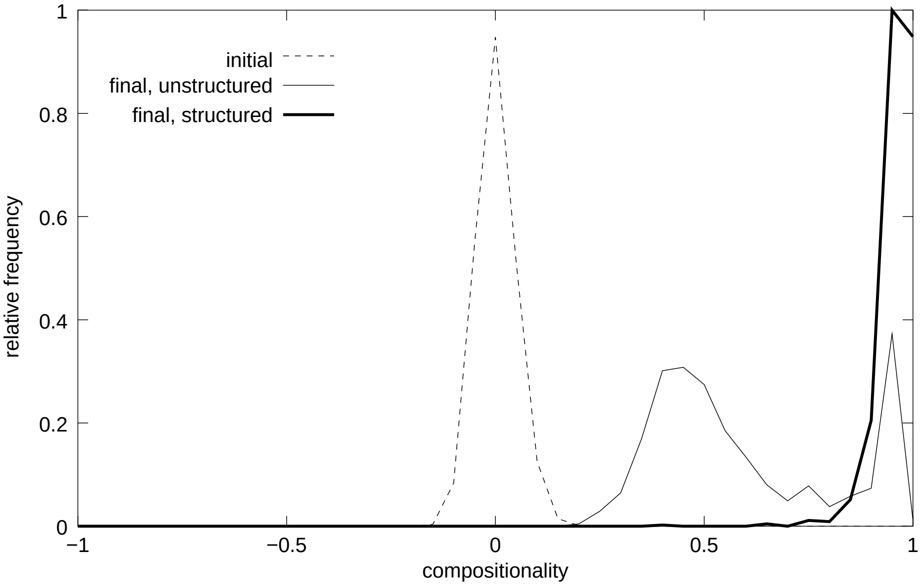


Figure 7: Frequency by compositionality when there is a bottleneck on cultural transmission. The results shown here are for the high-density environments given in Fig. 5 (c) and (d). The initial languages are holistic. The final languages are compositional, with highly compositional languages occurring frequently.

2. Highly compositional systems occur infrequently, and only when the environment is structured.

In the absence of a bottleneck on cultural transmission, the compositionality of the final systems is sensitive to initial conditions. The majority of the initial holistic systems are stable. This can be contrasted with the result shown in Fig. 3 (a), where compositional languages have a slight stability advantage for most meaning spaces when the transmission bottleneck is very wide ($b = 0.9$). When there is *no* bottleneck on transmission ($b = 1.0$) most holistic systems are perfectly stable. However, the initial system may exhibit, purely by chance, a slight tendency to express a given feature value with a certain substring. This compositional tendency can spread, over iterated learning events, to other parts of the system, which can in turn have further knock-on consequences. The potential for spread of compositional tendencies is greatest in structured environments — in such environments, distinct meanings are more likely to share feature values than in unstructured environments. However, this spread of compositionality is unlikely to lead to a perfectly compositional language.

Fig. 7 plots the frequency by compositionality of initial and final systems in 1000 runs of the Iterated Learning Model, in the case where there is a bottleneck on cultural transmission ($b = 0.4$). Learners will therefore only see a subset of the language of the previous generation. Whereas in the no-bottleneck condition each run proceeded to a stable state, in the bottleneck condition runs were stopped after 50 generations. There is no such thing as a truly stable state when there is a bottleneck on cultural transmission. For example, if all R utterances an individual observes refer to the same object then any structure in the language of the previous generation will be lost. However, the final states here were as close as possible to stable. Allowing the runs to continue for several hundred more generations results in a very similar distribution of languages.

Two main results are apparent from Fig. 7.

1. When there is a bottleneck on cultural transmission highly compositional systems are frequent.
2. Highly compositional systems are more frequent when the environment is structured.

As discussed with reference to the mathematical model, only highly compositional systems are stable through a bottleneck. The results from the computational model bear this out — over time, language adapts to the pressure to be generalisable, until the language becomes highly compositional, highly generalisable and highly stable. Highly compositional languages evolve most frequently when the environment is structured, because in a structured environment the advantage of compositionality is at a maximum — each meaning shares feature values with several other meanings, and a language mapping these feature values to a signal substring is highly generalisable.

Fig. 8 plots the compositionality by generation for three runs of the Iterated Learning Model. The behaviour of these runs is characteristic of the majority of simulations. Plots (a) and (b) show the development from initially random, holistic systems to compositional languages in structured and unstructured environments. In both these runs a partially compositional, partially irregular language rapidly develops, resulting in a rapid increase in compositionality. This partially compositional system persists for a short time, before developing into a highly regular compositional language where each feature value maps consistently to a particular subsignal. The transition is more rapid in the structured environment. In the structured environment, distinct meanings share feature values with several other meanings and as a consequence compositional languages are highly generalisable. Additionally, distinct meaning vary along a limited number of dimensions, which facilitates the spread of consistent, regular mappings from feature values to signal substrings. In plot (c) a partially compositional language develops from the initial random mapping, but fails to become fully compositional. The lack of structure in the environment hinders the development of consistent compositional mappings and allows unstable, idiosyncratic meaning-signal mappings to persist.

Conclusions

Language can be viewed as a consequence of an innate language organ. This view of language has been advanced to explain for the near-universal success of language acquisition

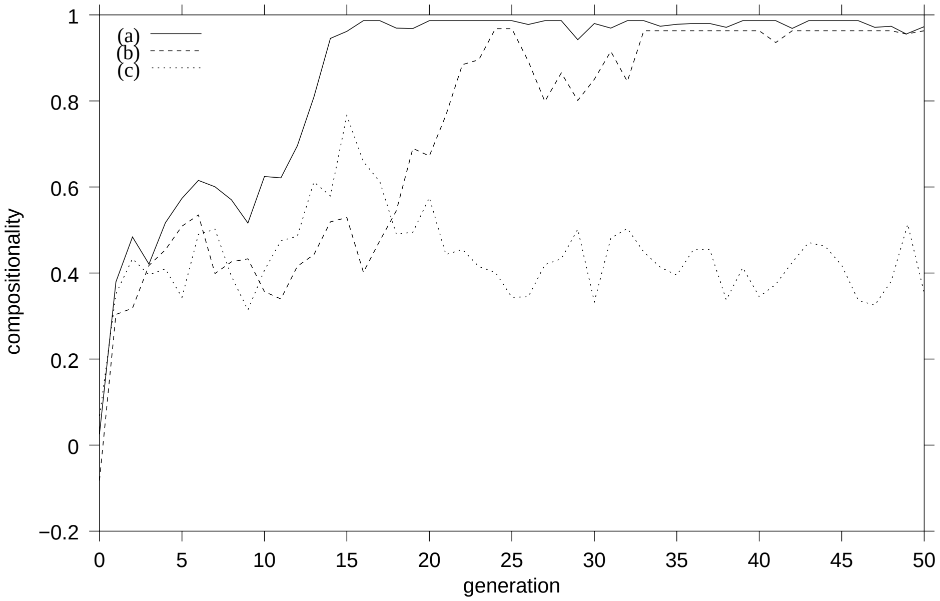


Figure 8: Compositionality by time (in generations) for three runs. (a) shows the development from an initially holistic system to a compositional language for a run in a high-density structured environment. (b) and (c) show the development of systems in high-density unstructured environments. The language plotted in (b) eventually becomes highly compositional, whereas the system in (c) remains partially compositional. Only the first 50 generations are plotted here, in order to focus on the development of the systems from the initial holistic state.

in the face of the poverty of the stimulus available to language learners. The innatist position solves this apparent conundrum by attributing much of the structure of language to the language organ — an individual's linguistic competence develops along an internally-determined course, with the linguistic environment simply triggering the growth of the appropriate cognitive structures. If we take this view, we can form an evolutionary account which explains linguistic structure as a biological adaptation to social function — language is socially useful, and the language organ yields a fitness payoff.

However, we have presented an alternative approach. We focus on the the cultural transmission of language. We can then form an account which explains much of linguistic structure as a cultural adaptation, by language, to pressures arising during repeated production and acquisition of language. This kind of approach highlights the *situatedness* of language-using agents in an environment — in this case, a socio-cultural environment made up of the behaviour of other agents. We have presented the Iterated Learning Model as a framework for studying the cultural evolution of language in this context, and have focussed here on the cultural evolution of compositionality. The models presented reveal two key factors in the cultural evolution of compositional language.

Firstly, compositional language emerges when there is a bottleneck on cultural transmission — compositionality is an adaptation by language which allows it to slip through the transmission bottleneck. The transmission bottleneck constitutes one aspect of the poverty of the stimulus problem. This result is therefore surprising. The poverty of the stimulus motivated a strongly innatist position on language acquisition. However, closer investigation within the Iterated Learning framework reveals that the poverty of the stimulus does not force us to conclude that linguistic structure must be located in the language organ — on the contrary, the emergence of linguistic structure through cultural processes *requires* the poverty of the stimulus.

The second key factor is the availability of structured semantic representations to language learners — Schoenemann's (1999) semantic complexity. The advantage of compositionality is at a maximum when language learners perceive the world as structured. If objects are perceived as structured entities and the objects in the environment relate to one another in structured ways then a generalisable, compositional language is highly adaptive.

Of course, biological evolution still has a role to play in explaining the evolution of language. The Iterated Learning Model is ideal for investigating the cultural evolution of language on a fixed biological substrate, and identifying the cultural consequences of a particular innate endowment. The origins of that endowment then need to be explained, and natural selection for a socially-useful language might play some role here. We might indeed then find, as suggested by Deacon, that "the brain has co-evolved with respect to language, but languages have done most of the adapting"(Deacon 1997:122). The poverty of the stimulus faced by language learners forces language to adapt to be learnable. The transmission bottleneck forces language to be generalisable, and compositional structure is language's adaptation to this problem. This adaptation yields the greatest payoff for language when language learners perceive the world as structured.

References

Batali, J. 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In (Briscoe 2002), 111–172.

Bloom, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Brighton, H. 2002. Compositional syntax from cultural transmission. *Artifical Life* 8.25–54.

Briscoe, E. (ed.) 2002. *Linguistic Evolution through Language Acquisition: Formal and Computational Models*. Cambridge: Cambridge University Press.

Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

—— 1980. *Rules and Representations*. London: Basil Blackwell.

—— 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Deacon, T. 1997. *The Symbolic Species*. London: Penguin.

Dunbar, R. 1996. *Grooming, Gossip and the Evolution of Language*. London: Faber and Faber.

Hurford, J. R. 1990. Nativist and functional explanations in language acquisition. In *Logical Issues in Language Acquisition*, ed. by I. M. Roca, 85–136. Dordrecht: Foris.

Jackendoff, R. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

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

—— 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation* 5.102–110.

—— 2002. Learning, bottlenecks and the evolution of recursive syntax. In (Briscoe 2002), 173–203.

Krifka, M. 2001. Compositionality. In *The MIT Encyclopaedia of the Cognitive Sciences*, ed. by R. A. Wilson & F. Keil. Cambridge, MA: MIT Press.

Livingstone, D., & C. Fyfe. 1999. Modelling the evolution of linguistic diversity. In *Advances in Artificial Life: Proceedings of the 5th European Conference on Artificial Life*, ed. by D. Floreano, J. D. Nicoud, & F. Mondada, 704–708. Berlin: Springer.

Pinker, S. 1994. *The Language Instinct*. London: Penguin.

——, & P. Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13.707–784.

Schoenemann, P. T. 1999. Syntax as an emergent characteristic of the evolution of semantic complexity. *Mind and Machines* 9.309–346.

SMITH, A. 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 J. Kelemen & P. Sosík, 381–390. Berlin: Springer-Verlag.

SMITH, K. 2002a. Compositionality from culture: the role of environment structure and learning bias. Technical report, Theoretical and Applied Linguistics, University of Edinburgh.

—— 2002b. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.

STEELS, L. 1998. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103.133–156.

——, F. KAPLAN, A. MCINTYRE, & J. VAN LOOVEREN. 2002. Crucial factors in the origins of word-meaning. In *The Transition to Language*, ed. by A. Wray, 252–271. Oxford: OUP.

WRAY, A. 1998. Protolanguage as a holistic system for social interaction. *Language and Communication* 18.47–67.

Bibliography

ACKLEY, D., & M. LITTMAN. 1992. Interactions between learning and evolution. In Langton *et al.* (1992), 487–509.

——, & ——. 1994. Altruism in the evolution of communication. In Brooks & Maes (1994), 40–48.

ANDERSEN, H. 1973. Abductive and deductive change. *Language* 40.765–793.

BÄCK, T. 1994. Selective pressure in evolutionary algorithms. In *Proceedings of the First IEEE Conference on Evolutionary Computation*, ed. by Z. Michalewicz, J. D. Schaffer, H. P. Schwefel, D. B. Fogel, & H. Kitano, volume 1, 57–62. Piscataway, NJ: IEEE Press.

BALDWIN, D. A. 1991. Infants' contribution to the achievement of joint reference. *Child Development* 62.875–890.

—— 1993a. Early referential understanding: Infants' ability to recognise referential acts for what they are. *Developmental Psychology* 29.832–843.

—— 1993b. Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language* 20.395–418.

BATALI, J. 1994. Innate biases and critical periods: Combining evolution and learning in the acquisition of syntax. In Brooks & Maes (1994), 160–171.

—— 1998. Computational simulations of the emergence of grammar. In Hurford *et al.* (1998), 405–426.

—— 2002. The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In Briscoe (2002), 111–172.

BATES, E., & J. ELMAN. 1996. Learning rediscovered. *Science* 274.1849–1850.

BELEW, R., J. MCINERNEY, & N. N. SCHRAUDOLPH. 1992. Evolving networks: Using the genetic algorithm with connectionist learning. In Langton *et al.* (1992), 511–547.

BEVER, T. G., & D. T. LANGENDOEN. 1971. A dynamic model of the evolution of language. *Linguistic Inquiry* 2.433–463.

BICKERTON, D. 1981. *Roots of Language*. Ann Arbor, MI: Karoma.

—— 1984. The language bioprogram hypothesis. *Behavioral and Brain Sciences* 7.173–221.

—— 1990. *Language and Species*. Chicago, IL: University of Chicago Press.

—— 1998. Catastrophic evolution: the case for a single step from protolanguage to full human language. In Hurford *et al.* (1998), 341–358.

—— 2000. How protolanguage became language. In Knight *et al.* (2000), 264–284.

—— 2002. Foraging versus social intelligence in the evolution of protolanguage. In Wray (2002), 207–225.

BLOOM, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

——, & L. MARKSON. 1998. Capacities underlying word learning. *Trends in Cognitive Sciences* 2.67–73.

BOYD, R., & P. J. RICHERSON. 1985. *Culture and the Evolutionary Process*. Chicago, IL: University of Chicago Press.

——, & ——. 2000. Memes: Universal acid or a better mousetrap. In *Darwinizing Culture: The Status of Memetics as a Science*, ed. by R. Aunger, 143–162. Oxford: Oxford University Press.

BRIGHTON, H. 2000. Experiments in iterated instance-based learning. Technical report, Language Evolution and Computation Research Unit.

—— 2002. Compositional syntax from cultural transmission. *Artifical Life* 8.25–54.

——, S. KIRBY, & K. SMITH. forthcoming. Situated cognition and the role of multi-agent models in explaining language structure. In *Adaptive Agents*, ed. by D. Kudenko, E. Alonso, & D. Kazakov. London: Springer.

BRISCOE, E. 2000a. Evolutionary perspectives on diachronic syntax. In *Diachronic Syntax: Models and Mechanisms*, ed. by S. Pintzuk, G. Tsoulas, & A. Warner, 75–108. Oxford: Oxford University Press.

—— 2000b. 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., & P. MAES (eds.) 1994. *Artificial Life 4: Proceedings of the Fourth International Workshop on the Synthesis and Simulation of Living Systems*. Cambridge, MA: Addison-Wesley.

BROWN, R. 1973. *A First Language*. Cambridge, MA: Harvard University Press.

BUCKLEY, C., & J. STEELE. 2002. Evolutionary ecology of spoken language: co-evolutionary hypotheses are testable. *World Archaeology* 34.26–46.

BULLOCK, S. 1997. An exploration of signalling behaviour by both analytic and simulation means for both discrete and continuous models. In Husbands & Harvey (1997), 454–463.

BURLING, R. 1992. *Patterns of Language: Structure, Variation, Change*. London: Academic Press.

CAMPBELL, D. T. 1965. Variation and selective retention in sociocultural evolution. In *Social Change in Developing Areas: A Reinterpretation of Evolutionary Theory*, ed. by H. R. Barringer, G. I. Blanksten, & R. W. Mack, 19–49. Cambridge, MA: Schenkman.

—— 1975. On the conflicts between biological and social evolution and between psychological and moral tradition. *American Psychology* 30.1103–1126.

CANGELOSI, A. 1999. Modelling the evolution of communication: From stimulus associations to grounded symbolic associations. In Floreano *et al.* (1999), 654–663.

——, & D. PARISI. 1998. The emergence of a 'language' in an evolving population of neural networks. *Connection Science* 10.83–97.

CANN, R. 1993. *Formal Semantics: an introduction*. Cambridge: Cambridge University Press.

CHENEY, D., & R. SEYFARTH. 1990. *How Monkeys See the World: Inside the Mind of Another Species*. Chicago, IL: University of Chicago Press.

CHOMSKY, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

—— 1971. *Problems of Knowledge and Freedom*. London: Fontana.

—— 1972. *Language and Mind*. New York, NY: Harcourt, Brace and World, enlarged edition.

—— 1975. *Reflections on Language*. New York, NY: Pantheon.

—— 1980. *Rules and Representations*. London: Basil Blackwell.

—— 1986. *Knowledge of Language: its nature, origin and use*. London: Praeger.

—— 1988. *Language and Problems of Knowledge*. Cambridge, MA: MIT Press.

—— 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

—— 2002. *On Nature and Language*. Cambridge: Cambridge University Press. Edited by A. Belletti and L. Rizzi.

CHRISTIANSEN, M. H., & C. M. CONWAY, 2002. The importance of hierarchical learning: A computational study of sequential learning in human and non-human primates. Presented at the 4th International Conference on the Evolution of Language.

CLARK, E.V. 1988. On the logic of contrast. *Journal of Child Language* 15.317–335.

—— 1990. On the pragmatics of contrast. *Journal of Child Language* 17.417–431.

—— 1993. *The lexicon in acquisition*. Cambridge: Cambridge University Press.

CORBALLIS, M. C. 2002. *From Hand to Mouth: The Origins of Language*. Princeton, NJ: Princeton University Press.

CROFT, W. 2000. *Explaining Language Change: an evolutionary approach*. London: Longman.

DARWIN, C. 1859/1964. *On the Origin of Species*. Cambridge, MA: Harvard University Press, a facsimile of the first edition. Original published by John Murray, London.

DAWKINS, R., & J. R. KREBS. 1978. Animal signals: Information or manipulation? In *Behavioural Ecology: An Evolutionary Approach*, ed. by J. R. Krebs & N. B. Davies, 282–309. Oxford: Blackwell.

DE BOER, B. 2000. Self-organization in vowel systems. *Journal of Phonetics* 28.441–465.

—— 2001. *The Origins of Vowel Systems*. Oxford: Oxford University Press.

DEACON, T. 1997. *The Symbolic Species*. London: Penguin.

DI PAOLO, E. A., J. NOBLE, & S. BULLOCK. 2000. Simulation models as opaque thought experiments. In *Artificial Life VII: Proceedings of the Seventh International Conference on Artificial Life*, ed. by M. A. Bedau, J. S. McCaskill, N. H. Packard, & S. Rasmussen, 497–506. Cambridge, MA: MIT Press.

DOR, D., & E. JABLONKA. 2000. From cultural selection to genetic selection: a framework for the evolution of language. *Selection* 1.33–55.

DUNBAR, R. 1996. *Grooming, Gossip and the Evolution of Language*. London: Faber and Faber.

ELMAN, J.L. 1993. Learning and development in neural networks: The importance of starting small. *Cognition* 48.71–99.

EVANS, C. S., L. EVANS, & P. MARLER. 1993. On the meaning of alarm calls: Functional reference in an avian vocal system. *Animal Behaviour* 46.23–38.

FITCH, W. T. 2000. The evolution of speech: a comparative review. *Trends in Cognitive Sciences* 4.258–267.

FLOREANO, D., J. D. NICOUD, & F. MONDADA (eds.) 1999. *Advances in Artificial Life: Proceedings of the 5th European Conference on Artificial Life*. Berlin: Springer.

FROMKIN, V., & R. RODMAN. 1988. *An Introduction to Language*. London: Holt, Rinehart and Winston, 4th edition.

FUTUYMA, D. J. 1998. *Evolutionary Biology*. Sunderland, MA: Sinauer Associates, 3rd edition.

GENTNER, D. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. In *Language development*, ed. by S. A. Kuczaj, volume 2. Hillsdale, NJ: Erlbaum.

GESCHWIND, N. 1964. The development of the brain and the evolution of language. In *Monograph series on language and linguistics*, ed. by C. I. J. M. Stuart, volume 17. Washington, DC: Georgetown University Press.

GIBSON, E., & K. WEXLER. 1994. Triggers. *Linguistic Inquiry* 25.355–407.

GOLDIN-MEADOW, S., & C. MYLANDER. 1990. Beyond the input given: The child's role in the acquisition of language. *Language* 66.323–355.

GOPNIK, M. 1994. Impairments of tense in a familial language disorder. *Journal of Neurolinguistics* 8.109–133.

——, & M. B. CRAGO. 1991. Familial aggregation of a developmental language disorder. *Cognition* 39.1–50.

GOULD, S. J., & R. C. LEWONTIN. 1979. The spandrels of San Marco and the Panglossian program: A critique of the adaptationist programme. *Proceedings of the Royal Society of London* 205.281–288.

——, & E. S. VRBA. 1982. Exaptation – a missing term in the science of form. *Paleobiology* 8.4–15.

GRAFEN, A. 1990. Biological signals as handicaps. *Journal of Theoretical Biology* 144.517–546.

GREENBERG, J. H. 1966. *Universals of language*. MIT Press, 2nd edition.

GREENFIELD, P. M. 1991. Language, tools, and brain - the ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and Brain Sciences* 14.531–550.

HAEGEMAN, L. 1994. *Introduction to Government and Binding Theory*. Oxford: Blackwell, 2nd edition.

HAIMAN, J. 1980. The iconicity of grammar: Isomorphism and motivation. *Language* 56.515–540.

—— (ed.) 1985. *Iconicity in Syntax*. Amsterdam: Benjamins.

HARE, M., & J. L. ELMAN. 1995. Learning and morphological change. *Cognition* 56.61–98.

HARNAD, S. 1990. The symbol grounding problem. *Physica* D 42.335–346.

HARRIS, J. W. K. 1983. Cultural beginnings: Plio-pleistocene archaeological occurrences from the Afar, Ethiopia. *African Archaeological Review* 1.3–31.

HARTL, D. L., & A. G. CLARK. 1997. *Principles of Population Genetics*. Sunderland, MA: Sinauer Associates, 3rd edition.

HASHIMOTO, T. 1998. Dynamics of internal and global structure through linguistic interactions. In *Multi-Agent Systems and Agent-Based Simulation*, ed. by J. S. Sichman, R. Conte, & N. Gilbert, 124–139. Berlin: Springer-Verlag.

HASPELMATH, M. 1989. From purposive to infinitive — a universal path of grammaticalization. *Folia Linguistica Historia* 10.287–310.

HAUSER, M. D. 1996. *The Evolution of Communication*. Cambridge, MA: MIT Press.

——, N. CHOMSKY, & W. T. FITCH. 2002. The faculty of language: What is it, who has it, and how did it evolve. *Science* 298.1569–1579.

HAWKINS, J. A. 1990. A parsing theory of word order universals. *Linguistic Inquiry* 21.223–261.

HAZELHURST, B., & E. HUTCHINS. 1998. The emergence of propositions from the co-ordination of talk and action in a shared world. *Language and Cognitive Processes* 13.373–424.

HEWES, G. W. 1973. Primate communication and the gestural origin of language. *Current Anthropology* 14.5–24.

HINTON, G., & S. NOWLAN. 1987. How learning can guide evolution. *Complex Systems* 1.495–502.

HOCKETT, C. F. 1960a. Logical considerations in the study of animal communication. In *Animal sounds and communication*, ed. by W. E. Lanyon & W. N. Tavolga, 392–430. Washington, DC: American Institute of Biological Sciences.

—— 1960b. The origin of speech. *Scientific American* 203.88–96.

HOLLOWAY, R. L. 1995. Evidence for POT expansion in early *Homo*: A pretty theory with ugly (or no) paleoneurologicall facts. *Behavioral and Brain Sciences* 18.191–193. Commentary on Wilkins and Wakefield (1995).

HOLM, J. 1988. *Pidgins and creoles*, volume 1. Cambridge: Cambridge University Press.

HUDSON, G. 2000. *Essential Introductory Linguistics*. Oxford: Blackwell.

HURFORD, J. R. 1987. *Language and number: the emergence of a cognitive system*. Oxford: Basil Blackwell.

—— 1989. Biological evolution of the saussurean sign as a component of the language acquisition device. *Lingua* 77.187–222.

—— 1999. Language learning from fragmentary input. In *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*, ed. by K. Dautenhahn & C. Nehaniv, 121–129. Society for the Study of Artificial Intelligence and the Simulation of Behaviour.

—— 2000. Social transmission favours linguistic generalization. In Knight *et al.* (2000), 324–352.

—— 2002a. Expression/induction models of language evolution: dimensions and issues. In Briscoe (2002), 301–344.

—— 2002b. The roles of expression and representation in language evolution. In Wray (2002), 311–334.

——, M. STUDDERT-KENNEDY, & C. KNIGHT (eds.) 1998. *Approaches to the Evolution of Language*. Cambridge: Cambridge University Press.

HUSBANDS, P., & I. HARVEY (eds.) 1997. *Fourth European Conference on Artificial Life*. Cambridge, MA: MIT Press.

HUTCHINS, E., & B. HAZELHURST. 1995. How to invent a lexicon: the development of shared symbols in interaction. In *Artificial societies: the computer simulation of social life*, ed. by N. Gilbert & R. Conte, 157–189. London: University College London Press.

JACKENDOFF, R. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford: Oxford University Press.

JANIK, V. M. 2000. Whistle matching in wild bottlenose dolphins *Tursiops truncatus*. *Science* 289.1355–1357.

——, & P. J. B. SLATER. 1997. Vocal learning in mammals. *Advances in the Study of Behavior* 26.59–99.

JOHNSON-LAIRD, P. N. 1990. Introduction: What is communication? In *Ways of communicating*, ed. by D. H. Mellor, 1–13. Cambridge: Cambridge University Press.

JONES, S., M. MARTIN, & D. PILBEAM (eds.) 1992. *The Cambridge Encyclopedia of Human Evolution*. Cambridge: Cambridge University Press.

KAGAN, J. 1981. *The second year*. Cambridge, MA: Harvard University Press.

KATZ, J., & P. POSTAL. 1964. *An integrated theory of linguistic descriptions*. Cambridge, MA: MIT Press.

KEGL, J., & G. A. IWATA. 1989. Lenguaje de Signos Nicaraguense: A pidgin sheds light on the "creole?" ASL. In *Proceedings of the Fourth Meeting of the Pacific Linguistics Conference*, ed. by R. Carlson, S. DeLancey, S. Gildea, D. Payne, & A. Saxena, 266–294. Eugene, OR: University of Oregon.

——, A. SENGHAS, & M. COPPOLA. 1999. Creation through contact: Sign language emergence and sign language change in Nicaragua. In *Language Creation and Language Change*, ed. by M. DeGraff, 179–237. Cambridge, MA: MIT Press.

KELEMEN, J., & P. SOSÍK (eds.) 2001. *Advances in Artificial Life: Proceedings of the 6th European Conference on Artificial Life*. Berlin: Springer-Verlag.

KIRBY, S. 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 Knight *et al.* (2000), 303–323.

—— 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Journal of Evolutionary Computation* 5.102–110.

—— 2002. Learning, bottlenecks and the evolution of recursive syntax. In Briscoe (2002), 173–203.

——, & J. R. HURFORD. 1997. Learning, culture and evolution in the origin of linguistic constraints. In Husbands & Harvey (1997), 493–502.

——, & J. R. HURFORD. 2002. The emergence of linguistic structure: An overview of the iterated learning model. In *Simulating the Evolution of Language*, ed. by A. Cangelosi & D. Parisi, 121–147. Springer Verlag.

KNIGHT, C. 1991. *Blood Relations: Menstruation and the Origins of Culture*. New Haven, CT: Yale University Press.

——, C. POWER, & I. WATTS. 1995. The human symbolic revolution: a Darwinian account. *Cambridge Archaeological Journal* 5.75–114.

——, M. STUDDERT-KENNEDY, & J.R. HURFORD (eds.) 2000. *The Evolutionary Emergence of Language: Social Functions and the Origins of Linguistic Form*. Cambridge: Cambridge University Press.

KOLEN, J. F., & J. B. POLLACK. 1990. Back propagation is sensitive to initial conditions. *Complex Systems* 4.269–280.

KRAMER, S. N. 1963. *The Sumerians: Their History, Culture and Character*. Chicago, IL: University of Chicago Press.

KREBS, J. R., & R. DAWKINS. 1984. Animal signals: mind-reading and manipulation. In *Behavioural ecology: an evolutionary approach*, ed. by J. R. Krebs & N. B. Davies. Oxford: Blackwell Scientific Publications.

KRIFKA, M. 2001. Compositionality. In *The MIT Encyclopaedia of the Cognitive Sciences*, ed. by R. A. Wilson & F. Keil. Cambridge, MA: MIT Press.

KVASNIČKA, V., & J. POSPÍCHAL. 1999. An emergence of coordinated communication in populations of agents. *Artificial Life* 5.319–342.

LABOV, W. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

LAI, C. S. L., S. E. FISHER, J. A. HURST, F. VARGHA-KHADEM, & A. P. MONACO. 2001. A forkhead-domain gene is mutated in severe speech and language disorder. *Nature* 413.519–523.

LALAND, K., F. J. ODLING-SMEE, & M. W. FELDMAN. 2000. Niche construction, biological evolution, and cultural change. *Behavioral and Brain Sciences* 23.131–175.

LANGACKER, R. W. 1977. Syntactic reanalysis. In *Mechanisms of Syntactic Change*, ed. by C. N. Li, 57–139. Austin, TX: University of Texas Press.

LANGTON, C., C. TAYLOR, J. FARMER, & S. RASMUSSEN (eds.) 1992. *Artificial Life 2: Proceedings of the workshop on Artificial Life held February, 1990 in Santa Fe, New Mexico*. Reading, MA: Addison-Wesley.

LASS, R. 1980. *On explaining language change*. Cambridge: Cambridge University Press.

LEPAGE, R. B., & A. TABOURET-KELLER. 1985. *Acts of identity*. Cambridge: Cambridge University Press.

LEVIN, M. 1995. The evolution of understanding: a genetic algorithm model of the evolution of communication. *Biosystems* 36.167–178.

LIEBERMAN, P. 1984. *The Biology and Evolution of Language*. Cambridge, MA: Harvard University Press.

—— 2000. *Human Language and Our Reptilian Brain: the Subcortical Bases of Speech, Syntax and Thought*. Cambridge, MA: Harvard University Press.

LIGHTFOOT, D. 1979. *Principles of Diachronic Syntax*. Cambridge: Cambridge University Press.

—— 1999. *The Development of Language: Acquisition, Change, and Evolution*. Oxford: Blackwell.

LIVINGSTONE, D., & C. FYFE. 1999. Modelling the evolution of linguistic diversity. In Floreano *et al.* (1999), 704–708.

——, & ——. 2000. Modelling language–physiology coevolution. In Knight *et al.* (2000), 199–215.

LYN, H., & S. SAVAGE-RUMBAUGH. 2000. Observational word learning in two bonobos (*Pan paniscus*): ostensive and non-ostensive contexts. *Language and Communication* 20.255–273.

MACLENNAN, B., & G. BURGHARDT. 1993. Synthetic ethology and the evolution of cooperative communication. *Adaptive Behaviour* 2.161–187.

MACNAMARA, J. 1972. The cognitive basis of language learning in infants. *Psychological Review* 79.1–13.

—— 1982. *Names for things: a study of human learning*. Cambridge, MA: MIT Press.

MAŃCZAK, W. 1980. Laws of analogy. In *Historical Morphology*, ed. by J. Fisiak, 283–288. The Hague: Mouton.

MARKMAN, E. M. 1989. *Categorization and naming in children: problems of induction*. Cambridge, MA: MIT Press.

—— 1992. Constraints on word learning: speculations about their nature, origins, and domain specificity. In *Modularity and Constraints in Language and Cognition: the Minnesota Symposium on Child Psychology*, ed. by M. Gunnar & M. Maratsos, volume 25, 59–101. Hillsdale, NJ: Erlbaum.

——, & G. F. WACHTEL. 1988. Children's use of mutual exclusivity to constrain the meaning of words. *Cognitive Psychology* 20.121–157.

MARLER, P. 1957. Specific distinctiveness in the communication signals of birds. *Behaviour* 11.13–39.

MARTIN, R. 1992. Primate reproduction. In Jones *et al.* (1992), 86–90.

MARTINET, A. 1972. Function, structure and sound change. In *A reader in historical and comparative linguistics*, ed. by A. R. Keiler, 139–174. New York, NY: Holt, Reinhart and Winston.

MAYNARD SMITH, J. 1978. Optimization theory in evolution. *Annual Review of Ecology and Systematics* 9.31–56.

MCMAHON, A. 1994. *Understanding Language Change*. Cambridge: Cambridge University Press.

MCWHORTER, J. H. 1997. *Towards a New Model of Creole Genesis*. New York, NY: Peter Lang.

MILLIKAN, R. G. 1984. *Language, thought, and other biological categories: new foundations for realism*. Cambridge, MA: MIT Press.

MITCHELL, M. 1996. *An introduction to genetic algorithms.* Cambridge, MA: MIT Press.

MONTANA, D. D., & L. D. DAVIS. 1989. Training feedforward networks using genetic algorithms. In *Proceedings of the international joint conference on artificial intelligence*. Morgan Kaufmann.

NEWMEYER, F. J. 1998. *Language Form and Language Function*. Cambridge, MA: MIT Press.

NIYOGI, P., & R. C. BERWICK. 1997. Evolutionary consequences of language learning. *Linguistics and Philosophy* 20.697–719.

NOBLE, J. 1998. Evolved signals: Expensive hype vs. conspiratorial whispers. In *Artificial Life 6: Proceedings of the Sixth International Conference on Artificial Life*, ed. by C. Adami, R. Belew, H. Kitano, & C. Taylor, 358–367. Cambridge, MA: MIT Press.

—— 1999. Cooperation, conflict and the evolution of communication. *Adaptive Behavior* 7.349–370.

NOLFI, S., J. L. ELMAN, & D. PARISI. 1994. Learning and evolution in neural networks. *Adaptive Behavior* 3.5–28.

NOWAK, M. A., & N. L. KOMAROVA. 2001. Towards an evolutionary theory of language. *Trends in Cognitive Sciences* 5.288–295.

——, ——, & P. NIYOGI. 2001. Evolution of universal grammar. *Science* 291.114–117.

——, J. PLOTKIN, & D. KRAKAUER. 1999. The evolutionary language game. *Journal of Theoretical Biology* 200.147–162.

——, J. B. PLOTKIN, & V. A. A. JANSEN. 2000. The evolution of syntactic communication. *Nature* 404.495–498.

ODLING-SMEE, F. J. 1988. Niche-constructing phenotypes. In *The Role of Behavior in Evolution*, ed. by H. C. Plotkin, 73–132. Cambridge, MA: MIT Press.

O'GRADY, W., M. DOBROVOLSKY, & F. KATAMBA. 1996. *Compemporary Linguistics: An Introduction*. London: Longman, 3rd edition.

OLIPHANT, M. 1996. The dilemma of saussurean communication. *BioSystems* 37.31–38.

——, 1997. Formal approaches to innate and learned communication: Laying the foundation for language. Doctoral Dissertation, University of California, San Diego.

—— 1999. The learning barrier: Moving from innate to learned systems of communication. *Adaptive Behavior* 7.371–384.

—— 2002. Rethinking the language bottleneck: Why don't animals learn to communicate? In *Imitation in Animals and Artifacts*, ed. by K. Deutenhahn & C. L. Nehaniv, 311–325. Cambridge, MA: MIT Press.

——, & J. BATALI. 1997. Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter* 11.

OUDEYER, P-Y. 2002. Phonemic coding might be a result of sensory-motor coupling dynamics. In *Proceedings of the 7th International Conference on the Simulation of Adaptive Behavior*, ed. by B. Hallam, D. Floreano, J. Hallam, G. Hayes, & J-A. Meyer, 406–416. Cambridge, MA: MIT Press.

PEREIRA, M. E., & J. M. MACEDONIA. 1991. Ringtailed lemur anti-predator calls denote predator class, not response urgency. *Animal Behaviour* 41.543–544.

PIATTELLI-PALMARINI, M. 1989. Evolution, selection and cognition: From "learning" to parameter setting in biology and in the study of language. *Cognition* 31.1–44.

PINKER, S. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.

——, & P. BLOOM. 1990. Natural language and natural selection. *Behavioral and Brain Sciences* 13.707–784.

POPPER, K. R. 1977. *The self and its brain: an argument for interactionism*. New York, NY: Springer International.

PREMACK, D. 1983. Representational capacity and accessibility of knowledge: The case of chimpanzees. In *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, ed. by M. Piattelli-Palmarini, 205–221. London: Routledge and Kegan Paul.

PULLUM, G. K., & B. C. SCHOLZ. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19.9–50.

QUINN, M. 2001. Evolving communication without dedicated communication channels. In Kelemen & Sosík (2001), 357–366.

RAEMAEKERS, J. J., P. M. RAEMAEKERS, & E. H. HAIMOFF. 1984. Loud calls of the gibbons (*Hylobates lar*): repertoire, organization and context. *Behavior* 91.146–189.

RAGIR, S. 2002. Constraints on communities with indigenous sign languages: clues to the dynamics of language origins. In Wray (2002), 272–294.

ROLLS, E. T., & S. M. STRINGER. 2000. On the design of neural networks in the brain by genetic evolution. *Progress in Neurobiology* 61.557–579.

ROSENBLATT, F. 1958. The perceptron: A probabistic model for information storage and organization in the brain. *Psychological Review* 65.368–408.

RUMELHART, D.E., G. E. HINTON, & R. J. WILLIAMS. 1986. Learning representations by back-propagating errors. *Nature* 323.533–536.

SAMPSON, G. 1997. *Educating Eve: The 'Language Instinct' debate*. London: Cassell.

SAVAGE-RUMBAUGH, S., K. MCDONALD, R. A. SEVCIK, W. D. HOPKINS, & E. RUBERT. 1986. Spontaneous symbol acquisition and communicative use by pygmy chimpanzees (*Pan paniscus*). *Journal of Experimental Psychology: General* 115.211–235.

——, S. G. SHANKER, & T. J. TAYLOR. 1998. *Apes, Language, and the Human Mind*. Oxford: Oxford University Press.

SCHOENEMANN, P. T. 1999. Syntax as an emergent characteristic of the evolution of semantic complexity. *Minds and Machines* 9.309–346.

SEYFARTH, R. M., & D. L. CHENEY. 1986. Vocal development in vervet monkeys. *Animal Behaviour* 34.1640–1658.

SHERMAN, P. W. 1977. Nepotism and the evolution of alarm calls. *Science* 197.1246–1253.

SINGLER, J. V. (ed.) 1990. *Pidgin and creole tense-mood-aspect systems*. Amsterdam: John Benjamins.

SLOBIN, D. I. 1973. Cognitive prerequisites for the development of grammar. In *Studies of child language development*, ed. by C. A. Freguson & D. I. Slobin. New York, NY: Holt, Rinehart and Winston.

—— 1977. Language change in childhood and history. In *Language Learning and Thought*, ed. by J. Macnamara, 185–221. London: Academic Press.

—— 1985. Crosslinguistic evidence for the language-making capacity. In *The Crosslinguistic Study of Language Acquisition*, ed. by D. I. Slobin, volume 2, 1157–1249. Hillsdale, NJ: Lawrence Earlbaum Associates.

SMITH, A. D. M. 2001a. Establishing communication systems without explicit meaning transmission. In Kelemen & Sosík (2001), 381–390.

——, forthcoming. Intelligent meaning creation in a clumpy world helps communication. To appear in *Artificial Life*.

SMITH, K., 1998. Learners are losers: natural selection and learning in the evolution of communication. MA Dissertation, Department of Linguistics, University of Edinburgh.

—— 2001b. The importance of rapid cultural convergence in the evolution of learned symbolic communication. In Kelemen & Sosík (2001), 637–640.

—— 2002. The cultural evolution of communication in a population of neural networks. *Connection Science* 14.65–84.

——, in press. Natural selection and cultural selection in the evolution of communication. To appear in *Adaptive Behavior*.

——, H. BRIGHTON, & S. KIRBY, submitted. Language evolution in a multi-agent model: the cultural emergence of compositional structure. Submitted to *Discrete Dynamics in Nature and Society*.

——, 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, L. 1997. Constructing and sharing perceptual distinctions. In *Proceedings of the European Conference on Machine Learning, ECML '97,* ed. by M. van Someren & G. Widmer, 4–13. Berlin: Springer-Verlag.

—— 1998. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103.133–156.

STRINGER, C. B. 1992. Evolution of early humans. In Jones *et al.* (1992), 241–251.

TALLAL, P., & J. SCHWARTZ. 1980. Temporal processing, speech perception and hemispheric asymmetry. *Trends in neuroscience* 3.309–311.

THOMPSON, D'A. 1961. *On Growth and Form*. Cambridge: Cambridge University Press. Abridged edition, edited by J.T. Bonner.

TOMASELLO, M. 1990. Cultural transmission in the tool use and communicatory signalling of chimpanzees? In *"Language" and Intelligence in Monkeys and Apes: Comparative Developmental Perspectives*, ed. by S. Parker & K. Gibson, 274–311. Cambridge: Cambridge University Press.

—— 1993. The interpersonal origins of self-concept. In *The perceived self: Ecological and interpersonal sources of self knowledge,* ed. by U. Neisser, 174–184. Cambridge: Cambridge University Press.

—— 1996. Do apes ape? In *Social Learning in Animals: The Roots of Culture,* ed. by C. Heyes & B. Galef, 319–436. San Diego: Academic Press.

—— 1997. *Primate Cognition.* Oxford: Oxford University Press.

—— 1999. *The cultural origins of human cognition*. Harvard: Harvard University Press.

——, & M. BARTON. 1994. Learning words in non-ostensive contexts. *Developmental Psychology* 30.639–650.

——, J. CALL, J. WARREN, G.T. FROST, M. CARPENTER, & K. NAGELL. 1997. The ontogeny of chimpanzee gestural signals: A comparison across groups and generations. *Evolution of Communication* 1.223–259.

——, R. STROSBERG, & N. AKHTAR. 1996. Eighteen-month-old children learn words in non-ostensive contexts. *Journal of Child Language* 23.157–176.

TRASK, R. L. 1995. *Language: The Basics*. London: Routledge.

TRAUGOTT, E. C. 1965. Diachronic syntax and generative grammar. *Language* 41.402–415.

TRUDGILL, P. 1972. Sex, covert prestige, and linguistic change in the urban British English of Norwich. *Language in Society* 1.179–196.

TURKEL, W. J. 2002. The learning guided evolution of natural language. In Briscoe (2002), 235–254.

TYACK, P. 1986. Whistle repertoires of two bottlenosed dolphins, *Tursiops truncatus*: mimicry of signature whistles? *Behavioral Ecology and Sociobiology* 18.251–257.

VARGHA-KHADEM, F., K. WATKINS, K. ALCOCK, P. FLETCHER, & R. PASSINGHAM. 1995. Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. *Proceedings of the National Academy of Science* 92.930–933.

VENNEMANN, T. 1978. Phonetic analogy and conceptual analogy. In *Readings in Historical Phonology*, ed. by P. Baldi & R. N. Werth, 258–274. University Park, PN: Pennsylvania State University Press.

VON FRISCH, K. 1974. Decoding the language of the bee. *Science* 185.663–668.

WANNER, E., & L. R. GLEITMAN. 1982. Language acquisition: the state of the art. In *Language acquisition: the state of the art*, ed. by E. Wanner & L. R. Gleitman, 3–50. Cambridge: Cambridge University Press.

WERNER, G., & M. DYER. 1992. Evolution of communication in artificial organisms. In Langton *et al.* (1992), 659–687.

WEXLER, K., & P. W. CULICOVER. 1980. *Formal Principles of Language Acquisition*. Cambridge, MA: MIT Press.

WILKINS, W. K., & J. WAKEFIELD. 1995. Brain evolution and neurolinguistic preconditions. *Behavioral and Brain Sciences* 18.161–226.

WOOD, B. A. 1992. Evolution of australopithecines. In Jones *et al.* (1992), 231–240.

WRAY, A. (ed.) 2002. *The Transition to Language*. Oxford: Oxford University Press.

——, & M. R. PERKINS. 2000. The functions of formulaic language: an integrated model. *Language and Communication* 20.1–28.

YAMAUCHI, H. 2001. The difficulty of the Baldwinian account of linguistic innateness. In Kelemen & Sosík (2001), 391–400.

ZAHAVI, A. 1975. Mate selection — a selection for a handicap. *Journal of Theoretical Biology* 53.205–214.

—— 1977. The cost of honesty (further remarks on the handicap principle). *Journal of Theoretical Biology* 67.603–605.

ZUBERBUHLER, K. 2001. Predator-specific alarm calls in Campbell's monkeys, *Cercopithecus campbelli*. *Behavioral Ecology & Sociobiology* 50.414–422.

——, R. NOË, & R. SEYFARTH. 1997. Diana monkey long-distance calls: messages for conspecifics and predators. *Animal Behaviour* 53.589–604.