2007

The Evolution of Language Universals: Optimal Design and Adaptation

Turner, Huck

The Evolution of Language Universals: Optimal Design and Adaptation

Huck Turner

A thesis submitted in partial fulfilment of the degree of Doctor of Philosophy.

Centre for Theoretical and Computational Neuroscience, University of Plymouth, UK.
May, 2007

The Evolution of Language Universals: Optimal Design and Adaptation

by Huck Turner

Abstract

Inquiry into the evolution of syntactic universals is hampered by severe limitations on the available evidence. Theories of selective function nevertheless lead to predictions of local optimality that can be tested scientifically. This thesis refines a diagnostic, originally proposed by Parker and Maynard Smith (1990), for identifying selective functions on this basis and applies it to the evolution of two syntactic universals: (1) the distinction between open and closed lexical classes, and (2) nested constituent structure. In the case of the former, it is argued that the selective role of the closed class items is primarily to minimise the amount of redundancy in the lexicon. In the case of the latter, the emergence of nested phrase structure is argued to have been a by-product of selection for the ability to perform insertion operations on sequences – a function that plausibly pre-dated the emergence of modern language competence. The evidence for these claims is not just that these properties perform plausibly fitness-related functions, but that they appear to perform them in a way that is improbably optimal.

A number of interesting findings follow when examining the selective role of the closed classes. In particular, case, agreement and the requirement that sentences have subjects are expected consequences of an optimised lexicon, the theory thereby relating these properties to natural selection for the first time. It also motivates the view that language variation is confined to parameters associated with closed class items, in turn explaining why parameter conflicts fail to arise in bilingualism.

The simplest representation of sequences that is optimised for efficient insertions can represent both nested constituent structure and long-distance dependencies in a unified way, thus suggesting that movement is intrinsic to the representation of constituency rather than an 'imperfection'. The basic structure of phrases also follows from this representation and helps to explain the interaction between case and theta assignment. These findings bring together a surprising array of phenomena, reinforcing its correctness as the representational basis of syntactic structures.

The diagnostic overcomes shortcomings in the approach of Pinker and Bloom (1990), who argued that the appearance of 'adaptive complexity' in the design of a trait could be used as evidence of its selective function, but there is no reason to expect the refinements of natural selection to increase complexity in any given case.

Optimality considerations are also applied in this thesis to filter theories of the nature of unobserved linguistic representations as well as theories of their functions. In this context, it is argued that, despite Chomsky's (1995) resistance to the idea, it is possible to motivate the guiding principles of the Minimalist Program in terms of evolutionary optimisation, especially if we allow the possibility that properties of language were selected for non-communicative functions and that redundancy is sometimes costly rather than beneficial.

To *Brett Jewell* (1974-2001) for presenting me with a vision of my future, which I adopted as a template, a vision forged around a campfire and under a remote street lamp on a warm summer's night a long time ago.

To *Loretta Cadenaro* for being prepared to take a bullet for me.

And to *Crunchy*, a surprisingly well-fed stray cat, who recently taught me a thing or two about how natural selection works in practice.

Contents

List of Tables and Illustrations

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Acknowledgements

The ideas presented in this thesis have been inspired by and enriched through discussion with many people. In particular, I am indebted to the many people who provided me with helpful feedback when I presented aspects of this work at the *Evolution of Language* conference at the Max Plank Institute for Evolutionary Anthropology in Leipzig in March-April 2004, the *Evolution of Language* conference at Harvard in March 2002, and the *Language, Brain, Culture* conference at the University of Sydney in December 2001. I would particularly like to acknowledge discussions with Michael Arbib, Robert Berwick, Noam Chomsky, Morten Christiansen, Terrence Deacon, Daniel Dennett, Tecumseh Fitch, Chris Golston, Txuss Martin, and Willem Zuidema. In a separate category, I would also like to thank the members of the Language Evolution and Computation group at the University of Edinburgh, especially Jim Hurford, Simon Kirby, Anna Parker, Andrew Smith, Kenny Smith and Monica Tamariz for their incisive comments both at the aforementioned conferences and when I travelled to Edinburgh to present parts of my work in talks delivered to their group in November 2001 and March 2006. I especially thank Anna who read chapter three and provided a very helpful critique. I'm also grateful to the members of the Centre for Theoretical and Computational Neuroscience and the School of Computing at the University of Plymouth, who have drawn my attention to many important subtleties during seminars I've given to them.

I am grateful to Roman Borisyuk, Paul Turner and André Grüning for checking over the mathematical formalisations I've used both in the current and earlier incarnations of my work, to Lynn Richards for providing helpful comments on an early draft of the language acquisition section, to Sana Murrani-Cooke for producing the beautiful spandrel illustration in section 3.1.1, to Loretta Cadenaro for proof-reading early sections and spotting more logical errors than anyone else, to Angelo Cangelosi

for giving me the opportunity and funding to undertake this research in the first place, and to my parents for all their support along the way too.

There is also a long list of people who have engaged me in far-reaching discussions about fundamentals on many occasions. Special mentions must go to Davide Marocco, John Dylan-Haynes, Andrew Hennell, Michael Norris, Michele Burigo, Theo Kyriacou, Linda Lanyon, Eduardo Coutinho, Joao Martins, Genoveva Gonzalez-Mirelis and Stalin Munoz.

I am especially grateful to my supervisors Dr. Sue Denham and Prof. Chris Harris for their guidance and patience. The enthusiasm they've shown towards my work has helped motivate me to see it through.

Naturally, I take full responsibility for the shortcomings that persist in this thesis and none of the aforementioned deserve to be tainted by their association with it.

This research was partly financed with a grant from the Engineering and Physical Sciences Research Council (grant number GR/N01118).

Author's declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee. This study was financed with the aid of a studentship, provided as part of a grant from the Engineering and Physical Sciences Research Council (grant number GR/N01118). It was not carried out as part of a collaborative project of any kind. Relevant scientific seminars were regularly attended and work was presented at a number of relevant international conferences. External institutions were visited for consultation purposes and a number of papers prepared for publication.

Details of publications and presentations can be found on the following page.

Word count of main body of thesis: 61989

Signed .....

Date 18/4/08

Publications and presentations

Publications:

Turner, H. 2002. An introduction to methods for simulating the evolution of language. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language*. London: Springer-Verlag.

Cangelosi, A. & Turner, H. 2002. L'emergere del linguaggio. In A.M. Borghi & T. Iachini (Eds.), *Scienze della mente*. Bologna: Il Mulino (in Italian).

Conference presentations and invited talks:

Turner, H. & Cangelosi, A. 2001. *The role of memory load in syntactic processing and language evolution*. Paper presented at Language, Brain, and Culture Conference. University of Sydney, Sydney.

Turner, H. 2001. *Implicating working memory in the representation of constituent structure and the origins of word-order universals*. Invited talk delivered to the Language Evolution and Computation Group, University of Edinburgh.

Turner, H. & Cangelosi, A. 2002. *Implicating working memory in the representation of constituent structure and the origins of word-order universals*. Paper presented at The 4th International Conference on the Evolution of Language, Harvard University, Boston.

Turner, H. 2004. *The appearance of design in grammatical universals as evidence of adaptation for non-communicative functions*. Paper presented at The 5th International Conference on the Evolution of Language, Max Plank Institute for Evolutionary Anthropology, Leipzig.

Turner, H. 2006. *The evolution of language universals: Optimal design and adaptation*. Invited talk delivered to the Language Evolution and Computation Group, University of Edinburgh.

1

Introduction

It is one thing to ask why a language capacity evolved in the hominid line and another to ask why it evolved with the specific properties that it has. Why for instance, does modern human language have properties like those in (1)?

1. a. case

   (e.g., the property that requires us to use *he* and *him* in different contexts)

   b. agreement

   (e.g., *I am happy* is acceptable but *I are happy* is not)

   c. displacement/movement

   (e.g., *the man* takes the same semantic role with respect to the event described by the main verb in paraphrases like *The dog bit the man* and *The man was bitten* despite appearing in different structural positions)

   d. binding constraints that determine when noun phrases can and cannot co-refer

   (e.g., *him* must refer to someone other than *George* in a sentence like *George attacked him*, but may co-refer with *George* in a sentence like *George's father attacked him*)

These and many other properties show up again and again in all of the world's language families (Chomsky, 1986; Croft, 1990; Greenberg, 1966) and in newly emerging creole languages (Bickerton, 1977; Kegl, Senghas & Coppola, 1999). Regardless of whether these properties are strictly universal, their ubiquity warrants an explanation in terms of

what makes them the most stable outcome of language development under typical genetic and environmental conditions and whether the emergence of these traits conferred a selective advantage to ancestors. Despite this, much of the work on the evolution of language has focussed on the selective advantage of the whole and ignored the parts, but it is far from obvious how functional descriptions at these different levels can be reconciled.

The situation is not unlike descriptions of the function of a winter coat. As a whole, the main function of a winter coat is to keep its wearer warm, but we would be missing something by attributing the same function to its buttons, pockets and colour. Although the buttons will indirectly serve to keep the wearer warm by allowing the coat to be fastened shut, some consideration of the fastening function is crucial for explaining properties of buttons that are not shared by other parts of the coat such as the lining and so on. By also considering its pockets, we are forced to broaden the definition of the coat's function to include less obvious things like the ability to carry small items. As for the coat's colour, this may be designed or it may be a by-product of the materials used – those materials being chosen on the basis of other things like their thermal characteristics and cost. The individual properties of the language capacity may of course fall into a variety of analogous categories.

The question of how we might inquire into the evolution of these properties is a formidable one, since it is far from obvious how we could test our theories. The case of language is particularly challenging because many of the sources of evidence that evolutionary biologists usually rely on are unavailable. There is, as yet, no evidence of properties like those listed in (1) in other species to form the basis of comparisons and very little can be inferred from the fossil record about changes in linguistic or, for that matter, any other kind of cognitive ability. Despite this, Botha (2003) argues that the main obstacle to advancing our understanding of the evolution of language is not the

paucity of evidence as such, but the paucity of restrictive theory, "restrictive to the extent that it makes it possible to distinguish in a non-arbitrary way between entities that are instances of a specific kind of evolutionary event, process or product and entities that are not" (Botha, 2003: 115).

One of the targets of Botha's criticisms is the approach adopted by Pinker and Bloom (1990), which involves making inferences about the adaptive significance of language by looking at whether or not it reveals any functional pressures in its design. Pinker and Bloom (1990: 709) argue that language reveals such pressures because it exhibits what they call "adaptive complexity", which they define as a property of "any system composed of many interacting parts where the details of the parts' structure and arrangement suggest design to fulfill some function". They say very little about how adaptations could be identified systematically, but some attempts have been made within mainstream evolutionary biology to address exactly this question. These ideas have not yet filtered through into discussions of language evolution, thus providing a major motivation for the present work. In integrating some of these ideas, I attempt to establish a more rigorous and restrictive diagnostic of adaptive function based on Parker and Maynard Smith (1990), which, although broadly compatible with Pinker and Bloom (1990), dispenses with the problematic notion of 'complexity' in favour of considerations of 'design optimality'.

The present work applies this diagnostic to two separate examples of syntactic universals. In chapter five, I use this diagnostic to argue that the selective role of the closed classes is primarily to minimise the amount of redundancy in the lexicon. In chapter six, I argue that the emergence of nested phrase structure is a by-product of selection for the ability to perform certain kinds of operations on sequences – a function that plausibly pre-dated the emergence of modern language.

Before getting to these arguments, there are three chapters examining the background literature – one about language, one about evolution, and one about the evolution of language. Specifically, chapter two introduces the linguistic background that is necessary to characterise the universals under consideration. Chapter three then examines the kinds of evolutionary explanations that can be applied and what kind of evidence would allow us to distinguish between them in a principled way. Chapter four reviews studies that are specifically about the evolution of language, thus placing the current work in context and providing a point of contrast to illustrate how the current approach departs from the methods employed in that body of work.

The optimality diagnostic is introduced and developed in chapter three whilst confronting a number of obstacles that stand in the way of providing a selectionist account of language evolution. Some of these obstacles apply to evolutionary argumentation generally and some apply to inquiry into the evolution of language specifically. The general problems concern whether there are categories of evolutionary explanation other than natural selection that can account for the existence of a trait, as argued by Stephen Jay Gould and others (Gould & Lewontin, 1979; Gould & Vrba, 1982; Gould, 1991; Lightfoot, 2000; Piattelli-Palmarini, 1989; Uriagereka, 1998). I argue that there are certainly important lessons to be drawn from Gould's observations that are important for constructing rigorous evolutionary arguments, but while he and his colleagues argue that there are a plurality of *forces* acting in addition to natural selection, I argue that they only succeed in demonstrating that natural selection has a plurality of *products*.

Chapter three also addresses a number of problems specific to inquiry into language evolution. Chomsky (2002) and others argue that language is too elegant in its design to have been the result of evolutionary 'tinkering'. Lightfoot (2000) and others argue that many properties of grammar are actually maladaptive. Geschwind (1980)

questions how any mutation of the language capacity could ever be favoured if the mutant is not conforming to the rest of the linguistic community. Hauser, Chomsky and Fitch (2002: 1574) argue that properties of grammar have a "tenuous connection to communicative efficacy". I attempt to show that none of these arguments are compelling, thus leaving the door wide open for selectionist explanations and hence the proposals to follow in chapters five and six.

To take the last of these arguments, the observation that properties like those in (1) have a "tenuous connection to communicative efficacy" (Hauser *et al*, 2002: 1574) is only evidence that they were not selected for *communicative* functions, not that they weren't selected for *non-communicative* functions. For instance, a universal could be selected for improving language learnability or for its effects on reducing the costs associated with the language faculty in terms of metabolic energy or other neural resources. The proposals in chapters five and six are of this character.

Chapter five examines closed-class lexical items (i.e., grammatical function words and inflections) and concludes that they have the effect of minimising the amount of redundancy in the lexicon. By encapsulating lexical categories, closed class items can mediate grammatical relations so that lexical entries can remain extremely economical in terms of the number of formal features they need to contain. For instance, learning that a noun is associated with determiners allows a noun, encapsulated within a determiner phrase, to be used as either the subject or object of a sentence or as the object of a preposition and so forth. The language learner does not have to learn all of the contexts in which a new noun can be used because this information is encoded in the few words that constitute the closed class of determiners. So long as the proportion of closed-class items in the lexicon is small relative to the open-class items, the additional representational burden that they impose will be more than offset by the reduction in redundancy in the very many more open-class items. This has the effect of minimising

the storage requirements of the lexicon, and would thereby presumably translate into savings of metabolic and neural resources – savings which we can expect natural selection to favour. A number of interesting findings also follow when examining the consequences of the theory. In particular, it appears that important properties like case (1a), agreement (1b) and the requirement that sentences have subjects are expected consequences of an optimised lexicon, the theory thereby relating these properties to natural selection for the first time.

Chapter six applies the optimality diagnostic to the representation of sequences. A number of proposals for how sequences could be represented are evaluated with respect to the efficiency with which items can be inserted into them. The worst case would be if each token is associated with an index marking its absolute order in the sequence, since the insertion of a new token anywhere before the end would require the remainder to be re-indexed. Other representations are considered that encode relative order rather than absolute order and are compared with respect to the level of redundancy that insertion operations cause. The optimal representation allows an insertion to be made with a single operation and without introducing redundancy. The surprising feature of this representation is that it also allows nested phrase structure to be represented, which suggests that the mechanisms involved were co-opted rather than selected for their role in language, having been originally selected for manipulating sequences. Some further consequences of the theory are explored by looking at a lower-level description of the same kind of representation. Strikingly, this representation appears to unify the representation of phrase structure with representations of the relation that is also implicated in movement (1c) and binding (1d). This relation, called *c-command*, is reviewed in chapter two. The theory also motivates other fundamental properties of language including the specifier-head-complement structure of phrases and sheds new light on the relationship between case and argument roles.

The optimality diagnostic, as applied to these domains, has served to focus inquiry in interesting ways. As well as providing evidence of the original adaptive function of the properties in question, it has served to relate various phenomena that until now were not thought to be related, and to draw attention to other phenomena that have mostly escaped attention. The approach has therefore not only been informative about the origins of these properties, but also their nature. By focussing inquiry into the nature of language, it has many parallels with the Minimalist Program in linguistics, which proceeds according to the working assumption that the computational system that implements grammar is a perfect solution to the constraints imposed upon it by the systems with which it interfaces (Chomsky, 1995). At first glance, the assumption of perfection would appear to be entirely compatible with the optimising influence of natural selection, but this interpretation is rejected by Chomsky (1995) for reasons I examine in chapter three. There, I argue that Chomsky's rejection of this position is nevertheless premature and conclude that selectionist accounts are not only compatible with the assumption of perfection but actually provide independent motivation for why we should expect inquiry based on it to succeed in generating theories with better empirical coverage. Additionally, by further specifying that this perfection be of a type that plausibly relates to fitness (generally in terms of minimising metabolic or other costs rather than maximising communicative efficacy), the approach can be made even more restrictive. The number of substantive results generated in the present work would appear to reinforce this view.

In terms of the scope of the present work, it is worth clarifying that the focus here is on the evolution of the computational system that implements grammar rather than the articulatory-perceptual or conceptual-intentional systems. Although questions about the evolution of these other systems are interesting in their own right, they are separable. The articulatory-perceptual systems can be separated because language is not

tied to any particular modality of expression, the same signal being expressible via speech, writing, hand signs for the deaf, braille for the blind, and many other conceivable systems. Thinking in words does not even require a signal to be externalised, but is still language. The distinction between the grammar system and conceptual-intentional systems is more difficult to motivate, but it is probably desirable to differentiate between the capacity for grammatical language and more general capacities for communication and thought which rely on the same conceptual knowledge but take non-linguistic forms.

The final chapter summarises the methodological and empirical contributions of the present work and highlights further challenges. The central methodological contribution lies in illustrating how optimality considerations can be applied profitably to focus inquiry into both the nature and function of evolved traits. Its success as a methodology is reinforced by the empirical contributions of the present work which deepen our understanding of the interrelations between many of the aforementioned linguistic properties. I conclude with some considerations about how the optimality diagnostic could be generalised to co-evolutionary dynamics involving dynamical optima.

2

Syntactic Universals

This chapter provides the necessary background to characterise the syntactic universals that are the subject of evolutionary claims in chapters five and six.

Ongoing disagreements about the nature of the language faculty mean it is something of a moving target for evolutionary theory, but there are nevertheless certain properties that most rational observers agree should be attributed to it even if questions remain about whether the principles that linguists use to describe them are formulated correctly, whether they are specific to language or the species, and what genetic and environmental factors influence their development. The present review covers areas of consensus as well as some areas of current research. Naturally, the focus is on characterising the phenomena that are the subject of the substantive claims of the thesis so much more could of course be said about the language faculty than can justifiably be covered here.

2.1 Approaches to the study of universals

2.1.1 The typological approach

Cross-linguistic comparisons are clearly an important source of data about language universals, revealing patterns that cannot be observed by analysing languages in isolation. This is the guiding principle of what is known as the typological approach.

According to Croft (1990), linguistic typology is more than the taxonomical classification of languages into types. The patterns that are discovered by cross-

linguistic comparison lead to theoretical claims about what a person knows when he or she knows a language and hence to claims about the analyses of individual languages.

The typological approach grew largely out of the work of Joseph Greenberg in the 1960s. In one of his earliest studies, Greenberg (1966) identified 45 universals from analysis of 30 languages from all over the world. Some of these universals were absolute as in (1) and some were statistical as in (2).

1. Languages with dominant VSO[1] order are always prepositional (Greenberg, 1966: 78).
2. With overwhelmingly greater than chance frequency, languages with normal SOV order are postpositional[2] (Greenberg, 1966: 79).

As in these examples, most of the universals that Greenberg identified were implicational, having the form "if a language has property *x*, then it will also have property *y*".

Another important kind of universal is the hierarchical universal, which is equivalent to a chain of implicational universals. An example is Keenan and Comrie's (1977) relative clause accessibility hierarchy which says that if a language allows possessors to take relative clauses, then it will also allow non-direct objects to take them and if non-direct objects can take them, then so can objects, and if objects can, so can subjects. The hierarchy is summarised using the notation in (3). If a language can relativise a particular type on this hierarchy then it can also relativise every type that appears to its left.

3. subject > direct object > non-direct object > possessor

[1] VSO is short for verb-subject-object.
[2] Postpositions are the same as prepositions, except they appear after the noun phrase they select.

Linguistic typology is often contrasted with the generative approaches, which seek to provide a more explicit theory of what a language user knows when he or she knows a language.

2.1.2 The generative approach

Under the generative approach, particular analyses are often argued on the basis of examples from within a single language, which form the basis of hypotheses about universals that are then tested with respect to other languages. To do this, researchers in the generative tradition rely heavily on potentially very subtle judgments about the contrasting grammaticality of carefully constructed examples. This is a very powerful method which is why researchers in this tradition are able to formulate such explicit hypotheses, but because of the difficulty of making such judgements for languages other than those that the researcher speaks with native proficiency, claims about universals need to be tested by eliciting judgements from a suitable sample of informants from different language-speaking backgrounds.

Early work in the generative tradition from the 1950s focussed on the *descriptive adequacy* of grammars. The goal was to produce grammars that generated all of the well-formed sentences of a language and none of the ill-formed sentences and these grammars were produced using rules specific to each language. With significant progress in this aim, an additional aim came to the fore, that of *explanatory adequacy* – explaining what a person knows when he or she knows a language and how this knowledge is acquired. This gave rise to what is known as the Principles and Parameters approach, an early version of which was articulated by Chomsky (1981).

Central to this approach is the notion of *universal grammar*, which Chomsky (1986) defined as the initial state of the language acquisition device prior to linguistic experience. This state is taken to be universal to the species in the sense that differences

between individual languages do not result from differences in the initial state. The term *universal grammar* is potentially misleading since the language acquisition device, in its initial state, is not a grammar in the sense of a complete description of a language nor is the linguistic knowledge that it embodies limited to grammar in a narrow sense since it encompasses principles of phonology and semantics (although in practice Chomsky's focus has been syntax). Another problem with the term arises in the context of discussions about the evolution of the language acquisition device where hypotheses about natural selection favouring some variant of universal grammar might be entertained, but to suggest that different variants of a universal could co-exist is at odds with the usual sense of 'universal'. For these reasons, the more transparent term, *language acquisition device* will be used in place of *universal grammar* throughout the present text. Note that the term *universal grammar* is also often used to refer to specific theories of Chomsky's about the nature of the initial state and the knowledge it embodies, but a connectionist model of the initial state, for instance, would also qualify as a theory of universal grammar in the sense in which the term was coined.

Under the Principles and Parameters approach, typological universals are taken to result from the fact that the language acquisition device places constraints on the form that grammars can take. These constraints are formalised in terms of principles that must hold for all languages,[3] and parameters that allow languages to vary in limited respects. A particular grammar is characterised by a particular setting of these parameters (rather than by a set of language-specific rules), and the acquisition of a language is viewed as the process by which these parameters are set. The role of language data is as a kind of triggering experience analogous to that of visual input in the maturation of the visual system.

[3] Here and throughout, 'language' will be used in the sense of I-language in Chomsky (1986), an I-language being the *internalised* grammar of an individual adult speaker as opposed to the more problematic notion of a 'language' as spoken by a speech community.

Chomsky (1965; 1986) argues that the existence of these constraints is necessary to explain how a language can be learnt from the evidence available to the language learner.

> [L]anguage poses in a sharp and clear form what has sometimes been called "Plato's problem," the problem of "poverty of stimulus," of accounting for the richness, complexity, and specificity of shared knowledge, given the limitations of the data available (Chomsky, 1986: 7).

Of particular interest is the finding that children appear to acquire a language in the virtual absence of negative evidence (Brown & Hanlon, 1970). Children are rarely corrected when they make mistakes, and even indirect evidence arising from the failure of parents to comprehend ungrammatical sentences is insufficient to draw the appropriate generalisations. Brown and Hanlon found that parental replies indicating a lack of understanding followed about as many grammatical utterances (42%) as ungrammatical ones (47%) and replies indicating understanding followed grammatical and ungrammatical utterances with equal frequency (45%). Marcus (1993) defends Brown and Hanlon's position against criticisms that parental replies are not so sharply distinguished and that a kind of noisy feedback is available. He analysed these claims statistically and concluded that "a child would have to repeat a given sentence verbatim more than 85 times to eliminate it from his or her grammar." (Marcus, 1993: 57). Furthermore, even if negative evidence is available in certain contexts, there is a great deal of anecdotal evidence that children simply fail to understand the point of corrections. Braine (1971: 160f) recounts a particular instance:

> One case was use by my two-and-a-half-year-old daughter of *other one* as a noun modifier. Over a period of a few weeks I repeatedly but fruitlessly tried to

persuade her to substitute *other* + N for *other one* + N. With different nouns on different occasions, the interchanges went somewhat as follows: "Want other one spoon, Daddy" – "You mean, you want THE OTHER SPOON" – "Yes, I want other one spoon, please, Daddy" – "Can you say 'the other spoon'?" – "Other...one...spoon" – "Say...'other'" – "Other" – "Spoon" – "Spoon" – "Other...spoon" – "Other...spoon. Now give me other one spoon?" Further tuition is ruled out by her protest, vigorously supported by my wife. Examples indicating a similar difficulty in using negative information will probably be available to any reader who has tried to correct the grammar of a two- or three-year-old child.

A number of mathematical proofs have been proposed in support of learnability claims (Gold, 1967; Nowak, Komarova & Niyogi, 2001) based on various assumptions about the learning procedure, but these are often based on worst-case learning conditions. Zuidema (2003) argues that a language should be easy to learn because the primary linguistic data is itself the output of a language learning process. But even if language acquisition generally occurs under conditions that are more favourable than researchers have previously thought, there are many documented cases in which languages are acquired despite the available data being much worse than what is normally available to a child and even in cases where children do not appear to have been exposed to *any* relevant data at all. The most striking examples of this occur in the process of *creolisation*, which will be discussed in section 2.2.2.

2.1.3 The Minimalist Program

During the 1990s, Chomsky (1991; 1993; 1995; 1999) instigated a new program of research within the generative tradition that focuses inquiry on theories that predict the system that implements grammar to be an extremely elegant, non-redundant solution to

the demands placed on it by other systems of the brain. The decision to focus inquiry in this way arose from the observation that, in the past, when principles have been postulated with overlapping coverage

> Repeatedly, it has been found that these are wrongly formulated and must be replaced by non redundant ones. The discovery has been so regular that the need to eliminate redundancy has become a working principle in inquiry. (Chomsky, 1995: 5)

This mode of inquiry is known as the *Minimalist Program* and has led to a number of interesting theoretical developments such as the bare phrase structure theory of Chomsky (1995: ch4) to be discussed in section 2.5.

It is far from obvious why the language faculty should exhibit the kind of elegance that Minimalist theories predict. Indeed, Chomsky (1995) regards this perfection to be a surprising feature of a biological system. But the success of the Minimalist Program does not rely on knowing why it yields results – only that it does. The situation is not unlike a gold prospector restricting his search for gold to quartz-rich rock. It isn't necessary to know why gold is very often to be found in quartz to exploit the regularity. Nevertheless, I will examine how the language faculty could come to exhibit this kind of elegance in the course of chapter three and use this explanation to justify a suggested refinement of the Minimalist approach that will be pursued for the remainder of the thesis.

2.2 Language acquisition

The study of syntactic universals necessarily involves the study of the properties of the language acquisition device. Hence, we should expect the way language develops in infancy to be instructive about its nature. Of particular interest are the cases of language

development that occur under exposure to extremely sparse or inconsistent linguistic input. In such cases, we observe the language acquisition device imposing order on the data to create new languages. Before discussing this process, called *creolisation*, I review the normal course of language development.

2.2.1 Stages in language development

Under normal circumstances, language acquisition proceeds through a number of distinctly recognisable stages marked by the kind of speech that infants produce. The first stage, characterised by prelinguistic vocalisations (babbling, etc.), is followed by a single word stage beginning at around 12 months. Infants begin to combine words in utterances from around 18 months and are usually producing completely adult constructions by about school age. Each of these stages will be examined in this section.

Stoel-Gammon and Menn (1997) divide the prelinguistic period into a number of sub-stages. In the first four months after birth, vocalisations are strongly associated with emotional states. In the first two months, these are usually limited to cries, but infants soon begin to express other emotional states by cooing and chuckling. From four to six months, infants begin to modulate the pitch and volume of vocalisations learning to yell, whisper, squeal and growl. During this period, they may also produce raspberries and sustained vowels. From around six months, infants enter the stage known as *canonical babbling* which is characterised first by the appearance of repetitive consonant-vowel sequences such as 'dadada' and 'mamama', and later by sequences in which the consonants and vowels vary within the repeating consonant-vowel pattern. From around ten months, infants begin to use adult-like intonation and stress in their vocalisations both while interacting with other people and in solitary play. Although they are still not using recognisable words during this period, vocalisations (often accompanied by adult-like gestures) provide a very convincing impression of adult

speech. The infant's first words appear at about 12 months, but they continue to babble for several months after this.

The first words that appear in an infant's productive vocabulary have been studied extensively with respect to their phonological and pragmatic properties. On the phonological side, the sounds used in early words have been found to be the same as those favoured in babbling (Stoel-Gammon & Menn, 1997). Pragmatically, early words tend to label aspects of the world that are important in the infant's daily life (Pan & Gleason, 1997).

The earliest single-word utterances include performatives such as 'hi' and 'bye' as well as *holophrastic* utterances that perform the function of complete sentences in adult language. For example, an infant might use 'drink' on its own as a way of requesting a drink.

When infants start to combine words at around 18 to 24 months, some basic properties of their syntactic knowledge begin to reveal themselves in that their two-word utterances usually conform to adult word order (Tager-Flusberg, 1997).

Braine (1963) also noted that early two-word utterances conform to paradigms with one word (the *pivot*) remaining constant while an *open* element varies. For example, one of the children in Braine's study was able to use 'see' as a pivot with a number of nouns such as 'boy' and 'sock' as the open word thus producing utterances such as 'see boy' and 'see sock'.

Infants soon begin producing longer utterances, but these early 'sentences' are characterised by a lack of function words and inflections. As a result, this stage has been labelled telegraphic speech because of its resemblance to the abbreviated language of telegrams (Brown, 1973). Brown and his colleagues found that during this stage, infants omit grammatical function words and inflections even in repetition tasks. For instance,

when prompted to repeat the sentence "I am drawing a dog", one of Brown and Fraser's (1963) subjects (aged 28½ months) responded with "I draw dog".

The last stage is the transition to producing adult-like constructions, which is usually complete by school age, but children continue to acquire vocabulary at a phenomenal rate until adolescence and to a lesser extent throughout life.

The stages in development that have been reviewed here have been described in terms of the productions of infants, but Elliot (1981: 81) cites a number of problems with using this kind of evidence to gauge a child's level of competence:

> Children come out with a lot of language which appears to indicate a more sophisticated level of language development than they have actually attained. For example, they sometimes go through a period of echoing the last parts of utterances addressed to them, often after a delay, so that it is difficult to tell whether an utterance has been spontaneously created by the child or is an imitation of an adult utterance. They often learn stock phrases ... without being able to segment the phrase and recombine its parts.

Studies that seek to test comprehension (e.g., Benedict, 1979) rather than production are also difficult to assess because children can often rely on contextual cues to infer intended meanings. Careful experiments have nevertheless been carried out using a preferential looking paradigm and sentence pairs that can only be disambiguated using word order (Hirsh-Pasek & Golinkoff, 1993). Golinkoff and Hirsh-Pasek (1995) found that children were able to reliably understand two word utterances before they were producing them at 17 months.

2.2.2 Creolisation and its implications

New languages occasionally form under circumstances in which adults of different language-speaking backgrounds are brought together in a community, often as slaves. At first, adults spontaneously develop a language that is sufficient for rudimentary communication. This initial language is known as a *pidgin* and lacks many of the properties we usually associate with natural languages like movement, recursive structure and functional categories (i.e., function words like determiners and auxiliaries, and morphological inflections like markers of tense and number). Utterances of pidgin languages could be regarded as being like those of children in the telegraphic stage of language acquisition at least insofar as they lack more-or-less the same properties.

In time, pidgins develop into *creoles*, which are full-fledged languages with structures characteristic of English or any other natural language. Given this, it is surprising that the transition from pidgin to creole can occur within a very short time, arguably even within a single generation (Bickerton, 1977).

DeGraff (1999) distinguishes between three different theoretical positions with respect to how this process works. These are the universalist, substratist and superstratist positions. Universalists such as Bickerton (1977) believe that creoles are invented by the children who acquire them natively under exposure to the pidgin input, their innate capacity for language allowing them to fill in those aspects of the language that are missing from the primary linguistic data. Substratists and superstratists argue that the properties of creoles result from the influence of other languages. The former (e.g., Lefebvre & Lumsden, 1989) take the substrate languages (i.e., the ancestral languages of the adults) to be influential and indeed the children usually acquire the native language of their parents in addition to the creole. The latter (e.g., Chaudenson, 1979) argue that the superstrate language (i.e., the socially dominant language spoken in the community) provides the model on which properties of the creole are patterned.

Indeed, the superstrate language is usually the lexifier language (i.e., the language from which vocabulary items are mostly drawn).

It is conceivable that each of the influences described by universalists, substratists and superstratists are active in the formation of creoles to varying degrees, but only the universalist claims are interesting in the context of the present study. To assess these, it is prudent to look at those instances in which the process was least likely to have been influenced by language contact. Bickerton (1999) attempted to do this by limiting his observations to plantation creoles formed after the displacement of populations from their ancestral communities thus minimising any influence from continued contact with substrate languages. Hawaiian creole is one such language and since this language formed relatively recently (about a hundred years ago), many records exist of utterances before, during and after its development. Bickerton (1999: 53) describes the early form of that language as follows:

> [A]ll utterances show similar limitations (if not a complete absence) of grammatical structure: an almost complete absence of grammatical items (including a complete absence of tense, modality, and aspect (TMA) markers), a virtually complete absence of embedded structures, and frequent ellipses of arguments and even verbs.

Within a decade or so, a recognisably distinct creole had emerged among the children of the community. This language had a developed system of grammatical morphemes. For instance, in the auxiliary system, speakers used the word 'bin' to mark tense, 'go' to mark modality, and 'stay' to mark aspect. Bickerton (1999: 57) describes how these words come to be used with these functions:

> In the normal case, a child of four or five will have acquired a wide range of grammatical items – enough to satisfy the structural requirements (in terms of government, anaphora, and so on) imposed by the innate syntax. In the creole case, for most of these requirements the child simply cannot find appropriate grammatical items in the pidgin. Grammatical items therefore have to be created by recruiting lexical items and bleaching them of their normal lexical meaning.

Less data exists about the emergence of creoles that emerged earlier, but Bickerton (1999) cites a number of grammatical similarities between them and the Hawaiian case suggesting that the language acquisition device falls back on default options in the setting of parameters when the primary linguistic data is lacking.

Another important creolisation event, which is much less likely to have been influenced by substrate and superstrate languages has been directly witnessed by Judy Kegl and her colleagues in the last two decades in Nicaragua (Kegl, *et al.*, 1999). They have found that prior to the revolution of 1979, deaf individuals in Nicaragua were virtually isolated from one another and hence had no shared sign language. Following the revolution, public schools were established that allowed large-scale contact between deaf children. Prior to that contact, the only mode of communication open to deaf individuals was the idiosyncratic gesture systems they used in the home. These 'homesigns' were rudimentary and almost never passed from generation to generation since deaf individuals rarely had (deaf) children due to their social isolation. Following contact in public schools, the first deaf students began to communicate with a more elaborate and conventional system known as Lenguaja de Senñas Nicaragüense (LSN), which Kegl *et al.* (1999) equate with a pidgin or jargon. Soon after, a more sophisticated language known as Idioma de Senñas Nicaragüense (ISN) emerged, which

they argue is a creole in the sense of Bickerton (1991). Kegl *et al.* (1999: 187) recount the emergence of the latter language as follows.

> Groups entering the school at later dates had the benefit of exposure to both other homesign systems and to the more elaborated communication (LSN) that had developed among the first-generation homesigners. When very young children acquiring LSN surpassed their models, a new language (ISN) was added to the mix. Thus, the "language pool" to which each new group of signers becomes exposed is itself dynamic and ever-changing.

The Nicaraguan case is a very important one because it appears to be an example of language emergence in the absence of any relevant influences from pre-existing languages (be they superstrate or substrate languages).

> First, the only potential superstrate was Spanish, a language inaccessible to Deaf people via the auditory modality and whose transmission via the visual modality is seriously compromised by the ineffectiveness of lipreading as well as the lack of literacy. Second, and more important, the only candidates for substrates were not languages but homesign systems (Kegl *et al.*, 1999: 205).

The phenomenon of creolisation presents a particularly vivid demonstration of the poverty of the stimulus argument, but as Kegl *et al.* (1999: 203) stress, the normal circumstances in which children acquire a language are the same in fundamental respects.

In typical child language acquisition, children hear only a fraction of the possible sentences in their native language, yet they are still able to master its complex rules and use them productively. In other words, children's input is not logically sufficient to lead them to a full grammar unless they are aided by an innate language faculty. So the child creolizing a language is doing nothing different from any child acquiring language. It's just that there is more information available from a natural-language model to determine the choice of more marked, yet still universally available, grammatical options. And thus, the typical case of child language acquisition will lead to a greater degree of linguistic conformity to the input language.

2.3 The autonomy of grammar

In the speech of foreigners and children we frequently hear ungrammatical utterances that are nevertheless readily intelligible. Uriagereka (1998: 65) provides the example in (4). As is the usual convention, asterisks are used throughout this text to indicate grammatically problematic utterances.

4. *What have you discovered the fact that English is?

Given that the status of such examples cannot be explained in terms of their interpretability, we must look to processes that are autonomous of semantic considerations for an explanation.

It is also possible to produce sentences that are unintelligible, yet grammatically well-formed. Pinker (1994) cites Lewis Carroll's poem *Jabberwocky* as an example, the first four lines of which are reproduced in (5).

5. Twas brillig, and the slithy toves

Did gyre and gimble in the wabe:

All mimsy were the borogoves,

And the mome raths outgrabe.

The presence in this text, of inflectional morphology and grammatical words like *the*, *and*, *did*, and so forth is sufficient to identify the nouns, verbs and adjectives it contains, yet none of them are real English words. The poem has been translated into a number of other languages with similar results. The French and German translations of the first four lines are shown in (6) and (7) respectively (from Hofstadter, 1979: 366).

6. *French*

Il brilgue: les tôves lubricilleux

Se gyrent en vrillant dans le guave.

Enmîmés sont les gougebosqueux

Et le môme rade horsgrave.

7. *German*

Es brillig war. Die schlichten Toven

Wirrten und wimmelten in Waben;

Und aller-mümsige Burggoven

Die mohmen Räth' ausgraben.

Examples (4-7) suggest that grammar is autonomous from those aspects of cognition that are concerned with the semantic and pragmatic interpretation of language. Examples (5-7) also support the view, advanced by Borer (1984) and Fukui (1995), that

language variation is limited to parameters associated with functional elements such as determiners, auxiliaries and prepositions, and morphological markers like tense and number. This is called the *Functional Parameterization Hypothesis* and one of the contributions of chapter five will be to demonstrate that it is an expected feature of an optimised lexicon.

Pinker and Bloom (1990) argue that selective impairments affecting language without other aspects of cognition, or vice versa provide further support for autonomy claims, but as Deacon (1992) points out, this does not mean that language is necessarily implemented in a qualitatively different kind of neural 'circuitry' – it may be processed in a separate area, but with the same or similar types of circuits as those that occur elsewhere.

Autonomy considerations are particularly important in studying language acquisition, where some researchers (e.g., Dromi, 1999) have placed great theoretical significance on the distribution of early words into different syntactic or semantic classes. Unfortunately, Dromi (1999) and others often use the term *noun* interchangeably with *object word*, and the term *verb* interchangeably with *action word*, but such definitions, based on semantic properties of word classes, are highly problematic and represent a basic misconception about syntax. Brown and Miller (1991: 236f) make this point in their introductory textbook:

> A moment's thought uncovers many forms that are syntactically nouns but do not 'signify a person or thing' – *action*, *activity*, *movement*, and so on. Indeed, nouns like this 'signify an activity', supposedly the criterion for verbs.

They also cite examples like the words *ripe* (adjective), *ripen* (verb) and we could add to this *ripeness* (noun), which have related meanings, but manifest themselves in different word classes.[4]

The association of verbs with predicates, and nouns with arguments is similarly inadequate. The proposition *studies(George, politics)* can be expressed with either a verb or a noun as the 'predicate' as illustrated in (8).

8. a. George *studies* politics.

   b. George is a *student* of politics.

Terms such as 'noun' and 'verb' are *syntactic*, and as such the preferred definitions are in terms of their formal syntactic properties rather than their semantic correlates. Nouns and verbs can only be distinguished on semantic criteria insofar as those criteria are syntactically represented. For example, count nouns label countable 'things' whether they are objects, abstract ideas or events and this fact is marked syntactically by the fact that these nouns can appear in both singular and plural forms. Nevertheless, there are many nouns (e.g., *scissors, spectacles, trousers, tweezers, entrails*) that have plural forms[5] even when they refer to single items and exhibit the usual agreement characteristics of any other plural as demonstrated by the contrasts in (9).

9. a. *This scissors is blunt.

   b. These scissors are blunt.

[4] The differences in the meanings of these words are clearly illustrated by defining them in terms of each other. To *ripen* means to become *ripe*, while *ripeness* is the dimension over which something varies as it *ripens*.

[5] These words do appear in singular form when modifying another noun (e.g., '*scissor* kick', '*spectacle* frames', '*trouser* pockets', etc.).

In all such cases, there is some element in the meaning that makes the plurality non-arbitrary – some multiplicity of parts (scissors have two blades, spectacles have two lenses, etc), yet the label can only be used to refer to the whole object and not the individual parts.

It is for reasons like these that linguists often define word classes in terms of formal morphological and distributional criteria (i.e., the kinds of inflections they take and the syntactic contexts where they occur) rather than in terms of notional or semantic criteria. This is not to say that there is no relationship between syntactic categories and semantic categories. For instance, the category of nouns appears to include terms for concrete objects in all languages (Maratsos, 1988: 127).

Problems also arise in determining the syntactic category of words used by infants who do not possess a complete grammar. Some researchers (e.g., Tomasello & Brooks, 1999) advocate determining the syntactic category of infant holophrases by looking to the word's category in adult usage. There are two obvious problems with this. First, there is no a priori reason to believe that an infant requires knowledge of a word's syntactic category to use it holophrastically. Second, words of different syntactic categories often have the same form, especially in English. For instance, 'drink' can be used as a noun or as a verb in adult usage and so, used holophrastically, there is no way to determine its syntactic categorisation even if we could be sure that there is a fact of the matter from the infant's point of view. This will also be true of its semantic categorisation in many contexts because when an infant uses 'drink', he or she may be labelling an *object* (a drink), labelling an *action* (the act of drinking), or requesting a drink in which case the word may conflate such things as the *state* of wanting a drink, the *actions* associated with its preparation or consumption, and the *object* that is the drink itself. The trouble is determining what 'drink' means to the child over and above what the word is used to achieve.

This kind of indeterminacy is a typical property of adult language as well, and linguists reserve the terms *sentence meaning* and *utterance meaning* to distinguish between the literal, invariant aspects of an utterance's meaning and those which are specifically tied to the pragmatic context of its use (e.g., Hurford & Heasley, 1983).

Citing a number of contradictory studies on the classification of early words Kuczaj (1999: 142) is conservative in his conclusions:

> Rather than arguing about whether young children find it easier to learn nouns or verbs, it seems more important to remember that young children's early words are based on aspects of the world that they can directly experience, regardless of whether the words are nouns, verbs, or adjectives.

2.4 Some syntactic universals

The purpose of this section is to provide some specifics about the kinds of universals that language exhibits. The discussion in this section will avoid a commitment to any particular grammatical formalism, but subsequent sections pursue the question of how we might capture the regularities described here.

2.4.1 Constituent structure

2.4.1.1 A BRIEF REVIEW OF THE EVIDENCE

The phrase structure of all languages is characterised by the hierarchical nesting of phrases as revealed by various kinds of evidence that syntactic operations are sensitive to this structure rather than just contiguous strings of words. The evidence comes from various kinds of so-called *constituency tests*. I will examine three of these here, but

several other kinds are also used and are discussed in most introductory texts on syntax (e.g., Radford, 1988: 65ff).[6]

One common constituency test involves substitution, where a *proform* (a word with an interpretation that is recoverable from context) is used in place of a longer phrase. The most familiar examples of proforms are pronouns like *he*, which can be used in place of noun phrases (NPs) such as *the professor*, but the same thing is possible with other types of phrases including verb phrases (VPs) and preposition phrases (PPs). Some examples follow. In each case, the (a) sentence provides a context for the interpretation of an italicised proform found in the corresponding (b) sentence.

Substitution using an NP proform (i.e., a pronoun):

10. a. The professor will arrive tomorrow.

    b. *He* will stay at the casino.

    (*he* is interpreted as *the professor*)

11. a. The student will gamble at the casino.

    b. The professor will stay at *it*.

    (*it* is interpreted as *the casino*)

Substitution using a VP proform:

12. a. The student will stay at the casino.

    b. The professor will *do so* (too).

    (*do so* is interpreted as *stay at the casino*)

Substitution using a PP proform:

[6] Incidentally, it might be possible to perform vaguely analogous tests to assess the structure that exists in other kinds of sequential representations such as music. We subjectively group sequences of musical notes into chunks that can be used as a theme that is repeated throughout a piece of music. That we see these sub sequences as somehow the same suggests they comprise a meaningful unit of structure.

13. a. The student will gamble at the casino.

b. The professor will stay *there*.

(*there* is interpreted as *at the casino*)

Each of the (b) sentences above is interpreted with the meaning of the sentence in (14), the different sets of brackets indicating each of the sub-sequences that were replaced by proforms in examples (10-13).

14. [The professor] will (stay {at [the casino]}).

It is important to note that these brackets are perfectly nested inside each other. That is, when two bracketed sections overlap, one is always completely contained within the other. Indeed, there are no interpretation-preserving proforms that could be used in place of sequences that would not be nested like this. For instance, there is no proform that could be substituted for sequences such as *professor will*, *will stay*, *stay at* and so on.

A second kind of constituency test concerns the availability of 'movement' operations which also appear to be sensitive to the same structural units. In each of the examples in (15), the bracketed segments are interpreted as if they appear in the positions marked with # symbols where they are found in similar sentences such as (14).

15. a. She saw [the casino] he will stay at #.

b. She wants him to stay at the casino and [at the casino] he will stay #.

c. She wants him to stay at the casino and [stay at the casino] he will #.

But you can't move strings such as *stay at* under any circumstances presumably because they aren't phrasal constituents. Hence, the closely analogous sentence in (16) is ungrammatical (as indicated, in accordance with the usual convention, by the asterisk).

16. *She wants him to stay at the casino and [stay at] he will # the casino.

A third type of constituency test involves coordinating conjunctions on the assumption that only constituents can be coordinated:

17. a. [The professor] and [his wife] will stay at the casino.

b. The professor will stay at [the casino] and [the hotel].

c. The professor will [stay at the casino] and [gamble all night].

d. The professor will stay [at the casino] and [at the hotel].

But consider the following instance of coordination.

18. [The professor will stay at], and [his wife will gamble in], the casino.

This example appears to indicate that *the professor will stay at* is a constituent of our basic template sentence (14), repeated here as (19), but if *stay at the casino* is also a constituent then, as indicated by the bracketing, they overlap without being nested one inside the other.

19. [The professor will {stay at] the casino}.

The proform substitution and movement diagnostics never indicate that constituents overlap in this way, so there appears to be something different about coordination that poses interesting questions. This challenge has met with a number of different responses (Phillips, 2003). One obvious move has been to deny that coordination is a genuine diagnostic of constituency, there being such a thing as *non-constituent coordination* as well as *constituent coordination*. Another has been to argue that a sentence like (18) has the structure of the unproblematic sentence in (20), but despite the italicised material being represented, phonological processes prevent it from being pronounced, with the advantage of eliminating stylistically awkward repetition.

20. [The professor will stay at *the casino*] and [his wife will gamble in the casino].

Another alternative is the proposal by Phillips (2003) in which coordination is still taken as a genuine diagnostic of constituency, but with the apparent conflicts being the result of the facts of constituency changing during the course of a sentence's derivation. The details of this proposal will be elaborated further under section 2.5.

The evidence obtained from constituency tests suggests that something about our mental representations of sentences has a nested structure at some level of description. I now turn to the question of how we might describe the constraints that apply to this structure in different languages.

2.4.1.2 X-Bar Theory

Before the principles and parameters approach was widely embraced, generative grammarians (e.g., Akmajian & Heny, 1975) described grammars using long lists of language-specific phrase-structure re-write rules like those in (21) below. Starting with the *S* or sentence symbol, phrase structure could be generated top-down by replacing a symbol on the left of an arrow with the terms on its right (with optional replacements in

brackets) until none of the symbols in the string could be found on the left side of any rule. These remaining terms would be labels such as *N*(oun) and *V*(erb) and could be replaced with words of their category.

21. a. $S \rightarrow NP\ AUX\ VP$

b. $NP \rightarrow (Det)\ N'\ (PP)$

c. $N' \rightarrow (AP)\ N'$

d. $AP \rightarrow (Deg)\ A$

e. $PP \rightarrow P\left(\left\{\begin{matrix} NP \\ PP \end{matrix}\right\}\right)$

f. $VP \rightarrow V\left(\left\{\begin{matrix} NP \\ PP \\ S' \end{matrix}\right\}\right)$

g. $S' \rightarrow Comp\ S$

Such lists not only contained a great many rules, but different lists had to be generated to describe the grammars of different languages. However, careful analysis revealed underlying patterns in these rules, which prompted the development of what is known as *X-Bar Theory* (Chomsky, 1970; Jackendoff, 1977). Under the version of X-Bar Theory presented in Haegeman (1994), these rules are generalised to meta-rules such as the following, which apply cross-linguistically and restrict structures to those that are binary-branching.

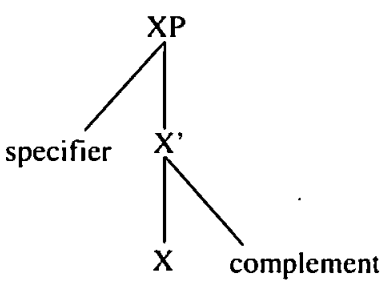
22. a. XP → X', (specifier)

b. X' → X, (complement)

The commas between terms on the right of the arrows indicate that no precedence is implied and, as before, brackets indicate optional constituents. The ordering of constituents was taken to be determined differently for different languages according to word order parameters. For example, rule (22b) would be realised as either $X' \rightarrow X$ (*complement*) or $X' \rightarrow$ (*complement*) X depending on whether, in the given language, X-level terms (i.e., *heads*) precede their complements (as in English) or follow their complements (as in Japanese). The claim was that a child is born with knowledge of this rule and so only needs to learn whether heads precede or follow complements in their language in order to use it (i.e., they simply set the *head parameter*).

The meta-rules in (22) are implemented in English to generate phrases of the form schematised in (23). As shown, specifiers precede heads and heads precede complements.

23.



The *head* of a phrase is an individual word or morpheme that characterises the whole phrase. For example, a noun is the head of a noun phrase (NP) and a verb is the head of a verb phrase (VP). In general, an X is the head of an X phrase (XP) and the XP is termed the *maximal projection* of X while X' (pronounced *X-bar*) is an *intermediate projection*.

The complement and head combine to form an intermediate projection, and an intermediate projection combines with a specifier to form a maximal projection. The

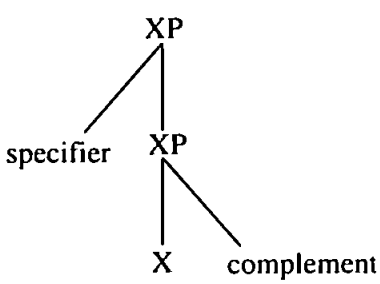
specifier and *complement* phrases are both instances of maximal projections themselves and so can have their own specifiers and complements internal to them ad infinitum.

Aside from specifiers and complements, X-bar theory allows adjunction which is captured with the following meta-rule, where Y can be an X, X' or XP:

24. $Y \rightarrow Y$, adjunct

Under the version of the theory presented by Kayne (1994), there are no intermediate projections, so a head and its complement form a maximal projection, and the specifier position is generated as an instance of adjunction to that. Hence, the schema in (23) would instead be as in (25) where the two XP terms are said to be different *segments* of the same *category*.

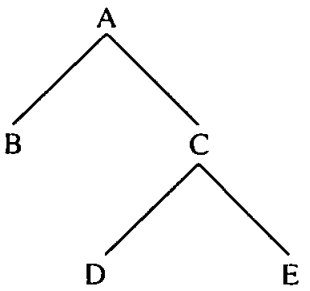
25.



As well as adjunction at the XP level, adjunction to heads is also possible accounting for structure within words (i.e., inflectional morphology).

Before concluding this section on phrase structure, it is worth introducing some terms used to describe the important structural relationships that hold between phrase markers in a tree. These will be defined with reference to the phrase markers in (26).

26.



A phrase marker α is said to *dominate* another phrase marker β if β is a constituent of α. For example, A dominates every other node in the tree in (26), while C dominates D and E, but not A or B. A phrase marker α that *immediately dominates* another β is said to be the *mother* of β, and β the *daughter* of α. So in (26), A is the mother of B and C while B and C are its daughters.

A phrase marker α is said to be the *sister* of another phrase marker β if α and β are both immediately dominated by the same phrase marker. For example, B is the sister of C and vice versa since they are both immediately dominated by A in (26).

Another important relation is traditionally known as *c-command* (*constituent command*) and now often simply *command*. In the pattern of the terms *mother* and *sister* this relation approximates the equivalent of an *aunt* relation (or great aunt, or great-great aunt etc.). A more formal definition follows where C_{HL} refers to the computational system of human language, and α and β are phrase markers as before:

27. *Command*

Where α and β are accessible to C_{HL}, α commands β if and only if

a. α does not dominate β,

b. $\alpha \neq \beta$, and

c. every category dominating α also dominates β.

(Uriagereka, 1998: 515)

Command is defined here in terms of dominance, but it is also possible to define dominance in terms of command (Frank & Kuminiak, 2000) for a restricted set of tree structures. Interestingly, Frank and Kuminiak have found that the set of tree structures for which this is possible is similar to the set of tree structures permitted by the constraints of X-Bar theory, thus suggesting that much of it can be derived from command. A variation on this proposal is adopted as part of the theory developed in chapter six.

Another proposal by Kayne (1994), embraced in a modified form within the Minimalist Program (Chomsky, 1995), is that linear ordering of constituents derives from the command relation. Kayne labelled this regularity the *Linear Correspondence Axiom* and under the theory he developed, all phrases exhibit the same order of specifier, head and complement, with the apparent language variation in overt word-order resulting from differences in the availability of processes that reorder constituents during derivations.

Command plays a central role in many aspects of grammar including theories of movement and binding, discussed in some of the following sections.

2.4.2 The lexicon

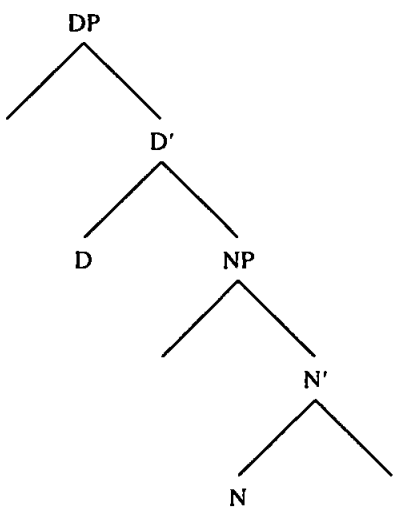
The lexicon is the mental dictionary and encodes the meaning, phonetic features and grammatical properties of all the words and word fragments that a person has acquired. A comprehensive discussion of the lexicon is beyond the scope of the present review, but a number of its properties are relevant for the discussion in chapter five.

The most important of these properties is the distinction between open and closed classes. The open classes are the nouns, verbs and adjectives of a language – the words that carry most of the meaning. The closed-class items are the frequently-used, but relatively small number of grammatical words such as determiners and auxiliary verbs, and other grammatical morphemes such as number and tense markings. The

former are always 'open' in the sense that new items can be added to them freely. An adult speaker can, for instance, add a new noun to his or her vocabulary without any difficulty. By contrast, the latter are 'closed' in the sense that an adult speaker will have great difficulty acquiring a novel item of their kind. For example, it would be difficult for an adult English speaker to accept the coining of a new tense marking to indicate a different kind of temporal relation such as the remote past.[7] Closed-class items clearly are acquired during language acquisition, but the capacity to do so declines severely by adulthood in a way that the capacity to acquire open-class items does not.

Within a sentence, the items of the open classes always appear to be encapsulated within phrases headed by items belonging to closed classes (Abney, 1987; Grimshaw, 2005; Larson, 1988). A typical analysis will involve nouns being included within determiner phrases, verbs within tense phrases and adjectives within degree phrases as illustrated in (28). These encapsulating phrases are called *functional projections*.[8]

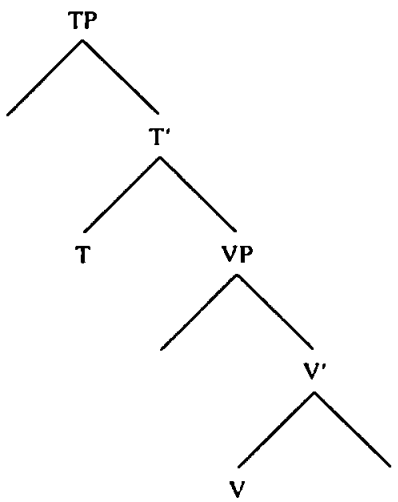
28. a.



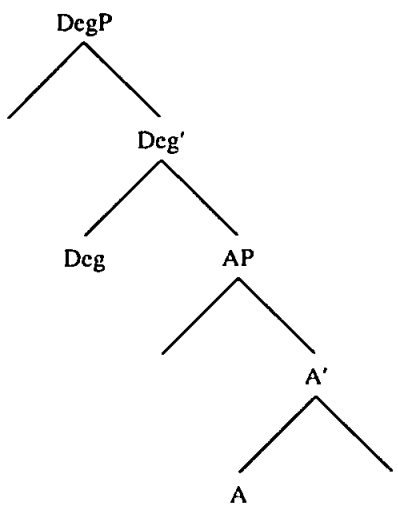
[7] A distinction between past and remote past is found for instance in Italian, so there can be no conceptual reason why a language could not make use of it.

[8] The meaning of the term *functional* here relates to grammatical 'functions' so should not be confused with evolutionary functions.

b.



c.



2.4.3 Movement

Movement theory is motivated by the observation that certain elements take the same interpretation despite appearing in different structural positions. Consider the pair of sentences given in (29).

29. a. The dog bit the man.

    b. The man was bitten.

In both sentences, *the man* is interpreted as the person who was bitten despite appearing in different structural relationships with respect to the verb *bite*. To explain this, some part of the syntactic description should encode this. In generative approaches to

grammar, examples like this are traditionally explained in terms of movement, the idea being that *the man* originates in the same position in the derivations of both sentences, this position being where it receives its interpretation as an argument of the verb. Subsequent steps in the derivation cause it to move to the subject position where it appears in (29b).

Within the earliest versions of Principles and Parameters Theory (Chomsky, 1981), the underlying level of representation was called D-structure (deep structure) and this was related to S-structure (surface structure) by transformations. In earlier approaches to generative grammar, there were many different types of construction specific transformations, but from Chomsky (1981), these were united under a single operation called move-α (move anything). A condition on movement is that the constituent in the landing site must command its *trace*, the trace being an unpronounced marker of the item's base position at D-structure.

The following are some examples of sentences that are argued to involve movement in their derivations. Co-indexed traces are included to indicate the base position of the moved items.

30. a. The book$_i$ was lent t_i to her.

    b. She$_i$ was lent the book t_i.

    (cf. Someone lent the book to her.)

31. Whom$_i$ will$_j$ Mary t_j see t_i?

    (cf. Mary will see whom?)

32. She$_i$ seems t_i to like books.

    (cf. It seems that she likes books.)

Other approaches to the formal description of grammar take a different stance towards movement phenomena. For instance, in Lexical-Functional Grammar (Kaplan & Bresnan, 1982) and Head-Driven Phrase Grammar (Pollard & Sag, 1994), 'movement' is not captured in terms of derivational processes, but in terms of lexical alternations that allow variations in word order. However, the observation that the 'landing site' commands the 'base position' is difficult to reconcile with these approaches.

2.4.4 Binding

Binding theory (Chomsky, 1981) is the part of the grammar responsible for explaining the interpretation of pronouns such as 'her' in (33) and anaphors such as 'herself' in (34) where the co-indexing denotes co-reference and the asterisks indicate the impossibility of the interpretation specified by the co-indexing.

33. a. *Mary$_i$ bit her$_i$.

    b. Mary$_i$'s dog bit her$_i$.

34. a. Mary$_i$ bit herself$_i$.

    b. *Mary$_i$'s dog bit herself$_i$.

In (33a), 'her' cannot refer to Mary, but in (33b) it can, but needn't. In (34), 'herself' must refer to Mary in the (a) example, but cannot in the (b) example. The command relation is central to explaining restrictions of this kind just as it is with the relationship between moved elements and their traces.

2.4.5 Case

The morphological marking of case is not very rich in English compared to languages like Latin and Greek, but its presence is revealed through pronominal forms such as the italicised constituents in the following sentences.

35. a. *He* bit the dog.

b. The dog bit *him*.

The pronoun in (35a) is marked for *nominative* case, while the pronoun in (35b) is marked for *accusative* case. In English, noun phrases such as *the man* do not mark this distinction overtly as the comparable examples in (36) illustrate.

36. a. *The man* bit the dog.

b. The dog bit *the man*.

Nevertheless, there are reasons to believe that they are assigned case abstractly in these positions, and indeed Chomsky (1981) proposed that all (pronounced) noun phrases must be assigned case. This universal principle is called the *case filter* and it is argued that the inability for a sentence to satisfy the case filter will lead to reduced acceptability. The following examples illustrate the kinds of data it can be used to explain.

37. a. [To give to the poor] is good.

b. *[He to give to the poor] is good.

c. *[Him to give to the poor] is good.

d. [For him to give to the poor] is good.

e. I've known [him to give to the poor].

Nominative case is associated with finite tense, while accusative case is associated with verbs (with the exception of passive and unaccusative verbs) and, in English, with

prepositions. In each of the above examples, the verbs are not tensed in the embedded clause (shown in brackets) and so cannot assign nominative case to their subjects. In (a) there is no overt subject, so there is no violation of the case filter, but in (b), there is a violation because *he* cannot receive nominative case. *Him* cannot receive accusative case in (c), so this too is ungrammatical. In (d), *for* is in the appropriate structural relationship to assign accusative to *him* so this sentence passes the case filter. The subject of the embedded clause in (e) is assigned accusative case by *known* despite not being one of its arguments.

Examples (d-e) illustrate that nominative case is not always associated with the subject position and that accusative case is not always associated with the object position. Example (e) also illustrates a dissociation between case and argument roles as do the examples in (38).

38. a. Someone lent the book to *him*.

    b. *He* was lent the book.

The (b) example in (38) is the passive sentence associated with the active sentence in (a). In both sentences, the pronoun has the same argument role (the recipient), with respect to *lend* but in (a) it is assigned accusative case and in (b) it is assigned nominative case. In (39) below, *it* does not refer to anything and is not an argument, yet it is in a position to receive case so the sentence passes the case filter.

39. *It* seems that the student borrowed a book from the teacher.

These examples serve to illustrate that case forms are not directly associated with structural relations like subject or object, or with thematic roles like agent and patient.

2.4.6 Theta theory

Theta theory (Chomsky, 1981) is concerned with the relationships between a *predicate* and its *arguments*. *Arguments* are constituents that take part in some kind of relation specified by the *predicate*. For example, in the following sentence, the predicate *hit* has arguments *Mary* and *John*.

40. Mary hit John.

Each argument of a verb has a different role called a *thematic role* (or *θ-role*). In the above example, the thematic role of *Mary* would typically be analysed as the AGENT of *hit*, while John would typically be analysed as the PATIENT of *hit*. Linguists have advocated the use of other terms for thematic roles such as THEME, EXPERIENCER, BENEFACTOR, GOAL, SOURCE and LOCATION, but there is little agreement about how to apply these terms. In the lexical entries for predicates, thematic roles are often simply numbered instead to avoid such controversies.

Theta roles are assigned to arguments in their base positions rather than their derived positions. This is suggested by the comparison in (41) where we see that 'the lie' is what is believed in both sentences despite being displaced some distance from its base position (i.e., the direct object position of 'believe') in (b).

41. a. Everyone believed *the lie*.

b. *The lie*$_i$ seems t_i´ to be believed t_i by everyone.

To rule out sentences like those in (42), which have either too many or too few arguments, Chomsky (1986) proposed a principle called the *theta criterion*.

42. a. *Mary slept the baby.

b. *Mary hit.

The theta criterion is formulated in terms of movement *chains* which mark each of the positions that a constituent occupied in a derivation. For example, in the case of (40b), a chain is formed with the three co-indexed positions t_i, t_i', and *the lie*$_i$ because the moved element moved twice leaving two separate traces. The chain is conventionally written with the *head* (or final position) first and the *foot* (or base position) last, hence in this case the chain would be <the lie$_i$, t_i', t_i>. We can now make sense of the definition of the theta criterion which is as follows.

43. **Theta criterion**

Each argument A appears in a chain containing a unique visible theta position P, and each theta position P is visible in a chain containing a unique argument A.

(Chomsky, 1986: 97)

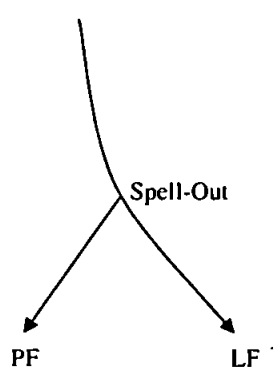
For well-formed argument chains, the theta role is assigned at the base position and case is assigned at the head position. The references to 'visibility' in the above definition relate to this property of having a case-marked position. Chains involving non-arguments are also formed, but are of little relevance to the present review.

2.5 Minimalist syntax

With the background of the preceding section, we can now look at developments that have occurred within the Minimalist Program (Chomsky, 1995) to simplify the theory of grammar. Minimalist syntax brings with it a new conception of phrase structure, movement and parametric variation, which are now reviewed.

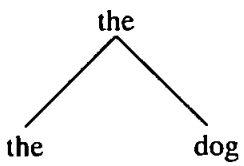
Within Minimalist syntax (Chomsky, 1995), X-Bar Theory is replaced with the theory of *Bare Phrase Structure* in which there is no equivalent of D-structure as such, only an indexed array of syntactic objects that are inserted into the derivation as it proceeds. These can be combined using an operator called *Merge* to form larger syntactic objects that can themselves enter into mergers. At some point (the equivalent of S-structure), an operation called *spell-out* causes the derivation to split into two parts (conceptualised in (44)). One part of the derivation, consisting of only the features that encode phonological information, proceeds to the level of *phonetic form* (PF) which interfaces with the articulatory-perceptual system. The rest of the derivation proceeds to the level of *logical form* (LF) which interfaces with the conceptual-intentional system of the mind. These interfaces are what define the system as one that mediates sound and meaning. A derivation must meet conditions at both of these interface levels to be valid.

44.



The Merge operator takes a pair of syntactic objects α and β and combines them to form a larger syntactic object, which takes a label derived from either α or β depending on which is the head. The resulting object can then enter into mergers itself. In (45), the phrase *the dog* is the result of merging *the* and *dog* and projecting the head *the*.

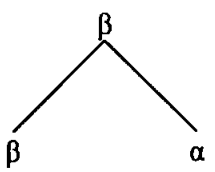
45.



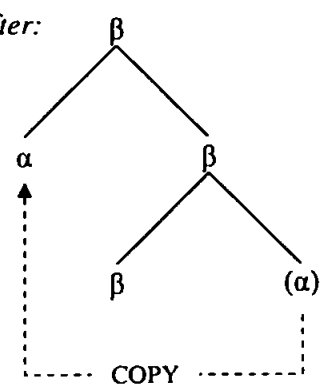
Within the terms of bare phrase structure, a maximal projection is defined as a constituent that does not project to a higher level. In (45), the maximal projections are *dog* and *the dog* but since *the* was projected when the merger took place, it is no longer a maximal projection. As this example illustrates, what counts as a maximal projection at one point in the derivation will not necessary count as a maximal projection at a later stage. This is important because projections that are neither maximal nor minimal (i.e., intermediate projections) are taken to be invisible to operations.

Movement phenomena are explained in terms of a second operation called *Move*, an example of which is illustrated in (46). A copy is made of a syntactic object α, which is part of a larger syntactic object with the head β. The copy of α then targets β merging with it to form a new syntactic object which again projects β.

46. *before:*



after:



In earlier theories of movement, it had to be stipulated that the unpronounced trace marking the base position of the moved element had to be commanded by the copy, but in Minimalist syntax, it follows from the nature of the Move operation that the copy will command its trace, so this no longer has to be specified as an explicit condition.

Merge and Move operations operate in the derivation to combine all of the lexical items of the sentence into a single syntactic object. When and how these operations occur is determined by features specified in the lexical entries of these items. In Stabler's (1997) derivational formalism, the features relevant for Merge are distinct from those relevant for Move. Those relevant for merging are of two kinds: the *base* features which indicate basic syntactic categories, and the *select* features that base features are checked against. Stabler indicates base features in lowercase as in (47).

47. *base* = {n, v, a, p, d, c, t, ...}

The features in this set are associated with the syntactic categories *noun*, *verb*, *adjective*, *preposition*, *determiner*, *complementiser*, *tense* and so on.

Select features have types corresponding to base features, but the way they are annotated depends on how the Merge is to occur. The select feature that checks the base feature *n*, is written =*n* for a simple merge, =*N* for the case where the head containing the *n* feature should be adjoined to the head containing the select feature with its phonetic content suffixed, and *N*= when it is to be adjoined but with its phonetic content prefixed to the selecting head. This means that when the select feature is lowercase, the selecting head will acquire either a complement or specifier phrase via the Merge, and when the select feature is uppercase, the item containing the select feature is a bound

morpheme which the Merge operation will attach to a morphological stem. To summarise, select features are any of the following:

48. *select* = {=x. =X, X=} | x ∈ *base*

The Move operator checks another pair of feature types, which Stabler calls *licensors* and *licensees*. These are always checked in a specifier-head relationship, specifiers being the only possible landing site for phrasal movement. The phrases that undergo movement to specifier positions have negatively specified features like those in (49).

49. *licensees* = {–case, –wh, ...}

The heads to which they attach have licensor features that are positively specified. Licensor features can also be either strong or weak, which determines whether the phonetic content they are associated with moves overtly or not. Strong features are indicated in uppercase, and weak features are indicated in lowercase. Hence, the possible licensors corresponding to the licensees in (49) are as in (50).

50. *licensors* = {+case, +CASE, +wh, +WH ...}

In Stabler's formalism, lexical entries contain ordered lists of these features, which determine the sequence in which items are merged and moved. As an illustration, Stabler derives the sentence *Some linguist speaks every language*, specifying the feature content of each item in the derivation as in (51), not all of these items having phonetic content (those bracketed).

51. *every* =n d –case

some =n d –case

language n

linguist n

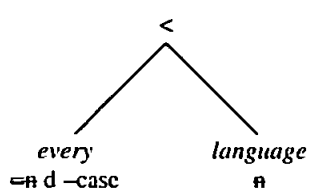
speaks =d +case =d v

(T) =v +CASE t

(C) =t c

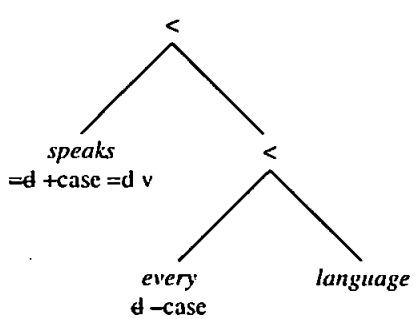
The derivation proceeds as in (52). In the first step, the initial *=n* feature of *every* and the initial *n* feature of *language* allow them to merge. When these features are checked they are then deleted from their respective feature lists making the next feature visible. In the case of *language*, there are no further features so it will no longer enter into any more operations. The item carrying the select feature is always assigned the status of the head of the phrase. This is indicated by the arrow < in (52a).

52. a. Merge



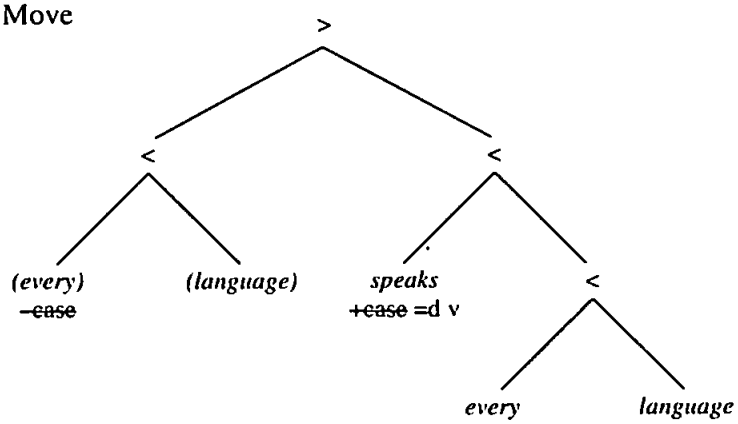
A second Merge operation allows the syntactic object created in (a) to combine with *speaks* this time checking the *=d* and *d* features to create the structure in (b).

b. Merge



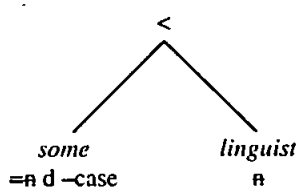
The initial feature of *speaks* is now the licensor feature *+case* so triggers a Move operation. For this operation to be valid, there must be exactly one syntactic object with an initial *–case* feature within its complement phrase. There is indeed such a phrase, *every language*, which will move to check the *+case* feature and form (c). However, since the *+case* licensor is weak, the phonetic content of the moved phrase is not carried with it. There is no clear consensus that the direct object actually moves covertly in this way, but Stabler assumes it does.

c. Move

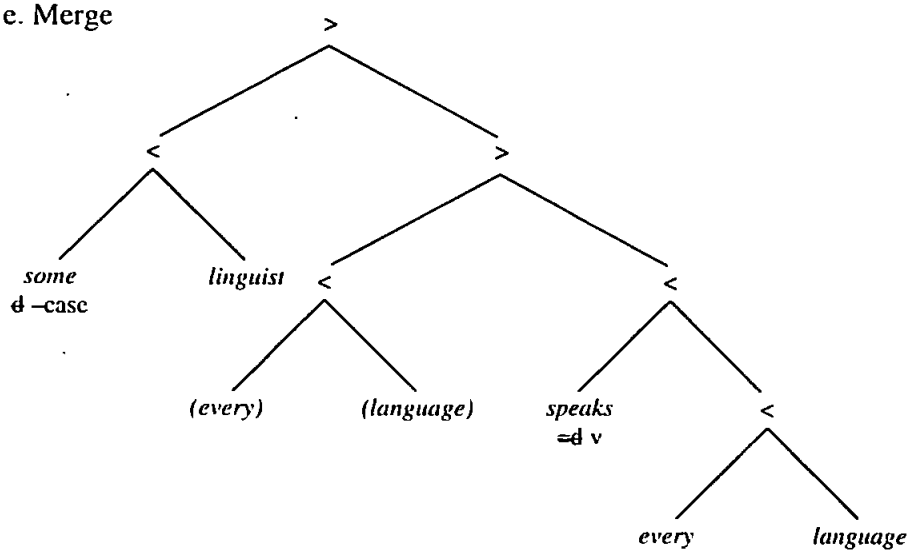


Some and *linguist* merge to form the syntactic object in (d) which in turn merges with the structure in (c) to form (e).

d. Merge

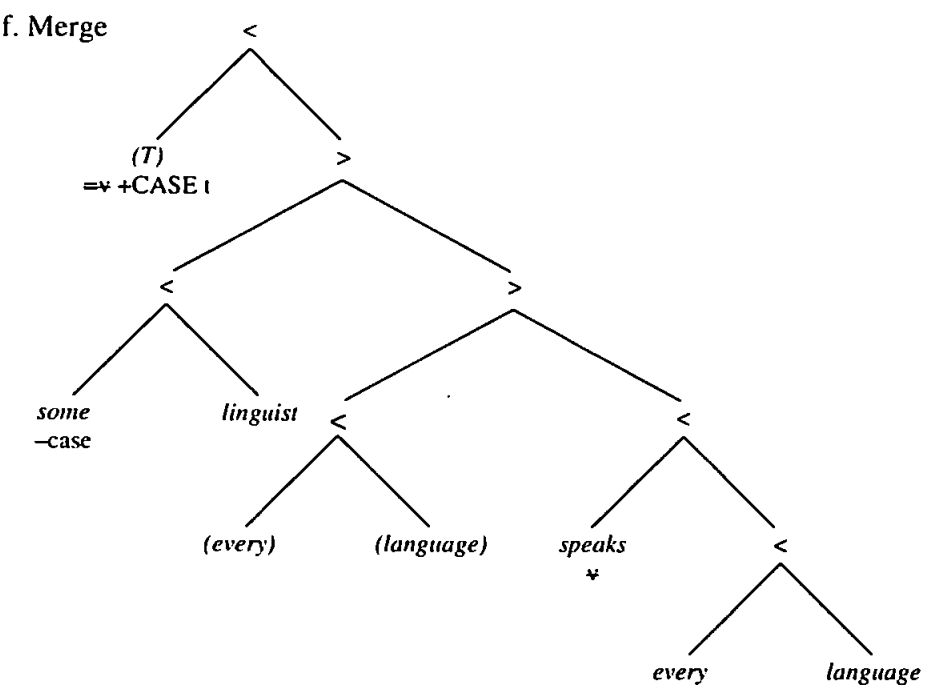


e. Merge

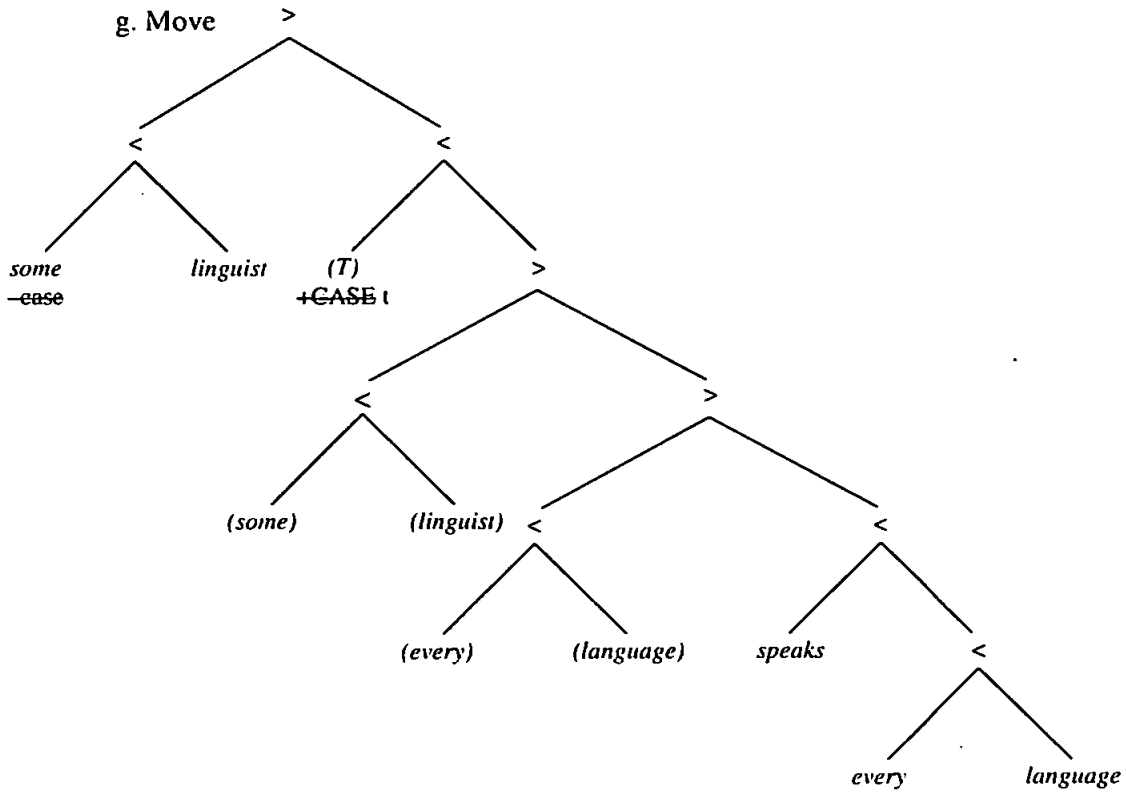


The phrase in (e) still has the verb *speaks* as its head and has two specifiers. In Stabler's example, it then merges with a tense element to generate (f), which in turn triggers an overt Move operation to check the *–case* feature of the subject resulting in (g). A final Merge occurs to join the unpronounced complementiser, which Stabler takes to be a kind of terminating symbol. At this point, the only unchecked feature is the terminating *c* feature of the complementiser.

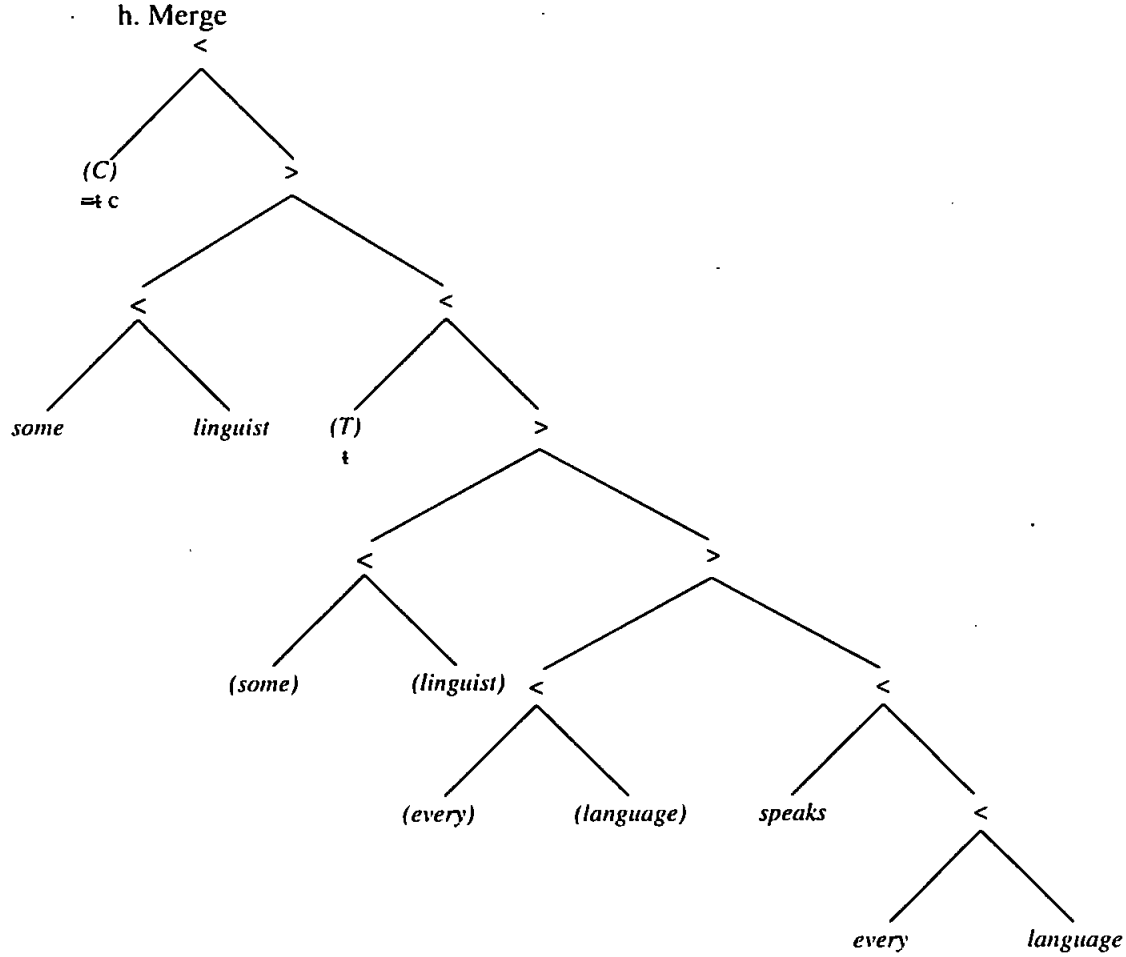
f. Merge



g. Move



h. Merge



Minimalist grammars provide a very economical way of accounting for word order variation between different languages. For instance, the above derivation could be altered to produce an SOV word order simply by making the case-licensing feature of the verb strong (i.e., *+CASE* instead of *+case*). This would trigger overt rather than covert movement of the object. Similarly, VSO languages could be generated by having strong verb selecting features (i.e., $=V$ instead of $=v$) that trigger head movement when verbs are merged with tense so that the verb head moves across the subject position (the specifier of tense). These aspects of language variation are thereby confined to the lexicon with no need to specify global parameters on grammars to account for them. The view that the lexicon is the only source of language variation is embraced by Chomsky (1995) within the Minimalist tradition, but has its roots in Borer (1984) and Fukui (1995).

The approach of Stabler (1997) differs from that of Chomsky (1995) in a number of ways. First, it doesn't have a single point at which Spell-Out occurs to split the derivation into its PF and LF parts. For Chomsky, strong features must be checked before spell-out which is what forces overt movement, covert movement occurring after spell-out in the LF-component where it cannot have consequences for the PF part of the derivation. Another important difference is that Chomsky takes chains generated by movement to be explicitly represented at the LF interface where they are necessary for the interpretation of argument structure.

On the question of chains, Brody (1995) argues that the derivations that produce them and the resulting representations end up duplicating one another's functions. Hence, by the logic of the Minimalist Program, the grammar should only need one or the other device. Brody developed a representational version of Minimalist syntax, which does away with derivations entirely, but it is unclear whether it has the same empirical coverage with respect to certain kinds of phenomena such as so-called

reconstruction effects that provide evidence for the existence of earlier stages in a derivation.

Phillips (2003) has provided some evidence that suggests that actually both representational and standard derivational approaches to Minimalist syntax may be improperly formulated. The evidence concerns sentences like those in (17-19) and summarised here in (53), which involve coordination. As we saw earlier, if coordination is used as a test of constituency, these examples suggest it leads to conflicting results as in (53c) where the elements that can be coordinated as in (a) and (b) clearly overlap.

53. a. The professor will [stay at the casino] and [gamble all night].

b. [The professor will stay at], and [his wife will gamble in], the casino.

c. [The professor will {stay at] the casino}.

Phillips (2003) argues that it is actually possible to maintain that coordination is a test of constituency without challenging the view that constituents are nested if sentence structures are built incrementally from left to right. The apparent conflicts arise because the facts of constituency actually change as the structure is being built. Hence, before the phrase *the casino* is added to the end of the sentence in (53c), the phrase *the professor will stay at* is a constituent, but once it is added, it ceases to be one.

This kind of derivational approach therefore accounts for the data quite elegantly, thus suggesting the nature of the Merge and Move operators have to be reconsidered. To argue that the facts of constituency change during the course of a derivation, one must also necessarily eschew a purely representational view. The issue of representational, derivational and linear structure-building processes will be relevant in chapter six.

2.6 Summary

This chapter served two main purposes. Firstly, it served to introduce some of the detail that theories of the evolution of language universals need to explain. These details include the course of development in first language acquisition, the capacity for infants to acquire language under the imperfect conditions that they do, and the universals that exist in constituent structure, the lexicon, movement theory, binding theory, case theory and theta theory. Although the emphasis has been on syntactic universals, there are also universals of phonetics, phonology and semantics that demand evolutionary explanations.

Secondly, the chapter served to introduce some of the specific phenomena that are the subject of the theories developed in chapters five and six.

3

Evolutionary explanations

Evolutionary biology, like geology, cosmology and other sciences concerned with reconstructing past events, cannot proceed by manipulating experimental variables, but must instead rely on inferences made from whatever traces these events leave in their wake. The case of language is particularly challenging because many of the sources of evidence that evolutionary biologists usually rely on are unavailable. There is, for instance, no evidence of properties like relative clauses or subject-verb agreement in the communication systems of other species to form the basis of comparisons and very little can be inferred from the fossil record about changes in linguistic or, for that matter, any other kind of cognitive capacity. Despite this, Botha (2003) argues that the main obstacle to advancing our understanding of the evolution of language is not the paucity of evidence as such, but the paucity of restrictive theory, "restrictive to the extent that it makes it possible to distinguish in a non-arbitrary way between entities that are instances of a specific kind of evolutionary event, process or product and entities that are not" (Botha, 2003: 115). It may be more accurate to say that there is a paucity of both evidence and restrictive theory, but that it is only the latter that we can do anything about.

This chapter attempts to address this challenge by considering (1) the categories of evolutionary explanations that could be applied to explaining the emergence of any kind of trait, (2) what evidence should lead us to prefer one type of explanation over another, even in the problematic case of human-specific cognitive capacities such as those associated with language, and (3) specific objections that have been raised about

the possibility of attributing selective functions to linguistic properties. In the process, I outline a diagnostic for identifying selective functions, which, following Parker and Maynard Smith (1990), is based on design optimality. This, I argue, is restrictive in Botha's sense and overcomes a number of shortcomings associated with vague diagnostics of the sort applied by Pinker and Bloom (1990), which are instead based on design *complexity*.

3.1 Non-selectionist categories of explanation

Stephen Jay Gould and his colleagues (Gould & Lewontin, 1979; Gould & Vrba, 1982; Gould, 1991, 1997, 2002) argue that there are a number of other factors aside from natural selection that account for properties of organisms. In the context of language evolution, Lightfoot (2000: 236) follows in this tradition and adopts its terminology, categorising theorists as either *singularists*, who believe "that every evolutionary change of any importance is due to the shaping effects of natural selection", or *pluralists*, who "allow that forces in addition to natural selection may be at work in guiding evolutionary developments". The following sub-sections are an attempt to review and clarify the kinds of alternatives the 'pluralists' have in mind. I will argue that these are not alternatives to natural selection as such, but that there are lessons to be drawn from these considerations for developing restrictive theories that would allow us to make legitimate inferences about evolutionary origins.

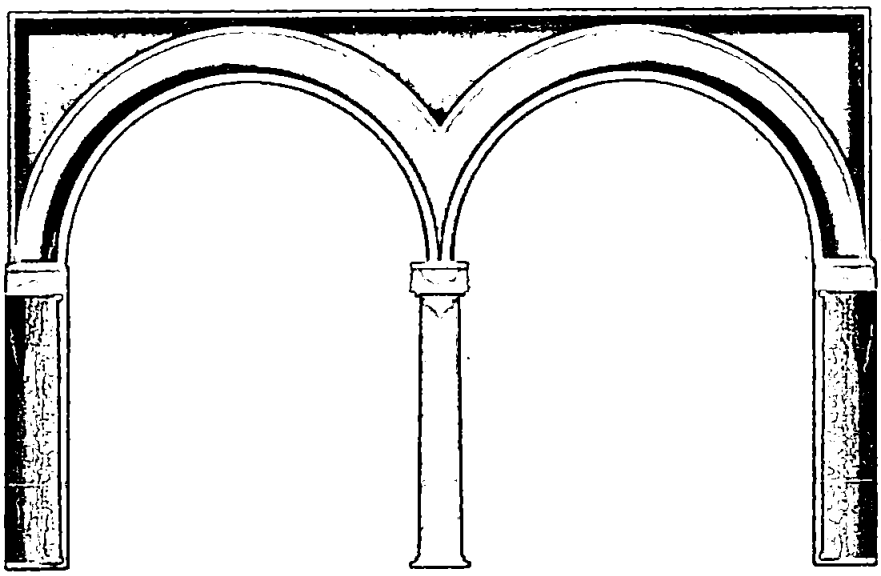
3.1.1 The lesson about spandrels: Concomitant changes

We cannot assume that all changes that occurred in the evolutionary history of a given trait were the result of selection for advantages it alone conferred to ancestors. Aside from the obvious case of selectively neutral genetic drift, some changes in a trait occur as necessary consequences of changes occurring in other traits with which they are inextricably linked. The redness of blood is often cited as an example. Blood is red

because it contains haemoglobin, a molecule that carries oxygen and waste gases around the body. Haemoglobin was presumably refined under selection for these useful properties, and it may have become redder as a result, but unlike its gas-transporting properties, colour probably didn't play any part in it being selected over other haemoglobin-like molecules because whatever colour variation existed was probably irrelevant for fitness. It is not completely inconceivable that colour was a factor in the selection of haemoglobin, but the point is merely that this needn't have been the case for a change in colour to occur. Such a change could have occurred simply as a by-product of selection for gas-transporting properties.

Gould and Lewontin (1979) use the architectural metaphor of a 'spandrel' to illustrate this point, a spandrel being the roughly triangular space found in the shoulder of an arch or in the shoulders of a pair of adjoining arches as in (1).

1.



Spandrels (darkly shaded) are the triangular spaces bounded by an arch, wall and ceiling or by two arches and a ceiling.

Specifically, they discuss the four spandrels that appear between the arches under the central dome of the Basilica di San Marco in Venice, noting that these spaces are elaborately decorated and that

> [t]he design is so elaborate, harmonious, and purposeful that we are tempted to view it as the starting point of any analysis, as the cause in some sense of the surrounding architecture. But this would invert the proper path of analysis. The system begins with an architectural constraint: the necessary four spandrels and their tapering triangular form. They provide a space in which the mosaicists worked (Gould & Lewontin, 1979: 581).

For Gould and Lewontin, spandrels are architectural by-products of constructing arches, arches being the analogue of adaptations. The biological analogues of spandrels, according to Gould (1997: 10750), are traits that "arise nonadaptively as secondary consequences ... but then become available for later cooptation to useful function in the subsequent history of an evolutionary lineage".

Gould and Lewontin argue that many properties of organisms are also inextricably linked in the way that spandrels and arches are, which means that it isn't always possible for an organism to be atomised into distinct traits with a different adaptive explanation applying independently to each. Other comments of theirs suggest that they believe that organisms can *never* be atomised into distinct traits regardless of the choice of ontology, but this is a much stronger claim and they advance no argument in support of it.

By defining spandrels as traits that arose as necessary by-products of adaptations, Gould and Lewontin (1979) also obscured at least three important generalisations. Firstly, there is no necessity that, when a pair of traits are linked, there

is an asymmetry such that one is functional and the other non-functional from the outset. It may be that both are originally functional or neither.[9] Therefore, knowing that the existence of one trait necessary implies the existence of another is not sufficient evidence to conclude that current utility of one or other is irrelevant for any explanation of why it was initially selected. Current utility may be relevant in some cases and not in others. To avoid any presuppositions about functional asymmetries I will simply refer to traits that are linked in this way as *concomitant traits*. Secondly, Gould and Lewontin define spandrels in terms only of origins rather than change generally, thereby giving us a term for a trait that *originates* as a by-product of the *origin* of another, but no term for the very similar concept of a trait that *changes* as a by-product of *changes* in another. In their architectural example, a change of this kind would correspond to a refurbishment of the building where a spandrel's shape is altered, say to accommodate a different kind of decorative design, with the shape of the arch being warped in the process. Since it is useful to refer to changes that are correlated in this way, it would be desirable to fill this terminological gap. I will do so here by referring to *concomitant changes*. Thirdly, the requirement that these linkages be *necessary* misses the generalisation that some linkages are not strictly necessary, but are just, in some sense, very likely, as when the frequency of a gene increases in the gene pool simply because it appears alongside another that confers an advantage on the same chromosome.

Despite citing spandrels in support of a plurality of forces at work in evolution, Lightfoot (2000: 237) himself tentatively acknowledges that a feature "might have arisen as a by-product of something else that was selected for" and cites the aforementioned example of the redness of blood as a by-product of selection acting on haemoglobin. In this example, and in general, the 'force' that produces these by-

[9] 'Functional' is to be understood here as enhancing some component of fitness *relative* to genetically similar variants. The functionality of a trait can only be understood in relative terms. For instance, nipples are no advantage to males relative to smooth uninterrupted skin, but would presumably be an advantage over many other imaginable alternatives such as long protruding spikes appearing in their place.

products is still natural selection. The fact that spandrels are only an indirect result of natural selection does not change that. The meaningful questions raised by the spandrel analogy are not about whether natural selection is required to produce them, but about (1) the extent to which a given trait was shaped under selection for functions served by traits that are concomitant with it, and (2) how we could reliably determine which traits are linked in this way.

3.1.2 The lesson about exaptation: Current utility and historical origins

We cannot assume that a given trait has been selectively shaped for present functions, since its form may owe much more to ancestral uses that are no longer relevant (Gould & Vrba, 1982). The arrangement of bones in a bird's wing, for instance, is essentially the same as in the forelimbs of its flightless ancestors and of tetrapods generally. The shapes and sizes of these bones have been modified under selection for flight, but the basic architecture evolved much earlier and persists either because it is maintained under selection for its new role or because to change it would require taking a radically different course at a very early stage in the growth of the embryo, a change that would have unmanageable consequences for all subsequent stages. Given that the particular number and arrangement of bones were determined during an earlier stage of evolution, it would clearly be a mistake to seek an explanation for the original selection of these properties in terms of selection pressures associated with flight.

To stress the importance of the dissociation between the historical origins and current utility of features, Gould and Vrba (1982: 4) coined the term *exaptation* for "features that now enhance fitness, but were not built by natural selection for their current role". They intended this definition to cover not only traits that previously had functions that they may or may not continue to have, but also traits that previously had

no function at all.[10] The term is widely used to refer to a trait that has been modified in some way to accommodate its new uses, but if we adopt this interpretation then

> according to orthodox Darwinism, every adaptation is one sort of exaptation or the other... if you go back far enough, you will find that every adaptation has developed out of predecessor structures each of which either had some other use or no use at all (Dennett, 1995: 281).

The distinction between adaptations and exaptations remains meaningful only in the case where a trait acquires a new function in the complete absence of any change in the actual form of the trait. If the label *exaptation* is reserved for such cases, adaptations and exaptations will typically refer to different properties of the same structures. The properties that remain the same will be the *exaptations* and the properties that are refined under selection will be the *adaptations*. Applied to the example of the bird's wing, we would say that the arrangement of its bones was *exapted*, while their specific shapes and sizes were *adapted*, both for flight. The arrangement of a set of bones (understood in terms of their number and the pattern of connectivity between them) is a rather abstract property, but it was literally the same before and after the forelimbs were co-opted for flight, the conversion of function only being possible because of the simultaneous adaptation of the shapes and sizes of the same bones.

There are a few things to note here. Firstly, like adaptation, the process of exaptation can be gradual with a trait being increasingly utilised for a new function over successive generations. Secondly, since the wing is not the same as the forelimb from which it evolved, it cannot, as a whole, be labelled an exaptation (under the narrow

[10] This is essentially how Gould (1991a) distinguishes the term from *preadaptation* (a concept familiar to Darwin), but Gould's definition of preadaptation as a trait that "performed a different function in ancestors" (p. 144) is much narrower than the definitions typically found in the literature which are often compatible with the wider sense that he reserves for *exaptation*.

usage applied here) even if the arrangement of the bones in it can be. Thirdly, whether we call something an *exaptation* depends on how generally we define its function in the ancestor and descendant species. For example, if the function of the arrangement of bones in a bird's wing is defined in very general terms as *providing a skeletal framework for the forelimbs*, then it would not have changed at all from the time when the forelimbs were legs rather than wings. With the function defined in these terms, there is literally no change. Hence, it is not an instance of exaptation even though it would qualify as one under a narrower characterisation of function in terms of flight. Finally, the definition of exaptation runs into conceptual difficulties when a trait reverts to the function it was originally selected for after a period in which it served other functions. The wings of penguins for example, were presumably originally selected for swimming in their earliest tetrapod ancestors, and although the wings also serve that function now, they were not utilised for this function by more recent ancestors who used them for walking and flying. Since the forelimbs of penguins (or aspects of them) were "built by natural selection for their current role", this case wouldn't technically count as an instance of exaptation even though there have been changes in function.[11]

These complications do not impact on Gould and Vrba's central point which was that we cannot assume that all traits emerged under selection for their current roles. This is not to say that we have no evidence that would bear on this. A possible hint about how we could distinguish original functions from exapted ones comes from the observation by Pinker and Bloom (1990: 710) that, when a trait has not been refined under selection for a function that it serves, this function is likely to be very 'simple'. For instance,

[11] Another instance of functional change that is not covered by this term is the simple loss of function which defines vestigial properties.

> A wing used as a visor is a case where a structure designed for a complex engineering task that most arrangements of matter do not fulfill, such as controlled flight, is exapted to a simple engineering task that many arrangements of matter do fulfill, such as screening out reflections ... When the reverse happens, such as when a solar heat exchanger is retooled as a fully functioning wing in the evolution of insects ... natural selection must be the cause.

A few pluralists, such as Piattelli-Palmarini (1989), cite exaptation in support of an alleged plurality of evolutionary forces acting in addition to natural selection, but exaptation is, by definition, change in a trait's function rather than in the trait itself, hence does not involve a change in an organism's morphology at all. To the extent that there are morphological changes associated with exaptation events, these are still the result of natural selection or genetic drift. Explanations in terms of exaptation and natural selection are therefore not mutually exclusive. The co-opting of a trait for a new function is nevertheless relevant for evolutionary explanations insofar as it says something about what selection pressures were at play. A change in one trait that enables another trait to be utilised for a valuable new function will have positive consequences for selection.

3.1.3 The lesson about the physical channel: Constraints on natural selection

Chomsky (2002: 141) observes that evolutionary explanations cannot ignore the fact that there are constraints acting on natural selection:

> Natural selection can't work in a vacuum; it has to work within a range of options, a structured range of options; and those options are given by physical law and historical contingency.

It is useful to think about this "structured range of options" by visualising the range of possible genotypes on an imaginary 'fitness landscape' where the elevation of the terrain at each point indicates the fitness of the variant that the point represents, nearby points being genetically similar and the highest peaks being associated with the fittest variants.

Many of the variants in the space of genetic possibilities will possess genetic 'instructions' that are incoherent from the point of view of development, causing the process to fail before an organism can reach maturity. Such variants cannot therefore reproduce and so no lineage could have them as ancestors. Without the possibility of having descendants, non-reproductive variants constitute boundaries within the genetic space within which natural selection acts,[12] and in terms of the fitness landscape, will be represented by low-lying plains with zero fitness, perhaps covering most of its area. These topological features of the fitness landscape are a function of the physical laws that determine what kind of development can occur and hence the fitness of different variants within a given environmental niche. This terrain is sometimes described as the *channel* within which natural selection acts (e.g., Gould, 1991; Chomsky, 2002).[13]

The constraints imposed by past evolution can also be understood in the vocabulary of the fitness landscape. In these terms, natural selection is a 'hill-climbing' process, but because it lacks the foresight to descend from small peaks in order to scale even larger ones, it will tend to get stuck in locally optimal regions. Hence, the range of options available for the future evolution of a variant depends on which peaks can be

[12] These boundaries will account for various constraints such as 'scaling laws' that relate the size of an animal to its metabolic rate, and so forth. As Lightfoot (2000: 238) notes, a mouse the size of an elephant "wouldn't have enough surface area to dissipate the heat generated by the superactive mouse metabolism, and it would cook itself to death in short order".

[13] Logic also imposes boundaries on what is possible. For example, an organism that simultaneously does and does not have some trait is impossible for reasons that have nothing to do with physical law.

scaled from its present position, its present position being the result of the "historical contingencies" of past evolution to which Chomsky refers.

Historical contingencies are also important for understanding the effect of mutations. The variation generated by mutations is the raw material on which natural selection acts, but their effects on a phenotype can be counterintuitive. The number of petals that appear on the flowers of different species of plant illustrate this point. We might expect that a mutation causing the average number of petals to change from eight to nine would be more likely to occur than one causing a change from eight to a larger number like thirteen. But even a cursory examination of the extant variation in flowering plants suggests that this would be wrong. There are many plants whose flowers normally have eight petals, and there are plants whose flowers normally have thirteen, but intermediates that normally have an average of nine are rare or nonexistent. Interestingly, the number of petals appears to be limited to a choice among the Fibonacci numbers (or numbers trivially related to the Fibonacci sequence). The restriction to the Fibonacci numbers appears to be the result of a genetic commitment to a particular type of growth function. To produce a variant with a non-Fibonacci number of petals would require a radical redesign of flower architecture, meaning a large and improbable leap through the space of genetic possibilities. By comparison, the existing variation in flowering plants suggests that mutations leading from one Fibonacci number to another are relatively common, the result of relatively small mutations.

Given the complexities involved, claims about which evolutionary changes have taken place in a structure (or behaviour) need to be supported by evidence that the hypothesised pathways are actually open to evolution. Existing variation within a species or between closely-related species is one source of this kind of evidence (Parker & Maynard Smith, 1990). The fossil record is another, though one which isn't very informative about behaviour and soft tissues like the brain. A third source of evidence

comes from the finding that the course of an organism's development often reflects the order in which evolutionary developments must have taken place. The evidence for this comes from observations by Karl Ernst von Baer made a few decades before Darwin and Wallace published on evolution (Gould, 1977). Von Baer observed that more species share earlier stages of development than later ones. This is usually explained in terms of the likelihood that adjustments to any given stage of development will impact on all subsequent stages that depend on it, making changes to features that appear earlier in development less manageable, and therefore less likely, than those that occur later. The earliest stages of development will therefore generally be the most ancient ancestrally and most widely observed in different species.[14]

It is possible to overstate the extent of the constraints that historical contingencies impose. The case of mammals is instructive on this point given that a commitment to the mammalian body plan has not stopped them from spreading into an enormous range of niches. In addition to all the land-going mammals inhabiting environments as diverse as deserts, polar regions, jungles and cities, evolution has produced dolphins with fins like fish, and bats with wings like birds. One of the reasons that historical contingencies do not always prevent certain forms appearing is that similar forms are often attainable by variants that are very far apart in genetic space.

In the spirit of pluralism, Lightfoot (2000: 237) considers the physical channel as yet another alternative to explanations in terms of natural selection.

> Some properties of organisms are not selected for and are not accidental by-products, but emerge because of deep, physical principles which affect much of life. For example, organisms as diverse as robins, redwoods and rhinos obey exactly the same mathematical laws governing the way size affects structure,

[14] This is the acceptable version of the discredited biogenetic law which says that 'ontogeny recapitulates phylogeny', originally proposed by Haeckel in the 19th century. See Slobin (2002) for a critical look at language evolution studies that have applied this kind of argument.

> physiology and life history. Those laws, the 'scaling relations', are a near-universal feature of life. They reflect fundamental limits on the kinds of things that evolution can make, and they arise from the interaction of a few simple physical principles.

But Lightfoot's position deviates from the that of other pluralists such as Gould and Lewontin (1979) who acknowledge that the physical channel is not so much an alternative to natural selection as the structured 'space' that it operates within.

> [This thesis] does not deny that change, when it occurs, may be mediated by natural selection, but it holds that constraints restrict possible paths and modes of change so strongly that the constraints themselves become much the most interesting aspect of evolution.

Discussions of physical constraints are effectively discussions about selection pressures in another guise. To attribute the existence of some property to natural selection without any discussion of these pressures (which determine the shape of the fitness landscape) is indeed to assert something so obvious as to be uninteresting. In Chomsky's (1972: 97) words, "it amounts to nothing more than a belief that there is some naturalistic explanation for these phenomena".

3.1.4 The lesson about laws of growth and form: Non-adaptive elegance

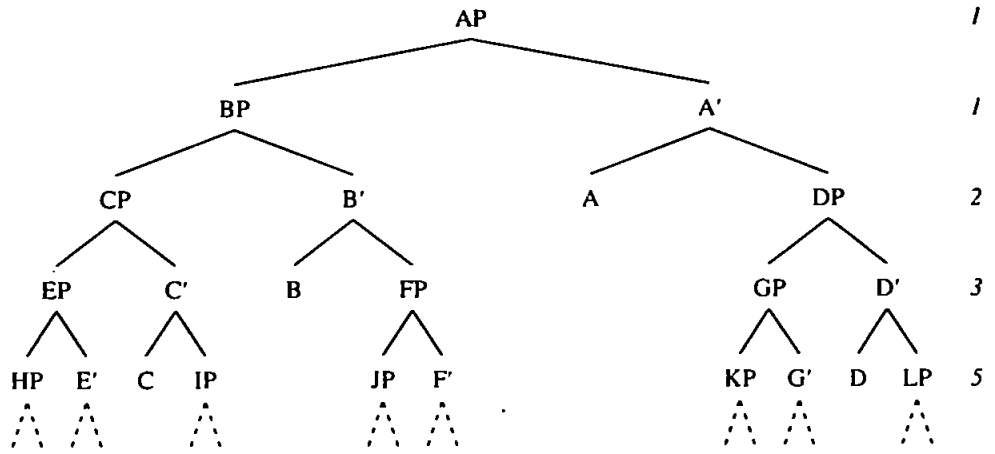
The recurrence of certain kinds of structures in biological systems can be explained in at least two ways aside from the obvious case of common ancestry. Firstly, natural selection might lead to the independent discovery of good or optimal designs (convergent evolution) as judged against the physical laws that determine what makes a

good wing, fin, eye, and so on. Secondly, certain ubiquitous patterns and features like logarithmic spirals, Fibonacci numbers and hexagonal tiling patterns show up in inorganic contexts as well which means that it is unnecessary to invoke a process of cumulative natural selection to explain them. When these properties show up independently in distantly related species, it may be simply because some very basic formal properties are shared by the relevant developmental processes or by the relevant environmental conditions under which the traits evolved. It isn't so surprising that a piece of mathematics developed to describe one phenomenon will be applicable to others if the formal properties that the phenomena share are very simple. On the other hand, if a pattern requiring a mathematically complex formal description were found throughout nature, then this would require an explanation perhaps in terms of convergent evolution. If such a pattern also appeared regularly within inorganic systems then this would challenge our understanding of the laws of physics. The Fibonacci sequence, for instance, is very much in the category of simple mathematics and so it shouldn't be surprising to find it showing up in all kinds of circumstances without any adaptive reason.

Fibonacci numbers even show up in language. Carnie and Medeiros (unpublished) have pointed out that, in a phrase structure tree like (2) below, which has every 'specifier' and 'complement' position filled, the number of maximal terms in each line of the tree (i.e., the XP terms where X is variable) is 1, 1, 2, 3, 5... (i.e., the Fibonacci sequence). The number of X'-level terms in each line also conforms to the sequence, but this time starting from 0 to give 0, 1, 1, 2, 3... Similarly, the number of X-level terms is 0, 0, 1, 1, 2... which conforms to the sequence except for the initial 0.[15] Uriagereka (1998: 483ff) has also made some interesting suggestions about how Fibonacci patterns might also show up in phonology and pragmatics.

[15] This tree can be generated using the pair of rewrite rules: (i) XP → YP X'. (ii) X' → X ZP (where X, Y and Z are variable). For some more general comments about how rewrite rules can be used to generate Fibonacci patterns in domains outside linguistics, see Uriagereka (1998: 192f).

2.



Uriagereka (1998) and Gould and Lewontin (1979) present laws of growth and form as an alternative to explanations in terms of natural selection citing the work of Thompson (1961) who examined many of the mathematical patterns that recur in the organic and inorganic world. However, unlike these authors, Thompson did not see these laws and natural selection as mutually exclusive modes of explanation. Paraphrasing Aristotle, he compared the situation to how we might explain the existence of a house: "the house is there that men may live in it; but it is also there because the builders have laid one stone upon another" (p. 5). In other words, developmental explanations and evolutionary explanations are not mutually exclusive.[16] However, we do not need to account for very simple mathematical patterns in terms of cumulative natural selection. The developmental processes that give rise to properties like logarithmic spirals are so simple that they are effectively the atoms of variation on which natural selection acts and so could plausibly arise via a single mutation rather than a series of them.

[16] Thompson (1961: 3) did, however, express reservations about doing "natural history" which "deals with ephemeral and accidental, not eternal nor universal things" preferring to study physical forms from a mathematical perspective involving "truths remote from the category of causation". This stance strongly resonates with Chomsky's (1980; 2000b) views on the need for idealization in naturalistic inquiry, typical of what he refers to as the 'Galilean style' of inquiry.

3.1.5 The lesson about the role of genes and the environment

To inquire into the contributions of genetic and environmental factors in the evolution of language, we need to ask, if it is meaningful to do so, which linguistic phenomena should be considered 'innate' and which are not. There are disagreements about this, but there are two points for which there is a broad consensus.

3. a. Human infants develop a type of linguistic competence that does not develop in any other existing species exposed to the same linguistic input.[17]
   b. The result of language development depends to some extent on the type of linguistic input that an infant is exposed to (i.e., which language).

The first point is ultimately an acknowledgement of genetic effects while the second is an acknowledgement of environmental effects. A difference in either could lead to a different developmental outcome.

Unlike discussions of genetic and environmental effects, it is often less clear what researchers mean when they claim that a given property is *innate*. Synonyms such as *genetically programmed*, *instinctive*, and *hard-wired* suffer similar short-comings. As an example, consider the oft-repeated assertion that the propensity to fall to the ground does not have to be 'genetically programmed'. One of the characters in Uriagereka's (1998: 14f) philosophical dialogue, expresses his version of it in the following terms:

> [N]ature is clever enough not to programme the shape of an organism into its genetic code, if the laws of physics or chemistry produce it anyway... [T]hink of the two distinct phases in the aquatic/aerial behaviour of the common salmon.

[17] Since language can be expressed in any modality (speech, signing, writing, etc.) the generalisation expressed here should therefore not be confused with a statement about the ability of other species to produce the relevant kinds of vocalisations. The ability to modulate sounds that could be used for speech is present in some species and absent in others.

> The first phase, the impulse to jump out of the creek, is programmed into the fish's genes. But the second stage, its falling back to water, is a consequence of gravity. Nature needn't specify that in the genes![18]

Chomsky (2002: 143) has also made similar comments:

> [N]obody thinks there are genes that [during mitosis] tell the breaking cell to turn into spheres, just as you do not have a gene to tell you to fall if you walk off the roof of a building. That would be crazy, you just fall because physical laws are operating, and it is probably physical laws that are telling the cells to break up into two spheres.

If we understand a 'genetic programme' in terms of genetic effects, then Uriagereka's claim would be that no genetic difference would enable the salmon to do anything other than fall, but if the genetic difference was such that the salmon was actually a sea bird, then this would clearly not be the case. It wouldn't have to fall back into the water because, after leaping into the air, it could fly away. There can be a genetic effect in this sense. Uriagereka and Chomsky might not be speaking of genetic 'programmes' and 'specifications' with this meaning, in which case this criticism would not apply. There is, however, another fundamental reason for rejecting their view. Given that all processes are governed by physical laws, the assertion that such processes are not 'genetically programmed' leaves nothing that is. We could argue that the salmon's leap does not have to be genetically encoded, only the flapping of its tail since physical law takes care of the rest. Or we could go further and say that the flapping of its tail does not have to be genetically encoded either, only the muscles and the system that generates

[18] Gould (1991) also uses the falling fish example, but merely as an example of a non-adaptation. It is originally due to Williams (1966).

the electrical impulses that control them since the flapping is inevitable once you have these things in place. And we could carry this reasoning down to the level of protein synthesis where we could argue that proteins do not have to be genetically encoded, only DNA, since the chemical properties of DNA in the context of a cell with its cellular machinery will inevitably lead to protein synthesis for reasons having only to do with the laws of chemistry. It should be clear that reasoning of this kind strips the notion of a 'genetic programme' of any useful meaning.

More commonly, a distinction is made between 'innate' and 'acquired' characteristics, and this is still pervasively seen as a strict dichotomy such that a given trait can only be one or the other, or at best, a mixture of components of one or the other type each of which is wholly innate or wholly acquired but not both simultaneously. At the same time, 'innate' properties are taken to have genetic causes while 'acquired' characteristics are taken to have environmental causes. Now, to say that a genetic factor is the 'cause' of some phenotypic property, it presumably means that if it were absent or altered, then the relevant phenotypic property would not be expressed in the same way or with the same likelihood. Similarly, a discussion of environmental 'causes' necessarily implies a comparison of environments in which the relevant property will and will not be expressed with the same likelihood. Depending on whether we are studying environmental or genetic differences, we may hold one of these variables constant to look at the effect of the other, but we cannot escape the conclusion that both factors will be 'causes' of one and the same property if the development of that property is simultaneously contingent upon the presence both of certain genes and of certain environmental conditions.[19] Given that all development depends upon environmental factors such as the presence of nutrients, and that there are very few properties that are necessarily shared by genetically dissimilar organisms (the exceptions being properties

[19] This treatment of causation omits many details but suffices for present purposes. For a recent review of some of the complexities involved see Dennett (2003: 83ff).

like *existing* that obtain of non-organisms as well), it is the norm for any given property to be subject to both genetic and environmental effects. In short, the dichotomous position must be abandoned, not just for unusual cases, but for all cases. Given that an acceptance of this dichotomy is usually implied by the use of the term *innate* and its synonyms, I will confine myself to speaking in terms of genetic and environmental effects, understood (in the standard way) in terms of differences.[20]

Restated in terms of effects, claims about innateness are usually not so much about genetic effects as a surprising lack of specific kinds of environmental effects. For instance, an infant acquiring language does not need to be exposed to any evidence about contrasts in grammaticality status for it to develop the ability to make the relevant distinctions (Brown & Hanlon, 1970; Marcus, 1993). That language development is insensitive to whether or not this data is present in the linguistic environment is also supported by the observation that infants are inattentive to it even when it is available, often simply failing to understand the point of parental corrections (Braine, 1971).

The genetic and environmental contributions to development can be viewed in terms of the range of phenotypes that can be expressed on the basis of a given genotype with environmental conditions determining which developmental pathway is actually followed in specific cases.[21] This is, in essence, the principles and parameters view adopted by Chomsky (1981), Piattelli-Palmarini (1989) and others, who describe the role of the environment as triggering or selecting between possible parameter settings that are within the range imposed by an individual's genetic endowment. For example, genetically identical ants will develop different morphologies (the body plan of a queen, soldier or worker) depending on environmental stimulation. Environmental factors such

[20] 'Innate' properties are also frequently referred to as *hard-wired*, which suggests a slightly different dichotomy (that between fixed and variable traits) that is also problematic. Even so-called 'hard wiring' has to get wired up during development so there is a problem with drawing the line between that which is fixed and that which is variable. If it is the immutability of something after it has been 'wired up' that is important, then anything that is 'learned', but never 'unlearned', would fit into that category.

[21] An equivalent formulation involves considering the range of phenotypes that can potentially be expressed under a given set of environmental conditions with genetic variables determining which of these phenotypes is actually expressed.

as temperature, photoperiod and diet trigger the setting of the body plan 'parameter' from among a restricted set of possibilities (Abouheif & Wray, 2002). Within the domain of cognition, contingent development of this kind is usually referred to as *learning*, but development is still bounded within a restricted (though possibly infinite) set of possibilities.[22]

The results of contingent development are another category of traits that Gould and Lewontin (1979: 591) argue are not produced by natural selection. They note that the "phenotypic plasticity that permits organisms to 'mold' their form to prevailing circumstances during ontogeny" produces characteristics that "are not heritable", though they qualify this by saying "the capacity to develop them presumably is". They conclude that characteristics that are not inherited could not have been selected for, but the impossibility of Lamarkism (the inheritance of 'acquired' characteristics) follows from the impossibility of inheriting *any* characteristics of the phenotype, not just those whose development is contingent on certain environmental conditions. What is inherited is always "the capacity to develop them" (in other words, an individual's genes), which is the reason why a person with an amputated leg does not have one-legged children. The claim that a given characteristic was selectively favoured can only mean that individuals who possessed "the capacity to develop" that characteristic (e.g., legs, a suntan, etc.) had more offspring than others on average by virtue of that characteristic being expressed in the phenotypes of those individuals (perhaps only intermittently). This idea is neatly captured in Ridley's (2003) phrase "nature via nurture".

[22] Even if the set of expressible phenotypes excludes most alternatives, there is no necessity that this set be finite in size contrary to what is often assumed about the principles and parameters approach. The presence or absence of each triggering stimulus will lead development down a different pathway (i.e., set a distinct developmental parameter) and by definition, there will be as many parameters to set as there are environmental influences capable of effecting development.

3.1.6 The lesson about cultural evolution

Some cognitive tasks are more intuitive than others due to the biases inherent in 'learning' mechanisms.[23] Deacon (1997: 105) argues that these biases will mean that, as well as the brain evolving to support language, we should expect that "language itself" adapts to the pressures governing its cultural transmission from generation to generation, the properties that flourish being those that are passed on with highest fidelity during language acquisition. In parameter-setting terms, the notion of a "language itself" being transmitted from one generation to the next can be interpreted as a statement about how inferences from the utterances of adult speakers will tend to induce in the next generation, a close approximation to adult parameter settings (including the lexicon).[24] Parameter settings that are more reliably inferred on the basis of primary linguistic data will tend to be more highly conserved from one generation to the next than alternatives.

The cultural transmission of linguistic parameter settings is just one example of natural selection acting in the cultural domain. Within culture, the replicating entities are not genes, but pieces of information or what Dawkins (1976) termed *memes*, and like a person's genes, a person's memes can be passed on to subsequent generations with modifications. Some of those memes will, by their nature, be passed on more readily than others depending on the consequences they have for their own replication.

Deacon (1997) and others have noted that the rate of cultural evolution is much faster than genetic evolution, but as Pinker (1994: 151f) stresses, the syntax of languages does not vary without limit:

[23] We could describe this situation in terms of cognitive biases favouring certain kinds of hypotheses about the world or equivalently, in terms of biases disfavouring others (cf. Deacon, 1997).
[24] Though see Chomsky (2002) who argues that Deacon's (1997: 105) remarks are unintelligible because the latter appears to entertain the notion that language exists "outside brains".

> Beyond a time depth of about a thousand years, history and typology often do not correlate well at all. Languages can change from grammatical type to type relatively quickly, and can cycle among a few types over and over; aside from vocabulary, they do not progressively differentiate and diverge.

Given that the environment that determines the fitness of a culturally evolving language (i.e., the language acquisition device) is comparatively stable, we should expect cultural evolution to continue only until languages are optimal with respect to the fidelity of transmission from generation to generation. At this point, we can expect that they will cease to improve although they may drift neutrally between equally fit alternatives thus accounting for the cyclicity to which Pinker refers. Prolonged periods of stasis are the norm in genetic evolution as well when environmental conditions are stable. This is because after a species has adapted to fit its niche, it has no further hill to climb in the fitness landscape. Many species have remained essentially unmodified for hundreds of millions of years, presumably because any minor deviation from their design would be a disadvantage. In other cases, such as the co-evolution of predators and prey or where there are interactions between different adaptive strategies in a population, the dynamics are more complex so the endpoint will not necessarily be a fixed point attractor, but may be some other kind of attractor such as a limit cycle or strange attractor. But if the cultural evolution of language has an optimal endpoint given some set of learning biases, then this endpoint is as subject to genetic effects as the learning biases themselves and could feed back to affect genetic selection. When genetic mutations alter the learning biases, the effect on cultural evolution would be delayed by a number of generations while the parameter settings are optimised for cultural transmission under the conditions the mutant language-learners impose, but the attractors in the space of cultural evolution are not thereby any less subject to genetic effects than the learning

biases themselves. In a literal sense, a culturally stable state can be regarded as an extended phenotypic property (cf. Dawkins, 1982) and hence, from a genetic perspective, the proper treatment of cultural evolution is alongside ontogenetic development but with the developmental process potentially extending beyond the lifetime of individual organisms.[25]

Gould and Lewontin (1979: 591) observe that "[t]he mere existence of a good fit between organism and environment is insufficient for inferring the action of natural selection" (i.e. natural selection in the biological domain) because this fit can also be the result of processes like learning and cultural evolution, a point also made by Kirby (1999). As I have attempted to show, there is no contradiction in saying that a trait that has arisen via either of these processes was also favoured by selection, but they would be right to warn us that characteristics that humans never exhibited before modern times cannot have been relevant for the selection of the capacities that give rise to them. Language universals are not just a modern phenomenon so do not fit into this category, but there remains a possibility that some of them exhibit fit to functions for which they were never indeed selected for in biological evolution.

3.1.7 Some conclusions about allegedly non-selectionist mechanisms

There is a certain perception, fostered particularly by Gould and his colleagues, that phenomena such as spandrels, exaptations, the physical channel, laws of growth and form, 'acquired' characteristics, and cultural evolution represent radical alternatives to

[25] The view that cultural evolution is open-ended is perhaps an illusion engendered by the rapid technological and social changes of our recent history, but it is far from obvious that even these aspects of culture will continue to change for much longer. We are, for instance, brushing up against the physical limits of computer miniaturisation, limits imposed by the Uncertainty Principle that mean the indeterminacy in the location of signal-carrying electrons becomes important. The reflex of miniaturisation has been advances in computing speed since signals take less time to propagate over smaller distances. Various developments may delay the inevitable, but it appears that advances in computing speed will soon grind to a halt as a result. Entire fields of scientific inquiry will also inevitably cease to be fruitful when the limits of what can in principle be discovered about them are eventually reached (cf. Horgan, 1996).

neo-Darwinian orthodoxy. This has led to claims, for instance by Piattelli-Palmarini (1989), that we need to replace the neo-Darwinian orthodoxy based on natural selection with one based on 'extra-adaptive' mechanisms. Nevertheless, representatives of the 'orthodoxy' do not accept that these ideas (insofar as they are intelligible) are radical at all (Dawkins, 1991; Dennett, 1995).

Those who argue that the existence of these phenomena is evidence of the existence of evolutionary 'forces' other than natural selection are guilty of a confusion between a plurality of 'forces' and a plurality of the products of a 'force' as well as between natural selection and the developmental constraints imposed upon it. There are interesting questions about which kind of product and what kind of constraints, but not generally which 'force'.

Gould and Lewontin (1979: 583) argue that the rhetorical strategy of the "evolutionist" (by which they mean researchers who prefer to see everything in terms of adaptation) is to admit the existence of a given mechanism of "nonadaptive evolution", but "circumscribe its domain of action so narrowly that it cannot have any importance in the affairs of nature". Gould and Lewontin list a number of apparently non-adaptive alternatives along the lines expressed in the preceding discussion, but almost all of these are not 'processes' at all but either the indirect results of natural selection or physical constraints acting on it. One of the mechanisms they cite is, however, a genuinely non-adaptive kind of change: *genetic drift* (selectively neutral change in the frequency of genes in the gene pool). But there are some sound reasons to doubt the importance of genetic drift even for those who refuse to adopt the kinds of entrenched positions to which Gould and Lewontin allude. The main reason is that genetic drift is the antithesis of a 'force', non-directional by definition, leading to the accumulation of selectively neutral mutations, which, among other things, results in the degeneration of vestigial

structures, structures on which selection has ceased to act.[26] Drift is in many ways the evolutionary equivalent of the second law of thermodynamics, an increase in entropy that occurs in the absence of selection. To describe it as a source of creativity would be misleading at best.

3.2 Optimality as a diagnostic of selective function

With the conceptual clarifications of the proceeding discussion some more precise questions (4) can now be formulated about the emergence of a given trait.

4. a. To what extent could it have been shaped under selection for functions served by concomitant traits?
   b. Does the trait result from developmental processes that are so simple that it could plausibly have arisen through a single mutation rather than being refined by a process of cumulative natural selection?
   c. Under which environmental conditions will the trait be reliably expressed?
   d. Does the trait only emerge after a period of cultural evolution?
   e. What selection pressures / physical constraints were relevant for the original selection of the trait?
   f. Have the selection pressures that have shaped it remained the same over evolutionary time?
   g. What limits have the historical contingencies of past evolution imposed on its variation?

Questions (a)-(d) are essentially about ontogenetic development so can be explored by observing processes that occur today. For instance, (a) can be addressed by determining

[26] It is worth stressing here that natural selection continues to act even after it has affected a change. It acts to preserve advantageous traits, weeding out mutations that would otherwise accumulate through genetic drift and which can accumulate when a property ceases to be relevant for reproductive success.

everywhere a given gene is expressed in an organism and how different traits are linked in developmental processes. In the case of language, answers would be suggested by looking at where else in the body genes implicated in language development are expressed (e.g., Enard, Przeworski, Fisher, Lai, Wiebe, Kitano, Monaco & Paabo, 2002) and what other kinds of effects are associated with heritable language deficits (e.g., Gopnik & Cargo, 1991). Question (d) can be evaluated by (1) looking at properties of language that are not acquired by imitation, (2) looking at properties that are found in spontaneously emerging creole languages like that which emerged among deaf children in Nicaragua (e.g., Kegl, *et al.*, 1999), and by (3) exploring the dynamics of cultural transmission in computational models (e.g., Kirby, 2001).

Questions (e)-(g) require us to make inferences about the past so are inherently more challenging to answer, but these issues may be informed by examining whether a trait exhibits optimality with respect to candidate functions. This is essentially a variation on the argument from the appearance of design, most famously presented by William Paley, who used it as the basis of an argument for the existence of a deity.

> In crossing a heath, suppose I pitched my foot against a stone, and were asked how the stone came to be there; I might possibly answer, that, for anything I knew to the contrary, it had lain there for ever: nor would it perhaps be very easy to show the absurdity of this answer. But suppose I had found a watch upon the ground, and it should be inquired how the watch happened to be in that place; I should hardly think of the answer which I had before given, that for anything I knew, the watch might have always been there. (Paley, 1828, cited in Dawkins, 1991: 5)

Paley goes on to argue that given the likelihood that an object such as a watch could form by chance is so remote, the existence of such objects can only be explained by attributing design to them by a designer. He then extended this argument to living things, which he observed to be even more intricately crafted. In the absence of an alternative explanation, he argued that living things must have also been designed.

Of course, since Darwin we are aware of another non-random process that can give rise to such objects, and so the appearance of design can now be viewed as evidence of the action of natural selection. Pinker and Bloom (1990: 707) make this point in relation to language:

> Evolutionary theory offers clear criteria for when a trait should be attributed to natural selection: complex design for some function, and the absence of alternative processes capable of explaining such complexity. Human language meets this criterion: grammar is a complex mechanism tailored to the transmission of propositional structures through a serial interface.[27]

The intuition is that complex functional designs are unlikely to come into existence by chance. Pinker and Bloom's discussion owes much to Dawkins (1986/1991: 9) who is more explicit about this.

> Complicated things have some quality, specifiable in advance, that is highly unlikely to have been acquired by chance alone. In the case of living things, the quality that is specified in advance is, in some sense, 'proficiency'; either proficiency in a particular ability such as flying, as an aero-engineer might

[27] In this quote, Pinker and Bloom are discussing language as a whole rather than individual properties of it, but the same logic applies in both cases.

admire it; or proficiency in something more general, such as the ability to stave off death, or the ability to propagate genes in reproduction.

It should be clear from the discussion in the preceding sections that non-trivial questions arise about how to determine what evolutionary pathways are available, and these present serious difficulties for determining the likelihood of evolutionary transitions, but for a selective explanation to be minimally favoured over a non-selective one, it need only be argued that the likelihood of the transition occurring is greater via a process of cumulative natural selection than via either genetic drift or a single mutation. We will only fail to have any preference for one variety of explanation over another when we know absolutely nothing about what variations mutations can produce, whether the resulting organism has greater fitness than the ancestral form, and so on.

The emphasis on complexity that Pinker and Bloom inherit from Dawkins reflects the particular concerns of the latter, whose aim was to illustrate the explanatory power of natural selection. Dawkins took seriously Paley's observation that the existence of extremely complex objects demands an explanation in a way that the existence of simple ones does not and sought to demonstrate that natural selection could provide this explanation even in the most remarkable cases of complex design. What Dawkins had no need to stress is that natural selection is also the best explanation for the existence of even mildly complex designs.[28]

A more fundamental reason to doubt the relevance of complexity as a diagnostic of adaptation is that it can decrease for adaptive reasons as well as increase. The real

[28] Paley's style of argument can also be applied to objects much simpler than watches. This is nicely illustrated in a short story by Arthur C. Clarke entitled *The Sentinel* and the film it inspired, Stanley Kubrick's *2001: A Space Odyssey*. In both the short story and the film, explorers discover an unusual object on the Moon, which its discoverers judge to be of alien construction based on the perfection of its geometry. The object exhibits the appearance of design, not because it is complex, but because it is *perfect*. In the film, this object is just a large, black, rectangular monolith. It is hard to imagine a simpler kind of object, yet despite its simplicity, if such an object really were found, the conclusion that it was of artificial construction would be almost inescapable.

issue is the likelihood of a useful property arising rather than the complexity of the result relative to some putative ancestor.

A clearer picture emerges of adaptation when we replace the emphasis on complexity with considerations of optimality. However, the view that natural selection is an optimising process is not without controversy. Simon (1957) coined the term 'satisficing' to capture an alternative conception of natural selection as producing solutions that are good enough for survival without necessarily being optimal. But as Dawkins (1982: 45f) notes "[t]he trouble with satisficing as a concept is that it completely leaves out the competitive element which is fundamental to all life." If an individual with a capacity that is merely *good enough* came into competition with another individual with a capacity which was *better*, natural selection would, by definition, favour the latter over the former. Few biologists would argue that natural selection finds perfectly optimal solutions, but it is still an optimis*ing* process, the designs attained generally being *locally* rather than *globally* optimal. With this rationale, assumptions of optimality guide inquiry in some strands of evolutionary biology much as they do within Minimalist inquiries in linguistics (e.g., McCleery, 1978; Orzack & Sober, 2001).

Parker and Maynard Smith (1990) call this approach *Optimality Theory*[29] and outline how evidence of optimality can be used to understand adaptations and their functions. A basic requirement is that a descriptive theory be constructed of the purported adaptation as a (locally) optimal solution to the constraints imposed upon it by the environment and the other systems with which it interfaces. In essence, this involves (a) identifying what they call the *strategy set*, which is the local range of genetically attainable variants that differ with respect to the trait in question, (b) identifying the function over the strategy set for which the observed trait is optimal

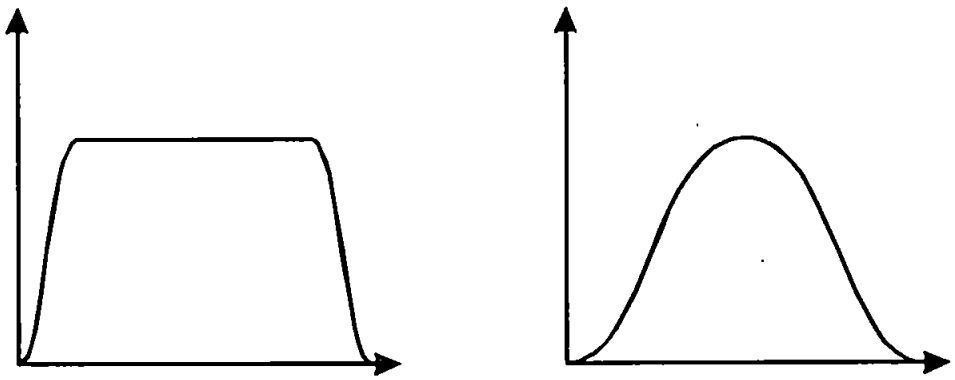
[29] Not to be confused with the linguistic theory of Prince and Smolensky (1993) with the same name.

among the possible choices, and (c) relating this function to fitness (even if only fulfilling a fitness-related role at some time in the past). If the trait is optimal for a function unrelated to fitness, it cannot have been selected for it. Attaining some plausible view of the strategy set requires some evidence about what local variations are possible given the limitations imposed by physical law and the contingencies of past evolution. Parker and Maynard Smith use existing variation in closely-related species as one source of evidence about this. Variation during ontogeny is another. On this issue, Parker and Maynard Smith (1990: 27) note that when constructing models

> [t]ypical biological constraints usually define some obvious boundary conditions for the strategy set, but strategic possibilities that have never in fact been observed are included unless there are reasons of this kind for leaving them out.

We can justify including possibilities that we cannot be certain are physically attainable on the grounds that if the observed value is optimal even with these possibilities included then it is also optimal without these possibilities included. However, an inflated strategy set could give us a false impression of how improbable the optimal value is. It may be that the observed state represents the only possibility or one of very few, in which case, it would be unnecessary to invoke a process of cumulative natural selection to account for it. Similarly, a function which can be served optimally in a large number of ways is not something that we can confidently conclude was a factor in the selection of the observed variant. This point is illustrated in the diagrams below. In (5a), a large proportion of variants correspond to the highest possible fitness value, while in (5b) only a small proportion are optimal. The presence of 'improbable' optimality is hence more likely to be a sign of the function's relevance to selection.

5. a. Probable optimality b. Improbable optimality



In its present form, the optimality diagnostic requires us to assume that functionally relevant properties of the environment are static and that evolution has an opportunity to scale peaks in a fitness landscape without its topography changing. This is an assumption that will not hold in cases where co-evolutionary dynamics exist between different variants such that changes in one will affect the other's chances of survival. In predator and prey species for example, the fitness landscape of each species will be continually changing as each innovation in predation is matched with an innovation for evading capture in the prey species. More will be said about how the optimality diagnostic could be generalised to more complex dynamics in chapter seven. Until then, the environment will be idealised as static, since none of the phenomena to which the optimality diagnostic is applied in the present thesis appear to be subject to co-evolutionary dynamics. This appears to be an appropriate idealisation for present purposes.

The discussion in the previous section concerned whether there are alternatives to natural selection for explaining the existence of traits. A related, though distinct question concerns whether there are processes other than natural selection that could account for the existence of a good fit between trait and function. The existence of such

mechanisms might of course lead to problems for applying the optimality diagnostic. Indeed, this problem does arise when there are optimally evolved biological mechanisms that themselves perform optimising functions such as learning mechanisms that generally seek to minimise certain kinds of errors (say between predictions and observations). The neural mechanisms involved in action planning and the immune response are also examples of optimising processes that sit atop the biological substrate. Cultural evolution is in a slightly different category, but is also an optimising process that could lead to problems for interpreting the optimality diagnostic.[30]

To resolve these issues, we need to find some way to distinguish between the optimising effects of genetic selection on the one hand and those of these other processes on the other. One way to do this is to examine whether the properties still emerge when the action of other processes can be ruled out. In the case of language, we can for instance observe whether a putative learning process has the necessary feedback available to it to form appropriate generalisations about the linguistic input. We can also observe which properties emerge in the earliest stages of creolisation before cultural evolution can be implicated. On the other hand, genetic evolution can be ruled out as an explanation for functions that are only optimally served in modern times (e.g., noses being optimally suited to supporting spectacles). For one of the other processes to be implicated, it would also be necessary to at least show that the trait in question optimises a function of the relevant kind. In the case of cultural evolution, the function would have to be some kind of selection pressure that applies to cultural transmission. To optimise this function, the trait would have to be something that can actually be passed on culturally and which replicates more successfully than similar but incompatible traits because of factors like the nature of the learning bottleneck and the

[30] Optimising processes that occur outside of the domain of biological evolution may also have an essentially selectionist nature. Proposals of this kind have been made to explain cultural evolution (Dawkins, 1976), certain kinds of learning (Edelman, 1987), planning and intentionality (Calvin, 1996; Dennett, 1991), and the immune response (Jerne, 1955). For an overview examining many of these ideas, see Cziko (1995).

extent to which people are predisposed to acquiring it. These predispositions may in turn have been genetically selected for the same function. So while it may be the case that a good fit between trait and function is insufficient to conclude that a function was relevant for genetic selection, evidence of cultural selection for the same function is not in itself an indication that it wasn't.

Considerations of optimality also allow the function for which a property was selected to be identified without making any specific claims about the form of earlier variants. Given that all variants in the proximity of an optimal form will be less fit (by definition), evolution could have preceded from any direction towards the current state. At the same time, this mode of inquiry severely restricts the kinds of adaptationist theories that can be proposed by requiring that a fitness-related function be identified for which the trait is locally optimal. Such restrictions rule out certain kinds of adaptive explanations that might otherwise seem plausible making it harder to tell unfalsifiable 'just-so' stories about which functions were relevant for the selection of a trait. The approach advocated here could be demonstrated to be invalid if it fails to differentiate between different adaptive stories even while adhering to the principles outlined above.

The discussion so far has been an attempt to place appropriate boundaries on evolutionary theorising and to develop a more principled set of diagnostics for identifying functional pressures. I have devoted considerable space to these issues because of their importance, but I now turn to the specific arguments relating to language.

3.3 Reconciling linguistics with evolutionary biology

The question of what kind of evolutionary explanation could account for language universals is something which should be asked separately for each property individually. Several varieties of explanation have been proposed, each implicating natural selection in a different way. A given property might have been exapted for its

new function from pre-existing neural structures that were selected for previous functions (Piattelli-Palmarini, 1989; Gould, 1991; Uriagereka, 1998; Chomsky, 2002; Hauser *et al.*, 2002), or from structures that became available as a result of adaptive increases in the size or complexity of the brain (Chomsky, 1972, 2002; Uriagereka, 1998) or the prefrontal cortex in particular (Deacon, 1992). Alternatively, Pinker and Bloom (1990: 721) argue that the emergence of language must have involved an adaptive reorganisation in the brain rather than simply an increase in size given that "mere largeness of brain is neither a necessary nor a sufficient condition for language" citing in support of this Lenneberg's (1967) studies of nanencephaly and individual craniometric variation.

It is not impossible to determine which of these or any other conceivable hypotheses is more accurate. If a property was originally selected for reasons other than language, then we should not expect it to serve its language-related functions optimally. The existence of such properties might nevertheless be explainable in terms of their optimal fit to previous or other functions of the relevant neural structures, or to functions of concomitant traits. Considerations of optimality therefore present themselves as a way of systematically comparing the relevant hypotheses. If a property is improbably optimal for a language-related function, we should conclude that it was selected for this function. If it is improbably optimal for a function that pre-dated language, then we should conclude that it was selected for that function and only later exapted for language.

Considerations of optimality already manifest themselves in interesting ways within linguistic theorising. Within the 'Minimalist Program' (Chomsky, 1995), inquiry proceeds with the null hypothesis that the computational system responsible for grammar is a more-or-less perfect solution to the demands that are placed on it by the cognitive systems with which it interfaces. In particular, it is hypothesised to have only

two interfaces, one with what is called the *articulatory-perceptual system* and the other with what is called the *conceptual-intentional system*. In less formal terms, it can be thought of as providing a mapping between a system dealing with signals and a system dealing with meanings. A Minimalist will assume that this mapping is achieved in an extremely economical way without redundancy thus severely limiting the kinds of theories that could be proposed to explain it. The legitimacy of this mode of inquiry is reinforced to the extent that explanatory theories can in fact be discovered this way and the approach has indeed led to important developments in many areas. If this is a realistic view of the language faculty, it appears to be entirely compatible with the optimising influence of natural selection, which, we might expect, would explain why it works. This interpretation is nevertheless rejected by Chomsky (1995) for reasons I will examine in section 3.3.1.

Another pattern in theorising that is in apparent harmony with the view that linguistic properties were selected for linguistic functions concerns learnability. Chomsky (1986) argues that the language faculty must be heavily constrained if we are to account for the swiftness of language acquisition, and linguists routinely appeal to learnability considerations in support of proposed universals. For instance, Chomsky (1986) uses learnability considerations to argue against vacuous movement (i.e., movement that does not result in a change of word order or pronunciation), Pinker (1994) argues for the head parameter (i.e., a single parameter that determines whether verbs precede or follow their objects, whether there are prepositions rather than postpositions, and so on), Kayne (1984) argues that syntactic trees are at most binary-branching, and Radford (1988) uses this kind of argument to support the view that words belong to different categories since assuming this allows a language learner to generalise syntactic knowledge acquired about one word to every other member of the

class to which it belongs.[31] The appeal to learnability considerations is a typical one and if swift acquisition confers a selective advantage it is one that is perfectly compatible with adaptive explanations because we should expect all constraints to be functional in limiting the range of hypotheses that need to be entertained in language acquisition.

Despite the apparent compatibility of these claims with selectionist accounts of language evolution, some linguists have presented a number of arguments that challenge the view that aspects of language could be adaptations. I summarise four of these here, which I take to encompass the most serious objections and for which I have never seen any convincing refutation. The objections are (a) that the 'tinkering' process of natural selection could not have crafted something as elegant as the language faculty, (b) that certain properties of language actually appear to be maladaptive, (c) that a beneficial language mutation could not succeed because the first mutant would not be able to benefit from it if he or she was the only individual in his or her linguistic community who possessed it, and (d) that grammatical universals are unlikely to be adaptations because there is very little evidence of links to communicative efficacy. I tackle each of these points by illuminating counterexamples and hidden assumptions, thus leaving the door wide open for selectionist explanations and hence the proposals to follow in chapters 5 and 6.

3.3.1 The messiness of evolution and the elegance of language

Chomsky (1995; 2002), Piattelli-Palmarini (1989) and others have expressed scepticism about the possibility of explaining the elegance of the language faculty in terms of what Jacob (1977) described as the 'tinkering' of evolution. Chomsky has particularly emphasised the apparent lack of redundancy in the principles of the grammar. To the extent that descriptions have had overlapping coverage he says

[31] These claims can be motivated in other ways as well and not all of them are currently accepted. The head parameter for instance was dispensed with by Kayne (1994) by deriving precedence relations in terms of the structural relation of asymmetric c-command, a proposal adopted in a modified form within the minimalist program of Chomsky (1995: ch4).

> Repeatedly, it has been found that these are wrongly formulated and must be replaced by non redundant ones. The discovery has been so regular that the need to eliminate redundancy has become a working principle in inquiry. (Chomsky, 1995: 5)

He regards this as surprising noting that

> language is a biological system, and biological systems typically are "messy," intricate, the result of evolutionary "tinkering," and shaped by accidental circumstance and by physical conditions that hold of complex systems with varied functions and elements. Redundancy is not only a typical feature of such systems, but an expected one, in that it helps to compensate for injury and defect, and to accommodate to a diversity of ends and functions. (Chomsky, 1995: 29)

But despite claims to the contrary, redundancy is not an expected feature of biological systems if it is costly to the organism. Naturally enough, redundancy is retained if it helps to "compensate for injury and defect" and so on but, under conditions that impose heavy demands on processing or other neural resources, the situation can be very different.[32]

Part of the difficulty in imagining how extremely elegant structures could be crafted by natural selection may be associated with seeing evolution exclusively in

[32] Chomsky blurs the distinction between the notion of redundancy in the functioning of a system and redundancy in the description of a system, but this makes a big difference. Many internal organs like the lungs come in pairs, suggesting a certain amount of functional redundancy, but a description of a pair of organs needn't itself contain superfluous elements. We could for instance describe the bifurcating bronchial structure of the respiratory system in terms of a derivational procedure that combines 'respiratory objects' (much as the merge operator of minimalist syntax combines syntactic objects) thus producing a very economical and non-redundant description of a system possessing redundancy.

terms of the adding of parts, but natural selection can also strip away unnecessary components established in connection with previous uses of structures. Structures can be stripped away under selection for streamlining (the loss of limb structures and bodily hair in aquatic mammals for instance). Features can also degenerate when they cease to play an important role in selection. This occurs as a consequence of the accumulation of mutations that would otherwise be selected out and can explain the deactivation of genes like those responsible for manufacturing products that have become readily available in an organism's diet such as vitamin C in the case of apes (Pauling, 1986), and the existence of vestigial organs like the eyes of cave fish, wings of flightless birds, and so on. Redundancy and streamlining are consequences of different design priorities, like those that govern the design of passenger airliners versus those that governed the design of experimental aircraft used to break the sound barrier. Redundancy is only inevitable in aircraft design if safety overrides other concerns.

A related issue concerns the view that a 'complex' adaptation necessarily consists of many different parts that could have evolved relatively independently, but it is not clear that this is true of language at the level of grammar or its neural representation. Indeed, within Minimalist inquiries (Chomsky, 1995), there is a trend towards a conception of the language faculty requiring fewer and fewer descriptive principles. These principles are so interrelated that language may be 'irreducibly complex' in the sense that no subset of its parts could be functional for any purpose. But even if this is true, it would only present a problem for an adaptive account of language evolution if it is assumed, incorrectly, that natural selection always proceeds via the adding of parts.

It is conceivable that when certain neural structures were exapted for language, they were very complex and cumbersome from the point of view of its new function, perhaps having the ability to encode an unnecessarily broad range of grammatical

possibilities. Natural selection might have worked to streamline them, making them more efficient in various ways. One result of streamlining may have been to limit the scope of hypotheses that each new generation of language learners needed to entertain about the structure of the input, a result that would have increased both the speed and accuracy of acquisition.

As an analogy, we know that Michelangelo's *David* could not have been assembled by adding one part at a time because it is a sculpture consisting of a single piece of marble. Nevertheless, we know that it was the result of a very slow process that 'released' it from an initially amorphous block of stone. There is a similar 'sculpting' process in the ontological development of the brain that involves the selective pruning of neurons and the connections between them. The structural consequences are dramatic with more than half of the neurons that develop being killed off by adulthood (Oppenheim, 1991). This phenomenon, called *programmed cell death*, also occurs in the development of other organs including the hands and feet where digits are sculpted as a result of the death of the tissue between them (Saunders, 1966).

The loss of structure also occurs in genomes. The human genome for instance, has fewer chromosomes and, according to the Animal Genome Size Database, is slightly smaller (by weight) than the genomes of all of the other great apes. The loss is probably mostly associated with the amount of non-coding rather than coding DNA, but there may be an adaptive reason for it. Piattelli-Palmarini (1989) uses the existence of non-coding DNA as evidence that evolution is wasteful and messy, but there is evidence to suggest that the optimal amount of non-coding DNA is not nil. In many cases, differences in genome size between closely related species might be explained in terms of their consequences for the rate of cell division. Cell division rates in turn have consequences for developmental rates and hence the kinds of seasonal niches that a species can occupy (Gregory, 2002).

Given that evolutionary mechanisms can and do reduce the redundancy of systems either by removing structures or by disabling their functions, the generality of Chomsky's assertion about the level of redundancy that should be expected in biological systems should be called into question. The language faculty (narrowly construed in terms of the computational system) may or may not be unusual, but it is what it is – a biological system with apparently very little redundancy.[33] The question is whether this can be explained in terms of the kinds of processes that are known to reduce redundancy in biological systems or whether a new category of explanation is required to account for how such an elegant system could emerge.

A further complication is that it is not clear whether a language faculty without a given property would be simpler in all cases. Take the property of movement for instance. Is a language without movement phenomena simpler than a language with movement phenomena? The answer is not obvious. Suppose that to realise a certain syntactic dependency between a pair of constituents, they need to appear in some specified structural configuration. If this structural configuration is defined in sufficiently broad terms then it may be that *the man* bears the same structural relationship to *bite* in both of the sentences below, thus accounting for the uniformity in interpretation (the man being understood to have been bitten in each case).

6. a. The dog bit the man.
   b. The man was bitten.

If the same configuration is implicated in both sentences, then the flexibility in word order would be a result of a *lack* of constraints that would prevent it rather than of the computational system having properties like a MOVE operator to generate it, so it is

[33] I will assume that Chomsky (1995) is correct about the lack of redundancy in language. This premise could perhaps be challenged, but I'm attempting to show that even if we accept it, a selectionist account is not ruled out.

possible in principle that a language faculty that allows movement would be less complex than one that disallows it. Rather than ask what accounts for this flexibility, it might be more sensible to ask what limits it. Such examples highlight the difficulties inherent in distinguishing figure from ground, or of what requires an explanation versus what does not.

3.3.2 Maladaptive consequences of grammatical universals

Some commentators argue that there are grammatical universals that are actually maladaptive (Piattelli-Palmarini, 1990; Uriagereka, 1998; Lightfoot, 2000). These arguments follow a general pattern of which Lightfoot (2000) is typical. He makes the point with reference to the constraint that blocks the extraction of subjects from tensed clauses under certain conditions. This constraint is illustrated by the comparison in (7). In (7a), *who* is moved from the embedded object position (leaving the unpronounced trace *t*). By contrast, extraction from the embedded subject results in the unacceptable sentence in (7b).

7. a. Who$_i$ do you think [that Ray saw t_i]?

   b. *Who$_i$ do you think [that t_i saw Fay]?

Lightfoot notes that this constraint appears to apply universally, but that there is variation in the strategies that are used to overcome it in different languages. In English, extraction from both the object and subject positions is permitted when the overt complementiser *that* is not used to introduce the embedded clause, as in (8).

8. a. Who$_i$ do you think [Ray saw t_i]?

   b. Who$_i$ do you think [t_i saw Fay]?

Lightfoot argues that since different languages employ different strategies to circumvent this constraint, speakers have some use for expressing what the constraint disallows. This, he argues, makes the constraint maladaptive, from which he concludes that selectionist accounts of such properties are misguided.

Assuming for a moment that Lightfoot is justified in concluding that the constraint is non-functional or maladaptive, then there are only two ways to account for its existence. Either it exists despite itself – by chance surviving the processes that tend to weed out maladaptive properties, or it is a concomitant of some other trait that does have adaptive consequences. Indeed, if a property is costly in some ways, then its continued existence is puzzling and so could actually be used as evidence for it being associated with a counterbalancing benefit as when the increased choking hazard apparently associated with a descended larynx is used as evidence of it being an adaptation for speech (Lieberman, 1984).[34] Lightfoot (2000: 244) takes this approach, arguing that "the restriction on the movement of subjects is a by-product of the more general condition on movement traces" and therefore, "a spandrel", noting that this "general condition may well be functionally motivated, possibly by parsing considerations". If it is a by-product of a constraint that evolved by natural selection, it is, by virtue of this, a by-product of natural selection, which means that selectionist accounts are not, as he suggests, irrelevant, although these cases do serve to warn us that when we are attempting to discover what selective changes have taken place, we cannot assume that a given pair of traits exist independently of one another (the lesson of §3.1.1).

The logically prior conclusion that the constraint on subject extraction is maladaptive is not adequately established by Lightfoot either. Firstly, it doesn't follow

[34] Fitch and Reby (2001) argue that a descended larynx could be an adaptation for something other than speech on the basis that it has also been observed in other species. A possibility they raise is that a descended larynx, by lowering formant frequencies, was selected for as a way of exaggerating apparent body size. This may be a matter of controversy, but the logically prior conclusion that the existence of a choking hazard would suggest a counterbalancing benefit is not generally disputed.

that a constraint that rules out an utterance phrased in a particular way would present a disadvantage to language users if its meaning can easily be expressed in other ways. The intended meaning of (7b) can, after all, be expressed by (8b) so a language faculty that allowed both wouldn't necessarily be better adapted. It may actually be desirable from a computational point of view to eliminate the kind of arbitrary choices that would allow exactly the same meaning to be expressed in more than one way.

Secondly, even if we grant that a given constraint is a disadvantage from the point of view of certain functions in certain contexts, this disadvantage may be offset by (a) the function being adaptive in other contexts that are more relevant, or (b) the constraint also serving functions other than those for which it is considered maladaptive. There are disadvantages associated with most traits, even the vertebrate eye, which many regard as the archetypal adaptation. Eyes are easily irritated, and provide a gateway to the body and the brain for foreign bodies and infections, so if we evaluated their functional significance in the environment of a sandstorm, we might conclude that they too are maladaptive. We would also come to this conclusion if in more typical contexts we only considered the occasional problems of eye infections and ignored the advantages associated with vision, but all of the consequences of a trait in combination contribute to whether it will be selected. Evidence that the constraint on subject extraction prevents speakers from communicating certain kinds of useful meanings tells us nothing about its evolution other than the mundane conclusion that it wasn't selected for communicating those specific meanings. It doesn't tell us that there was nothing it *was* selected for.

This kind of conceptual error might have something to do with the pervasive tendency to define traits in terms of their functions or effects. It is tempting to conclude that the only effect associated with the property that blocks extraction from subjects is to block extraction from subjects, but the function of a trait cannot simply be stipulated

in its definition – it must be empirically discovered. Unfortunately, the strategy of defining traits in terms of their functions is often necessary when dealing with cognition since very little is known about how cognitive capacities are physically realised in the brain. The situation is akin to someone with no knowledge of aerodynamics discussing, in functional terms, the property of birds that gives them the power of flight. It is possible to discuss this property without having any understanding of the role that wings play in it, and from this perspective, it may seem implausible that other consequences of wings, such as those having to do with thermal regulation, could be associated with it. Likewise, the property responsible for preventing the extraction of subjects might have various other effects that, from our present state of knowledge, appear unrelated.

A broader variant of Lightfoot's argument concerns the autonomy of syntax from cognition. Uriagereka (1998) argues that the possibility of producing sentences that are ungrammatical but comprehensible (speech errors, speech of foreigners, speech of infants under three years old, etc.) and sentences that are grammatical but incomprehensible (garden path sentences, sentences with complex embedding, difficult technical language, etc.) suggests that usability considerations are orthogonal to grammaticality considerations, but this is not necessarily the case. Such examples have traditionally been used by Chomsky (1980) and others merely to argue that grammatical and semantic representations are dissociable, not that the well-formedness of each type of representation is uncorrelated. Grammatical sentences appear to be easier to process and comprehend than ungrammatical ones so a functional role is not ruled out. If the computational system responsible for grammar is autonomous from the rest of cognition, its constraints may operate even when they are unhelpful, but nevertheless be favoured by selection on the basis of their utility in other contexts.

3.3.3 Mutants would have no one to talk to

Another potential problem with an adaptive view of language evolution arises from the necessity that it proceed via a sequence of plausible stages such that each step represents an increase in fitness over the last. The problem, noted by Geschwind (1980), Hurford (1999b) and others, concerns how any mutation of the language faculty could ever be favoured if an individual's ability to communicate relies on conformity to the conventions of a linguistic community. Any departure from grammatical conventions on the part of the mutant, it is argued, would present a problem for mutual comprehension.

A similar question arises in the familiar case of language change that is cultural rather than genetic like the gradual sequence of changes that led from Latin to French. These changes do not have a biological basis. Instead they are the result of language learners 'misanalysing' structures, and of speakers inventing new words, applying old ones to new contexts and so forth. We could argue that these departures from conformity impair communication too, but it would be manifestly untrue to claim that this kind of gradual change is impossible and if the conformity argument doesn't work as an argument against gradualism in this specific case, then it cannot be used in general.

Assuming that the evolution of the language faculty did proceed via a number of small steps, the problem would still remain as to how changes initially arising in one individual could come to dominate in the broader population. If the change was also associated with other effects that were a selective advantage then these could have offset any disadvantage associated with nonconformity. There are at least three kinds of arguments that could be made along these lines.

Firstly, the disadvantage associated with nonconformity might have been buffered by a tendency for people to value novel, creative use of language. Unconventional language can draw attention to itself and provoke curiosity in the way

that puzzles do, a fact long exploited in literature, advertising and politics. Minor deviations from standards might therefore be advantageous precisely *because* of their nonconformity, thus outweighing any costs associated with occasional difficulties in comprehension.

The second possibility is that communicative functions were outweighed by other functions of language that don't require conformity to the conventions of the linguistic community. As Chomsky (2002: 148) notes, language also has non-communicative functions.

> Actually you can use language even if you are the only person in the universe with language, and in fact it would even have adaptive advantage. If one person suddenly got the language faculty, that person would have great advantages; the person could think, could articulate to itself its thoughts, could plan, could sharpen, and develop thinking as we do in inner speech, which has a big effect on our lives.

The third possibility relates to the observation by Sperber (1990) that adaptive changes can actually occur without breaking the conventions of the linguistic community. Introducing grammatical constraints that improve the efficiency of language acquisition, and that reduce costs associated with computation and representation would make the language faculty better adapted to acquiring and using the kinds of languages that it is already exposed to without leading to nonconformity.

3.3.4 The lack of convincing adaptive explanations

The previous arguments fail to rule out adaptive functions, but Hauser *et al.* (2002) point out that there isn't much in the way of compelling evidence to rule them in either. They observe simply that many linguistic properties have a "tenuous connection to

communicative efficacy" (p.1574), and argue that the most promising explanation for this lack of fit is that the language faculty "evolved for reasons other than language" (p.1578) later being co-opted for its present function without much modification.

This view is open to criticism on the basis that what is generally meant by "communicative efficacy" is an unnecessarily narrow view of linguistic function so seeing selection for prior non-linguistic functions as the only alternative is to ignore many categories of explanations. Misleading connotations of the word *function* may be largely to blame given that the word usually suggests an increased benefit rather than a reduced cost while *non-functional* misleadingly lumps neutral and costly traits together. Some general categories of functions that might be associated with linguistic properties are listed below, beginning with some that could come under the category of "communicative efficacy," and leading on to others that wouldn't normally be described in these terms.

9. increased expressiveness; reduced ambiguity; better information theoretic signal properties; reduced processing load (in production and comprehension); distribution of processing load facilitating simultaneous activities like hand movements; increased biases facilitating language acquisition making it faster and more accurate; reduced cost associated with memory consumed by the lexicon; reduced nutritional and energy requirements (growth/acquisition; maintenance and running costs); reduced dependency on scarce resources (metabolic or processing resources); reduced risk associated with faults, instability and other negative side-effects; increased capacity for reasoning and structuring of thought

Costs associated with language may have been more relevant in the selection of its specific properties than communicative functions. Properties could have been selected for their effects on reducing representational and processing costs ultimately understood in terms of metabolic energy or other neural resources, or reducing the costs associated with language acquisition in terms of time, effort and reliability. The proposals presented in chapters 5 and 6 are of this kind.

3.4 Summary

There are various naïve paths to a selectionist account of language evolution which we would be right to criticise, but there is arguably also an *informed* path. The naïve paths are pursued without adequately appreciating the facts in (10).

10. a. Not every trait has an independent selective function (cf. §3.1.1).
    b. A trait that currently serves a given function may not have originally arisen under selection for that function (cf. §3.1.2).
    c. The forms that natural selection is capable of producing are constrained by physical law and the historical contingencies of past evolution (cf. §3.1.3).
    d. Elegant structures can emerge as a result of simple, self-organising, developmental processes (cf. §3.1.4).
    e. Some properties emerge as a result of learning (cf. §3.1.5).
    f. Some properties emerge as a result of cultural evolution (cf. §3.1.6).

The informed path involves recognition of the facts in (10), but also those in (11) and (12).

11. a. Demonstrating that a property is an incidental by-product of selection for a concomitant trait does not allow us to conclude that the property was not refined under genetic selection (cf. §3.1.1).

b. Demonstrating that a property serves functions that it wasn't selected for does not allow us to conclude that the property was not refined under genetic selection (cf. §3.1.2).

c. Demonstrating that there are physical constraints on the forms that natural selection can produce does not allow us to conclude that a given property was not refined under genetic selection (cf. §3.1.3).

d. Demonstrating that there is a self-organising developmental explanation for the emergence of a given property in ontogeny does not allow us to conclude that the property was not refined under genetic selection (cf. §3.1.4).

e. Demonstrating that a property will be acquired under some conditions and not others during the lifetime of an organism does not allow us to conclude that the property was not refined under genetic selection (cf. §3.1.5).

f. Demonstrating that a property can or did emerge after a period of cultural evolution does not allow us to conclude that the property was not refined under genetic selection (cf. §3.1.6).

12. a. Complexity can decrease as well as increase for adaptive reasons (cf. §3.3.1).

b. Selection against redundancy is expected to occur under certain circumstances (cf. §3.3.1).

c. Maladaptive consequences of a property are not evidence of its non-optimality (cf. §3.3.2).

d. It is not necessarily a disadvantage for an individual to exhibit non-conformity with respect to the conventions of his or her linguistic community (cf. §3.3.3).

e. Mutations that affect language competence will not necessarily result in the mutant exhibiting non-conformity with respect to the conventions of his or her linguistic community (cf. §3.3.3).

f. Language universals were not necessarily selected for communicative functions (cf. §3.3.4).

A recurrent obstacle to theoretical progress appears to be the acceptance of a number of false dichotomies. We see this in the 'debates' about functional versus physical modes of explanation (cf. §3.1.4), genetic versus environmental causes (cf. §3.1.5), and genetic versus cultural evolution (cf. §3.1.6).

Another obstacle appears to be the use of misleading terminology. I have argued here that the term *spandrel* should be abandoned in favour of the terms *concomitant traits* and *concomitant changes* (cf. §3.1.1), that the term *exaptation* should be reserved exclusively for changes in the function of a trait in the literal absence of any changes in its form (cf. §3.1.2), and that the term *innate* and synonyms such as *genetically programmed*, *instinctive* and *hard-wired* should be set aside in favour of discussions in terms of *genetic and environmental effects* or the absence thereof (cf. §3.1.5).

I have also attempted to outline and extend a diagnostic of selective functions based on optimality considerations (cf. §3.2). This diagnostic allows theories of selective functions to be compared on the basis of the extent to which the relevant trait performs a hypothesized function optimally and the extent to which this optimality is improbable, thus providing a principled way of distinguishing between candidate hypotheses. The approach potentially enriches the Minimalist Program of Chomsky (1995) in which inquiry into the nature of the language faculty proceeds with the working assumption that it is a perfect, non-redundant solution to the constraints imposed upon it by the systems with which it interfaces. This working principle has

been adopted simply because it appears to yield results, but has lacked independent motivation. I have attempted to show that traditional arguments against motivating this principle in terms of the optimising effect of evolution are not compelling, and that perfection and elegance are indeed compatible with the process of natural selection under certain conditions (cf. §3.3.1). To demand that the perfection of language be related to adaptive functions would further restrict Minimalist theorising. As such, assumptions of evolutionary optimality may help to facilitate inquiry into the nature of the language capacity as well as its origins.

4

The Evolution of Syntactic Universals

Chapters two and three looked at syntactic universals and evolutionary explanations respectively. The current chapter examines the intersection of these two fields within the context of the language evolution literature more broadly, setting the concerns of the current thesis against this backdrop.

4.1 Broader issues in the language evolution literature

Much of the work on the evolution of language has ignored or postponed the issue of syntactic universals and concentrated on issues like dating the origin of language, the evolution of performance systems, the social function of language, or the neuroanatomical changes that accompanied its emergence. Work on the evolution of syntactic universals occurs within the broader context of this literature and similar methodological issues arise within it so it is reviewed here before focusing on the literature that touches most closely on the particular concerns of the current thesis. At each stage, it will be instructive to review the kinds of evolutionary arguments used to address these questions and categorise them in the vocabulary developed in chapter three.

4.1.1 Dating the origin of language

In dating the origin of language, Lieberman's (1984) work on the evolution of the vocal tract has been very influential. Of particular interest are the developments that begin with archaic Homo sapiens from around 250,000 years ago, which involved the lowering of the larynx in the vocal tract. These changes, Lieberman argues, provide

evidence of speech because they had the effect of increasing the number of speech sounds that could be produced. These changes are interesting in light of an observation originally noted by Darwin about the choking hazard associated with the anatomy of the human vocal tract. Cziko (1995: 182) summarises this point:

> [U]nlike mammals that maintain separate pathways for breathing and feeding, thus enabling them to breathe and drink at the same time, adult humans are at a much higher risk for having food enter their respiratory systems; indeed, many thousands die each year from choking ... The risk of choking to which we are exposed results from our larynx being located quite low in the throat. This low position permits us to use the large cavity above the larynx formed by the throat and mouth (supralaryngal tract) as a sound filter ... We thus see an interesting trade-off in the evolution of the throat and mouth, with safety and efficiency in eating and breathing sacrificed to a significant extent for the sake of speaking.

In other words, we should expect these vocal tract modifications to be selected against without a counterbalancing pressure favouring them. Archaic Homo sapiens must have been using their vocal tracts in ways that provided the necessary selection pressure to power their adaptation and the changes probably occurred soon after the selection pressure was introduced. Speech is assumed to be the most likely source of this pressure, but language must have reached a level of sophistication that made it reasonably important for reproductive success in order to counterbalance the choking hazard associated with the modern shape of the vocal tract.

The vocal tract evidence suggests a lower limit on the timing of some form of spoken language at around 250,000 years ago. It is a lower limit because, as Pinker (1994: 389) has noted, "...e lengeege weth e smell nember ef vewels cen remeen quete

expresseve, so we cannot conclude that a hominid with a restricted vowels space had little language". Indeed, there appear to be modern languages with a comparable number of vowels. In Yimas for instance, "90 percent of all tokens of vowels are central vowels" (Foley, 1997: 71). Such languages may be possible, but that does not necessarily make them as efficient. Having access to a larger repertoire of sounds gives a speaker the ability to produce a greater number of distinct signals of a specified length so it is conceivable that such pressures drove the evolution of the vocal tract once speech had gained some evolutionary value. However, Fitch and Reby (2001) argue that a descended larynx could also be an adaptation for other things since it is not uniquely human. They also observe it in red and fallow deer and argue that in these cases a descended larynx, by lowering formant frequencies, was selected for as a way of exaggerating apparent body size.

Changes in the vocal tract may or may not be evidence of the emergence of speech, but further evidence would be required to link this evolutionary event to the emergence of language itself. Manual sign languages do not make use of the vocal tract at all so could have predated spoken language as Corballis (1991) and others have suggested. Alternatively, the changes in the vocal tract might have occurred merely as a response to increased vocalisations that were more like the call systems of other species than human language, with a fully syntactic language coming later and making use of speech organs that were pre-adapted to the task. The vocal tract is part of the articulatory-perceptual system which interfaces with the computational system, hence the constraints imposed by it may have also influenced the evolution of the latter. This is to be expected of performance systems generally, a point which will be pursued in more detail later in this chapter.

Lieberman (1984) was attempting to use the vocal tract evidence to date the origin of language, but in so doing, needed to establish that the descended larynx was

selected for speech. He appealed to design properties to do this, one of those properties being the choking hazard originally highlighted by Darwin (a concomitant trait associated with the lowering of the larynx), which he argued implies the existence of a counterbalancing benefit. To establish that this counterbalancing benefit was associated with speech, he argued that a descended larynx allows a greater variety of sounds to be produced.

In principle, it would be possible to evaluate this hypothesis more rigorously by appealing to optimality considerations. If the descended larynx not only increases the variety of speech sounds that can be produced, but also satisfies the functions of the vocal tract optimally, then it would be surprising if these were not the functions it was selected for. The optimal solution in this case would be some kind of compromise between the speech function and other functions such as breathing and digestion, but without further study, it is far from obvious what the ideal vocal tract characteristics would be for any of these functions, let alone what we should expect the perfect compromise to be like.

4.1.2 The evolution of the performance systems

As discussed in chapter two, a working hypothesis within Chomsky's (1995) Minimalist Program is that the computational system is an optimal solution to the constraints imposed on it by the performance systems with which it interfaces. If this is correct, then an understanding of the functional pressures on the performance systems is likely to provide some important insights into the evolution of the computational system. A number of studies have looked at the evolution of the performance systems, though not necessarily from this perspective. These studies have looked at the articulatory-perceptual side as well as the conceptual-intentional side, and also memory, which is necessarily used by both the 'performance' and computational systems. Indeed, any limitations on resources available to the computational system can be regarded as

arising from performance rather than competence issues, competence having only to do with the knowledge that the system embodies rather than how it is applied, though the question of how knowledge can be distinct from the processes that apply or extract it is ultimately a metaphysical one.

On the articulatory-perceptual side, the aforementioned studies of the evolution of the vocal tract are examples. De Boer's (1997) work on the self-organisation of vowel systems is another. De Boer uses computational models to optimise the acoustic distinctiveness of vowel systems for communication over a noisy channel. In this model, pairs of simulated agents in a population engage in "imitation games" in which one member of the pair called the *initiator* produces a vowel sound chosen randomly from its repertoire which the other member attempts to match with one of the acoustic prototypes in its own repertoire. The closest match is then repeated back to the initiator and in turn matched against its acoustic prototypes. If the closest match corresponds to the sound that this agent originally uttered, the imitation game is judged to be successful, but if it is different, it is judged to be a failure. The agents use failures to modify the formant[35] characteristics of their acoustic models, to add entirely new vowels or to remove those that consistently lead to imitation failures. The signals produced by the agents also have some random noise added, which has the effect of causing failures when an agent possesses more than one acoustic model that comes close to matching the properties of the signal. Over time, the acoustic models of agents self-organise such that they conform to those of the population at large while also being maximally distinct within the vowel space. The optimal dispersion of vowels throughout the vowel space is also typically observed in real languages. Hence, de Boer (1997) argues that something analogous to the process he models could explain it,

[35] Formants are bands of high energy in the frequency spectra of vowel sounds. Differences in formant frequency relationships correspond to differences in the perception of vowel quality (Ladefoged, 1993).

contrasting this explanation with ones that would account for the optimal distinctiveness as a genetic adaptation.

However, as discussed in chapter three, demonstrating that there is a self-organising explanation for the emergence of some property in development (either within an individual's lifetime or over a number of generations via cultural evolution) does not allow us to conclude that the property was not refined under genetic selection. In this case, a genetic mutation that for instance increased the fidelity of transmission so that less noise was introduced in exchanges between speakers would allow an agent to represent acoustic models that were more alike. In turn, this would mean that a larger number of acoustic models could be squeezed into the available vowel space. Hence, a mutant allele that increased fidelity could be viewed at a higher level as a gene controlling the number of vowels in the language. Likewise, a population of mutants who responded to successes in the way that de Boer's agents respond to failures would presumably not succeed in producing the optimal dispersion of vowels. Hence, the genes that control what happens when an imitation game fails can be viewed as genes for the optimal dispersion of vowels. In the real world, such genes may have been selected for this role. There are a wide range of functionalist explanations which seek to explain the existence of language universals in terms of semantic or pragmatic pressures rather than in terms of 'innate' or 'genetically determined' characteristics, but they all insist on the same dichotomy that de Boer embraces, and for that reason, prematurely exclude a role for genetic selection.

De Boer's use of optimality appears to raise important questions for the optimality diagnostic if optimality can arise for reasons other than successive genetic refinements. The difficulty arises because the mechanisms underlying learning and development also perform optimising processes. But since these mechanisms are themselves optimised by evolution, the outcomes are nearly always in alignment with

genetic fitness. Learning mechanisms that optimised behaviour in a way that routinely did more harm than good for instance, would not be selected. In this way the genes controlling the optimisation mechanisms can also be said to be genes for the traits those optimisation mechanisms produce. There is in fact no sense to a dichotomy of the kind that says a given trait T is either the result of developmental optimisations or genetic optimisations, but not both.

Performance systems have also been studied on the conceptual-intention side, particularly with reference to the conceptual capacities of non-human species. Heyes (1998) for instance, reviews the evidence for whether nonhuman primates possess a theory of mind, which it would seem would be a crucial prerequisite for being able to reconstruct the intentions of an interlocutor in communicative exchanges of the kind that characterise modern human language.

4.1.3 The social functions of communication

4.1.3.1 VERBAL GROOMING

Dunbar (1993; 1996) has observed a linear correlation between the amount of time nonhuman primates spend grooming each other and mean group sizes. Given the time allocated to social grooming (as much as 18.9% of the day for Papio papio baboons) that could otherwise be utilised for feeding and other survival-related activities, he argues that it serves an important adaptive function and suggests that this function is to maintain group cohesion. This would explain why the time budget allocated to grooming increases with group size since the larger the group, the greater the number of relationships an individual needs to maintain.

Dunbar argues that if humans were to use the same strategy to maintain group cohesion, the group sizes observed in hunter-gatherer and traditional horticulturalist societies could only be achieved by placing prohibitively excessive demands on time. He argues that the emergence of linguistic communication could have arisen as a more

efficient strategy for maintaining group cohesion since speech can accompany other activities and can be directed at more than one individual at the same time. Language also allows individuals to learn about third parties without having to engage with them directly. Dunbar (1993) cites a number of pieces of evidence in support of this including a study of the content of conversations in a university refectory in which he found that a large proportion of conversation was devoted to a kind of 'gossip' used to reinforce social relationships. 38% of conversation concerned personal relationships while a further 24% concerned personal experiences. Both types of content, he regards as important in developing social knowledge and bonding.

Dunbar extrapolates from a correlation between neocortex size and group size in non-primates to predict the expected group sizes of primitive humans. He uses this group size to calculate how much time humans would need to devote to grooming to maintain group cohesion using the same strategy. There are interesting questions about the validity of claims about group sizes and expected grooming time allocations, but I will set these aside and concentrate on the specific claims about the role of language in usurping the social function of grooming.

Dunbar (1993: 689) summarises what he sees as the link between the content of modern conversations and the evolutionary function of language as follows.

> The acquisition and exchange of information about social relationships is clearly a fundamental part of human conversation. The implication, I suggest, is that this was the function for which language evolved.

There are a number of serious problems with this inference. Firstly, as we saw in the previous chapter, historical origins cannot be directly inferred from current utility. People who can be found conversing in university refectories allocate a vanishingly

small proportion of their time to the defining activities of hunter-gatherers so we can hardly expect the content of modern conversation to reflect whatever ancient concerns it may have been selected for. Secondly, if we were to take time allocations to be an indicator of selective functions generally, then we would frequently find ourselves drawing rather absurd conclusions. In making inferences about the role of genitalia for instance, the amount of time they are utilised for sex relative to the time that they aren't would lead us to dramatically underrate their principal function. For the same reasons, we cannot conclude that because most conversations are of a certain type, these conversations served an important selective function. Gossip may simply be what the language faculty does in its downtime and hence be as frivolous as it appears to be.

It is also far from obvious that language is a more efficient means of reinforcing social bonds. As Hauser, Gardner, Goldberg and Treves (1993) suggest, the bonding function of grooming may actually rely on it being costly to the groomer, since expending effort demonstrates personal commitment. Dunbar offers no particular explanation why a communication system with the specific properties of human language would be especially suited to this function either. Social relationships could in principle also be reinforced by engaging in various other social activities that have the social benefits of spoken language without being as complex. Simple calls that allow a hearer to infer whether a speaker is a friend or foe such as the hisses and purrs of a cat would do. These calls could be produced while carrying out other tasks and could be directed towards multiple individuals simultaneously. A communication system as powerful as language with the ability to symbolise, communicate propositional content, ask questions, speak about objects that are not present or hypothetical events, and so on would only be necessary to serve social functions that are not served by grooming such as gaining knowledge about other individuals without having to observe or interact with them directly.

It is also far from obvious what kind of evidence could falsify Dunbar's theory. As Deacon (1993: 699) notes, Dunbar's claim

> is yet another in a long line of reverse logic, "just-so" stories about language evolution of the form: "Language makes X more efficient, therefore selection for X explains the origins of language." Substitute your favourite fashionable X from a large range of possible alternatives (more efficient foraging, better transmission of past experience to offspring, stronger social cohesion for intergroup competition, more subtle and devious social-sexual manipulation, closer bonds between kin and sexual partners, etc.)... The generic quality of this argument excludes few alternatives and offers little in the way of explanation for the remarkable structural complexity and semiotic uniqueness of language as compared with other forms of communication.

Indeed, other explanations for the correlation between group size and grooming are also easy enough to imagine. As the size of a group increases, an individual comes into contact with more conspecifics, who could potentially be carrying lice or other parasites. The increase in the amount of time spent grooming may be an adaptive response to this threat, regular and thorough checks allowing an infestation to be curtailed before it spreads out of control. The number of parasites found during grooming sessions might often be none, which again shouldn't be any indication that this is not its function. The number of weapons found by x-ray machines in airports is also extremely low, but that is still their purpose. The consequences of missing a weapon are serious, hence the need for vigilance. Likewise, a serious infestation of blood-sucking lice can lead to anaemia or the spread of various blood-borne diseases. Many mammals have special adaptations for grooming which highlight the seriousness

of the threat and while grooming may have a social function amongst primates, it is important to note that they are actually grooming, rather than just stroking, cuddling or fondling each other. Lemurs, like other primates, groom one another (Nakamichi & Koyama, 1997), but whatever social functions this behaviour serves will not directly explain why they have specially adapted grooming claws or why many also have lower incisor teeth that act as a comb (Rosenberger & Strasser, 1985).

The social cohesion theory and the hygiene theory do in fact make different predictions. If the former is correct and grooming time increases with group size because individuals have to service more relationships rather than because they need to be more vigilant about parasite infestations, then we would not only predict more grooming in larger groups, but more variety in grooming *partnerships*. Dunbar (1993: 687) himself cites evidence against this:

> The distribution of the data suggests that grooming does not necessarily function in such a way that each individual grooms with every other group member; rather ... it suggests that the intensity of grooming with a small number of "special friends" (or coalition partners) increases in proportion to increasing group size.

Social grooming may service relationships between conspecifics, but the evidence just cited suggests that its primary function has little to do with the specific social function of improving the cohesion of the broader group. Dunbar regards it as mysterious how this kind of grooming functions to integrate large primate groups, but fails to view it as counterevidence against his theory. On the other hand, if grooming time were determined by hygiene requirements, then we wouldn't necessarily expect the number of grooming partners to increase with group size because it wouldn't matter which

individuals were grooming each other so long as all members of the group were adequately checked.

Whatever the link between grooming time and group size, the correlation is necessarily broken in modern humans. If grooming served to reinforce social bonds, language might have been able to fulfil this role for group sizes beyond the practical limits of the grooming strategy, but a similar question also arises about how those practical limits would be exceeded if the primary function of grooming relates to hygiene. If the hygiene theory is correct, we should expect to observe novel adaptations for avoiding parasite infestations when group sizes get as large as those of hunter-gatherer societies. Indeed, this may be the long-sought explanation for what triggered human hairlessness.[36]

---

[36] The suggestion that human hairlessness is an adaptation to avoid parasites is not new, but as Morris (1967: 42) argued, the theory has lacked a clear explanation for why our ancestors were more susceptible to this problem than mammals that retain their fur coverings:

> One explanation is that when the hunting ape abandoned its nomadic past and settled down at fixed home bases, its dens became heavily infested with skin parasites. The use of the same sleeping places night after night is thought to have provided abnormally rich breeding-grounds for a variety of ticks, mites, fleas and bugs, to a point where the situation provided a severe disease risk. By casting off his hairy coat, the den-dweller was better able to cope with the problem. There may be an element of truth in this idea, but it can hardly have been of major importance. Few other den-dwelling mammals – and there are hundreds of species to pick from – have taken this step.

Pagel and Bodmer (2003) argue that hair loss was a response to the problem of parasites, but only once humans acquired technological methods of regulating body temperature such as clothing and fire that would compensate for the lack of insulation. However, it appears that humans have evolved a layer of insulating subcutaneous fat to compensate for the loss of hair, as have aquatic mammals such as dolphins and whales (Morris, 1967). If technological innovations provided a substitute for hair, we would not expect hair loss to continue to the point that a compensating adaptation of this kind would be required.

Cross-infection risks associated with larger group sizes may partly explain hairlessness, but this could not be the only factor at play given the existence of the very many mammalian species that are untroubled by their fur coverings despite living in much larger groups. Wildebeests for instance migrate in herds numbering in the hundreds of thousands. However, Morris (1967: 31) notes that most herding animals are not parasitized by fleas because the flea "lays its eggs, not on the body of its host, but amongst the detritus of its victim's sleeping quarters" and "for at least the first month of its life a flea is cut off from its host species". This means that animals of a nomadic species will leave any flea eggs behind and will generally be untroubled by them. By contrast, den-dwelling animals such as humans are parasitized by fleas and, as Morris (1967: 36) notes,

> we have our own special kind of flea – one that belongs to a different species from other fleas, one that has evolved with us. If it had sufficient time to develop into a new species, then it must have been with us for a very long while indeed, long enough to have been an unwelcome companion right back in our earliest hunting-ape days.

The evidence of a correlation between grooming time and group size is no more supportive of the social cohesion theory than the hygiene theory and the kinds of evidence Dunbar (1993) uses to determine the selective functions of social grooming and language are too weak to have any confidence in his conclusions. The weaknesses become more apparent if we apply the optimality diagnostic described in the previous chapter. If grooming behaviour was primarily selected to reinforce social bonds, we should expect it to be optimal among similar strategies, but it is unclear why this function would be better served by grooming than other social activities observed in primates including other forms of physical contact such as cuddling. Likewise, there is nothing particularly remarkable about language that suggests it was specifically selected for the exchange of socially relevant information. This function does not lead us to predict specific qualities like hierarchical phrase structure, movement, case systems, agreement and so on. Dunbar's theory also requires that language would only replace grooming when the cost of the latter exceeds a certain threshold at which language becomes the more efficient strategy, but no explanation is given for why language would not also be a more efficient strategy for reinforcing social bonds in much smaller groups. By contrast, a hygiene theory of grooming would predict the transition to hairlessness to occur where the cost associated with parasite infestations and the grooming time required to combat them exceeds the thermal benefits of having a thick covering of hair, but there is no reason to expect the costs to tip the benefits in the same way all over the body. Indeed, the areas that retain a thick covering are the most vital for survival and reproduction so are precisely the areas that need the greatest protection from the cold.

4.1.3.2 CULTURAL INHERITANCE

The capacity for humans to imitate and communicate means that members of each new generation benefit from the accumulated wisdom of past generations without having to

rediscover everything for themselves. Isaac Newton captured this idea in a letter to Robert Hooke in 1676 with the words "If I have seen farther, it is by standing on the shoulders of giants". Interestingly, the 'giants' who preceded Newton also deserve most of the credit for the substance of this very phrase itself. Earlier versions of it are known to date back at least to Bernard of Chartres in the twelfth century with the wording "We are like dwarfs standing upon the shoulders of giants, and so able to see more and see farther than the ancients" (Merton, 1993). Here we see that even Newton's task of expressing this idea was itself made easier by exposure to the cultural legacy he'd inherited, and by re-appropriating the idea as it had been passed down to him, he was unconsciously illustrating it.

Dawkins (2006) argues that the reason humans are easily deceived into believing falsehoods during childhood is because our capacity to take cultural knowledge on trust is adaptive more often than it is maladaptive for the simple reason that it allows us to rapidly acquire a vast reservoir of knowledge accumulated by previous generations without having to expend the time and effort to rediscover it for ourselves. This argument parallels that of the choking hazard being an indicator of a counterbalancing benefit associated with a descended larynx, but in this case, the cost is associated with being vulnerable to deception and the counterbalancing benefit is argued to be access to the accumulated wisdom of previous generations. In the terminology adopted in the previous chapter, vulnerability to deception and access to accumulated wisdom would be 'concomitant traits'.

Language isn't necessary for all forms of cultural transmission. It is possible to learn various customs and technologies simply by observing conspecifics. Nevertheless, language enables us to transmit knowledge about events and objects that are not directly present and which might be difficult or costly to observe under normal circumstances. As mentioned in the previous chapter, Pinker and Bloom (1990) argued that the ability

to transmit propositional content was the selective function of grammar, but their argument relied on a diagnostic of selective function that was based on the complexity rather than the optimality of its design, and it is far from obvious how their argument could be successfully recast in terms of the latter. Logicians have explored countless formalisms for encoding propositional content that differ quite substantially from the grammars of natural language, lacking characteristic properties such as case, agreement, movement, and tense, none of which appear to be especially important for expressing propositional content. Certain properties of natural language are nevertheless found within predicate calculus style formalisms. These include nested structure, quantifiers, pronoun-like variables and the distinction between predicates and arguments. Any attempt to argue that the transmission of propositional content was a relevant selective function would, at the very least, have to focus on these properties of grammar to the exclusion of others that contribute nothing to this end.

If language was selected for the capacity to transmit knowledge from speaker to listener, a number of authors have raised the question of what adaptive benefit this would confer on the speaker (e.g., Ackley & Littmann, 1994; Batali, unpublished; Cangelosi & Parisi, 1998; Hurford, 1999b). The benefit to the listener is taken to be obvious if it allows knowledge to be acquired without having to discover it for oneself, but as Cangelosi and Parisi (1998: 85) ask, "[w]hat is the advantage of producing the signal to the individual that produces it? Why should an individual that produces the appropriate signals live longer and have more offspring than other individuals that fail to do so?"[37]

[37] Interestingly, Catania (1990: 730) makes virtually the opposite argument, that the primary function of language is to change the behaviour of others, thus explaining the benefit to speakers but not listeners:

> By talking, we can change what someone else does. Sometimes what gets done involves nonverbal consequences, as when we ask someone to move something or to bring something to us. Sometimes it involves verbal consequences, as when we change what someone else has to say about something.

We could approach this from the point of view of kin selection, in which case providing information to listeners would confer an advantage to the speaker's genes if the listeners are close kin because they would likely share copies of the speaker's speaking genes. Another way to account for it would be to invoke the idea of a social contract (helping others with the expectation of being helped in return). The benefits of this kind of cooperation would presumably outweigh the very low costs associated with providing information to others. Talk is cheap.

But even a cursory examination of the public relations industry suggests that such theoretical manoeuvres are unnecessary. It simply isn't the case that speaking benefits the listener to the exclusion of the speaker. The daily flood of advertising messages to which people in developed countries are exposed, provide an obvious counterexample. It is also the speaker, rather than the listener, who is the primary beneficiary in acts of boasting and other forms of propaganda, so the question of the speaker's advantage may not be as paradoxical as some have suggested. Nor is it the case that only one or the other benefits at any one time. In acts of persuasion and negotiation, both parties potentially benefit from the speaker's act simultaneously. The view that communication is a zero-sum game is simply an inappropriate model.

4.1.4 The evolution of the linguistic brain

The marked asymmetry in hominid brain anatomy is detectable from fossil endocasts of Homo habilis onwards (Tobias, 1987) and is argued to result from tool use in which the hands are used asymmetrically – the left hand being used to hold or stabilise an object, and the right hand being used to manipulate the object with a tool. This pre-existing hemispheric specialisation may provide clues as to why language also became one of the functions that is strongly lateralised, being processed predominantly in the left hemisphere for most people.

Deacon (1992) stresses that non-human primates do not lack regions found in the modern human brain. However, from the time of Homo habilis about two million years ago, the ratio of hominid brain to body size has been larger than for any other primate. The proportion of brain regions has also changed, the prefrontal region being disproportionately large in modern Homo sapiens. Considering the areas that are enlarged, Deacon (1992: 64) argues that

> This likely produced differences in functional dominance of prefrontal circuits over other cortical and subcortical circuits, enhancements of certain computational capacities (e.g., verbal short-term memory, combinatorial analysis, and sequential behavioural ability), and probably corresponding differences in predispositions to employ these functional strategies – but not qualitative differences in the kinds of neural calculations possible. These are very specific changes that can best be explained as a consequence of constant directional selection for the specific demands imposed by language processing.

Deacon (1992, 1997) argues that as well as the brain evolving to support language, language use probably adapted to the pressures governing its cultural transmission from generation to generation – the grammars and vocabulary items that flourished being those that were easiest to learn from available linguistic input. If so, the changes in neural anatomy associated with language would in some sense reflect both the cause and effect of language use.

Another area of research arguably relevant for understanding the neural substrate of language concerns an observation by Rizzolatti, Fadiga, Gallese and Fogassi (1995) that neurons in an area of monkey cortex, homologous to Broca's area in humans, fire both when grasping and when observing another individual performing the same action.

Because of this property, Rizzolatti et al. (1995) label these *mirror neurons* and experimentally distinguish them from other neurons in the same area of cortex including those that only fire when a monkey performs the action itself in response to the sight of a graspable object.

Arbib (2002) argues that the mirror system is important for understanding the evolution of language because some such system appears to be required to capture the fact that interlocutors share meanings. It is also compatible with gestural theories of the origin of language since the Rizzolatti et al. (1995) evidence concerns the control of hand movements in monkeys in a brain area crucial for language processing in humans. Arbib (2002) also argues that this system is important for imitation, a skill which isn't well developed in monkeys but plausibly developed out of the mirror system.

4.1.5 The evolution of the critical period in language acquisition

A child can acquire multiple languages with native proficiency if models of each are present in his or her linguistic environment in early life. But the capacity to acquire a language declines with age, the greatest success being generally achieved prior to puberty. Language learning in adulthood also progresses via quite a different course, suggesting that it relies on different processes (Lenneberg, 1967). The existence of a critical or sensitive period for language development presents a puzzle. If we assume that the capacity to acquire a language is adaptive, why shouldn't it persist throughout life?

Hurford (1991) argues that the question of decline is wrongly formulated and that we should instead ask what selective advantage there is for a language acquisition capacity to exist at each stage of life. His view is that the existence of the capacity at a given stage requires an explanation in a way that the lack of the capacity at another stage does not.

Hurford (1991) developed a computational model to attempt to evolve critical period effects under the assumption that genetic variables can determine the ease with which language acquisition can occur at each stage of life. In his simulations, genes affecting language acquisition in early life were found to be under strong positive selection pressure, while those affecting later life stages were neither strongly favoured nor disfavoured, the capacity to acquire a language being by then mostly irrelevant. The result was that entities evolved that possessed language acquisition capacities that were only active in the earliest stages of life.

In the model, the simulated 'genes' that determined the level of language acquisition capacity at each stage were shared to a large extent with those that determined it at earlier and later stages, ensuring that abrupt differences from one stage to the next did not occur, but the capacity under genetic control otherwise had no continuity of existence, as if individuals are disassembled and reassembled with a slightly modified plan at stage of life. The model did not allow the existence of the language acquisition capacity to be simply a vestige of earlier development in the way that a functionless womb persists in a post-menopausal woman. The model parameters were such that the language acquisition capacity would have to evolve independently for each life stage at which it contributed to fitness, with the exception that some smoothing would occur due to genes affecting development at one stage also affecting development at neighbouring stages.

The simulations illustrated when, during the course of an individual's lifetime, a language acquisition capacity would positively contribute to fitness. However, the question of whether this capacity would actually be present at any given stage of life is a separate one which must be answered with reference to the history of an individual's development. Hurford's simulation model could have captured the inertia of an individual's life history by reinterpreting the combined genetic contribution at each life

stage as the change in language acquisition capacity rather than as the absolute value. This interpretation seems more plausible than the one he stipulated, and under such conditions, it is hard to see any reason why the capacity would decline once present, unless the decline was driven by natural selection, but Hurford (1991: 172) "can think of no plausible circumstances in which it would be advantageous to lose, or to have lost, the capacity."

Elman's (1993) computational model of language acquisition provides one possible explanation. In this model, a recurrent neural network was trained to predict the next word in a text containing nested structures and other complex properties. Initial attempts failed, but in subsequent attempts, he found that appropriate generalisations could be obtained by effectively limiting the memory of the network. He did this by resetting the units in the context layer at semi-regular intervals, very frequently at first and then less so as the learning proceeded. This "starting small" approach had the effect of limiting the dimensionality of the optimisation problem, making it more tractable. Networks that started with large 'working memory' capacities were unable to acquire the structures contained in the input to anything like the same standard. Elman relates these findings to the observation that short-term memory capacity matures in childhood and argues that it is these limitations that account for the critical period in language acquisition.[38]

In evaluating this proposal, it is important to note that language development is far from alone in exhibiting a sensitive period. Other systems such as those involved in vision and audition also need to receive sensory input to develop normally. Such phenomena have been studied in a wide range of species. Elman's explanation for the sensitive period in language acquisition does not readily generalise to these other cases. Memory limitations would not explain why sensitive periods exist in the development

[38] Baddeley (1999) argues that it may be the rate of rehearsal in the phonological store that matures rather than memory span per se, but this would have the same effect of windowing the input so that fewer words are visible to the language system.

of vision for instance (at least, not in any obvious way). In principle, it is possible that the sensitive period in language acquisition exists for reasons that are entirely unlike those that account for the sensitive periods in other developmental contexts, but some scepticism about this is probably justified.

Another problem for the "starting small" theory is that it leads to the prediction that bilingualism should only be possible if both languages are literally acquired at a matched pace, since under the theory, 'working memory' capacity would limit the complexity of sentences that can be comprehended in one language to the same extent that it would in the other. But this is simply not the case – an infant can be at the stage of comprehending and producing quite complex structures in one language while just beginning to acquire the basic structures of another. So long as both languages are acquired within the sensitive period, there will generally be no problem attaining fluency in both, regardless of whether or not they are acquired exactly in parallel. It is far from obvious how such facts could be reconciled with a view of language acquisition as being dependent on the increase of 'working memory' capacity or indeed any other variable that would affect cognition globally.

A proposal of the kind Elman presents would nevertheless resolve one apparent paradox. It would explain how the decline in sensitivity could arise as a by-product of selection. Again, this would be a case of an apparently maladaptive consequence (losing the language acquisition capacity) being offset by a counterbalancing benefit (gaining a more tractable learning regime for first language acquisition and better 'working memory' capacity in adulthood). Nevertheless, there are alternatives that could allow us to attribute a selective benefit to the decline directly.

If a system ceases to be of use beyond a certain age, it would make no evolutionary sense to continue to devote resources to its maintenance. It may even make sense to actively disable the system so that it cannot consume metabolic resources that

could be better spent elsewhere. Examples of organisms shedding juvenile characteristics are not hard to find. One example is the sea squirt, which in its juvenile state,

> wanders through the sea searching for a suitable rock or hunk of coral to cling to and make its home for life. For this task, it has a rudimentary nervous system. When it finds its spot and takes root, it doesn't need its brain anymore so it eats it! (Dennett, 1991: 177)

It is conceivable that the language acquisition device receives similar treatment after its period of usefulness expires, the cells that comprise it literally dying off so that their nutrients can be reabsorbed into the body.

Another explanation might be that sensitive periods are simply the timeframe during which certain kinds of finite neural resources are consumed. As an analogy, we could imagine an employee who is on call, being paid wages regardless of whether he or she is actually called into service. The employee will continue to be paid while there is funding available, but once it runs out, he or she can no longer be called upon to perform the service. There may be a neural resource analogous to this funding, which simply dries up over time even if the relevant developmental programme is never enacted. This resource may relate to what is referred to as 'plasticity', although a proper treatment of what this deceptively simple term may mean is beyond the scope of the present review. A variation on the same idea would be that there is a finite resource, which is competitively consumed by different developmental processes in the brain so that if a particular process is prevented from making use of the resource, others will eventually end up consuming its budget. The competition might be for control over cortical area, which can be carved up and rewired to suit various different sensory-

motor or higher-level functions. Sensitive periods would necessarily exist if such resources are finite, and more noticeably so for those developmental processes that have higher budgetary demands.

Evolutionary simulations of the kind employed by Hurford (1991) are inherently compatible with a methodology based on optimality considerations. Simulations model the traits that undergo optimisation and the selective functions by which their fitness is evaluated. Running them will produce entities possessing traits that are optimised for these functions and if these traits are observed in the real world, then the results of such simulations can be used to support claims that the selective functions embodied in them have relevance for the evolution of corresponding real-world phenomena.

4.1.6 Stages in the evolution of language

Bickerton (1990) draws parallels between the language of infants in the telegraphic stage of language development, pidgin speakers, adults who were deprived of language as children and language-trained apes. He equates these with a hypothesised stage in the evolution of language which he dubs *protolanguage* and argues that it provides the foundation on which fully syntactic language is built. Bickerton argues that the earlier appearance of this kind of language in ontogeny is evidence that the development of fully syntactic language rests on its foundation and must also therefore have come later in evolution. He further supports this by arguing that people fall back onto it under various social conditions such as when forced to communicate outside of one's native language as well as when people suffer brain damage that results in Broca's aphasia.

Bickerton argues that the transition from protolanguage to fully syntactic language must have occurred quite abruptly given the abruptness of the analogous transitions observed in child language development and creolisation. This would be

possible if, in the space of genetic variants, they are actually quite close to each other despite the dramatic difference in the phenotypes.

As discussed in chapter three, optimality considerations will allow us to assume that variants sit atop peaks in the fitness landscape, but will say nothing about the direction from which evolution has scaled these peaks. Nevertheless, they can allow us to make inferences about functional pressures that may have existed in the past if a trait is optimal for a function that it no longer serves. In such cases, they may be a useful source of evidence for reconstructing evolutionary history. If an argument along these lines could be developed, it may reinforce Bickerton's protolanguage theory. I will return to these issues in chapter five, which concerns a key aspect of what could be interpreted as the transition from protolanguage to full syntax.

4.2 The evolution of syntactic universals specifically

Many studies have also looked specifically at the evolution of grammar and its features. The evidence available to these inquiries is generally even more limited than many of the aforementioned studies. As Botha (2003) observes, many of them also lack a restrictive theoretical framework with which to assess the hypotheses they entertain.

4.2.1 Sequential motor control representations exapted for syntax

Lieberman (1985, 1991) argues that the neural mechanisms used to encode syntactic structures were exapted from mechanisms used in serial motor control, from the motor control of hand movements to the motor control of speech and then to syntax.[39] Evidence in support of this is that the areas of the brain most associated with grammar also appear to be implicated in the motor control of speech (see also Deacon, 1997). However, Bickerton (1998), Hurford (1999b) and Botha (2003) regard Lieberman's

[39] Kimura (1979) and Calvin (1983) have also made similar suggestions.

emphasis on sequential processing to be simplistic, since it fails to appreciate that syntactic representations are not merely representations of word order, but of nested hierarchical structure. Lieberman's account could be strengthened if it could be demonstrated that syntactic representations are optimally suited to the representation of sequences, or that they are as suited as one might expect if the mechanisms involved were previously selected for sequential processing and only slightly modified under selection for their new function. Indeed, a novel proposal of this kind is presented in chapter six in which I argue that the nested hierarchical structure of syntactic representations is a by-product of selection for the ability to perform certain kinds of operations on sequences.

4.2.2 Conceptual structure exapted for syntax

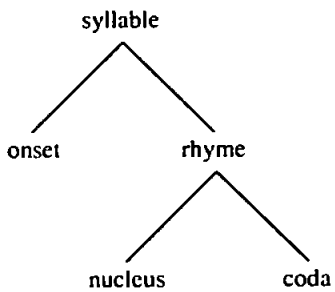
Wilkins and Wakefield (1995) argue that the brain utilises non-modality specific conceptual structures that are hierarchically organised and govern a wide variety of processes including those involved in motor control (see also Greenfield, 1991), arguing that it was these conceptual structures that were exapted for syntax. In a similar vein, Bickerton (1998) and Calvin & Bickerton (2000) argue that the particular aspects of conceptual structure involved in keeping track of the social calculus of who did what to whom provided the foundations of linguistically represented aspects of predicate-argument structure, namely thematic roles. These proposals are essentially speculative, relying mostly on analogy rather than evidence. As with Lieberman's claims, they could be strengthened if it could be shown that the traits in question are of something like the optimal form for serving their previous functions.

4.2.3 Syllable structure exapted for syntax

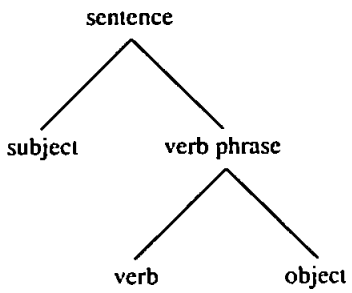
Carstairs-McCarthy (1999) identifies a number of similarities between the structure of syllables and the structure of simple sentences, which leads him to suggest that representations of the latter were co-opted from representations that previously evolved

for the former. One of the key parallels is that both have an asymmetrical structure, as illustrated in (1).

1. a.



b.



A syllable always has a nucleus, which is defined as a sonority peak, in most cases consisting of a vowel, but in some cases, a sonorous consonant. The nucleus may be preceded by one or more consonants comprising the onset. It may also be followed by one or more consonants comprising the coda. The nucleus and the coda together form a unit called the *rhyme* to the exclusion of the onset.

The structures illustrated in (1) were inferred on the basis of distinct evidence in the fields of phonology and syntax. The asymmetry in the structure of sentences was inferred from constituency tests of the sort reviewed in chapter two. In the case of syllables, the psychological reality of the asymmetry is universally reflected in the experience of alliteration (where syllables have identical onsets but different rhymes) and rhyming (where syllables have different onsets but identical rhymes). Similar effects are not obtained with pairs of syllables that possess patterns of similarity that are the mirror image of these combinations (i.e., where the syllables have identical codas but different onset-nucleus sections, or vice versa). The asymmetry is also suggested by the presence of spoonerism in speech, which are slips of the tongue such as "dat and cog" in place of "cat and dog". These involve exchanging the onsets of a pair of syllables, but the analogous situation involving codas does not seem to occur readily, if

at all (e.g., "cag and dot"). The implication is that the nucleus is more tightly bound to the coda than to the onset.

Carstairs-McCarthy (1999: 162) argues that "[i]f we reject the syllabic model for syntactic evolution, there is no obvious reason apart from historical accident why this subject-object asymmetry should exist". But there doesn't appear to be any obvious reason why the onset-coda asymmetry should exist either, so the question that would remain is of exactly the same order as the one this manoeuvre is designed to explain. Indeed, what is missing under his exaptationist account is any justification that the traits in question evolved under selection for phonology. It may be that the curious asymmetry of syllables is a vestige of representations that were themselves co-opted from yet another domain.

Aside from the asymmetry in (1), Carstairs-McCarthy (1999: 148ff) lists a number of other parallels between syllables and sentences. The central ones are summarised in (2) below.

2. a. Nuclei are obligatory in syllables and verbs are obligatory in sentences.

   b. The phonemes that can fill onsets and codas are drawn from a similar set (consonants), which does not include phonemes that generally fill the nucleus (vowels and sonorous consonants), and in the kind of sentences that Carstairs-McCarthy appears to have in mind, the elements appearing before and after the nucleus-like position are also drawn from a similar set (essentially noun phrases), while the nucleus-like position is filled by elements of a different kind (verbs).

   c. Onsets have a "privileged" status compared to codas that subjects also possess in sentences.

Carstairs-McCarthy (1999: 149) observes that in some languages, codas are not permitted on syllables, but that no such constraint is placed on onsets in any language. This, he says, gives onsets a privileged status in the syllable. As for the correspondence with sentences, he says "the onset-like position is ... privileged in that, unlike the coda-like position, it occurs in all languages, and most sentences contain it." The claim that there are languages that disallow coda-like positions (direct objects) in sentences is unreferenced, and I have been unable to find any evidence in support of it. A second reason he says onsets are privileged is because the phonemes that appear at troughs in sonority could in principle be analysed as part of the coda of one syllable or as part of the onset of the next, but are instead universally grouped with the onset. However, the "privilege" that onsets have over codas in determining syllable boundaries has no obvious analogue in sentences – at least, none that Carstairs-McCarthy makes explicit. The other aspect of privilege (quoted above) was that "most sentences contain [the onset-like position]". On many accounts, clausal subjects are indeed obligatory,[40] but the same is not true of onsets. If onsets and subjects are each "privileged" in some sense, they appear to be so for an essentially disjoint set of reasons.[41] Referring to both as "privileged" invites a comparison that does not appear to hold up under scrutiny.

As Carstairs-McCarthy notes, there are at least four ways to account for the similarities between syllables and sentences. First, the similarities may be coincidental, the properties in question having arisen independently. Second, syntactic representations may have influenced phonology. Third, phonological representations may have influenced syntax. And fourth, both syntactic and phonological representations may have co-opted representations that evolved in a third and as-yet unidentified domain.

[40] I present an alternative explanation for the obligatoriness of subjects in chapter five as one of the consequences derived from the theory presented there.

[41] This is at least true for the reasons Carstairs-McCarthy (1999) cites for assigning them this status (according to my reading of his work), but there may yet be homologous privileges that have so far escaped attention.

His case for eliminating the first option involves arguing that the parallels between syllables and sentences are too close to have arisen by chance and, since he does not see how the traits in question could be adaptive, he argues that convergent evolution could not explain them either. Of the remaining options, he argues for the priority of phonological representations on the basis that it meshes well with the broader theory of language evolution he presents, in which phonological developments play a central role.

Drawing on a broad range of evidence, Carstairs-McCarthy (1999) argues that a cascade of evolutionary events was triggered by the lowering of the larynx in the vocal tract. The newfound capacity to produce a larger range of speech sounds, coupled with a pre-existing disposition (discernable in language trained apes) to assign distinct meanings to distinct signs (synonymy avoidance) led, under this account, to a proliferation of new meanings being acquired.

There are immediate difficulties with this account. The number of distinct phonemes in a language is not a good indicator of vocabulary size since a modern language will possess only a very small set of them. The phonetic alphabet used in the *New Oxford English Dictionary* for instance, consists of only 23 vowels (including diphthongs and triphthongs) and 25 consonants, which is sufficient to encode the pronunciations of all of its 350,000 entries (with the exception of some words of foreign origin). This is a far greater number of lexemes than will typically be found in an English speaker's vocabulary, but average vocabulary size is still several orders of magnitude larger than the number of phonemes in any human language. This is because it is not so much the number of phonemes in a language, but the combinatorial possibilities that arise from being able to string them together into distinct words that is important. This ability to create meaningful signals by combining meaningless ones in different ways is called the *duality of patterning*. Having more speech sounds would

clearly increase the combinatorial possibilities, but the importance of the number of phonemes pales in comparison with the effect of the duality of patterning.

If an increase in vocabulary size conferred a selective advantage, then lifting a constraint on its enlargement would allow natural selection to take its course. If we assume for the sake of argument that the relevant constraint was indeed the limitation on the number of distinct speech sounds that could be produced, then the lowering of the larynx could have been the enabling condition. However, the enormous vocabularies possessed by modern humans are much larger than we appear to need. In this respect, we can sympathise with Premack's (1986: 133) conclusion that "[h]uman language is an embarrassment for evolutionary theory because it is vastly more powerful than one can account for in terms of selective fitness." Carstairs-McCarthy (1999: 132) does not attempt to provide an adaptive explanation for vocabulary expansion, instead arguing that synonymy avoidance drove it once the restriction imposed by vocal tract anatomy was lifted:

> In the partnership between meaning and sound, sound may sometimes take the lead, so as to stimulate the discovery of new meanings. I am not suggesting that, as language evolved, extralinguistic meanings were accumulated arbitrarily. We need to distinguish between the pressure for vocabulary expansion to take place and the directions in which this expansion might proceed. There is plenty of scope in this scenario for cognitive, social, cultural and technological factors to exert an influence ... All I am suggesting is that, thanks to inherited synonymy-avoidance principles, the capacity for a much larger repertoire of distinct vocalizations introduced a new kind of pressure for this expansion to take place.

Carstairs-McCarthy does not explain why synonymy avoidance would require the spoken modality to drive vocabulary expansion rather than language expressed in another modality such as hand gestures. Indeed, the evidence cited by Carstairs-McCarthy that led him to conclude that non-human primates also use synonymy-avoidance principles is actually based on non-verbal communication in chimpanzees.

The main source of evidence for his exaptationist claims is structural parallels, but as Botha (2003) observes, Carstairs-McCarthy does not present his claims within a restrictive theory of exaptation that clearly specifies the conditions under which we should attribute the status of exaptation to a trait. The question of whether structural parallels are sufficient for this purpose and how close they would have to be for something to warrant the label 'exaptation' are not adequately addressed.

By contrast, an optimality approach would test exaptation claims in a very specific way, namely by testing whether the properties in question are optimal for the functions they are hypothesized to have served in the past rather than those they serve in the present.

A key feature of Carstairs-McCarthy's approach is to enumerate various conceivable language systems to illustrate how things could have been different had evolution taken a different course. In this respect, his work resembles an optimality approach since enumerations of this sort are essentially what is required when reconstructing a plausible *strategy set*, which the reader will recall from chapter three is the set of imagined variants that are in the neighbourhood of the observed type on the fitness landscape. However, Carstairs-McCarthy uses these enumerations to different ends, namely to illustrate that certain properties of language, which seem to be intrinsically important are not in fact necessary for serving certain functions.

Although Carstairs-McCarthy's (1999) work tends to raise more questions than it answers, the questions it raises are nevertheless interesting.

4.2.4 Compositionality

Kirby (2001) uses computational models to explore the effect of language learning under conditions that impose strict limits on the amount of linguistic data to which simulated agents are exposed. The model involves the transmission of linguistic knowledge from simulated parents to simulated children and demonstrates that if data available for reconstructing the language of the teacher are limited, languages become optimised over many generations for transmission fidelity, with the best language systems being those that can be most reliably reconstructed under the constraints imposed by the "learning bottleneck". The optimal languages are compositional, meaning that their expressions have discernable parts that can be syntactically combined in different ways to generate expressions.[42] In a compositional language, the number of symbols that can be combined to make expressions is much smaller than the number of possible expressions, but in a non-compositional language, the number of symbols required is as many as there are possible expressions. A compositional language is therefore more likely to be transmitted faithfully from generation to generation than a non-compositional one because it reduces the burden on the language learner.

To consider a real example, consider the large set of distinct terms used to identify baby animals, a selection of which are summarised in (3). The problem with having a distinct vocabulary item for each type of these becomes particularly apparent for the animals we rarely discuss such as *leverets* (baby hares) and *elvers* (baby eels). A compositional way of referring to the young of a species S would be to apply the template "baby S" and this is indeed the solution that many of us will fall back on when unable to summon one of the more obscure terms.[43]

[42] Similar results have been obtained in a computational model by Batali (2000) and a mathematical model by Nowak, Komarova and Niyogi (2001).

[43] The same situation arises for collective terms such as *herd*, *flock*, *shoal*, *colony*, *gaggle*, *pride* and so on, that could otherwise be referred to using the compositional template "group of S".

3. **Terms for baby animals and their compositional equivalents**

ANIMAL	*ACTUAL TERM*	*COMPOSITIONAL TERM*
cat	kitten	baby cat
dog	puppy	baby dog
cow	calf	baby cow
sheep	lamb	baby sheep
goat	kid	baby goat
kangaroo	joey	baby kangaroo
duck	duckling	baby duck
goose	gosling	baby goose
swan	cygnet	baby swan
hare	leveret	baby hare
eel	elver	baby eel

The learning bottleneck makes rare terms particularly difficult to learn. However, possessing a compositional language comes at the cost of having to produce longer expressions. Incorporating this cost into his model, Kirby (2001) found that short non-compositional expressions emerged for the meanings that were most frequently used, while less frequent meanings were expressed compositionally. This accords well with the observation that irregularity and frequency of use are indeed highly correlated in natural languages.

Like de Boer's (1997) model discussed in section 4.1.2, Kirby's (2001) model is intended to show that natural selection in the genetic domain is unnecessary to explain the optimality of properties of language (in this case, of compositional syntax). But, as with de Boer's model, it is easy to imagine the model's language learners being different in some way that would make it impossible for them to acquire compositional

languages. In the model, they are equipped with an induction algorithm that associates meanings with expressions. Part of this algorithm involves a procedure to make generalisations over expression-forming rules, which has the effect of replacing non-compositional rules with compositional ones. Clearly, if the agents lacked this aspect of the algorithm, compositional languages would fail to emerge. Therefore, if there is a genetic difference such that variants possessing a certain allele had the generalisation algorithm and variants without it didn't, then we would be justified in referring to this allele as a gene for compositional syntax and one which may have been genetically selected for this capacity. The observation that a compositional language only emerges after a period of cultural evolution would not make it any less a phenotypic effect of this gene.

That compositional languages appear to be optimal for transmission fidelity provides strong support that they have been selected for this function in cultural evolution. This follows from the logic of the optimality diagnostic, which is just as applicable in the cultural domain as it is in the genetic domain. However, we cannot conclude that since it is optimised under cultural selection that it is not also optimal under genetic selection.

4.3 Summary

The purpose of this chapter was to review the methods used in studies of the evolution of language and position the concerns of the current thesis within the context of other kinds of inquiries in this area, such as those concerned with dating the origin of language, the evolution of the performance systems, the social functions of language, the evolution of the linguistic brain, and the evolution of the critical period for language acquisition. It also examined a number of studies concerned with the evolution of syntactic properties specifically.

From a methodological point of view, the studies differed both with respect to the types of evolutionary explanations proposed and the types of evidence that were utilised. The studies included claims about exaptation, concomitancy and cultural evolution as well as the selective functions of the traits they examined. The evidence used to support these claims drew on sources such as the fossil record, comparisons between related species, neuroanatomical studies, observations of the modern uses of traits, and in some cases optimality considerations, typically in the studies involving computational and mathematical modelling.

None of the studies reviewed discuss optimality as a diagnostic of selective function directly, but questions of design optimality are nevertheless lurking in many of them, even if not actually surfacing explicitly. For instance, in dating the emergence of spoken language, optimality considerations are central to Lieberman's reasoning. In this case, it was the *sub*-optimality of the design from the point of view of the associated choking hazard that was used to argue in favour of a counterbalancing benefit.

Claims about exaptation can also be couched in terms of optimality, considering that a claim about any given exaptation event can be decomposed into two distinct claims, as in (3).

3. a. Trait T was originally selected for function F.

   b. Trait T currently serves function F'.

Claims of the type in (3b) can be readily verified from present-day observations while claims of the form (3a) are indistinguishable from those about the selective function of an adaptation and hence, as with other claims about selective functions, lead to predictions about optimality. The situation would initially appear more complex if what is being claimed is that an exaptation event was followed by some adaptive

modifications that enhanced the new function, but whatever differences exist between the original trait and the modified variant cannot have any bearing on what was exapted, so exaptation claims needn't make reference to them.

Optimality considerations also arise in connection with learning, development and cultural evolution, the final states of each of these processes being obtained through optimisations. The optimality diagnostic can also be used to test which functions these optimising processes are optimising, but as discussed in connection with both de Boer's (1997) work on the evolution of vowel systems, and Kirby's (2001) work on the evolution of compositional syntax, evidence of optimality in these domains does not mean that an optimisation in the genetic domain has not also occurred.

Computational models are a very useful tool for assessing optimality claims allowing researchers to explore the complex effects of model parameters on evolutionary processes. For a more thorough review of computational methods used in language evolution studies, see Turner (2002).

This chapter also touched on some of the empirical concerns of the remainder of this thesis. The issue of large vocabulary sizes arose in reviewing Carstairs-McCarthy (1999) and the pressures that a large vocabulary places on neural resources plays a central role in chapter five. In chapter five, a theory is developed about the role of closed-class items which may also motivate aspects of what, in Bickerton's (1990) terms, would be the transition from protolanguage to full syntax. And in chapter six, the representation of sequences (discussed here in connection with Lieberman (1985, 1991)) has a central role to play in explaining the emergence of nested phrase structure.

5

Closed-class items and the lexicon

There are many thousands of distinct words in the vocabulary of a typical adult. We should therefore expect the representation of the lexicon to place a significant demand on metabolic and other kinds of neural resources that could otherwise be utilised for other purposes. This cost was presumably offset for our ancestors by some benefit associated with having a large vocabulary, but regardless of what that benefit was (a question that will not be addressed here), we should expect there to be a strong selection pressure to make lexical representations as efficient as possible. The pressure for streamlined lexical entries may introduce costs associated with the loss of redundancy, which as Chomsky (1995: 29) notes, may help "to compensate for injury and defect". In the absence of such costs, we should expect the optimal lexicon to be essentially just a repository of exceptions, something we should also expect on Minimalist assumptions (Chomsky, 1995). This chapter examines how the existence of closed-class vocabulary items could have improved the efficiency of lexical representations and applies the optimality diagnostic outlined in chapter three to assess whether closed classes were indeed selected for this role.

5.1 Closed-class items as an adaptation

In chapter three, I argued that plausible candidates for the selective function of a given trait would be those for which it exhibits improbable local optimality along available dimensions of variation. In the present context, the trait in question is closed-class items and hence the only dimensions of variation that need be examined are those affecting

their existence. If the currently instantiated characteristic is improbably optimal in terms of a given candidate function that varies over these dimensions, then we would not be able to rule it out as having a role in selection.

I will apply the optimality diagnostic of chapter three to two candidate functions. The first function follows predictably from the discussion up to this point and relates the strategy set to the level of redundancy in representations of syntactic features, with the optimal strategies being those in which redundancy is minimised. I will then turn to a second function which concerns the level of redundancy in representations of meaning, the optimal strategies being those in which no two lexical entries share the same meaning.

5.1.1 Redundancy in representations of syntactic distributions

Recall from chapter two (§2.3) that word classes are most clearly differentiated on the basis of their syntactic distributions rather than semantic criteria. In this sense, a set of words can be judged to form a grammatical class if one member can be substituted for any other in sentences without altering their grammaticality status. In these terms, the question of how a lexeme's category is encoded in its lexical entry reduces to one of how its syntactic distribution is encoded.

In principle, this information could be encoded in a variety of different ways. It may be that each lexeme encodes its own syntactic distribution independently of every other and that it is only because some encode similar or identical distributions that we take them to form a class. Alternatively, each member of a class might instead derive its membership via a reference to a single instance of the information that describes the syntactic distribution of the class as a whole.[44] Radford (1988: 63) argues for the latter view on learnability grounds:

[44] The difference between these two possibilities is not without consequences. Under the latter view, it ought to be possible to construct a taxonomy of categories such that they are neatly nested inside one another rather than freely intersecting, the latter being a possibility if categories simply derive from

> [I]f every word had its own utterly idiosyncratic set of linguistic properties, then the task of acquiring competence in a language would be an impossible one (within the constraints that the child operates under). By contrast, if words are grouped into a small finite set of categories, and if phonological, morphological, syntactic and semantic rules are all category-based, then the child's acquisition task is enormously simplified.

This is because the language learner who assumes that words belong to categories can generalise syntactic knowledge acquired about one word to every other member of the class to which it belongs. This means that the full syntactic behaviour of a class of words can be determined without having to observe every member of that class in every allowable context.

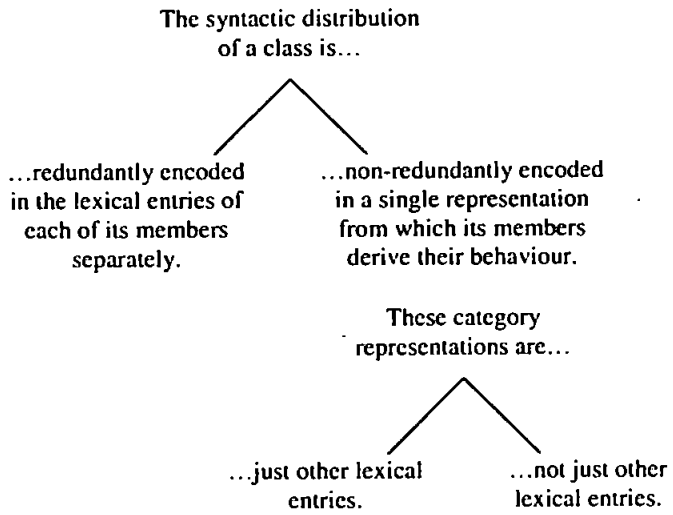
The category-based view would also be preferred on Minimalist grounds since, by offloading the distributional information repeated in every lexical entry to a separate category representation, the system would eliminate redundancy. At first glance, this theoretical manoeuvre would appear to require positing an additional and qualitatively different kind of representation that pairs a category label with associated information about its syntactic distribution in a kind of lookup table, but this is not necessary. All that is required is that the distributional information be offloaded to a separate representation, but representations of the required type are already among the theoretical apparatus available. The separate representation could simply be a further lexical entry that has an association with lexical entries belonging to the class in question. For instance, the distributional information for the class of nouns could be offloaded to the lexical entry for a closed-class item that heads a functional projection encapsulating

---

feature specifications. It is not clear that this is necessarily the case, but it is not possible to explore the details of this prediction here.

noun phrases, the distribution of noun phrases thereby being determined indirectly by the distribution of the encapsulating projection. The diagram in (1) summarises the various theoretical positions described.

1.



The relative parsimony of the lexical entry theory of category representations compared with its alternative is also evident in the way in which each theory must deal with subdivisions within a grammatical class between sets of words that have overlapping but non-identical syntactic distributions. There are, for instance, subdivisions within the class of nouns such as between those that are countable and uncountable and within the class of verbs between those that are transitive and intransitive, as well as many other such distinctions. Verbs in particular have many subtypes. If each subclass has a separate entry in something akin to a lookup table for categories (as required by a non-lexical theory), then some of the information contained in these entries would be redundantly repeated. This would be the information that gives each subclass its characteristically verb-like properties such as the ability to take tense and agreement morphology – properties that hold true for every type of verb. This redundancy could of course be offloaded to a further representation in a further lookup

table for categories of categories and so on ad infinitum, but the conceptual problems with this solution should be fairly evident. By contrast, under the lexical entry theory, this redundancy could be avoided simply by positing a lexical entry for the head of a further encapsulating projection that groups subclasses within a broader class. This is consistent with the view – adopted within the Minimalist tradition and theories ancestral to it (e.g., Abney, 1987; Grimshaw, 2005; Larson, 1988) – that there are layers of functional projections stacked inside one another.[45] This structure has been postulated for reasons that are independent of the present concerns about the evolutionary role of these encapsulating projections, but the fact that this structure could be used to minimise redundancy in representations of distributional information suggests an answer to the evolutionary question that I will now elaborate in detail.

Selection against redundancy in representations of syntactic distributions is plausibly relevant for explaining the existence of closed-class items. By encapsulating lexical items, projections headed by closed-class items mediate grammatical relations so that fewer syntactic features need to be represented for each open-class entry in the lexicon. For instance, learning that a noun is associated with determiners could allow a noun, encapsulated within a determiner phrase, to be used wherever a determiner phrase can be used, including the subject or object of a sentence and so forth without having to learn and encode this distributional information separately for each noun, the information instead being encoded in the lexical entries of the small number of determiners.

The introduction of entries for closed-class items into the lexicon must introduce a certain amount of cost, but so long as the number of closed-class items is small, their cost would presumably be more than offset by the reduction in the amount of

[45] For some detailed proposals within a Minimalist framework, see Rizzi (1997).

information that needs to be encoded in the entries for the very many more open-class items.

For closed-class items to be of any use in this way, it would have to be the case that there is a positive correlation between the amount of information a lexical entry contains and the number of distinct syntactic environments it can appear in (i.e., the items it can be in construction with and checked against).[46] The simplest hypothesis that could be made along these lines is that this syntactic behaviour is determined by a set of formal features contained in a lexical entry, with the inclusion of each formal feature contributing to its representational cost.[47]

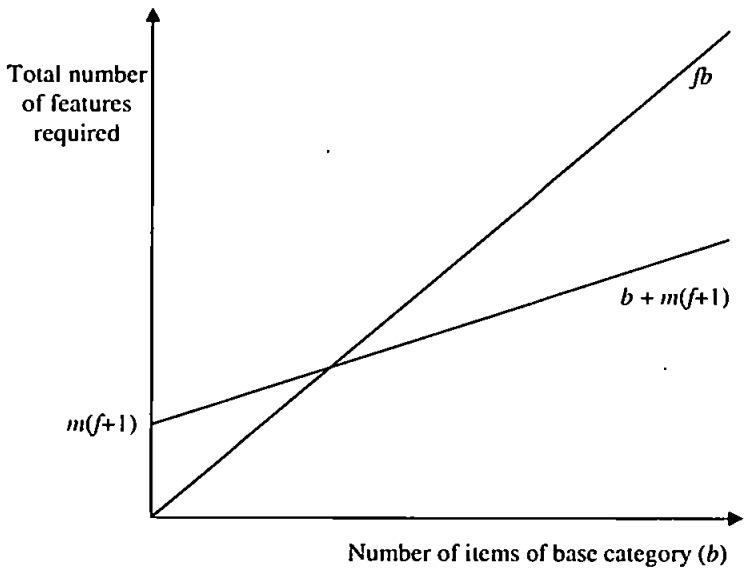
With these points explicit, it is possible to specify precisely the conditions under which the existence of a certain number of closed-class mediating items in the lexicon would be an advantage. Take f to be the number of formal features needed to represent the distribution of a lexical item of a given category (let's call this the *base* category). If there are b members in the base category and f features are needed to encode their distribution, then the total cost of representing this information separately for every member would be the product fb. If all of the features of the base items are instead redistributed to a set of m items of a closed-class mediating category (where m is constant), then each of the mediating items will contain those f features and the base items will contain none. However, given that the association between mediating and base items also has to be encoded in lexical entries, I will assume an additional feature contained in both mediating and base items is required to achieve this. Hence, each mediating item will possess $f + 1$ features and each item in the base category will

[46] A negative correlation could conceivably result from having lexical items list all of the environments they *cannot* appear in.

[47] The term *formal feature* is used here in approximately the sense that it is used within Minimalist syntax and various other syntactic formalisms – a sense that is not entirely compatible with usage in other contexts. Within Minimalist syntax, features can themselves have secondary properties (e.g., checked or unchecked, strong or weak) and they can move independently of the things that they are features of. This latter quality is particularly unlike the notion of a 'feature' as applied to ordinary objects. There are no known processes that could for instance, move the colour of a shirt without the shirt moving, or the shape of a ball without the ball, or the size of a house without the house, etc.

possess just one, the one that associates it with items of the mediating category (the equivalent of a reference to a lookup table). In this case, the number of features required to represent all members of the base and mediating categories combined would be $b + m(f + 1)$. For large values of b, it will always be the case that $b + m(f + 1)$ is less than fb as illustrated in (2). Hence, a lexicon that has mediating items will be more economical than one lacking them for all but the lowest values of b.

2.



To illustrate, suppose there are 1000 nouns to represent and that only two formal features are required to represent all of the syntactic environments nouns can appear in. If this information is stored separately for each noun, then a total of 2000 features would need to be represented. If, on the other hand, these two formal features were not contained in the lexical entries of every noun, but were contained within the lexical entry for a single mediating item that encapsulates nouns, then although a small amount of complexity will be introduced to have features that link nouns and the mediating item (one for each noun and one for the mediating element), the total number of features that

would be required will be only 1003 (each of the thousand nouns having one feature and the mediating item having three). This would clearly be an enormous saving.[48]

I will provide a more comprehensive discussion about why there is typically more than one mediating item for each open class as I refine the proposal in the following section, but it is worth pointing out at this point that having a small number of them does not lead to a dramatic departure from optimality. In the above example for instance, if there are ten items in the mediating category then they would each have to possess three features bringing the total number to 1030, which is still much closer to the optimum than the 2000 that would be required if the same features had to be represented for each noun.

5.1.2 Redundancy in representations of meaning

In the space of conceivable variants in the neighbourhood of the lexicon that has actually evolved, there are various imaginable lexicons that lack the closed-class items that we actually observe in language. One variety, already discussed, is those in which lexical entries of the same class all redundantly encode the same distributional information independently. Let's call this type the *Redundant Distribution Lexicon* and contrast it with the *Observed Lexicon*. Another conceivable alternative to the Observed Lexicon is one in which redundancy in distributional information is not offloaded to closed-class items, but which still achieves minimal redundancy in representations of syntactic distributions at the cost of disallowing items to be used in more than a single syntactic position. For a language with a lexicon like this to be as expressive as one with the Observed Lexicon, unique lexical entries would be required for each of the syntactic positions in which a given meaning is to be expressed. For instance, if the meaning that a noun expresses needs to be used in both subject and object positions, then two distinct

[48] The reasoning applied in the above paragraphs is similar to that applied to 'normalize' relational databases within computer science. This process eliminates redundancy in an analogous way and thereby reduces the amount of memory that databases consume as well as making it more manageable to update the information they contain.

nouns with the same meaning would be required – one for use in each position.[49] This method would simply shift the problem of redundancy from representations of syntactic distributions to representations of meaning, hence we could call it the *Redundant Meaning Lexicon*. Both kinds of redundancy would be more costly compared to an equivalent lexicon in which distributional information is non-redundantly encoded in closed-class items, as I will argue is the case with the Observed Lexicon.

If a closed-class item that encapsulates noun phrases has to encode the fact that it can appear in both subject and object positions, then there would have to be some way of representing the equivalent of an exclusive-or operator in its lexical entry. An unstructured set of formal features is not sufficient to represent this or any kind of optionality so the equivalent of logical formulae along the lines of (3) would have to be used. In (3), it is simplistically assumed, for the sake of illustration, that the determiner *the* encapsulates noun phrases (checking the N feature) allowing them to appear in either the subject or object position (checking either the S or O feature).

3. *the*: (S *XOR* O) *AND* N

There are various conceivable variations on how to represent a formula like this. Two distinct logical operators needn't be employed since *AND* and *XOR* can both be expressed in terms of the *NOR* operator (or alternatively in terms of the *NAND* operator), or in the streamlined notation of boundary logic (Meguire, 2003) as illustrated in the following table.[50]

[49] This option appears to be realised for pronouns in English giving us pairs such as *I* versus *me*, *we* versus *us*, and so on, but the morphological form of these words is not conclusive evidence that they have separate lexical entries. If they are listed as separate lexical entries, then this raises the question of why they are exceptions to the rule that you can't have two words with exactly the same meaning. If they are derived from the same lexical entry, then we have to explain how it is that they can be realised in such different forms.

[50] In boundary logic, concatenation is the equivalent of the OR operator so that the formula "*p q*" means "*p* OR *q*". Brackets negate the truth value of their contents so that "(*p q*)" means "NOT (*p* OR *q*)".

4. Some possible ways of representing syntactic distributions as feature formulae

	In terms of NOR	*Boundary Logic*
p AND q	(p NOR p) NOR (q NOR q)	((p) (q))
p XOR q	((p NOR (q NOR q)) NOR ((p NOR p) NOR q)) NOR ((p NOR (q NOR q)) NOR ((p NOR p) NOR q))	(p (q)) ((p) q)

It should not escape attention that formulae like those in (3) and (4) have a kind of syntax themselves, relying on hierarchical representations, the linear order of symbols, or both. Hence, these representations are essentially just as complex as the grammatical properties that we are trying to explain. Rewrite rules as well as lexical representations of the kind used in various other descriptive frameworks such as Lexical Functional Grammar (Bresnan, 2001), Head-Driven Phrase Structure Grammar (Pollard & Sag, 1994) and Link Grammar (Sleator & Temperley, 1991) suffer the same shortcomings, as does the Minimalist notation of Stabler (1997). Using such representations leads to an infinite regress for theories that seek to explain properties like hierarchy and linear order in terms of the content of lexical entries.

If we pursue the possibility that featural information in lexical entries is encoded more simply, consisting of just an unordered set of features, then mutual exclusivity – and indeed optionality in general – could not be encoded. With a Redundant Meaning Lexicon, this problem would be overcome by having several distinct lexical entries with the same meaning so that one could be used in each syntactic position that it needed to be used in. For instance, there would be distinct words for 'dog' when used as a subject

Boundary logic is a more typographically convenient notational variant of the *laws of form* algebra devised by Spencer-Brown (1969).

versus an object, and so on. However, a lexicon with an open/closed class distinction could confine this kind of duplication to the relatively small number of closed-class items. For instance, the apparent ability for all nouns to appear in both subject and object positions would be achieved by having one closed-class item performing the mediating role in the subject position and another performing the role in the object position. The information contained in the formula in (4) could be encoded in terms of the unordered sets of features of two separate lexical entries for the word *the* as in (5), one having a subject feature and the other an object feature.

5. the_1: {S, N}
   the_2: {O, N}

In many languages, articles have a different case form depending on whether they appear in the subject or object position. The German equivalent of (5) is illustrated in (6) with the masculine definite article.

6. *der*: {S, N}
   den: {O, N}

Hence, one of the consequences of stipulating that no optionality be encoded in lexical entries is that we predict the existence of something consistent with case-marking. This appears to be an extremely promising result.

Under this view, the existence of closed-class items in the lexicon allows the open class items to be free of redundancy in representations of meaning while still permitting a given meaning to be expressed in more than one syntactic position. Hence the introduction of closed-class items eliminates the redundancy that would characterise

a Redundant Meaning Lexicon. As discussed in the previous section, the introduction of closed-class items also represents an optimal solution to the problem of redundancy in representations of distributional information, hence also avoids the problem of a Redundant Distribution Lexicon. The existence of closed-class items optimises two distinct functions, but these are not necessarily competing explanations for their existence. Demonstrating that a property is improbably optimal for a given function does not mean it was necessarily selected for that function alone. The following sections elaborate on the optimality questions in more detail.

5.2 Applying the optimality diagnostic

In chapter three, I introduced a diagnostic test, adapted from Parker and Maynard Smith (1990), for identifying adaptive functions in terms of optimality. This involved (a) identifying the strategy set, (b) identifying a function over the strategy set for which the trait is optimal among the possible choices (where the optimality is of the 'improbable' variety), and (c) relating this function to fitness. I will now apply these considerations to refine and evaluate the candidate functions outlined above.

If it turns out that the trait that has actually evolved is not the optimal possibility in the strategy set for a proposed function, it could be for several reasons. It may be that the function wasn't actually relevant for its selection, that the function is not an independent component of fitness (because the trait cannot be altered without compromising other functions that it has or that concomitant traits have), that none of the values that are more optimal than the observed trait value should have been included in the strategy set because they are not actually physically attainable, or that natural selection has not yet had sufficient time to ascend to the optimal value. If it turns out, on the other hand, that the observed trait is the optimal possibility in the strategy set for the proposed function, then we have grounds to reject such alternatives.

5.2.1 The strategy set

Up to this point, I have left imprecise the question of what kind of variation would physically account for the presence or absence of closed-class items in language. If the use of closed-class items was selected for in our ancestors, then the innovation may not have been the capacity to represent them as such, but rather the tendency to analyse input during language acquisition as though it contains them (and perhaps a particular set of them). This predisposition is most dramatically evident in studies of the rapid transition from pidgin languages, which lack such items and which develop among communities of adults lacking a common language, to creole languages, which possess them and which are spoken by the children and subsequent generations who were exposed to the original pidgin (Bickerton, 1977; Kegl, *et al.*, 1999). The implication is that during language acquisition, modern humans analyse sentences as containing closed-class items even when these items are not explicit in the input.

We know that certain variations are at least physically attainable since the capacity to acquire closed-class items without explicit teaching varies throughout the lifetime of an individual, being greatly enhanced during infancy but diminishing by adolescence and mostly disappearing by adulthood. This justifies including, in the strategy set, weak through to strong predispositions for infants to analyse input as containing closed-class items. That closed-class vocabulary items appear later than open-class items in ontogenesis is also some evidence that they appeared later in phylogenesis. During language development, infants go through a stage of producing what Brown (1973) called telegraphic speech, in which closed-class elements are omitted despite the fact that they occur with very high frequency in the infant's linguistic environment. Brown and his colleagues even found a tendency for infants to omit them in tasks requiring immediate repetition. In a typical example, one of Brown and Fraser's (1963) subjects (aged 28½ months) was prompted to repeat the sentence *I*

am drawing a dog and responded with *I draw dog*. Examples like this suggest that infants in the telegraphic stage are sensitive to closed-class items, but have not yet mastered them. As discussed in chapter four (§4.1.6), Bickerton (1990) draws parallels between the language of infants in the telegraphic stage, pidgin speakers, adults who were deprived of language as children and language-trained apes. He equates these with a hypothesised stage of protolanguage and argues that language in this mode is distinct from the capacity for fully syntactic language observing that only the latter has a critical period as attested in adult language learning by the relative difficulty with which the closed-class items of a second language can be acquired compared with the open-class items. The same consideration would also account for why closed-class categories are indeed closed in one's native language. Despite appearing more frequently in the linguistic environment, closed-class items are nevertheless attained later than the earliest open-class items in language acquisition. This could be because their use is dependent on the foundation laid by protolanguage or perhaps because feature representations for closed-class items are more complex than those for open-class items.

I will take the variation in the strategy set to encompass only the learning dispositions. Variation in representational properties would only be relevant if lexical entries of closed-class items are qualitatively unlike those of open-class items. Observational evidence doesn't appear to require making such a claim so it would be gratuitous to make it. In applying the optimality diagnostic, I will assume that the syntactic distribution of a lexical entry is encoded as a set of formal features along the lines discussed in section 5.1.2 such that there is no optionality and such that it must be possible to check all of its features in the position where it is used in a sentence.

5.2.2 Optimality

Redundancy is presumably costly except to the extent that it helps "to compensate against injury and defect" as Chomsky (1995: 29) notes. A completely non-redundant

solution would be to have only a single closed-class item for each syntactic context that members of a class can appear in. This would maximise the ratio of vocabulary size to the number of features that need to be represented for all but very low values for vocabulary size. Having more than one closed-class item for each syntactic context would cause a departure from optimality, though only amounting to a small perturbation away from the optimum (as discussed in §5.1.1). By contrast, an absence of closed classes would have catastrophic consequences for redundancy.

Given that the strategy set contains only two kinds of solutions – one in which closed classes are acquired by the language learner and one in which they are not – we could be forgiven for rushing to the conclusion that the optimum does not exhibit the necessary improbability to meet the demands of the optimality diagnostic. If there were many more solutions to choose from, then we might be more confident that it was not by chance that a solution arose that is optimal for this function. But it would be a mistake to conclude that the two categories are equally probable. It depends on the proportion of language acquisition devices in the space of all language acquisition devices possessing the particular algorithmic properties that enable them to acquire closed-class items. I would think there are far fewer ways of structuring a language acquisition device that acquires closed classes than there are of structuring a language acquisition device that fails to do so. There are also many conceivable ways in which closed classes could be acquired without conferring any benefit in terms of redundancy minimisation. The theory predicts a number of very specific predictions along these lines which correspond with observations beautifully. These are examined in the following section.

5.3 Some consequences of the theory

5.3.1 Language variation and parameter-setting

Within Minimalist syntax, Chomsky (1995) adopts the view that syntactic parameter-setting amounts to determining the formal features of closed-class items (see §2.3 of chapter two). By contrast, the learning of open-class items appears to have little relevance for grammar as attested by the ability to recognise the well-formedness of 'Jabberwocky' sentences in which none of the open-class items are real words (e.g., *'Twas brillig and the slithy toves did gyre and gimble in the wabe*). Furthermore, if all parametric variation is associated with the lexicon, then this explains why the infant does not set contradictory parameters when exposed to bilingual input.

This state of affairs is exactly what we should predict if closed classes evolved to minimise lexical redundancy. It follows trivially that closed classes will be the locus of parametric variation simply because they take most of the featural complexity away from the open classes. Hence, considerations of representational optimality are sufficient to account for it without any additional theoretical motivations.

The relative complexity of open and closed classes would also explain why the latter are more difficult to learn, but would not help to explain why the ability to acquire them degrades after the critical period.

5.3.2 Case theory

As discussed in section 5.1.2, a consequence of eliminating optionality in individual lexical entries is that the appearance of optionality has to be achieved by selecting between different lexical items that allow different structural options to be realised. Hence, we could have a number of different nouns in the lexicon with the same meaning, one for each syntactic context in which that meaning needs to be expressed (subject, object and so on). As discussed earlier, this would lead to massive redundancy

in semantic representations, which could be eliminated if the lexicon instead made use of a small number of closed-class items to mediate nouns. Distinct closed-class items would be needed in each context in which nouns appear. Hence, considerations of representational optimality also appear to derive the property of grammatical case, which, given the tenuous link between structural cases and semantic/pragmatic considerations, has traditionally been extremely difficult to reconcile with functional considerations.

5.3.3 Agreement

For a pair of open classes to be brought into a systematic relation with one another, the interface between them will have to reach a convergence point where they are connected via a single feature type that unifies all of the subcategories of both of the categories involved. The noun contained in the subject of a sentence for instance is associated with the verb regardless of whether the noun is singular or plural, or countable or uncountable, and whether the verb is transitive or intransitive and so on. However, the convergence point needn't be strictly at a phrasal boundary. The elements on either side of the convergence point would only be guaranteed to be constituents if the convergence point appears between a specifier and the intermediate projection to which it is attached. But if the convergence point appears between the head and the complement of the phrase, the element on one side will not be a constituent, consisting of the specifier and head of the phrase to the exclusion of the complement. If an inflected verb like *eats* originates as a structure that has a functional projection headed by a closed-class item associated with the third person singular agreement feature and a complement phrase headed by the verb *eat*, then a convergence point would occur between the head and its complement. This is illustrated in the English examples below. In these examples, any of the forms that appear on one side of the point indicated with

>< can freely co-occur with any of the forms on the other side, which is what makes it a convergence point.

7. a. Those men eat [(Those men) 3pl >< eat]

   b. That man eats [(That man) 3s >< eat]

8. a. Those men have eaten [(Those men) 3pl >< have eaten]

   b. That man has eaten [(That man) 3s >< have eaten]

9. a. Those men are eating [(Those men) 3pl >< be eating]

   b. That man is eating [(That man) 3s >< be eating]

In terms of the optimality functions discussed in the current chapter, it makes no difference whether convergence points occur at the boundary between specifiers and heads or at the boundary between heads and complements. Under this analysis, agreement occurs not because it is functional, but simply because it can, and for predominantly right branching structures, there are many more opportunities for convergence to occur between head and complement than between specifier and head, so in such structures agreement is not only possible, but likely. This should be true for the interface between any pair of meaningful elements and we also readily observe it in the concord between adjectives and nouns in many languages.

Note that there are other convergence points in the examples in (7-9) between the number morphology on the noun and the noun stem and between the participle morphology of the verb and the verb stem. The point to note about this is that there isn't always an open-class item enclosed between convergence points. This is unexpected if the role of closed-class items is purely to mediate open-class items. An additional motivation must therefore exist to permit this. This may be to modulate frequently-used meanings. The domain that appears between the convergence point at the noun and the

convergence at the verb is concerned with encoding the number feature of the noun and the tense feature of the verb, both of which have meanings. Likewise for the participial forms which encode meaning about the temporal structure of the event or state indicated by the verb. We should not expect to find instances of domains that are entirely composed of closed-class items that do not modulate meanings.

5.3.4 The timing of the acquisition of open and closed classes

As mentioned earlier, there is a difference in the age at which open and closed classes enter the productive vocabulary of infants. Despite being the most frequently occurring items in the linguistic input provided by adults, closed classes are nevertheless not the first items to be acquired, infants first proceeding through a stage of 'telegraphic speech' (as discussed in §2.2.1). However, if we recall the graph in (2), this is exactly what we should expect, since the introduction of closed-class items will only confer an advantage once the number of open-class items exceeds a certain threshold from which point the advantage they confer grows steadily more and more significant as vocabulary size increases. This is a rather unexpected consequence of the optimisation of the lexicon and a very encouraging finding.

A precise quantification of the vocabulary size at which we should expect this to occur will depend on how we assign values to the variables in (2), but can be estimated within certain limits. A lexicon with mediating items will only be more efficient than an equally expressive variant if the number of closed-class items that are needed to mediate all of the kinds of syntactic dependencies that need to be expressed is at least smaller than the number of open-class items that are present in the lexicon. According to Fenson, Dale, Reznick, Bates and Thal (1994), the average vocabulary size reached by 24 months is approximately 300 words with closed-class items gradually emerging in productive vocabularies from around this age, being more or less complete by between 27 and 48 months (Brown, 1973). This would be consistent with a language requiring

approximately 200-300 closed-class items and appears to be reasonably close to the actual number of closed-class items we observe.

5.3.5 Tense and obligatory subjects

Nearly all verbs have at least one noun phrase argument, so it would make sense on economy grounds to liberate individual verbs from the responsibility of licensing a position for it to occupy, leaving it to an encapsulating projection like Tense instead. This may be the motivation for the subject position and may explain why subjects are obligatory even for the few verbs like those in (10), which lack arguments of the type that can occupy the subject position. The requirement that clauses have subjects is fulfilled in such cases by using semantically null 'dummy' subjects (*it* and *there* in English).

10. a. *It* rains/snows/hails a lot in winter.
    b. *There* seems/appears to be a problem with the veal.
    c. *It* is amazing that you survived.

The requirement that all clauses have subjects is called the *Extended Projection Principle*. Subjects are only overtly required when the clause has finite tense as in (11a) but may also be present though unpronounced in non-finite subordinate clauses such as the bracketed portion of (11b).

11. a. He helped her.
    b. He_i tried [$_{IP}$ PRO_i to help her].

The tense element that clauses require may be the means by which the subject position is licensed and which would account for its existence as well, its use in conveying

temporal meaning being secondary. Indeed, there are languages such as Wintu in which the 'Tense' element does not have a temporal meaning, but is instead used to indicate whether the information being imparted was observed directly or acquired second hand (Sapir, 1921).

The details of how syntactic features may be used to license positions will be elaborated in much more detail in chapter six (§6.3.3).

5.3.6 Closed-class items with identical distributions

Some closed classes (like determiners in English) have several members, but if redundancy is the sole concern, we should predict that there are no two closed-class items that have exactly the same syntactic distribution in the sense that one could be exchanged for the other without affecting the grammaticality status of the sentences in which they appear. The fact that the pair of determiners *this* and *that* appear to be freely interchangeable (not semantically, but grammatically) suggests that redundancy is not the only concern. They are interchangeable in most contexts such as (12), but perhaps not all contexts as the contrast in (13) suggests.

12. a. I will give her *this* sock.
    b. I will give her *that* sock.
13. a. I will give her *this* or *that* sock.
    b. ?I will give her *that* or *this* sock.

To the extent that there are closed-class items with identical distributions, the explanation may be that once a system like this has emerged, these items may as well be co-opted to make semantic distinctions that are most frequently needed and since a lexicon containing one extra closed-class item of a certain category will be more costly by only one item, the additional cost would not undo the massive savings associated

with reducing the redundancy of all the members of the open class with which the encapsulating items are associated. In short, we should expect that where there is a choice of mediators, it is motivated by pressures for semantic compositionality of the kind examined by Kirby (2001). However, the fact that there is not always a choice and that mediators do not always have meanings suggests that semantic compositionality cannot on its own account for the existence of closed-class items. If there are completely meaningless closed-class items, we shouldn't expect to find that they can be exchanged with one another to modulate meaning in the way that *this* and *that* can. The only place they could exist then, would be at convergence points. Interestingly, there appears to be some support for this. The complementiser *that* in a sentence like (14) cannot be substituted for anything else. It can be omitted, but without changing the meaning.

14. I believe (that) corn grows in Illinois.

A similar case can be made for the meaninglessness of the inflections indicating verbal agreement, which are entirely predictable from the combination of number, person and/or gender features of the subject (depending on the agreement system of the language).

5.3.7 Measuring the information content of lexical entries

Up until this point, I have omitted a precise discussion of how to measure the information content of a lexical entry. The number of bits that are required to represent a feature will depend on how many features we wish to distinguish. A lexicon that makes use of only four features will require at least two bits to encode each of the four different values (00, 01, 10, 11). Three bits would allow eight values to be encoded (000, 001, 010, 011, 100, 101, 110, 111) and so on. Taking this into account, we should

not simply assume that a lexicon in which entries have fewer features is necessarily more economical than one with more features. It will also depend on the number of features that are distinguished. For instance, an entry with two features in a lexicon that has only four feature types will require two bits to encode each feature meaning a total of four (e.g., {00, 10}) which is just as many as would be required to represent an entry with only one feature in a lexicon with 16 feature types (e.g., {0010}). In general, n bits are required to represent 2^n distinct types.

The measurement becomes more complicated if we discard the assumption that each feature is encoded with a bit-string of the same length. Shorter codes can be assigned to the most frequently occurring symbols as in Morse Code where the letter E (the most frequently occurring character in English) is represented as a single ".", while the less common Q is encoded as the longer string "– – . –". Likewise, the brain might reserve the most parsimonious neural representations for its most frequently utilised codes.

A second, interesting possibility is that the same features are reused for different classes, much as two different homes in different places may have exactly the same local phone number, which is nevertheless distinguished via distinct area codes. This kind of solution would predict that the lexicon be organised such that all members of an open class will tend to be represented in a localised area of the brain – an area for each type of noun, verb and adjective with perhaps also distinct areas for each distinct language that a person speaks as well. The 'area code' would effectively translate the local code into a longer bit string so that the individual lexical entries can also be distinguished in a context global to the lexicon.

These comments are of course extremely speculative, but are intended to point to areas of future research.

5.4 Competing explanations

It is difficult to imagine a language without closed-class items so it is tempting to think that they exist because they are somehow necessary to express certain kinds of concepts, but most closed-class items appear to have very loose associations with meanings and there is no obvious reason why these notions couldn't also be expressed using open-class items. To take an example of closed-class items that have very little semantic content, consider case markers. Nominative and accusative case markings on nouns indicate specific thematic roles with respect to the verb in simple active sentences, but not in other sentences as illustrated in (15) and (16).

15. a. *He* bit the dog.
    b. *He* was bitten by the dog.
16. a. She remembers *him* biting the dog.
    b. The dog bit *him*.

He can refer to the person who did the biting as in (15a) or the person who was bitten as in (15b). *Him* can also take both roles, referring to the person who was biting (16a), or who was bitten (16b).

Other closed-class items have meanings, but the meanings can be expressed in other ways. Instead of past tense for instance, speakers could use an open-class element like the adverb *yesterday* to indicate that an event occurred in the past. Note also that the biting event described in (16a) is understood to have occurred in the past despite there being no overt past tense marking on either verb. It is easy to verify that other open-class alternatives are available (or could be invented) to encode the meanings associated with aspect, number, definiteness, and all of the other closed classes. There is therefore no general reason why they would be necessary in a language that is

expressive as ours though they may be an efficient way of encoding frequently utilised semantics. But if this is true, then we would struggle to motivate things like structural case, which as we saw in (15) and (16) has no consistent interpretation with respect to argument roles. Other words seem to be virtually devoid of semantic content too. The complementiser *that* serves a purely grammatical function in introducing declarative sentences and can often be omitted. This kind of evidence is at odds with an account that would seek to explain the existence of closed classes primarily in terms of semantic considerations.

5.5 Summary

This chapter attempted to provide an adaptive explanation for the open/closed class distinction. In sum, large vocabulary sizes were assumed to apply a strong pressure for non-redundant lexical representations. Distributional information that would otherwise have to be redundantly encoded in each distinct lexical entry can be offloaded onto a small, closed class of items used to mediate syntactic dependencies between the very much larger number of open-class items. This is why the closed-class items contain most of the information that constitutes a person's grammatical knowledge. The theory also predicts the existence of case, agreement and the extended projection principle as concomitant traits. For so many grammatical properties that have so far resisted convincing evolutionary explanations to fall out of this theory is a very promising result. It also follows from the theory that during language acquisition, the earliest closed-class items will not be acquired until an initial body of open-class items is already present in an infant's vocabulary. This is consistent with what appears to occur in language acquisition (Brown, 1973).

This chapter has illustrated how the optimality diagnostic can be applied. Predictions of optimality follow from claims about selective functions and can be tested in accordance with the scientific method. In this case, if it turned out that closed-class

items had properties that make them suboptimal for the hypothesised function of economising lexical representations, then the theory would either have to be discarded or modified. As we learn more about how lexical entries are represented in the brain, further opportunities to test the detail of the theory will arise. Regardless of whether the predictions of the theory continue to be confirmed, the value of the optimality diagnostic in allowing such claims to be tested is undeniable.

Many questions arise about the path evolution could have taken from a language lacking closed-class items to one possessing them such that each stage conferred an advantage to the mutants at the frontier of the innovations. One possibility is that the first mutants began analysing input as though it contained unpronounced closed-class items, which conferred an advantage without leading to a difference in the actual spoken form of the language. Speakers would have benefited by having a superior encoding of basically the same thing, giving them spare capacity to use on larger vocabularies or other things.

It is argued here that what was central to the advantage of closed-class items was the elimination of redundancy, but Chomsky (1995) has argued that redundancy is actually useful, and so we should expect evolution to favour it. Chomsky also argues that the computational system is surprisingly free of redundancy. Hence, we can construct the syllogism in (17).

17. *Premise 1*. If natural selection favours redundancy so as to "compensate against injury and defect", and

Premise 2. The computational system lacks this redundancy, then

Conclusion. We should expect the computational system to be vulnerable to injury and defect.

Logically, there are three different positions one could take on this issue. First, one could accept both premises, thereby concluding that the language faculty (narrowly construed in terms of the computational system) is vulnerable to injury and defect. This would lead to predictions that ought to be readily observable. Second, one could reject the second premise and hold that it does in fact possess redundancy, from which one would predict that the language faculty is robust in the face of injury and defect. Or third, one could hold that it is non-redundant, but reject the premise that redundancy is always necessary to create a robust system. The robustness of the language faculty can only be evaluated in a fairly arbitrary way, but language disorders are certainly not common. This suggests that one or other of the above premises is doubtful.

With respect to redundancy, the lexicon is peculiar in the sense that its size appears to be much larger than can easily be accounted for in terms of survival needs. Human memory capacity for other things such as faces, songs and events is also much larger than we might expect. These are deeper questions for further research.

6

Phrase structure and sequences

In the previous chapter, the optimality diagnostic was applied to hypotheses of selective functions, the rationale being that we can filter out implausible hypotheses by demanding that we consider only functions that are optimally satisfied by the trait under investigation. In such cases, inquiry begins with an observed trait for which a functional explanation is then sought. In other cases, we wish to inquire into the *nature* of a complex biological system rather than, or in addition to, its *selective functions*. In the process, we formulate hypotheses about the existence of properties that have not yet been directly observed, hypotheses which could also arguably be filtered by applying optimality considerations. This would work by demanding that the only properties postulated in our models are those that are optimal for fitness-related functions with the effect that hypotheses about unobserved details could be made less arbitrarily. As a guiding principle, this is almost certainly fallible, amounting only to a heuristic, but one which would allow us to focus inquiry into areas that we should expect to be the most fertile in terms of generating explanatory theories. This bears a strong resemblance to Chomsky's (1995, 1999, 2000a) Minimalist Program with its emphasis on perfection and economy, but is distinct from it insofar as the perfection that is assumed to exist is assumed, under the present approach, to relate to selective fitness, thus further restricting the scope of plausible theories to those that are compatible with an evolutionary account. This chapter presents a case study in reasoning of this kind, exploring both the selective functions of constituent structure representations as well as their nature.

There is another important difference between the previous and current chapters with respect to the explanatory role of the optimality diagnostic. Considerations of optimality can be informative about the selection pressures that shaped a trait in ancestors regardless of whether or not the selection pressures are still present. A trait that has been co-opted for a new function will still bear the hallmarks of its previous roles. Therefore, demonstrating that a trait is optimal for a function other than the one it currently serves is one source of evidence that an exaptation event has taken place.

In this chapter, I will provide some evidence that suggests nested phrase structure could have arisen under selection for functions relating to the representation and manipulation of sequences, functions which were plausibly relevant for other cognitive tasks before and after the emergence of language. Representations of sequences are presumably implicated in many cognitive systems including those governing motor control, planning, musical competence, and episodic memory. It is conceivable, though not necessary, that some of these systems depend on shared mechanisms, either in the sense of a shared system dedicated to processing sequences of various types, or in the sense of the same kinds of neural structures appearing in each of these different subsystems under the influence of the same genes. The alternative is that sequence representations evolved independently in each subsystem. This chapter therefore also bears on the question of whether the sequence representations used by the language system arose independently of other aspects of cognition.

By asking what would constitute an evolutionarily optimal design for sequence representations, the current chapter leads the reader to a theory of the representation of constituent structure, from which a long list of grammatical phenomena follow naturally. In particular, it would appear that movement, far from being a separate innovation, is inherent to the representation of constituent structure itself. The nature of the command relation and feature checking are also discussed in light of these

developments, leading to further refinements. The interplay between case and theta roles also gains a natural explanation.

6.1 The optimal representation of sequences

In this section, a number of methods for representing sequences are examined in abstract terms. These methods are taken to constitute the major types contained in the strategy set over which evolution may have operated during the evolution of sequence representations. The methods are compared using a number of different criteria of optimality, but the discussion remains neutral about whether these sequences could be related to language or other cognitive systems.

6.1.1 Proposal 1: Indexing

A simple approach to representing a sequence of tokens would be to assign each an index marking its absolute order. If the index corresponds to a natural number, insertions into the sequence could only be achieved at the expense of having to re-index every token that appears after the insertion point. If the index was a real number instead of a natural number, a new item inserted between say tokens 5 and 6 could be assigned a value like 5.5 and no re-indexing would be required, but as more and more tokens are inserted, nearby indices would end up being more and more alike which would make them more vulnerable to noise-induced error. This is the familiar pitfall of analogue as opposed to digital representations.

6.1.2 Proposal 2: Pairs

There are other ways to represent a sequence that don't have these shortcomings. The linear order could be represented in relative rather than absolute terms using a set L of ordered pairs (α, β) with the interpretation that token α immediately precedes token β. Call the set L the linearization set. To insert token γ into the sequence between tokens α and β, the new linearization set L' would contain the two pairs (α, γ) and (γ, β) not

included in the original set L and would not include (α, β) which was in L. The differences between the two sets are captured in (1) where M is the subset of pairs that are unaffected by the insertion.

1. a. $L = \{(\alpha, \beta)\} \cup M$

   b. $L' = \{(\alpha, \gamma), (\gamma, \beta)\} \cup M$

The pair (α, β) would not be included in L' because although α still precedes β, it no longer *immediately* precedes it. Alternatively, we could choose to interpret pairs in terms of simple precedence without the immediacy requirement, in which case (α, β) could remain in the set. This would remove the overhead of an operation to remove the pair but at the expense of creating redundancy since the precedence of α and β can be inferred from the inclusion of (α, γ) and (γ, β).

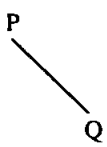
These kinds of representations are digital and never require potentially costly re-indexing operations when tokens are inserted. However, if we assume that the operations that add and remove pairs are also costly, then a preferable representation would allow an insertion to be made by adding only a single term to the linearization set without introducing redundancy. A representation that meets these criteria is described in what follows.

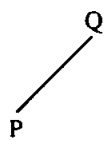
6.1.3 Proposal 3: Triples

Consider a set L of triples of the form (α, β, δ) where α and β are tokens with the interpretation that α precedes β (it may or may not also immediately precede it) and δ is a relation that specifies which of the two tokens 'dominates' the other. Let dominance be a transitive relation such that if α dominates β and β dominates γ, then α also dominates γ. Let us also stipulate that dominance is used to infer linear order such that

whatever order holds between the dominating token α and the dominated token β will also hold between α and everything β dominates in turn.[51]

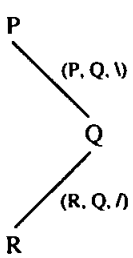
The relation δ can be either *right domination* (written '\'), or *left domination* (written '/'). Relations of precedence and dominance are most clearly illustrated in tree diagrams with each triple in L drawn as a diagonal arc connecting the dominating token at the top to the dominated token at the bottom. The symbols used for the ordering relations are intended to suggest the slant of arcs. For instance, the triple (P, Q, \) would be drawn as in (2a), while the triple (P, Q, /) would be drawn as in (2b).

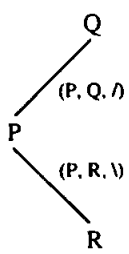
2. a. 

b. 

Note that in both graphs, token P is drawn preceding token Q. By contrast, the triples (Q, P, \) and (Q, P, /) would be drawn with Q preceding P.

Let us now look at how insertions would be handled in terms of a set of such triples. Consider how a token R could be inserted so that it appears between P and Q in the examples in (2). For reasons to be made clear, this would be achieved for (2a) by adding the triple (R, Q, /) and for (2b) by adding (P, R, \). The resulting graphs are depicted in (3) with labels indicating which arc corresponds to each triple.

3. a. 

b. 

[51] An equivalent ordering relation could be devised by interpreting the triple (α, β, δ) to mean that α dominates β and that the value of δ determines whether it also precedes it. A triple is simply a device for simultaneously representing the relations of both precedence and dominance.

Let us now look at how the set {(*P*, *Q*, \), (*R*, *Q*, /)}, illustrated in (3a), specifies an ordering of the tokens *P*, *Q* and *R*. The triple (*P*, *Q*, \) tells us that *P* precedes not only *Q* but every token that *Q* dominates. The triple (*R*, *Q*, /) tells us that *Q* dominates *R*. Hence, we can infer that *P* precedes both *Q* and *R*. The triple (*R*, *Q*, /) also tells us that *R* precedes *Q*. Hence we have an ordering for all three tokens, which is $P < R < Q$. The graph in (3b) represents the set {(*P*, *Q*, /), (*P*, *R*, \)}, which produces the same ordering as in (3a).

This kind of representation allows an insertion to be made anywhere in the sequence including the beginning and end by adding only a single element to the linearization set and this is possible without introducing any redundancy so that the set will only need to contain $n-1$ elements to represent a complete ordering of n tokens. However, not all possible sets of triples would produce a linear ordering of tokens. Only those that do should be considered well-formed linearization sets. In practice, it may be easier to check for conformity to certain constraints that will always imply linearizability than to check for linearizability itself. The following constraints would serve this purpose.

4. a. **Non-circularity constraint**

For all tokens α and β, if α dominates β, then β does not dominate α.

b. **Single parent constraint**

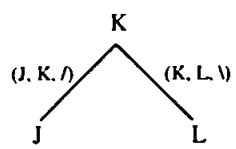
A tree has exactly one token (called the *root*) that is not dominated by any other token. For every token in the tree aside from the root, there is exactly one other token that immediately dominates it. In other words, for all tokens α, if α is not the root, there is always exactly one distinct token β such that *L* contains either (β, α, /) or (β, α, \), and if α is the root there are no such tokens β.

c. **Binary branching constraint**

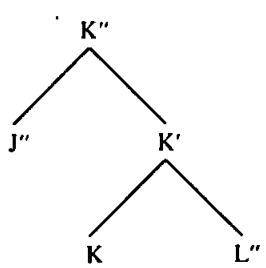
Each token can immediately dominate up to two other tokens – one to the left and one to the right. In other words, for all tokens α, there is at most one token β such that *L* contains (α, β, /) and at most one token γ, such that *L* contains (α, γ, \).

This method of representing sequences allows insertions to be made by adding a single term to the linearization set without introducing redundancy. An additional and rather surprising property of this representation is that it also allows constituent structure to be represented in the sequence without introducing any additional notation. The maximal projection of a head represented by a given token could be defined as the subsequence that includes this token and all of the tokens it dominates, the constituent dominated to the left being the specifier and the constituent dominated to the right being the complement. Hence, the linearization set {(*J*, *K*, /), (*K*, *L*, \)} with the structure (5) could equally be drawn with the equivalent X-Bar structure in (6).

5.



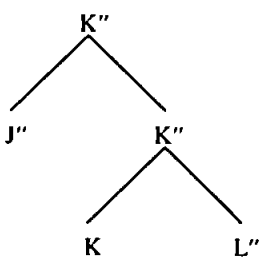
6.



Brody (2000) also rejects representations of the kind in (6) in favour of the kind in (5), but for reasons independent of present considerations. His justification is purely that (5) captures all of the facts that (6) captures, only more elegantly.[52]

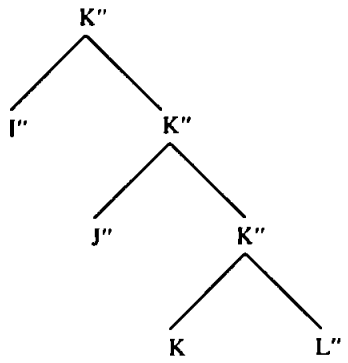
It isn't immediately obvious how to treat adjunction under this notation, but the approach of Kayne (1994) suggests at least one possible strategy. Under his theory, the distinction between specifiers and adjuncts is eliminated, specifiers being adjoined to generate a *multi-segment category* as in (7) where the category K″ has two segments. As discussed in chapter two (§2.4.1.2), Kayne's (1994) Linear Correspondence Axiom (LCA) derives a linear ordering of a tree's terminal nodes (leaves) from the command relation. Under Kayne's analysis, further adjunction to the same phrase is prohibited, effectively ruling out structures such as (8), and thereby deriving the assumption that a phrase can have at most one specifier. Kayne then invokes a proliferation of further functional projections to generate further specifier positions for the attachment of constituents at higher levels, an approach also adopted by Brody (2000).

7.



[52] Brody (2000) looks at many of the consequences of adopting this notation that won't be relevant here. In particular, he presents a solution to the issue of representing structure that is internal to minimal projections.

8.



Under Kayne's definition of command, a specifier of a phrase will always command its head and complement, while the head of a phrase will always command everything within its complement (though strictly speaking, not the maximal projection that is the complement itself). In short, this means that all of the terminals in a specifier will precede the head, which is a terminal, and the head will precede all of the terminals in the complement. Considering how this maps onto the austere representation in (5), we can see that each triple in the linearization set effectively corresponds to a command relation with *J* commanding *K* and *K* commanding *L*, and if we take this relation to be transitive, we can derive from these facts that *J* also commands *L*.

Although Chomsky (1995) adopts a variation on Kayne's Linear Correspondence Axiom in his theory of bare phrase structure, he not only maintains the distinction between adjuncts and specifiers, but also permits a phrase to have multiple specifiers. Chomsky's position is hence more difficult to reconcile with the current representations. For now, it is sufficient to conclude that, at least to a first approximation, constituent structure can be represented in terms of triples.

6.1.4 Proposal 4: Pairs reconsidered

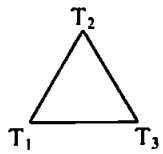
I now revisit a variant of Proposal 2 and show how it too can allow constituent structure to be represented.

Proposal 2 is in some ways simpler than Proposal 3 and in some ways more complex. It is simpler insofar as it uses pairs instead of triples. It is more complex insofar as it appears to require more operations to be performed on the linearization set to insert an item. The following revises the proposal so that no terms need be removed from L when inserting tokens. Insertions will however lead to redundancy in the linearization set from the point of view of information about linear order, but I will argue that this information is non-redundant under another analysis.

An immediate consequence of using pairs instead of triples is that a term cannot by itself encode both precedence and dominance simultaneously. The pair (α, β) could be interpreted as a precedence relation or as a dominance relation but cannot be used to specify both relations independently. If both relations are to be preserved, then another mechanism is required to infer whichever relation is not explicit in the interpretation of the pair.

This can be achieved in the following way. Firstly, let us interpret the pair (α, β) with the meaning α precedes and is adjacent (in the graph theoretical sense) to β. Secondly, let us redefine dominance relations in terms of three tokens α, β, γ that are all adjacent to one another in a graph, such that the middle one (in terms of linear order) dominates the other two. In other words, if the linearization set contains the subset {(α, β), (α, γ), (β, γ)}, then β dominates both α and γ and no dominance relation holds between α and γ. Under this interpretation, three adjacent tokens can be drawn as a triangle as in (9a) with the corresponding linearization set given in (9b).

9. a.

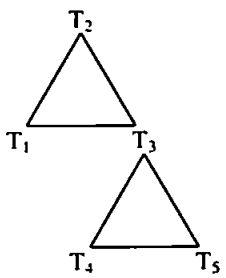


b. $\{(T_1, T_2), (T_1, T_3), (T_2, T_3)\}$

From the new definition of dominance, we can infer that, while there is no dominance relation between T_1 and T_3, both of these tokens are dominated by T_2. Let the other properties of dominance as defined earlier apply as well so that dominance remains a transitive relation and continues to be relevant for linear ordering such that if α dominates β, then whatever order holds between α and β will also hold between α and everything β dominates in turn.

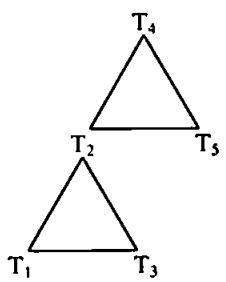
Further structure can be added to (9a) by introducing more triangles. So for example, we can build on this structure to produce the structures in (10) or (11).

10. a.



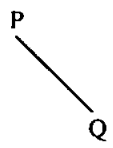
b. $\{(T_1, T_2), (T_1, T_3), (T_2, T_3), (T_4, T_3), (T_4, T_5), (T_3, T_5)\}$

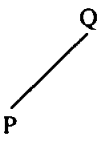
11. a.



b. $\{(T_1, T_2), (T_1, T_3), (T_2, T_3), (T_2, T_4), (T_2, T_5), (T_4, T_5)\}$

The point of interest here is that the structures in (10) and (11) resemble those in (2) repeated below, which you will recall were represented in the terms of Proposal 3 using the triples (*P*, *Q*, \) and (*P*, *Q*, /).

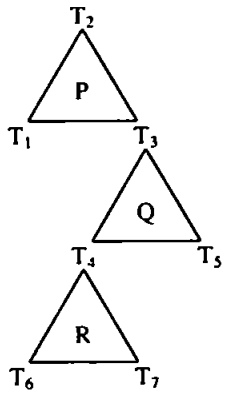
2. a. 

b. 

Indeed, the tokens of Proposal 3 are, in effect, higher level descriptions of the triangles of the current proposal. Let us call the individual tokens that compose triangles *micro tokens* to distinguish them from triangles themselves, which correspond to the tokens of Proposal 3, and which we'll call *macro tokens*. Note that under Proposal 3, macro tokens were linked by arcs, but under Proposal 4, macro tokens are linked by virtue of sharing a common vertex such as T_3 in (10a).

Let us now look at how insertions would be handled in terms of triangles. In (3a) a token *R* is inserted between *P* and *Q*. An analogous representation that uses triangles is illustrated in (12a) corresponding to the linearization set in (12b) and the linear ordering given in (12c).

12. a.



b. $\{(T_1, T_2), (T_1, T_3), (T_2, T_3),$

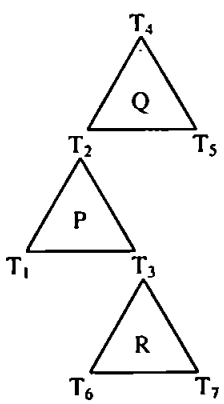
$(T_4, T_3), (T_4, T_5), (T_3, T_5),$

$(T_6, T_4), (T_6, T_7), (T_4, T_7)\}$

c. $T_1 < T_2 < T_6 < T_4 < T_7 < T_3 < T_5$

Similarly, the triangle representation of (3b) is given in (13).

13. a.



b. $\{(T_1, T_2), (T_1, T_3), (T_2, T_3),$

$(T_2, T_4), (T_2, T_5), (T_4, T_5),$

$(T_6, T_3), (T_6, T_7), (T_3, T_7)\}$

c. $T_1 < T_2 < T_6 < T_3 < T_7 < T_4 < T_5$

While micro tokens are ordered, the concept of linear order as it applies to triangles remains undefined. If triangles are to be equated with the tokens of Proposal 3, then the order of triangles would have to be $P < R < Q$ in both (12) and (13). This is achieved here and in general if we take the order of triangles to be defined by the relative order of the micro tokens that constitute their top vertices (indicated in bold in (12c) and (13c)).

It should be obvious that, as with Proposal 3, a triangle can always be inserted between any two macro tokens without requiring any terms to be removed from the linearization set. Although a pair-based representation using triangles will contain redundant information about linear order at the level of micro-tokens, this information is required to represent dominance relations. Viewed from the level of macro tokens, no redundancy is necessary since it is possible to represent any of the possible structural arrangements of n triangles simply by representing those n triangles. This is because the linear order and dominance relations that hold between triangles are completely determined by which vertices they share.

The constraints on allowable structures that were proposed in (4) can now be restated in terms of the triangle-based representation as in (14). The *binary branching constraint* of (4c) is replaced here with (14c) which expresses the same generalisation more simply as a constraint on the intersection of triangle vertices.

14. a. **Non-circularity constraint**

For all micro tokens α and β, if α dominates β, then β does not dominate α.

b. **Single parent constraint**

There must be exactly one micro token in the structure that is not dominated by any other. Every other micro token must be directly dominated by exactly one parent.

c. **Triangle constraint**

Every micro token must be a vertex of at least one triangle. If it is a vertex of more than one, it must be the top vertex of exactly one of them.

As we have seen, this pair-based representation appears to be ideally suited to representing sequences when insertion operations are carried out on them while remaining extremely simple. A sequence can be represented merely as a set of ordered pairs. By incorporating a derived concept of dominance in addition to precedence, the representation allows insertion operations to be carried out without requiring items be removed from the linearization set and without introducing redundancy. A side-effect of this is that it allows constituent structure to be represented without requiring any additional representational apparatus.

The question of why insertion operations should have been important in the evolution of sequence representations is unclear. Some alternative metrics for evaluating the optimality of sequence representations are pursued in the following section. A number of very interesting consequences of using a pair-based representation for syntactic structures will then be considered, shedding light on movement

phenomena, the nature of feature representations and the interplay between case and theta theory.

6.2 Alternative metrics

Aside from the cost of insertion operations, we could look at deletions or other operations that might need to be carried out on sequence representations or at processing costs associated with querying the linearization set to find the next token in the sequence.

6.2.1 Optimal representations for deletion operations

If, instead of inserting, we delete a token from a sequence, then this would lead, under an indexing scheme along the lines of proposal 1, to index values being non-consecutive. If the system requires that they be consecutive, every token after the deletion point would have to be re-indexed; but sequential order would still be adequately represented if no re-indexing were performed.

Under a simple pair-based representation like that in proposal 2, deleting a token R that appears between P and Q would require three operations (the same number as is required for an insertion), but in this case, the pairs (P, R) and (R, Q) would be deleted from, rather than added to the linearization set, and the pair (P, Q) would be added rather than deleted.

Under a representation using triples, deleting token R would require only a single operation in the case where R does not dominate either P or Q as in (15), but otherwise reduces to the pair-based case with two deletions and an addition as in (16).

15. {(*P*, *Q*, \), ~~(*R*, *Q*, */*)~~} → {(*P*, *Q*, \)}

[deletion of (*R*, *Q*, */*)]

16. {~~(*P*, *R*, \)~~, ~~(*R*, *Q*, \)~~} → {<u>(*P*, *Q*, \)</u>}

[deletion of (*P*, *R*, \) and (*R*, *Q*, \); addition of (*P*, *Q*, \)]

Under the modified pair-based representation of proposal 4, a triangle corresponds to the tokens of proposal 3 and can be removed with analogous costs.

The four proposals rank for deletion essentially as they do for insertions except under an indexing scheme when deletion is potentially very economical while insertions are potentially extremely costly. We should nevertheless expect insertions to be more important than deletions for the simple reason that there must logically be more of them since for something to be deleted at all, it must first be inserted.

6.2.2 Optimal representations for spelling out a linear sequence

Under an index-based representation, spelling out the tokens in a linear sequence will involve determining which token has the lowest index above an ever increasing threshold. Alternatively, index values could be represented in terms of activation levels in a connectionist model such that the earlier tokens, having greater activation, are able to inhibit subsequent ones until they are spelled out.[53]

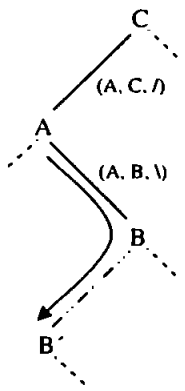
With a pair-based representation, determining the token that immediately follows token *A* in the sequence involves finding the token *x* such that the linearization set contains the pair (*A*, *x*).

For the triples representation, determining the next token in the sequence is much more complicated. To find the token that immediately follows token *A* requires traversing the tree structure in the following way. First, traverse the right child of *A* if

[53] For an early connectionist model that produces serial behaviour in this way see Rumelhart and Norman (1982).

one exists. Token *B* is the right child of *A* if the linearization set contains a triple of the form (*A*, *B*, \). If there is no such token *B*, we must then check to see if *A* is immediately dominated to the right by a token *C* which would be the case if the linearization set contains (*A*, *C*, /). If no *B* exists, but *C* does, then *C* is the token that immediately follows *A*. On the other hand, if *B* exists and *B* does not have a left child, then *B* is the token that immediately follows *A*. If *B* has a left child, it will precede it and hence can potentially be the token that immediately follows *A*. If *B*'s left child has a left child in turn, then it will appear even earlier in the sequence, but will still follow *A* like everything dominated by *B*. Hence, we must traverse the tree via the left descendants of *B* until we find a token *B'* that does not have a left child. This will be the token that immediately follows *A*. This traversal is schematised below.

17.



The processing involved in spelling out the tokens in sequence would appear to be most costly for a representation based on triples (and for the analogous triangle-based representations) and less costly using the pair-based and index-based representations. This means that if these alternatives were available during the evolution of sequence representations, the triple and triangle representations, which allow constituent structure to be represented, would not have emerged under selection for minimising the processing cost of spelling out tokens in linear order. It is clear that

constituent structure is represented in the case of human language and that its tokens are linearized, so there is nothing in the above that provides evidence that processing costs of the kind examined were instrumental in selecting these representational mechanisms.

6.2.3 Optimal representations for representing constituent structure

The ability to represent nested constituent structure may have been relevant in and of itself in the selection of the representations that support it, either instead of or in addition to selection for the ability to make efficient insertions into sequences. Without knowing more about how either of these functions could have conferred a selective advantage, there is very little to decide the issue without pursuing the question further, but by using optimality considerations, we have dramatically narrowed the scope of inquiry into the nature and function of constituent structure representations. Further evidence is necessary to decide between the remaining hypotheses.

Given that we are defining the object we would like to inquire about in functional terms (i.e., as the thing that represents constituent structure), it is unsurprising that it would serve the function it is defined in terms of optimally. Indeed, anything is optimal at being itself. A mountain is a perfect implementation of a mountain and a rock is a perfect implementation of a rock. It is only by relating properties of living things to functions that are within a narrow range of functions that can be specified in advance (such as those that relate to fitness advantages), that attributions of function can avoid being post hoc (Dawkins, 1991). That something is improbably optimal for a function other than simply being itself is something that should count in favour of the insertion hypothesis, even if it is uncertain why this function would be so important during the evolution of the representational mechanisms involved. However, this hypothesis would look much more favourable if independent evidence could be found for the reality of triangle-based representations. Consequences

of the triangle-based representation are examined in greater detail in the following section where very suggestive evidence of this sort is indeed presented.

6.3 Some consequences and further refinements

This section examines some syntactic properties that appear to be derivable from triangle-based representations, providing support for their existence.

6.3.1 Movement as by-product

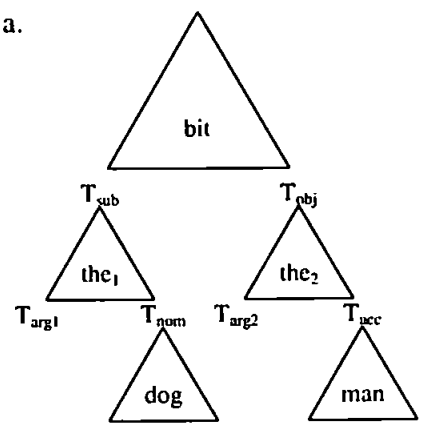
Consider the sentences in (18) below.

18. a. The dog bit the man.
    b. The man was bitten.

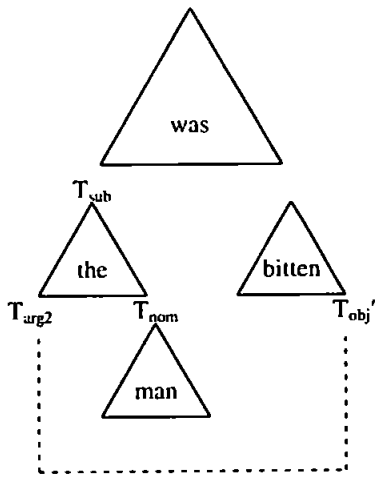
In both of these sentences, *the man* is interpreted as the person who was bitten despite appearing in different structural relationships with respect to the relevant verb. To explain this, some part of the syntactic description should encode this. In generative approaches to grammar, this and many other examples like it are traditionally explained in terms of movement as discussed in chapter two (§2.4.3), the idea being that *the man* originates in the object position in the derivations of both sentences, this position being where it receives its interpretation as an argument of the verb. Subsequent steps in the derivation of (18b) cause it to move to the subject position. The relationship between the base and landing sites of a movement operation is encoded in what is called a chain, which is often represented (in modern formulations) simply as a *pair* of positions (e.g., Chomsky, 1995: 252). Pairs are of course the basis of triangle representations so chains can be represented without requiring any revision. This is a rather striking result since it unifies the representation of chains/movement with the representation of constituent structure itself.

The structures in (19) illustrate the relevant aspects of the sentences in (18) though many of the details are left imprecise. In (19a) *the man* appears in the direct object position and is linked to *bit* by the pairs (T_{arg2}, T_{obj}) and (T_{obj}, T_{acc}). I adopt the usual assumptions that a passive verb is unable to assign accusative case to its object meaning that, in (19b), the second of these pairs will be blocked for *bitten*. The effect of this is that the argument has to be attached to an earlier case assigning position. In this instance, it is attached to the subject position where it receives nominative case. It is linked to *was* using the pairs (T_{arg2}, T_{sub}) and (T_{sub}, T_{nom}). The chain is encoded in (19b) by including the pair (T_{arg2}, T_{obj}'). Since no triangle is created by including this pair in the linearization set, it has no consequences for dominance relations within the structure and hence no consequences for dominance-related ordering effects.

19. a.



b.



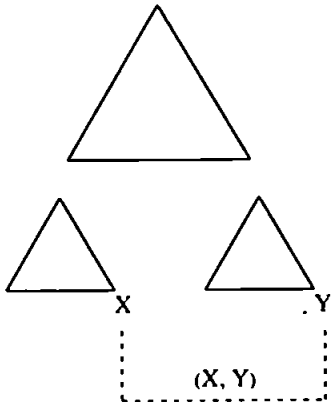
Although constituent structure and movement chains can both be represented using pairs, it is not clear that there are any implications for whether a derivational or representational approach is to be preferred. Recall from chapter two (§2.5) that, within Minimalist Syntax, the property that a moved constituent always commands its base-generated position can be derived from the nature of the Move operator. However, a representational approach must specify this explicitly as a well-formedness condition on

chains. If such a condition is placed on pairs that are interpreted as chains, then it would be instructive to see whether it also holds of the pairs that are implicated in the representation of constituent structure itself and indeed, for a triangle given by the set $\{(\alpha, \beta), (\alpha, \gamma), (\beta, \gamma)\}$, a command relation will hold between tokens of each pair. This leads to some suggestions pursued in the following section.

6.3.2 The status of the command relation

Up until this point, we have seen that dependencies traditionally described as arising through movement can be unified with phrase structure itself under a modified pair-based representation in which dominance relations are inferred from the existence of other pairs that we can interpret as forming triangles. However, nothing in this picture yet provides any motivation for the restriction that moved constituents must command their base positions. This is because there is nothing preventing dependencies existing between constituents that are not in a command relation. Given that the austere representation of phrase structure collapses the distinction between maximal and minimal projections, there are consequences for what phrase markers can be said to dominate others and hence some clarification of the notion of command is necessary. I will return to this question shortly, but irrespective of this, if we translated the austere representation in (20) back to the traditional tree structure notation, the dependency between X and Y would not satisfy the definition of command regardless of whether they are treated as maximal or minimal projections.

20.



The inclusion of the pair (X, Y) in the linearization set does not lead to a violation of any of the constraints in (14). Hence, some other mechanism is required to rule out the possibility of forming such dependencies.

This is achieved by treating the command relation as a primitive rather than derived relation. Instead of interpreting each pair (α, β) in the linearization set in terms of precedence, it will now be defined as representing an instance of the command relation such that α commands β. This will nevertheless imply precedence following Kayne (1994). A new definition of command is provided in (21) in terms of pairs and dominance.

21. α commands β iff

$((\alpha, \beta) \in L) \vee$

$\exists\gamma\ ((\alpha, \gamma) \in L \wedge (\gamma \text{ commands } \beta)) \vee$

$\exists\gamma\ (\alpha \text{ commands } \gamma \wedge \neg(\gamma \text{ dominates } \alpha) \wedge (\gamma \text{ dominates } \beta)$

Command relations are partly defined in terms of dominance, and dominance can be derived either by the transitivity of the dominance relation or by stipulating the triangle

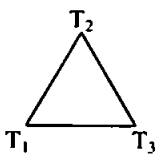
rule discussed earlier and restated in terms of command in (22).[54] In words, this rule says that for a linearization set containing the pair (α, β), a dominance relation will hold between α and β if and only if there is no intervening γ such that α commands γ and γ commands β.

22. if $(\alpha, \beta) \in L$ then

$(\alpha \text{ dominates } \beta \vee \beta \text{ dominates } \alpha)$ iff $\neg\exists\gamma$ (α commands $\gamma \wedge \gamma$ commands β)

This is a slightly more general version of the rule informally introduced earlier, in which dominance relations were inferred from pairs that formed closed loops of dependencies between sets of three tokens. The rule in (22) is more general in that it applies to closed loops of any size. The earlier rule allowed us to infer the dominance relations that hold in triangle structures as in (9), repeated below.

9. a.



b. $\{(T_1, T_2), (T_1, T_3), (T_2, T_3)\}$

Using the new rule in (22), we can derive exactly the same dominance relations as follows. First, we can infer the facts in (23) from the linearization set.

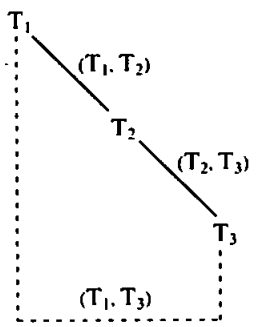
23. a. Either T_1 dominates T_2 or T_2 dominates T_1.

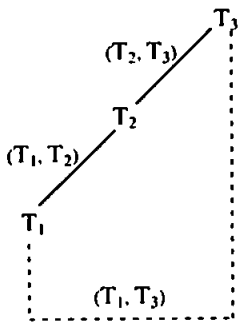
b. Either T_2 dominates T_3 or T_3 dominates T_2.

c. There is no dominance relation between T_1 and T_3.

[54] Since the definition of *command* refers to dominance and the definition of *dominance* refers to command, these are recursive definitions. At the most fundamental level, they both ultimately derive from the simplest type of command relation, which is represented by a pair in the linearization set.

These facts rule out the structures in (24) as a way of representing these three dependencies, thus leaving either the triangle structure of (9a) or an upside-down version of it as the only alternatives, the latter being ruled-out in turn by condition (14c), which says that no token can be dominated by more than one parent.

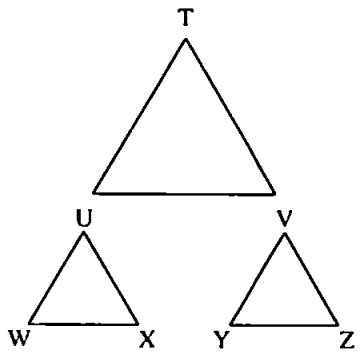
24. a. 

b. 

Making command a primitive relation was motivated by the inability to rule out structures such as (20). By redefining pairs as instances of the command relation and generalising the rule for inferring dominance relations, we find that no dependency linking X and Y can now exist in this structure without it having consequences for dominance. The lack of an intervening token that X commands and which in turn commands Y means that a dominance relation must necessarily hold between X and Y, but no such tree can be constructed without violating the *single parent constraint* of (14b) since it would result in either X or Y having more than one parent.

To show that these changes do not also rule out the formation of legitimate pairs, consider the tree in (25).

25.



Apart from the pairs implied by the triangles already in place in (25), other pairs could be added to encode long-distance dependencies without having any consequences for dominance. These are listed in (26). The addition of certain other pairs would have consequences for dominance that lead to violations of the *single parent constraint* (14b). These are listed in (27). The only other pairs that could be added are the mirror images of those already accounted for, each of which would violate the constraints in (14) if inserted.

26. (W, V), (W, Y), (W, Z), (U, Y), (U, Z)

27. (W, T), (X, V), (X, T), (X, Y), (X, Z), (T, Y), (T, Z)

In this example and in general, no pairs are ruled out that would be required to account for the long-distance dependencies implicated in phenomena such as movement, anaphoric binding and the scope of quantifiers.

The use of command as a primitive relation is not new. Frank and Kuminiak (2000) showed that the command relation can be used to define a subclass of the tree structures that can be defined using precedence and dominance relations, and that this subclass approximates the class that are permitted within X-Bar theory, thus deriving many of its constraints.

Within the Minimalist syntax of Chomsky (1995), the constraint that a moved constituent must command its base-generated position derives from the nature of the Move operator. However, the observation that the command relation is also relevant for binding and scope phenomena suggests that either these phenomena also involve movement in some subtle way or that the command relation actually derives from a more general feature of the computational system. The idea that the command relation would be implicated in distinct phenomena for unrelated reasons should be abhorrent to the Minimalist.

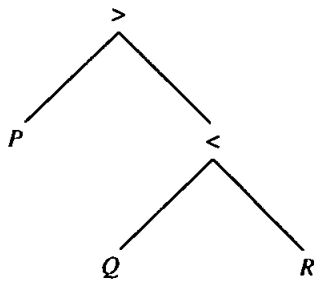
The prominence of the command relation is achieved under the current proposal by treating it as the primitive in terms of which all legitimate linguistic structures can be specified, the pairs in a linearization set being instances of the simplest type of command relation.

6.3.3 Feature checking

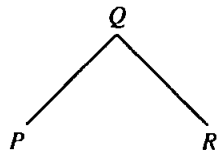
In chapter five (§5.1.2), I argued that the syntactic representations contained in lexical entries are free from both optionality and structure, and hence that syntactic properties can be encoded as an unordered set of features. Under this view, the appearance of optionality arises purely from lexical choice and no distinction of the kind advocated by Stabler (1997) need be made between 'selecting' and 'licensing' features, nor would features need to be linearly ordered in lexical entries as they are in his formalism (see §2.5 of chapter two). The syntactic distribution of a lexical item is then determined by a set of features which can be checked against each other to establish dependencies. To indicate which features can be checked against one another, we can represent features as being either positively or negatively specified. A treatment of the role of formal features in a triangle-based representation follows. Comparisons with Stabler's (1997) notation will serve to illustrate their formal equivalence.

Stabler's notation is equivalent to the austere representations of tree structures that collapse the distinction between maximal and minimal projections for simple phrases as in (28), but if multiple specifiers are permitted within the same phrase as in (29a), the austere representation requires positing a further token to mediate the dependency, hence the introduction of the token *R´* in (29b).

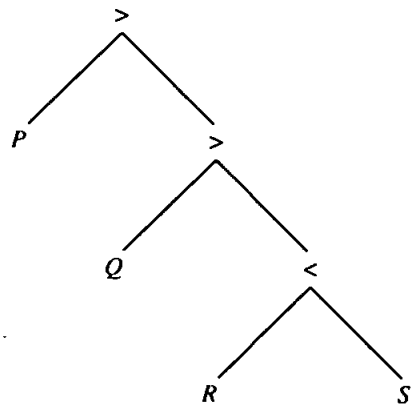
28. a. Stabler's notation



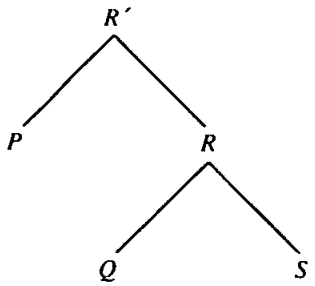
b. austere notation



29. a. Stabler's notation

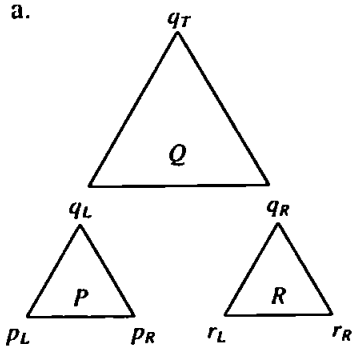


b. austere notation

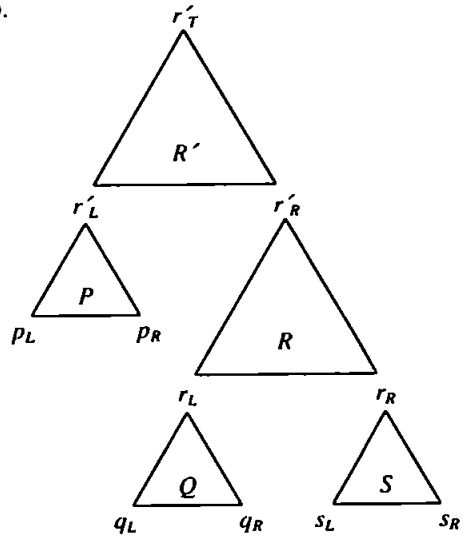


The triangle-based tree representations corresponding to (28b) and (29b) are as in (30a) and (30b) respectively.

30. a.



b.



Each vertex of a triangle is associated with a micro token of a specified type with its own lexical entry. For instance, the tokens in (30a) could be specified with the feature sets given in (31) to give the desired structure. Each pair is licensed by the association of a positively specified feature of one token with the corresponding negatively specified feature of the token it commands. For instance, the bottom left token p_L of the *P* triangle contains the +P feature that links to the –P feature of q_L to form the pair (p_L, q_L).

31. p_L {+P, +PLR}

p_R {–P2, –PLR}

q_T {–Q, +Q2}

q_L {–P, +QLR, +Q, +P2}

q_R {–R, –Q2, –QLR, +R2}

r_L {+RLR, +R}

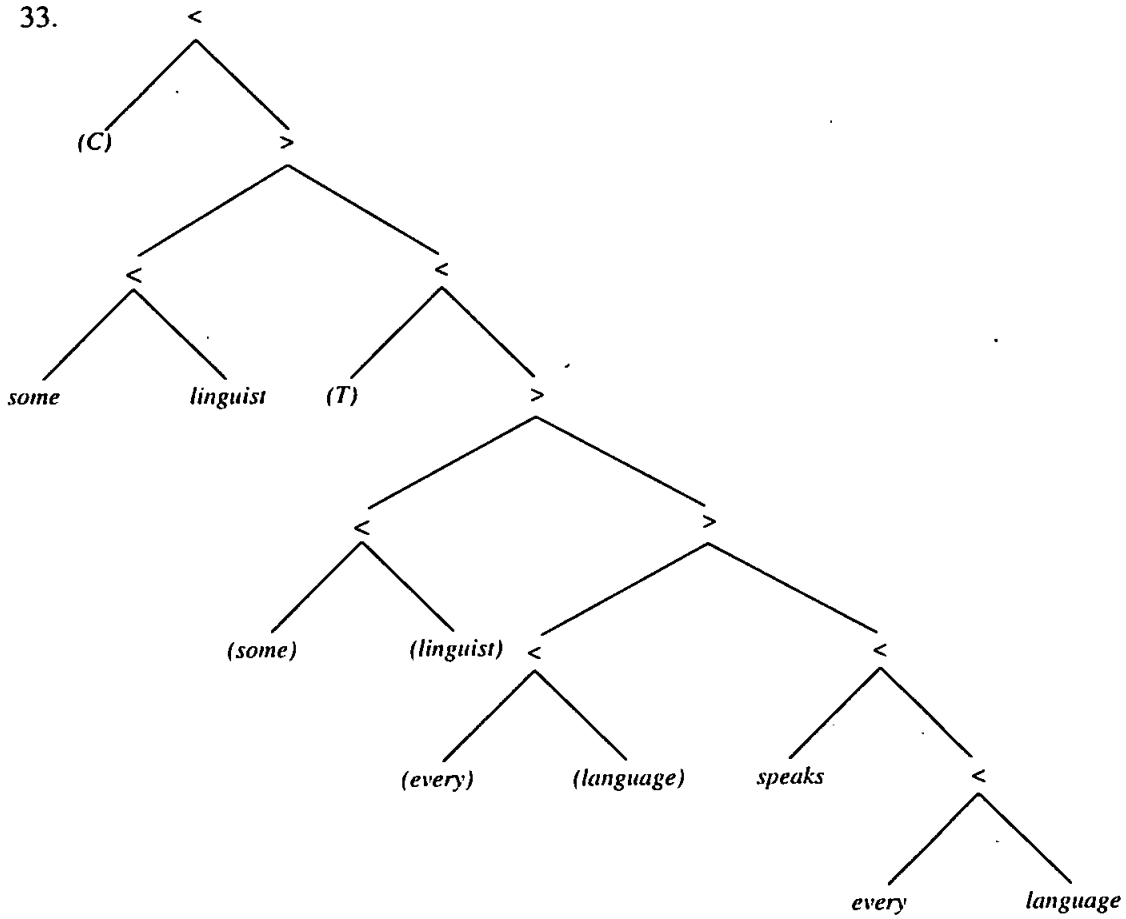
r_R {–R2, –RLR}

In chapter two (§2.5), we saw how we could derive the sentence *Some linguist speaks every language* using Stabler's formalism. I will now demonstrate how the same structural facts can be captured using a pair-based representation. In Stabler's derivation, the lexical items featuring in the derivation were specified as in (32) with the convention that the items lacking phonetic content are enclosed within brackets. The final step of his derivation produced the structure repeated here in (33).

32.

every	=n d –case
some	=n d –case
language	n
linguist	n
speaks	=d +case =d v
(T)	=v +CASE t
(C)	=t c

33.



The lexical entries for the triangle-based representation require that each of the lexical items listed in (32) be split into two to represent the bottom two vertices of each triangle. The resulting feature sets are listed in (34). Except for the token at the root of the tree, the top vertex of each triangle is always shared with another, so only one token in the tree will be unaccounted for once we have split each of the items in (32) into two parts. The root token is here assigned the label $(C)_R$ suggesting its role as the right hand vertex of an incomplete triangle representing a non-overt complementiser.

34. $every_L$ {+DLR, +ARG, +D}

$every_R$ {–N, –D2, –DLR, –CASE, +N2}

$some_L$ {+DLR, +ARG, +D}

$some_R$ {–N, –D2, –DLR, –CASE, +N2}

$language_L$ {+NLR, +N}

$language_R$ {–N2, –NLR}

$linguist_L$ {+NLR, +N}

$linguist_R$ {–N2, –NLR}

$speaks_L$ {+TVLR, +TV}

$speaks_R$ {–D, –ARG, –TV2, –TVLR, +D2}

$(ext)_L$ {–ARG, +CASE, +VLR, +V}

$(ext)_R$ {–TV, –V2, –VLR, +TV2}

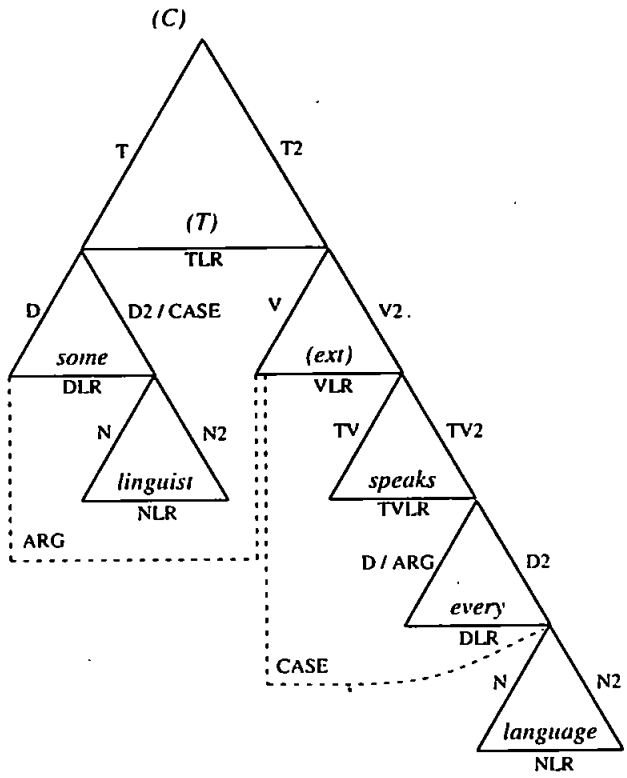
$(T)_L$ {–D, +TLR, +T, +CASE, +D2}

$(T)_R$ {–V, –T2, –TLR, +V2}

$(C)_R$ {–T, +T2}

The tree structure corresponding to (33) is given in (35) with dominance determined by the triangle rule specified in (22). The features that correspond to each dependency are also marked.

35.



Since the tree structure in (33) contains a verb phrase with two specifiers, an additional macro token was also needed in (34) and (35). This is labelled *(ext)* to be suggestive of its function in assigning a theta role to an argument that is external to the verb phrase. Chomsky (1995: ch4) also postulates a distinct functional projection for assigning an external thematic role which he denotes *vP*. Splitting the verb into two macro tokens in this way also allows us to capture the active/passive distinction of transitive verbs since the external role is not assigned to the subject when a verb like *speak* is passivised. The *(ext)* token would then be missing in the structure of a passive sentence and, if this token is also associated with the CASE feature that is checked with the object, then its absence in passive structures will also explain why passive verbs are unable to assign structural case to their objects and hence why the object must 'move' to the subject position to check its case features.

There were two instances of movement in the derivation of (33), one involving the overt movement of the subject and the other involving the covert movement of the

object. The long-distance dependencies associated with both are also represented in (35) with dashed lines. The subject is assigned case from a local token checking the CASE feature while its theta role (ARG feature) is checked by a remote token internal to the verb phrase as if having undergone movement. The object on the other hand, has its CASE feature checked remotely and its ARG feature checked locally capturing the effect of covert movement in Stabler's derivation. The spirit of the overt/covert distinction is therefore also preserved.

Many of the tokens in (34) have identical feature content, suggesting that they may not actually have distinct lexical entries. For instance, the left tokens of *every* and *some* contain the same features. This is also true of their right tokens.

36. $every_L$ / $some_L$ {+DLR, +ARG, +D}

$every_R$ / $some_R$ {–N, –D2, –DLR, –CASE, +N2}

As discussed in chapter five, this kind of redundancy, particularly if found in the lexical entries for the very many open class items, would be extremely costly. Therefore, it seems more likely that only one or the other of the left and right micro tokens is actually associated with the semantic and phonetic content of a macro token. For all of the nouns and verbs in the example in (34), the left token has fewer features. This would make economic sense if it is the left rather than the right token that is unique for every open class item. If the right hand tokens are not unique for each member of an open class, then no redundancy would necessarily result by offloading most of the feature content to them. One consequence of this would be that right hand tokens would mediate a greater number of dependencies compared to left hand tokens thus producing structures that are predominantly right-branching. The causal link might however be in the opposite direction, with whatever is motivating consistent branching to the right (in

English at least) being responsible for the lighter feature content of left hand tokens. This is an open question.

6.3.4 Lexical features and linearization sets

In this section, I will demonstrate that if each token in a sequence has a set of features associated with it, then the linear order of the tokens is the only other piece of information required to establish the contents of the linearization set. To see why this is the case, we have to make the method of translating from one type of representation to the other explicit. Consider the two ways that the *A* features in (37) could be checked against each other.

37. The man whom she punched died.

a. +A +A −A −A

b. +A +A −A −A

In principle, there are two ways to link these features, as in (37a) or as in (37b). But in all English constructions that involve this potential ambiguity, the dependencies never cross. Regardless of how many feature pairs are nested inside each other, the features form pairs like an onion skin with the outermost features forming a pair, then the next two outermost features forming a pair and so on down to the innermost features. Given this regularity, one can always translate between linearly ordered feature sets and pair-based representations. The onion skin regularity can be formalised as in (38).

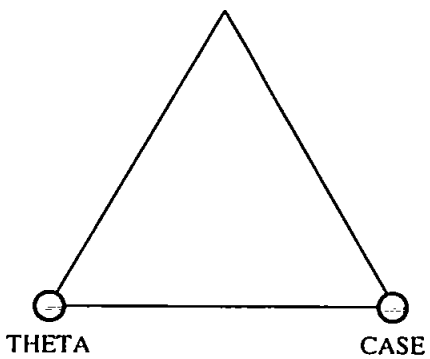
38. *The onion skin algorithm*

Start at the token whose feature set contains the positive feature, setting the feature counter to 1. Scan each token one by one, working rightwards. For each feature set encountered that contains a feature of the same type, increase the counter if the feature is positively specified, or decrease it if it is negatively specified. The negatively specified feature that reduces the counter to zero will belong to the matching token. A pair linking these two tokens will therefore need to exist in the linearization set.

6.3.5 Theta roles and case assignment

The spirit of Stabler's distinction between selecting and licensing features is also preserved in the distinction between features contained in the left and right tokens of a triangle. The left token will carry the ARG feature necessary for determining the theta role of a determiner phrase for instance, while the right token will contain its CASE feature as schematised in (39).

39.



What this means is that arguments that are traditionally analysed as having undergone overt movement will necessarily have long-distance theta dependencies and local case dependencies, which derives the observation that movement is always leftwards and the fact that a constituent appearing in the base-generated position can receive the same theta role as a constituent that has undergone 'movement'. Covert

movement will be the opposite with long-distance case dependencies and local theta dependencies.

The interplay between case and theta roles is a natural by-product of the symmetry of triangular structure and needn't have a functional motivation that is distinct from it. It is a concomitant trait in the sense defined in chapter three (§3.1.1). The ability to explain independent evidence like the case/theta interaction is an extremely promising result and strongly suggests that pair-based representations capture something fundamental about syntactic structure.

6.4 Summary

Optimality considerations can be used to constrain inquiry into the nature as well as the evolutionary functions of a trait, the logic being that we have much more reason to expect to encounter a trait that can be related to fitness than one that can't be. By focussing inquiry on hypotheses that are compatible with evolutionary concerns, we can expect to increase our chances of discovering explanatory theories. But if we pursue this further by focussing not only on hypotheses that relate an unobserved trait to fitness, but more narrowly on those that do so in such a way that the trait represents an optimal solution to a fitness-related problem, we can expect to find ourselves exploring the most fertile region in the space of possible hypotheses.

In the present case, the central hypothesis grew out of the observation that sequence representations that were optimal for performing insertion operations would also be capable of capturing the basic facts of constituency. Exploring this hypothesis in more detail led to the surprising conclusion that a representation that was optimal for performing insertion operations would also be capable of capturing long-distance dependencies like those that are traditionally analysed in terms of the movement of constituents in derivations.

This result presents a challenge to Chomsky's (1995, 1999, 2000a) view that movement is a kind of imperfection motivated by constraints external to the computational system:

> Speculations about [movement] invoked considerations of language use: facilitation of parsing on certain assumptions, the separation of theme-rheme structures from base-determined semantic (θ) relations, and so on. Such speculations involve "extraneous" conditions ... imposed on [the computational system] by the ways it interacts with external systems. That is where we would hope the source of "imperfections" would lie, on minimalist assumptions (Chomsky, 1995: 317).

If movement is intrinsic to constituent structure representations, a language faculty that lacked this phenomenon would not be any simpler or more elegant. Under the present analysis, movement is possible, not because the language faculty has been augmented with externally motivated mechanisms that enable it, but because it lacks constraints that would prevent it.[55]

In terms of the debate over derivational versus representational approaches to Minimalism (briefly reviewed in §2.5 of chapter 2), the current approach certainly comes down on the side of representations. In section 6.3.3, I demonstrated that it is possible to apply the representational approach developed here to phenomena that are usually captured in terms of derivations. The current approach was compared directly with Stabler's (1997) Minimalist formalism, and found to be sufficiently powerful to capture all of the same phenomena, but without requiring the four-way distinction he made between base, select, licensee and licensor features, and without needing to

[55] This point was foreshadowed in chapter three (§3.3.1) in relation to the problem of what does and does not require an evolutionary explanation.

specify an ordering for the formal features of lexical entries. Following the conclusions of chapter five, lexical entries simply contain an unordered set of formal features that are either positively or negatively specified.

Just as each lexical entry contains an unstructured set of formal features, the full constituent structure of a sentence is represented as an unstructured set of pairs. The use of these set-based representations may prove simpler to reconcile with neural representations than more traditional approaches.

If constituent structure representations were exapted, then we might expect to find the same representational mechanisms in other systems of the brain that are used for sequencing, and if properties like long-distance dependencies are inherent to these representations, then we should expect to find them in these other systems too. We should also expect to observe the triangle structure that gives rise to the specifier-head-complement structure at the level of macro tokens. As reviewed in chapter four (§4.2.3), the onset, nucleus and coda of syllables possess an analogous structure to that of the specifier, head and complement of phrases. Long-distance dependencies also appear to be attested in the phonology of some languages (Carstairs-McCarthy, 1999: 138f), which is at least suggestive that phonology is one such system.

Not all linearization sets correspond to legitimate trees. Those that are ruled out are ruled out by the constraints in (14). These are all motivated to allow the tokens in the tree to be linearized. The *non-circularity constraint* (14a) prevents contradictions arising in which one token both precedes and follows another. The *single parent constraint* (14b) is necessary because no order could be specified for unconnected tree fragments and no order is guaranteed to be specifiable for tokens dominating different parents of the same node. The *triangle constraint* (14c) is necessary because no order could be specified between triangles that share their top vertices. Since all of these constraints are motivated by linearization, they all serve the single overarching

constraint in (40), although there are a few special cases of structures that would still be linearizable without violating these constraints.

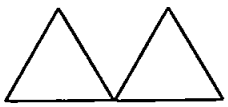
40. The linearization set must specify a consistent linear order for all of the tokens it references.

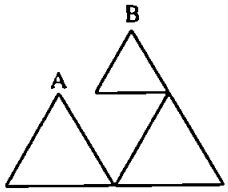
This constraint requires a definition of precedence, which can be given as in (41) in terms of command (21) and dominance (22),. the latter itself also being ultimately defined in term of command.

41. α precedes β iff

α commands $\beta \vee \exists\gamma$ (γ dominates $\alpha \wedge \neg$(γ dominates β) $\wedge$ γ commands β)

The constraints in (14) can be viewed as a particular solution to satisfying the constraint in (40). It may be more computationally tractable to check for conformity to certain constraints that will always imply linearizability than to check for linearizability itself, even if this is at the expense of ruling out a few special cases such as the structure in (42), which violates the *single parent constraint*. However, it is not easy to build on structures like (42) without immediately leading to violations of linearizability. For instance, the linear order of tokens A and B cannot be determined in the structure in (43).

42. 

43. 

To put the results of this chapter in perspective, it follows from the constraints in (14) alone that that valid linearization sets will be those that exhibit the properties listed in (44) below.

44. a. nested constituent structure

b. the specifier-head-complement structure of phrases

c. long-distance dependencies that are in a command relation such as those implicated in movement, anaphor binding and quantifier scope effects

d. theta/case interactions

This is a remarkable set of features to arise out of such simple constraints.

7

Discussion

If inquiry into the evolutionary functions of language universals is to proceed, it must do so without the luxury of fossil evidence and, for properties that are unique to human communication, without comparative evidence of the kind that Darwin utilised to study variation in the finches of the Galapagos Archipelago. Because of this, research in this field has to take full advantage of what little evidence is actually available. I have argued that optimality considerations are an important source of evidence that has been mostly overlooked in studies of language evolution. This led to the development of the optimality diagnostic presented in chapter three, which was applied to two different empirical questions in chapters five and six. The methodological and empirical contributions arising from this work are summarised in what follows. Some remaining issues and future challenges are then discussed.

7.1 Methodological contributions

The present thesis is sympathetic to many of the criticisms that have been levelled at selectionist accounts of language evolution. Chapter three devoted considerable space to these objections, but nevertheless concluded that none of them are in fact fatal. It is clear that many researchers hold naïve assumptions about the explanatory role of natural selection in language evolution, but with these weaknesses highlighted, a more principled method for evaluating claims about selective functions was developed, which centres on the expectation that natural selection will lead to improbable optimality. These points are summarised in more detail below.

7.1.1 The lack of non-selectionist categories of explanation

Stephen Jay Gould and others (Gould & Lewontin, 1979; Gould & Vrba, 1982; Gould, 1991; Lightfoot, 2000; Piattelli-Palmarini, 1989; Uriagereka, 1998) have argued that there are other forces that determine the properties of organisms aside from natural selection, but almost all of the forces these authors discuss are not mutually exclusive alternatives to it at all, but are either the indirect results of natural selection or the physical constraints acting on it. The only mutually exclusive alternative to explanations in terms of natural selection is genetic drift between different variants of equal fitness. Explanations in terms of drift are limited to the emergence of traits that do not alter fitness, but in terms of structure, genetic drift increases entropy and thereby leads to the deterioration of vestigial traits which have lost their functions and which are therefore no longer kept in check by natural selection. To describe genetic drift as a source of creativity would be misleading at best.

Gould and his colleagues have succeeded in raising a number of important issues, even if they have incorrectly presented them as an anti-selectionist critique. What they have highlighted instead are that selectionist explanations cannot assume that all traits are selected independently of one another (§3.1.1), that their current functions are the same as those they were selected for (§3.1.2), that all conceivable variants of a trait are physically attainable (§3.1.3), that elegant structures can emerge as a result of very simple developmental processes (§3.1.4), and that some properties emerge through other kinds of optimisation such as learning (§3.1.5) and cultural evolution (§3.1.6). I argued in chapter three that none of these issues, as they apply to a given trait, would rule out the possibility that it was refined under genetic selection. More will be said about the interaction between genetic selection and other optimising processes in section 7.3.1.

7.1.2 The irrelevance of communicative functions

If properties of language are adaptations, it is generally assumed they must be adaptations for communicative functions. This seems so obvious that it is rarely (if ever) even acknowledged as an assumption. Hauser *et al.* (2002: 1574) assume this for instance, when they express doubts that the evolution of syntactic universals can be exclusively explained in terms of adaptation, noting that many such constraints have a "tenuous connection to communicative efficacy", but even if a given universal is unrelated to communicative efficacy, it could still be an adaptation for a non-communicative function. It may be an adaptation that improves the efficiency of representations or computations for instance. An adaptive change of this kind could occur without having any consequences for expressiveness at all, while allowing the language faculty to operate at lower costs with respect to metabolic energy or other neural resources. Other properties may relate to communicative efficacy indirectly, much as a button on a winter coat contributes to the overall function of keeping its wearer warm, but only via the more specific function of fastening it closed, the fastening function being much more relevant for explaining its peculiar design characteristics. The design of coat buttons has, at best, a "tenuous connection" to thermal functions. In the case of language, there may be syntactic universals that have a "tenuous connection to communicative efficacy" for much the same reason, having been selected for more specific functions which account for their design properties more directly.

Initially, it may seem surprising, but adaptive explanations appear to be much more forthcoming if we drop the assumption that universals have functions that relate directly to communication. Chapters five and six explored adaptive hypotheses of this kind. In chapter five, I argued that closed-class items exist, not for any communicative purpose, but merely because their existence economises the representation of the

lexicon. Likewise, in chapter six, I argued that constituent structure representations were selected for being able to perform insertions into sequences.

7.1.3 The irrelevance of design complexity

At least since Dawkins (1986), the argument from the appearance of design has been borrowed from theology and applied to biology as evidence of selection for particular functions. To the extent that *optimality* considerations have been used to make such arguments, they have been used virtually interchangeably with *complexity*, but there are important differences. While complex adaptations are the best illustrations of the power of natural selection, natural selection is also capable of stripping away structure to sculpt features, much as it does in development via apoptosis (for instance, in killing off cells to make gaps between fingers in the development of an embryo). The loss of body hair and limb structures in exclusively aquatic mammals appears to be an example of this. Adaptive changes that reduce complexity (in the sense of removing structure) evidently occur, so there is no reason why it couldn't craft features that are non-redundant and elegant. These structures would not be more complex than those they evolved from, but we should nevertheless expect them to be *optimal* within the space of local variation for serving the function for which they were selected. In the case of aquatic mammals, what is optimised is presumably the drag associated with swimming, which would be less for a smooth and streamlined body. The thermal functions of hair that apply to mammals out of water are largely ineffective under water so are achieved by other adaptations such as increased levels of subcutaneous fat (Morris, 1967).

7.1.4 The optimality diagnostic

In chapter three, I introduced a diagnostic, adapted from Parker and Maynard Smith (1990), for identifying selective functions. This diagnostic dispenses with the notion of design complexity, being founded instead on the assumption that natural selection will produce traits that occupy local optima in the space of possible variants. It follows from

this that the only functions that could have been relevant for the selection of a given trait are those that describe fitness landscapes that have a local peak at the point the trait occupies.

The reliability of the diagnostic depends on the accuracy with which the local space of possible variants, which Parker and Maynard Smith call the *strategy set*, can be determined. It also relies on the ability to relate the hypothesised function to fitness, which is what excludes functions from being stipulated in a post hoc way. We should also be more confident that a hypothesised function had relevance for selection if only a small proportion of the strategy set exhibits optimality with respect to it. Because none of these things are trivial to determine, uncertainty about any of them will affect the confidence we should have in the result of the diagnostic as a whole. The diagnostic will not therefore always provide a crisp delineation between functions that were and were not relevant for selection. Even so, it is adequate for comparing competing theories of function and arguing from the best explanation.

In addition to claims about adaptation, the optimality diagnostic can be used to assess claims relating to exaptation by providing evidence that a given trait is improbably optimal for survival-related functions that it no longer serves.

Although the use of optimality as an explicit diagnostic tool is essentially novel within studies of language evolution, optimality considerations have previously been used to infer selective functions in other areas of evolutionary biology (Parker & Maynard Smith, 1990; Orzack & Śober, 2001).

7.1.5 A narrower view of perfection for the Minimalist Program

For a gold prospector, the observation that gold deposits are often found in veins of quartz suggests a strategy for how one should proceed with the search. Obviously, one should focus on rocks rich in quartz. Likewise for a scientist, it is sensible to focus

one's search for good theories on those that exhibit a regularity that is often observed to hold of good theories generally. This is the logic of the Minimalist Program, which arose from the observation that the principles of the computational system appear to be 'perfect' in some sense. On the basis of this observation, it made sense to direct inquiry towards theories that suppose this perfection is a general feature of the computational system.

A prospector needn't know why quartz and gold are found together for his strategy to work and the same is true for the Minimalist. However, a strategy may work better if the physical basis of the association is understood. In the case of the prospector, this knowledge may allow him to focus his search on the quartz that is most likely to have been deposited by processes of the relevant kind. Likewise, it may be possible to clarify what type of perfection we should expect to find in the computational system by gaining some understanding of why it is there.

In chapter three (§3.3.1), I argued that the type of 'perfection' that the computational system exhibits is entirely compatible with selection for streamlining in evolution. Specifically, I challenged Chomsky's (1995) assertion that redundancy is an expected property of a biological system by arguing that it will only be favoured if it overrides competing design priorities for factors such as efficiency in representations and computations. The notion of 'perfection' can of course be interpreted in terms of these priorities too, but if the computational system were perfect according to other criteria that happen to be incompatible with fitness-related functions, then this perfection could not have been directly selected for in evolution. Hauser, *et al.* (2002) may be right about the lack of a relationship between properties of grammar and communicative functions, but the properties typically cited as evidence of the perfection of language (economy principles, etc.) appear to be entirely compatible with natural selection for non-communicative functions. If natural selection is indeed the basis of

this 'perfection', then the search for explanatory theories should focus only on the kind of perfection that is compatible with optimisation under its influence.

Optimality considerations have been applied in the current thesis in two distinct ways. We have seen that they can be used to focus inquiry into selective functions by only entertaining functional hypotheses for which the observed trait is optimal. We have also seen that optimality considerations can be used to inquire into the *nature* of systems by only proposing the existence of traits that would serve fitness-related functions optimally. Chapter five was principally a case study in the former and chapter six was principally a case study in the latter, although neither was entirely one or the other.

To contrast the approach of chapter six with inquiry within the Minimalist Program, it is instructive to ask what would compel a Minimalist to consider the kind of perfection entertained there as a basis of a theory of representations. Without linking perfection to natural selection, there is no compulsion to entertain an exaptive hypothesis, and hence no particular compulsion to link insertion operations on sequences with constituent structure.

7.2 Empirical contributions

7.2.1 Closed-class items and the lexicon

In chapter five, I presented a theory of the selective function associated with closed-class items, where I argued that they serve to minimise the representational burden of the lexicon by eliminating redundancy. This is achieved by offloading what would be redundantly repeated feature content contained in the open-class items to the small set of closed-class items, which then take on the role of mediating dependencies between the open classes in syntactic configurations. By encapsulating open-class items, closed-class items also avoid the need to have different words with the same meaning for

different syntactic positions within sentences (e.g., having a different word to express a given meaning when used as a subject as opposed to an object, etc.). A lexicon that reduced redundancy of these kinds would presumably place less of a burden on metabolic and other neural resources, which we should expect natural selection to favour when all else is equal. Alternative languages are imaginable in which closed classes do not exist, or if they do exist, they exist without serving to mediate dependencies between open-class items. These would be suboptimal relative to the observed trait.

A number of interesting findings follow when examining the consequences of the theory. In particular, it appears that properties like case, agreement and the requirement that sentences have subjects are expected consequences of an optimised lexicon, the theory thereby relating these properties to natural selection for the first time. It also motivates the view that language variation is confined to parameters associated with closed-class items, in turn explaining why parameter conflicts fail to arise in bilingualism.

7.2.2 Phrase structure and sequences

In chapter six, I presented a theory of the selective function associated with constituent structure, arguing that this was a by-product of selection for performing efficient insertion operations into sequences. A number of conceivable sequence representations are described to populate a speculative strategy set, each of which is evaluated with respect to the efficiency of insertions. These include methods in which order is encoded using indexes that encode the absolute order of each token in the sequence, and representations that encode relative order using pair- and triple-based representations. The optimal representation in the strategy set allows an insertion to be made with a single operation and without introducing redundancy. A surprising feature of this representation is that it also allows nested constituent structure to be represented. This

suggests that constituent structure may have arisen under selection for performing efficient insertions into sequences and later co-opted for their use in language. Some further consequences of the theory are explored by looking at a lower-level description of the same kind of representation in which dominance relations are inferred from pair-based representations that form triangles. The ability to encode long-distance dependencies such as those associated with movement phenomena follows from the use of the pairs to encode the structure of dominance and precedence relations in the structure meaning that no additional theoretical apparatus is required to account for it. The representation of movement phenomena is thereby unified with the representation of constituent structure itself – a striking result. Other properties of grammar also follow such as the specifier-head-complement structure of phrases, which arises from the way triangles can share vertices, and the interaction between case and theta assignment, which also arises from the triangular structure. These findings bring together a surprising array of phenomena, reinforcing its correctness as the representational basis of syntactic structures, while also providing an excellent example of how optimality considerations can be useful for constructing hypotheses not just about the evolutionary functions of systems, but also their nature.

7.3 Remaining questions

7.3.1 Alternative sources of fit

Kirby (1999) is right to warn us that optimising processes that rest atop the biological substrate such as learning, cultural evolution and planning processes can produce a fit between traits and functions that is difficult to distinguish from the fit produced by natural selection in the genetic domain. To distinguish these sources of optimality, we need to look at what we expect from each.

When developmental processes occur under environment conditions that our ancestors were never exposed to, such as the zero gravity conditions of orbital space flight, we shouldn't necessarily expect the outcomes to be fitness-enhancing from a genetic point of view, but in some respects they should be. No environment will be exactly like those our ancestors were exposed to in detail, but the typical features of the environments our ancestors experienced are likely to have left their mark on our design. Some aspects of the environment, such as the laws of physics, are completely invariant, so a learning or developmental process that relied only on regularities of this kind would tend to succeed everywhere. For instance, a learning process that generally seeks to associate causes with effects or minimise certain kinds of errors between predictions and observations would always tend to produce a fit between traits and functions that we should expect to be fitness-enhancing. In such cases, the functions being optimised would be of a very general nature, applying as much to life in an orbiting spacecraft as to life on the ground. To the extent that more specific functions would be optimised by these processes, this is most likely to occur when they are special cases of the more general function in the way that the ability to tie a knot in a piece of string is a special case of the ability to tie knots generally.

Under typical circumstances, we should expect the effects of these optimising processes to be aligned with genetic fitness since a process that optimised a function that was generally maladaptive to perform would tend to be selected out. When looking at the evolution of these processes, we therefore need to look at what would constitute the optimal optimiser. To apply the optimality diagnostic correctly, we'd have to populate the strategy set with variations on the existing optimiser (or concomitant mechanisms) rather than the space of solutions the optimiser itself explores. It may of course require careful analysis to separate one from the other.

Many aspects of our environment are adapted to us rather than the other way around. This phenomenon is particularly evident in human artefacts and can be understood in terms of cultural evolution. To borrow an example from Gould and Lewontin (1979), spectacles were designed for noses rather than the other way around. The optimality diagnostic will help identify when an optimising process has occurred, but not necessarily which aspect of the system was optimised to fit which. In the case of human artefacts like spectacles, we have historical evidence to decide the issue, but this is not always available. Deacon (1997), Kirby (1999) and others have raised the possibility that languages, as cultural entities, have also evolved, at least to some extent, to fit the constraints of the language acquisition device rather than the other way around. Evidence from creolisation can bear on this issue by telling us which properties are in fact present from the very birth of a language and which emerge later. We can also expect both the strategy sets and the functions that are optimised to be different for different optimising processes. For instance, the functions cultural evolution will tend to optimise are things like transmission fidelity from person to person, and this optimisation will occur even when the changes are fitness-neutral from a genetic point of view. Hence, the functions that are optimised can be used to distinguish one process from the other under certain circumstances. The results of the optimality diagnostic will have to be supported by evidence like this before it can be concluded that the source of the optimality is genetic evolution rather than something else.

7.3.2 Optimality and stability

The view of optimality presented in this thesis is something of a simplification. I have used the term *optimum* in essentially the sense of what would be called a *stable point attractor* in dynamical systems theory, meaning that all possible solutions in the neighbourhood of the optimum will tend to migrate towards it over time and stay there once reached, but this assumes that the relevant species is evolving to fit a static niche.

In many cases, the environment of one genetic variant will include other genetic variants that are co-evolving with it in such a way that they will affect each other's chances of survival. Competitive interactions between these different variants will lead to complex evolutionary dynamics such as 'arms races' of the kind that occur between predators and prey in which each innovation in predation is matched with an innovation for evading capture in the prey species. This is a runaway process that will only terminate once fundamental physical limitations are reached or an extinction event occurs. In other cases, the interactions between variants will be such that stability will only arise when certain combinations of variants are present in the system. Maynard Smith (1982) labelled these *evolutionarily stable strategies*. This kind of stability is only possible when variants occupy local optima on their respective fitness landscapes, but since their fitness landscapes were changing as they co-evolved, each variant will come to rest atop a peak that it may never have actually scaled. When applying the optimality diagnostic to traits for which competitive interactions of this sort are relevant, we cannot simply model natural selection as optimising a variant to fit a static niche. The dynamics of the whole system will have to be taken into account instead of just a part of it, and more sophisticated models of stability will need to be applied. Modelling the niche as static is nevertheless appropriate for the purposes of the current thesis, since we have no reason to expect that the cost, for one genetic variant, of lexical representations or of insertions into sequences would depend on the qualities possessed by other genetic variants in its environment.

In constructing a strategy set, we also need to take into account that optimality will not always imply stability so that even if a given solution is the best in the space of local variants, nearby solutions may have a tendency to migrate away from it over time. Solutions of this kind would be *unstable point attractors* and we should not expect evolution to produce them even if they are locally optimal. The appreciation of such

possibilities would lead to a more sophisticated diagnostic in which the role of optimality is replaced by stable attractors of various kinds.

7.4 New horizons

Many of the questions that arise in the study of language evolution are not yet clearly formulated. The present work has attempted to apply corrective lenses to at least some of these questions, refocusing and recasting them. It also responded to a challenge presented by Botha (2003) to provide a restrictive theoretical framework that would allow inquiry into this area to proceed with rigor even in the face of severe limitations on the evidence available from traditional sources such as the fossil record and comparative studies. The result was the optimality diagnostic.

The optimality diagnostic has been applied to two questions reported in this thesis. There are of course, many other aspects of language that it may be useful to apply it to. A natural extension would be to look at how optimality principles can guide inquiry into the nature of language acquisition, where it may prove fruitful to examine theories in which this process is carried out in an optimal way. Infants may for instance attempt to assign features such that they are satisfied as locally as possible, a strategy that may be motivated by memory limitations and one which would explain the 'starting small' results of Elman (1993), who found that his neural networks were best able to acquire syntactic structures when the length of visible dependencies was initially highly constrained (see §4.1.5). The theory of representations presented in chapter six is also explicit in a way that lends itself well to computational implementations which would help to clarify many of the details.

The optimality diagnostic is also generalisable to other traits involving behaviour and soft-tissues, which are difficult to study from an evolutionary perspective because they don't fossilise. It would also be useful to study species-specific properties

which cannot benefit from the comparative method. Candidates would perhaps include aspects of musical ability, the capacity for religion and consciousness.

References

Abney, S. 1987. *The English noun phrase in its sentential aspects*, unpublished diss., MIT Cambridge, MA.

Abouheif, E. & Wray, G.A. 2002. Evolution of the gene network underlying the wing polyphenism in ants. *Science, 297*: 249-252.

Ackley, D.H. & Littman, M.L. 1994. Altruism in the evolution of communication. In R. Brooks & P. Maes (Eds.), *Proceedings of the Fourth Artificial Life Workshop.* Cambridge, MA: MIT Press.

Akmajian, A. & Heny, F. 1975. *An introduction to the principles of transformational syntax*. Cambridge, MA: MIT Press.

Arbib, M.A. 2002. Computational models of monkey mechanisms for the control of grasping: Grounding the mirror system hypothesis for the evolution of the language-ready brain. In A. Cangelosi & D. Parisi (Eds.).

Baddeley, A.D. 1999..*Essentials of human memory*. Hove: Psychology Press.

Barrett, M. 1999. *The development of language*. East Sussex, UK: Psychology Press.

Batali, J. 1995. *Small signalling systems can evolve in the absence of benefit to the information sender*. Unpublished manuscript.

Benedict, H. 1979. Early lexical development: Comprehension and production. *Journal of Child Language, 6*: 183-200.

Bickerton, D. 1991. On the supposed "gradualness" of creole development. *Journal of Pidgin and Creole Languages, 6*: 25-58.

Bickerton, D. 1977. Pidginization and creolization: Language acquisition and language universals. In A. Valdman (Ed.) *Pidgin and creole linguistics*. Bloomington: Indiana University Press.

Bickerton, D. 1990. *Language and species*. Chicago: University of Chicago Press.

Bickerton, D. 1998. Catastrophic evolution: the case for a single step from protolanguage to full human language. In J.R. Hurford, M. Studdert-Kennedy & C. Knight. (Eds.) *Approaches to the evolution of language.* Cambridge, UK: Cambridge University Press.

Bickerton, D. 1999. How to acquire language without positive evidence: What acquisitionists can learn from creoles. In M. DeGraff (Ed.).

Borer, H. 1984. *Parametric syntax.* Dordrecht: Foris.

Botha, R.P. 2003. *Unravelling the evolution of language.* Amsterdam: Elsevier.

Braine, M.D.S. 1963. The ontogeny of English phrase structure: The first phase. *Language, 39*: 1-14.

Braine, M.D.S. 1971. On two types of models of the internalization of grammars. In D.I. Slobin (Ed.), *The ontogenesis of grammar.* New York: Academic Press.

Bresnan, J. 2001. *Lexical Functional Syntax.* Blackwell.

Brody, M. 1995. *Lexico-logical form.* Cambridge, MA: MIT Press.

Brody, M. 2000. Mirror theory: Syntactic representation in perfect syntax. *Linguistic Inquiry, 31*, 29-56.

Brown, K. & Miller, J. 1991. *Syntax: A linguistic introduction to sentence structure* (2nd Ed). London: Routlege.

Brown, R. & Fraser, C. 1963. The acquisition of syntax. In C.N. Cofer & B. Musgrave (Eds.) *Verbal behavior and learning: Problems and processes.* New York: McGraw-Hill.

Brown, R. & Hanlon, C. 1970. Derivational complexity and order of acquisition on child speech. In J. Hayes (Ed.), *Cognition and the development of language.* New York: Wiley.

Brown, R. 1973. *A first language: The early stages.* London: George Allen & Unwin.

Calvin, W. 1983. A stone's throw and its launch window: timing precision and its implications for language and hominid brains. *Journal of Theoretical Biology, 104*: 121-135.

Calvin, W.H. & Bickerton, D. 2000. *Lingua ex machina: Reconciling Darwin and Chomsky with the human brain*. Cambridge, MA: MIT Press.

Calvin, W.H. 1996. *The cerebral code: Thinking a thought in the mosaics of the mind.* Cambridge, MA: MIT Press.

Cangelosi, A. & Parisi, D. (Eds.) (2002). *Simulating the evolution of language*. London: Springer-Verlag.

Cangelosi, A. & Parisi, D. 1998. The emergence of a 'language' in an evolving population of neural networks. *Connection Science, 10*: 83-97.

Carstairs-McCarthy, A. 1999. *The origins of complex language: An inquiry into the evolutionary beginnings of sentences, syllables and truth*. Oxford: Oxford University Press.

Catania, A.C. 1990. What good is five percent of a language competence? *Behavioral and Brain Sciences, 13*: 729-731.

Chaudenson, R. 1979. *Les créoles françias*. Paris:L'Harmattan.

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

Chomsky, N. 1970. Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in English transformational grammar*. Waltham, MA: Ginn.

Chomsky, N. 1972. *Language and mind: enlarged edition*. New York: Harcourt Brace Jovanovich.

Chomsky, N. 1980. *Rules and representations*. New York: Columbia University Press.

Chomsky, N. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. 1986. *Knowledge of language: Its nature, origin, and use*. New York: Praeger.

Chomsky, N. 1991. Some notes on economy of derivation and representation. In R. Freidin (Ed.), *Principles and parameters in comparative grammar*. Cambridge, MA: MIT Press.

Chomsky, N. 1993. A minimalist program for linguistic theory. In K. Hale & S.J. Keyser (Eds.), *The view from building 20*. Cambridge, MA: MIT Press.

Chomsky, N. 1995. *The minimalist program*. Cambridge, MA: MIT Press.

Chomsky, N. 1999. *Derivation by phase*. MIT Occasional Papers in Linguistics 18.

Chomsky, N. 2000a. Minimalist inquiries: The framework. In R. Martin, D. Michaels, & J. Uriagereka (Eds.), *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*. Cambridge, MA.: MIT Press.

Chomsky, N. 2000b. New horizons in the study of language and mind. Cambridge, UK: Cambridge University Press.

Chomsky, N. 2002. *On nature and language*. Cambridge, UK: Cambridge University Press.

Corballis, M.C. 1991. *The lopsided ape: Evolution of the generative mind*. New York: Oxford University Press.

Croft, W. 1990. *Typology and universals*. Cambridge, UK: Cambridge University Press.

Cziko, G. 1995. *Without miracles: Universal selection theory and the second Darwinian revolution*. Cambridge, MA: MIT Press.

Dawkins, R. 1976. *The selfish gene*. Oxford: Oxford University Press.

Dawkins, R. 1982. *The extended phenotype*. Oxford University Press: Oxford.

Dawkins, R. 1986. *The blind watchmaker*. Harlow: Longman.

Dawkins, R. 1991. *The blind watchmaker* (2nd ed). London: Penguin.

Dawkins, R. 2006. *The God delusion*. London: Bantam Press.

de Boer, B. 1997. Generating vowels in a population of agents. In P. Husbands & I. Harvey (Eds.), *Proceedings of the Fourth European Conference on Artificial Life*. MIT Press.

Deacon, T.W. 1992. Brain-language coevolution. In J.A. Hawkins & M. Gell-Mann (Eds.) *The evolution of human languages: Proceedings of the workshop on the evolution of human languages held August, 1989 in Santa Fe, New Mexico*. New York: Addison-Wesley.

Deacon, T.W. 1993. Confounded correlations, again. *Behavioral and Brain Sciences, 16*, 698-699.

Deacon, T.W. 1997. *The symbolic species: The co-evolution of language and the human brain*. London: Penguin.

DeGraff, M. (Ed.) 1999. *Language creation and language change: Creolization, diachrony, and development*. Cambridge, MA: MIT Press.

DeGraff, M. 1999. Creolization, language change, and language acquisition: A prolegomenon. In M. DeGraff (Ed.).

Dennett, D.C. 1991. *Consciousness explained*. London: Penguin.

Dennett, D.C. 1995. *Darwin's dangerous idea: Evolution and the meanings of life*. London: Penguin.

Dennett, D.C. 2003. *Freedom evolves*. London: Penguin.

Dromi, E. 1999. Early lexical development. In Barrett, M. (Ed.).

Dunbar, R.I.M. 1993. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences, 16*, 681-735.

Dunbar, R.I.M. 1996. *Grooming, gossip and the evolution of language*. London: Faber and Faber.

Edelman, G.M. 1987. *Neural Darwinism: the theory of neuronal group selection*. New York: Basic Books.

Elliot, A.J. 1981. *Child language*. Cambridge, UK: Cambridge University Press.

Elman, J.L. 1993. Learning and development in neural networks - The importance of starting small. *Cognition, 48*, 71-99.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., & Paabo, S. 2002. Molecular evolution of FOXP2, a gene involved in speech and language. *Nature, 418*, 869-872.

Fenson, L., Dale, P.S., Reznick, J.S., Bates, E., & Thal, D. 1994. *Variability in early communicative development*. Chicago: University of Chicago Press.

Fitch, W.T., & Reby, D. 2001. The descended larynx is not uniquely human. *Proceedings of the Royal Society, Biological Sciences, 268*, 1669-1675.

Foley, W.A. 1997. *Anthropological linguistics: An introduction*. Oxford: Blackwell.

Frank, R. & Kuminiak, F. 2000. Primitive asymmetric c-command derives X'-theory. In *Proceedings of NELS 30*. Amherst, MA: GLSA.

Fukui, N. 1995. *Theory of projection in syntax*. Standford: CSLI Publications.

Geschwind, N. 1980. Some comments on the neurology of language. In D. Caplan (Ed.), *Biological studies of mental processes*. Cambridge, MA: MIT Press.

Gleason, J.B. (Ed.) 1997. *The development of language* (4th ed.). Boston: Allyn and Bacon.

Gold, E.M. 1967. Language identification in the limit. *Information and Control, 10*, 447-474.

Golinkoff, R.M. & Hirsh-Pasek, K. 1995. Reinterpreting children's sentence comprehension: Toward a new framework. In P. Fletcher & B. MacWhinney (Eds.) *The handbook of child language*. Oxford: Blackwell.

Gopnik, M. & Cargo, M. 1991. Familial aggregation of a developmental language disorder. *Cognition, 39*, 1-50.

Gould, S.J. & Lewontin, R.C. 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, Series B, 205*: 581-598.

Gould, S.J. & Vrba, E.S. 1982. Exaptation: A missing term in the science of form. *Paleobiology, 8*: 4-15.

Gould, S.J. 1977. *Ontogeny and phylogeny*. Cambridge, MA: Harvard University Press.

Gould, S.J. 1991. Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues, 47*: 43-65.

Gould, S.J. 1997. Evolution: The pleasures of pluralism. *New York Review of Books.*

Gould, S.J. 2002. *The structure of evolutionary theory*. Cambridge, MA: Harvard University Press.

Greenberg, J.H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In J.H. Greenberg (Ed.) *Universals of language* (2nd Ed.). Cambridge, MA: MIT Press.

Greenfield, P.M. 1991. Language, tools, and brain: The ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and Brain Sciences, 14*: 531-51.

Gregory, T.R., 2002. Genome size and developmental complexity. *Genetica, 115*: 131-146.

Grimshaw, J. 2005. *Words and Structure*. Stanford, CA: CSLI Publications.

Haegeman, L. 1994. *Introduction to Government and Binding Theory* (2nd ed.). Oxford: Blackwell.

Hauser, M., Gardner, L., Goldberg, T. & Treves, A. 1993. The functions of grooming and language: The present need not reflect the past. *Behavioral and Brain Sciences, 16*, 706-707.

Hauser, M.D., Chomsky, N., & Fitch, W.T. 2002. The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*: 1569-1579.

Heyes, C.M. 1998. Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21*: 101-134.

Hirsh-Pasek, K. & Golinkoff, R.M. 1993. Skeletal supports for grammatical learning: What the infant brings to the language learning task. In C.K. Rovee-Collier (Ed.) *Advances in infancy research* (Vol. 10). Norwood, NJ: Ablex.

Hofstadter, D.R. 1979. *Gödel, Escher, Bach: An eternal golden braid.* London: Penguin.

Horgan, J. 1996. *The end of science: Facing the limits of science in the twilight of the scientific age.* New York: Addison Wesley.

Hurford J.R. & Heasley, B. 1983. *Semantics: A coursebook.* Cambridge, UK: Cambridge University Press.

Hurford, J.R. 1991. The evolution of the critical period for language acquisition. *Cognition, 40*:159-201.

Hurford, J.R. 1999. The evolution of language and languages. In R. Dunbar, C. Knight & C. Power (Eds.) *The evolution of culture.* Edinburgh: Edinburgh University Press.

Jackendoff, R. 1977. *X-bar Syntax.* Cambridge, MA: MIT Press.

Jacob, F. 1977. Evolution and tinkering. *Science, 196*: 1161-1166.

Jerne, N.K.1955. The natural selection theory of antibody formation. *Proceedings of The National Academy of Sciences, 41*: 849-857.

Kaplan, R.M. & Bresnan, J. 1982. Lexical-functional grammar: A formal system for grammatical representation. In J. Bresnan (Ed.) *The mental representation of grammatical relations.* Cambridge, MA: MIT Press.

Kayne, R. 1984. Connectedness and binary branching. Dordrecht: Foris.

Kayne, R.S. 1994. *The antisymmetry of syntax.* Cambridge, MA: MIT Press.

Keenan, E. & Comrie, B. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8*: 62-100.

Kegl, J., Senghas, A. & Coppola, M. 1999. Creation through contact: Sign language emergence and sign language change in Nicaragua. In M. DeGraff (Ed.).

Kimura, D. 1979. Neuromotor mechanisms in the evolution of human communication. In H. Whitaker & H.A. Whitaker (Eds.) *Current trends in neurolinguistics*. New York: Academic Press.

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

Kirby, S. 2001. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation, 5*: 102-110.

Kuczaj, S.A. 1999. The world of words: Thoughts on the development of a lexicon. In Barrett, M. (Ed.).

Larson, R.K. 1988. On the double object construction. *Linguistic Inquiry, 19*: 335-91.

Lefebvre, C. & Lumsden, J. 1989. Les langues créoles et théorie linguistique. *Revue Canadienne de Linguistique, 34*: 319-337.

Lenneberg, E.H. 1967. Biological Foundations of Language. Wiley.

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

Lieberman P. 1985. On the evolution of human syntactic ability: its pre-adaptive bases – motor control and speech. *Journal of Human Evolution, 14*: 657-668.

Lieberman, P. 1991. Uniquely human: The evolution of speech, thought and selfless behavior. Cambridge, MA: Harvard University Press.

Lightfoot, D. 2000. The spandrels of the linguistic genotype. In C. Knight, M. Studdert-Kennedy & J. R. Hurford (Eds.), *The evolutionary emergence of language: Social function and the origins of linguistic form*. Cambridge, UK: Cambridge University Press.

Maratsos, M. 1988. Crosslinguistic analysis, universals, and language acquisition. In F.S. Kessel (Ed.).

Marcus, G.F. 1993. Negative evidence in language acquisition. *Cognition, 46*: 53-85.

Maynard Smith, J. 1982. *Evolution and the theory of games*. Cambridge, UK: Cambridge University Press.

McCleery, R.H. 1978. Optimal behaviour sequences and decision making. In J.R. Krebs & N.B Davies (Eds.), *Behavioural Ecology*. Oxford: Blackwell Scientific.

Meguire, P.G., 2003. Discovering Boundary Algebra: A Simplified Notation for Boolean Algebra and the Truth Functors. *International Journal of General Systems, 32*: 25-87.

Merton, R.K. 1993. *On the shoulders of giants*. 3rd Edition. Chicago: University of Chicago Press.

Morris, D. 1967. *The naked ape: A zoologist's study of the human animal.* London: Jonathan Cape.

Nakamichi, M, Koyama, N. 1997. Social relationships among ring-tailed lemurs (*Lemur catta*) in two free-ranging troops at Berenty Reserve, Madagascar. *International Journal of Primatology, 18*: 73-93.

Nowak, M.A., Komarova, N.L. & Niyogi, P. 2001. Evolution of universal grammar. *Science, 291*: 114-118.

Oppenheim, RW. 1991. Cell death during development of the nervous system. *Annual Review of Neuroscience, 14*: 453-501.

Orzack, S.H. & Sober, E.R. (Eds.) 2001. *Adaptationism and Optimality.* Cambridge, UK: Cambridge University Press.

Pan, B.A. & Gleason, J.B. 1997. Semantic development: Learning the meanings of words. In J.B. Gleason (Ed.).

Parker, G.A. & Maynard Smith, J. 1990. Optimality theory in evolutionary biology. *Nature, 348*: 27-33.

Pauling, L. 1986. *How to live longer and feel better.* New York: W.H. Freeman and Company.

Phillips, C. 2003. Linear order and constituency. *Linguistic Inquiry. 34*: 37-90.

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.

Piattelli-Palmarini, M. 1990. An ideological battle over modals and quantifiers. *Behavioral and Brain Sciences, 13*: 752-754.

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

Pinker, S. 1994. *The language instinct: The new science of language and mind.* London: Penguin.

Pollard, C. & Sag, I.A. 1994. *Head-Driven Phrase Structure Grammar.* Chicago: University of Chicago Press.

Premack, D. 1986. *Gavagai! or the future history of the animal language controversy.* Cambridge, MA: MIT Press.

Prince, A. & Smolensky, P. 1993. *Optimality theory: Constraint interaction in generative grammar.* Rutgers University Center for Cognitive Science Technical Report 2.

Radford, A. 1988. *Transformational grammar: A first course.* Cambridge, UK: Cambridge University Press.

Ridley, M. 2003. *Nature via nurture: Genes, experience and what makes us human.* London: Fourth Estate.

Rizzi, L. 1997. The fine structure of the left periphery. In L. Haegeman (Ed.), *Elements of grammar*. Dordrecht: Kluwer.

Rizzolatti, G., Fadiga, L., Gallese, V. & Fogassi, L. 1995. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research, 3*: 131-141.

Rosenberger, A.L. & M.E. Strasser. 1985. Toothcomb origins: Support for the grooming hypothesis. *Primates, 26*: 76-85.

Sapir, E. 1921. *Language*. New York: Harcourt, Brace and World.

Saunders, J.W. 1966. Death in embryonic systems. *Science, 154*: 604-612.

Simon, H.A. 1957. Models of man. New York: Wiley.

Sleator, D.D.K. & Temperley, D. 1991. *Parsing English with a Link Grammar*, Carnegie Mellon University Computer Science technical report CMU-CS-91-196, October 1991.

Slobin, D.I. 2002. Language evolution, acquisition, diachrony: Probing the parallels. In T. Givón & B.F. Malle (Eds.), *The evolution of language out of pre-language*. Amsterdam: John Benjamins.

Spencer-Brown, G. 1969. *Laws of Form*. London: Allen & Unwin.

Sperber, D. 1990. The evolution of the language faculty: A paradox and its solution. *Behavioral and Brain Sciences, 13*: 756-758.

Stabler, E. 1997. Derivational minimalism. In C. Retoré (Ed.), *Logical aspects of computational linguistics*. Berlin: Springer-Verlag.

Stoel-Gammon, C. & Menn, L. 1997. Phonological development: Learning sounds and sound patterns. In J.B. Gleason (Ed.).

Tager-Flusberg, H. 1997. Putting words together: Morphology and syntax in the preschool years. In J.B. Gleason (Ed.).

Thompson, D.W. 1961. *On growth and form (abridged addition).* Cambridge, UK: Cambridge University Press.

Tobias, P.V. 1987. The brain of *Homo habilis*: A new level of organization in cerebral evolution. *Journal of Human Evolution, 16*: 741-762.

Tomasello, M. & Brooks, P.J. 1999. Early syntactic development: A construction grammar approach. In Barrett, M. (Ed.).

Turner, H. 2002. An introduction to methods for simulating the evolution of language. In A. Cangelosi & D. Parisi (Eds.).

Uriagereka, J. 1998. *Rhyme and reason: An introduction to minimalist syntax.* Cambridge, MA: MIT Press.

Wilkins, W. & Wakefield, J. 1995. Brain evolution and neurolinguistic preconditions. *Behavioral and Brain Sciences, 18*: 161-182.

Williams, G. 1966. *Adaptation and natural selection.* Princeton, NJ: Princeton University Press.

Zuidema, W. 2003. How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02).* Cambridge, MA: MIT Press.

Appendix A: Publications

The research undertaken during the preparation of this thesis resulted in the following publications. The full text of each is included here for reference.

Turner, H. 2002. An introduction to methods for simulating the evolution of language. In A. Cangelosi & D. Parisi (Eds.), *Simulating the evolution of language*. London: Springer-Verlag.

Cangelosi, A. & Turner, H. 2002. L'emergere del linguaggio. In A.M. Borghi & T. Iachini (Eds.), *Scienze della mente*. Bologna: Il Mulino (in Italian).

Chapter 2

Simulation Methods for the Evolution of Language

Huck Turner

Little is known about how language evolved in our species and we don't even really know what the alternatives are. Without taking a computational approach, a theorist can only guess about what is and is not possible and therefore which assumptions are and are not necessary. The present review aims to provide some perspective on computational modeling in this area in order to get a sense of what approaches are being taken and to categorize and describe the methodologies applied. In the second half, some of these models are then placed in the context of relevant theoretical issues to illustrate how they can and cannot be used to inform the debate.

Computational models have been applied to many different aspects of the evolution of language and in many different ways. The models covered here differ with respect to the linguistic subject they attempt to illuminate and the simulation methods they enlist for the task. In terms of the linguistic subject, the major models fall into a number of categories. Cangelosi and Harnad (in press), Cangelosi and Parisi (1998), MacLennan and Burghardt (1994) and Steels and Vogt (1997) have modeled the emergence of symbols and simple lexicons, while others have concentrated on the emergence of various syntactic properties. These include regular compositionality (Batali, 1998; Kirby & Hurford, 1997; Steels, 1998), recursion (Batali, 2000; Christiansen & Devlin, 1997; Kirby, 1999) and syntactic selection (Cangelosi, 1999). Steels (1998) has also produced a composite model in which both symbols and simple syntax emerge. Others have modeled the self-organization of sound-systems for communication (De Boer, 1997) and aspects of historical change such as the formation of dialects (Livingstone & Fyfe, 1999).

With some perspective on what issues are being explored, the following section will describe the major methodologies being used, explain how they are implemented and provide some examples of their application.

Simulation methods

All of the models discussed here involve populations of communicating agents. What varies between them is the means by which agents come to possess linguistic knowledge and these means are discussed here. Broadly, they include rule-based approaches, neural networks including recurrent neural networks, genetic algorithms and variants thereof, ecological simulations and robotic language games. Some introductory concepts will be covered in this section so it should be accessible to readers from a broad range of disciplines. Those familiar with these concepts can probably skip many of the details, but may nevertheless gain some perspective by following the discussion.

Rule-Based Inference

The term 'rule-based inference' is used here to refer to learning methods that involve symbolic (as opposed to sub-symbolic) manipulations. An example is Kirby's (in press) model of the emergence of stable irregularity in compositional syntax in which he uses a rule-based algorithm to induce and invent rules for relating meanings to signals. The following are some examples of the kinds of rules that Kirby uses to represent this mapping.

(1) a. $S : (a_0, b_0) \rightarrow$ abc
b. $S : (a_0, b_1) \rightarrow$ abd

The terms a_0, b_0 and b_1 are elements of the meaning to be expressed and the strings on the right of the arrow are the signals that are used to express the specified meaning components. These rules do not capture the similarities between the two mappings so rather than have a single rule for each, Kirby's system attempts to generalize. In this process, pairs of non-compositional rules like those in (1) are replaced with equally expressive compositional rules like those in (2) where x and y are variables that stand for elements of meaning.

(2) a. $S : (x, y) \rightarrow A : x\ \mathrm{B} : y$
b. $A : a_0 \rightarrow$ ab
c. $B : b_0 \rightarrow$ c
d. $B : b_1 \rightarrow$ d

Kirby's model involves the transmission of linguistic knowledge from generation to generation in a population of agents. The model demonstrates that given a limited number of opportunities for agents to learn the language of their parents, the languages that evolve exhibit the appearance of design for minimizing errors in transmission from generation to generation by being strongly compositional, and with pressure for shorter strings as well, the occurrence of non-compositional irregularity becomes stable. Kirby introduces a pressure for shorter strings by introducing noise that has the effect of corrupting longer signals more

often than shorter ones, but there are plausibly a number of other factors such as time constraints or least-effort principles that could create a pressure for shorter strings.

Applying rule-based methods can be very good for testing very specific hypotheses. In this case, Kirby's model was designed to test whether certain linguistic properties can emerge as a result of selection pressure on languages themselves rather than their speakers, which reinforces some of Deacon's (1997) ideas on the subject. More will be said about the theoretical context of this work in the second half of this paper and indeed in Kirby and Hurford (this volume).

Neural Networks

Artificial neural networks work in a way vaguely analogous to the neural networks that constitute nervous tissue. They are made up of a number of nodes (also called units or neurons) linked to one another via weighted connections that communicate activation levels between them. Each node has an activation level that is determined either by an external input or as a function of inputs from other nodes. Connection weights can be either excitatory or inhibitory and tuning these weights amounts to learning.

Neural networks are distributed representations meaning that memories are distributed over many nodes. This property has the advantage of making them less vulnerable to noise and localized lesions than rule-based approaches which rely completely on the accuracy of input. Noise or damage will cause a neural network to produce categorizations that are less accurate, but which still serve as approximate solutions. In other words, the quality of solutions degrades gracefully.

Graceful degradation is not a typical feature of rule-based systems. For example, in Kirby's (in press) model (discussed above), the introduction of noise caused longer signals to be corrupted more frequently than shorter ones. This is a fact about rule-based systems which is not generally true of neural networks. Consider the following pair of incomplete signals: 'p#t' and 'el#ph##t'. It is not possible to identify the first example because the information available is not sufficient to narrow it down to a single word. It could correspond to any of *pat*, *pet*, *pit*, *pot* or *put*. Being a longer word, English speakers can narrow the second example down to a specific word without any trouble at all presumably because their memory of it is distributed. Rule-based systems generally do not degrade gracefully so cannot make use of partially specified data. As a consequence, they are just as likely to fail when a small amount of noise is present as they are when a large amount is present.

The capacity to function despite noisy input makes neural networks good pattern classifiers, and it is this ability that Harnad (1990) argues is a basic prerequisite for symbolic communication. Without it we wouldn't be able to distinguish between different symbols or between the different objects and concepts they label.

There are many different neural network architectures, but in models of language evolution, the most commonly used variants are multi-layer feed-forward neural networks and recurrent neural networks (see Figure 2.1).

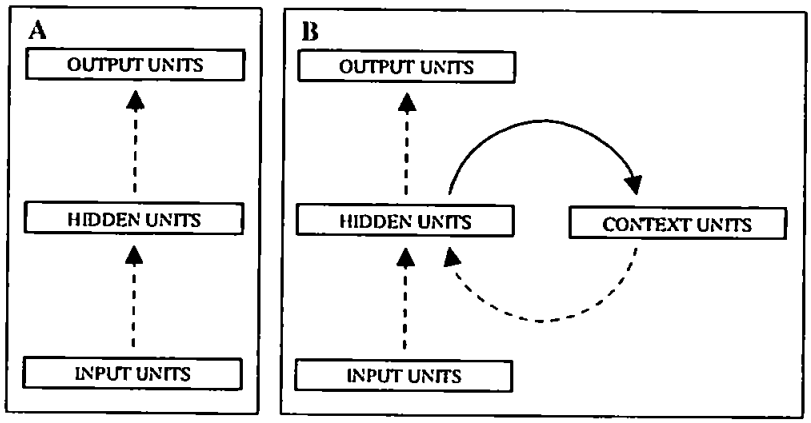


Figure 2.1 (a) A standard feed-forward neural network with one hidden layer, and (b) a recurrent neural network. Dashed arrows indicate sets of trainable connections where each unit in one layer is linked to every unit in the other. The solid arrow in (b) indicates fixed, one-to-one connections responsible for direct copying of activation levels from hidden-layer units to context-layer units.

Multi-Layer, Feed-Forward Neural Networks

As the name suggests, this network consists of a number of layers each of which feeds its unit activations on to the next. The input layer receives its activation values from the researcher or from some other system external to the network. This input will be a raw encoding of some kind of pattern that, when fed through the network, will generate a classification represented as activations of output layer nodes. For instance, a network might be used to generate a classification of speech sounds in terms of phonetic features (at the output layer) using an encoding of their frequency components as input (at the input layer).

The interpretation of inputs and outputs is usually pre-specified by the researcher, but this is not the case for nodes in hidden layers and for many problems no clear interpretation will be possible even after learning due to the distributed nature of the representations. The advantage of this is that the researcher does not have to make unnecessary assumptions about internal representations. A disadvantage is that it can make it more difficult to explain how a model does what it does and hence more difficult to extrapolate from it to the real world.

The most commonly-used algorithm for learning in feed-forward networks is back-propagation (Rumelhart, Hinton & Williams, 1986). In simple terms, it involves comparing actual output activation levels to target activation levels (often called teaching inputs) to calculate an error value for each output node. The error of a node is then propagated backwards through the network to every node in the previous layer (from which it has an input) and is apportioned according to the relative contribution each made to that error. These errors are then used to modify the weights of connections slightly.

The neurological and psychological plausibility of back-propagation is doubtful for a number of reasons. Firstly, it is not clear that real synapses can transmit error backwards. Secondly, we rarely have the opportunity to quantify the errors we make by comparing our output behaviors with correct target behaviors provided by a teacher. Thirdly, unlike learning in the real world, back-propagation typically requires exposure to a vast number of examples.

There are other ways learning can be achieved in a feed-forward architecture. For instance, connection strengths can be evolved using a genetic algorithm (Montana & Davies, 1991), which may or may not be more neurologically plausible. For a selectionist account of learning at the neurological level see Edelman (1987).

RecurrentNeural Networks

Recurrent neural networks (Elman, 1990) are a variation on standard feed-forward networks with the ability to learn temporal sequences. In terms of architecture, the only difference between them is that recurrent neural networks have a set of what are called context units which store the previous activation levels of the hidden units and feed them back as inputs to the hidden layer at the next time-step (see Figure 2.1).

A significant failing of standard feed-forward nets is that the number of inputs that can be presented to the network is fixed. If inputs could be presented to a network sequentially then it would be possible to process data such as words and sentences that can vary in length. This kind of sequential processing is achieved using recurrent neural networks. Elman (1990), who devised them, trained one on grammatical sequences of text (generated using a context-free grammar) using the next word in the sequence as the target output at each time step. After training, the text was presented to the network again and the activation levels in its hidden layer were compared at each time step. The hidden representations for each word clustered (in terms of similarity) into established word classes and subclasses and, for each, the network was able to estimate the approximate likelihood that the next word was a member.

Elman (1990) argues that the ability to predict the category of the next word indicates a knowledge of syntactic structure, but this is far from obvious. It might be fairly easy to predict the class of the next word in some cases, but it should be very difficult in contexts where the next constituent modifies, or is dependent on, subsequent constituents. For example, in (3) the word *quickly* modifies *disappearing* which appears later.

(3) It was quickly disappearing.

In fact, there are very few limitations on what can directly follow *was* in this context. Various nouns, verbs, adjectives, prepositions, adverbs, determiners, degree words, complementizers and infinitival *to* are all acceptable. Some of these alternatives will occur more frequently than others, but that does not make them more grammatical than others.

Constituents are *usually* constrained by their preceding context in English, but the reverse scenario is the norm in head-final languages such as Japanese where

modifiers precede the constituents they modify, objects come before verbs and instead of prepositions there are postpositions. At the very least, this means that language processing should be more difficult if it relies on the ability to make predictions from preceding context.

Batali (1994; 2000) has applied recurrent neural networks to studies of the Baldwin effect and more recently to the evolution of compositional syntax. Christiansen and Devlin (1997) have also applied recurrent neural networks to the evolution of consistent branching in linguistic structures.

Genetic Algorithms

Genetic algorithms apply natural selection to refine solutions progressively. Solutions are encoded as a population of coded strings analogous to chromosomes, which are reproduced to a greater or lesser extent according to a fitness function calculated with respect to how well each satisfies the constraints of a given problem. John Holland (1975) first developed and formalized this approach in the 1960s. In his original formulation he used strings of ones and zeros to represent chromosomes, but encodings involving larger alphabets (more than just ones and zeros), real numbers and tree-like structures are now also used. There has been a proliferation of terminology to describe such variants, but classes that are widely recognized include evolutionary strategies (Schwefel & Rudolph, 1995), genetic programming (Koza, 1990) and evolutionary programming (Fogel, Owens & Walsh, 1966).

The fitness function that is used to select the best solutions is analogous to the environment of a species in that the solutions that proliferate are those that best satisfy its constraints. The fitness function is not in any sense a description of a target solution; it is instead a specification of constraints that a solution must satisfy. The actual form that any solution assumes is no more directed than in the biological domain except in examples used in textbooks where fitness functions are defined in terms of the similarity to a target form, but these are useful only for exposition and are obviously without any practical value (if the target solution is already known, there is no need for it to be evolved). Perhaps as a result of such expository simplifications, misunderstandings abound about the directedness of evolution in genetic algorithms and they have attracted criticism on this basis (e.g. Berwick, 1996).

Genetic algorithms can be viewed at different levels of description as optimizing, search or learning algorithms. They are optimizing algorithms because they are used to maximize the fitness of solutions. They are search algorithms because optimization is a kind of search – a search for solutions with (near) optimal properties. Evolution also resembles operant conditioning insofar as it proceeds by penalizing mistakes and rewarding successes – a correspondence recognized by Skinner (1966, 1981) – so genetic algorithms can also be used as learning algorithms (by optimizing connection weights in a neural network for instance). Computational models that involve both evolution and learning (e.g. Cangelosi & Parisi, 1998) are often constructed using different optimization

algorithms for the two aspects (typically a genetic algorithm for the former and back-propagation[1] for the latter), but there is nothing necessary about this.

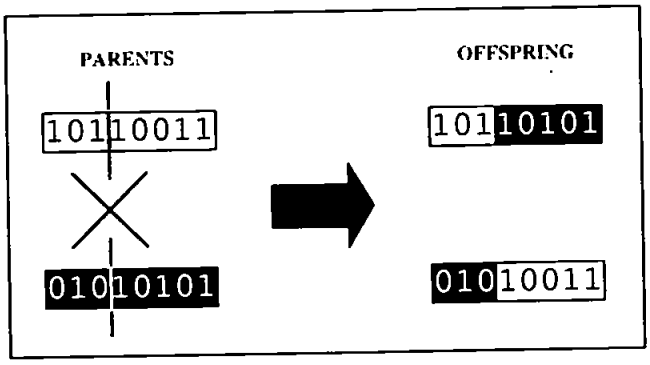


Figure 2.2 The crossover procedure. Two parents are combined to produce two offspring.

As in the biological example, genetic algorithms involve selection, recombination and mutation. The selection procedure scores each string within the population according to what is called the objective function, which it uses to determine its fitness relative to others. Strings are then duplicated with a probability dependent on their fitness to produce a new population. The recombination (or crossover) procedure mates pairs of strings by randomly choosing a crossover point and producing two offspring each sharing features of both parents (as in Figure 2.2). The usefulness of the crossover procedure will depend on the genotype encoding and the problem domain. As in biology, crossover is not strictly necessary for evolution, but organisms that reproduce this way tend to evolve more rapidly than those that reproduce asexually.

The mutation procedure acts with a very low probability to alter features randomly within the genotype. As in biology, mutations are usually deleterious but occasionally produce better solutions than can be generated through crossover alone. To determine a setting for the mutation rate, the need for variation in the population (preventing premature convergence) must be balanced with the need to prevent degrading the population by introducing too many mutants.

In the genetic algorithm literature, individual digits within a genotype encoding are called genes or features. For many problems to which genetic algorithms are applied, this use of gene conflicts with biological usage where genes are (often by definition) regarded as the unit of selection[2] (i.e. the thing that is replicating).

---

[1] Back-propagation is sometimes referred to as 'gradient descent learning' - an elusion to an error landscape (not unlike a fitness landscape) on which learning is viewed as the descent of a solution (i.e. the set of weights in a network) to progressively lower points (representing lower error) on the landscape.

[2] That genes are the unit of selection is the consensus view in biology, yet many researchers will (in the same breath) equate genes with *cistrons* (a length of chromosome coding for exactly one protein molecule). Dawkins (1989) shows that the one will not always

Dawkins (1989:28) for instance, defines a gene as "...any portion of chromosomal material that potentially lasts for enough generations to serve as a unit of natural selection". He argues that a section of chromosome counts as a gene if it leads to the organism being selected over others for possessing certain traits and if it is unlikely to be disrupted by crossover or mutation and thus likely to be passed on to its offspring. A section of chromosome has a higher likelihood of being split up by the crossover procedure if it is very long. Consequently, genes tend to be contiguous or closely-packed segments that are relatively short. This definition corresponds to the concept of a building block in the genetic algorithm literature while the biological counterpart of a digit in the genotype encoding is probably a codon or an individual nucleotide.

In models of language evolution, genetic algorithms have been used to evolve weights in populations of neural networks in the ecological simulations of Cangelosi and Parisi (1998), described below, and in Batali's (1994) study of the Baldwin effect where he used them to evolve weights in recurrent neural networks.

Ecological Simulations

A similar approach to genetic algorithms involves simulating an environment with which evolving agents interact. This kind of approach was developed by Parisi, Cecconi and Nolfi (1990) who used it to evolve neural network agents whose reproductive success was dependent on their ability to interact with a simulated world.

This kind of approach has been applied to modeling the evolution of language by MacLennan and Burghardt (1994) and more recently by Cangelosi and Parisi (1998). Cangelosi and Parisi (1998) evolved linguistic agents in a simulated environment containing edible and poisonous mushrooms that agents could eat and observe and about which they could communicate. The agents were essentially standard multi-layer feed-forward neural networks with some inputs dedicated to perceptual qualities of the environment and some outputs dedicated to the agent's motor response. The agents evolved to perceive qualities that helped them to distinguish between edible and poisonous mushrooms and thereby produce the appropriate behaviors of *approaching* and *avoiding* respectively. In addition to the perceptual inputs and motor outputs, each agent possessed some inputs and outputs dedicated to linguistic signaling, but the form and meaning of signals was not pre-specified allowing for a communication system to emerge spontaneously and with respect to the other categorization tasks that the agents were performing. The symbols that the agents used were therefore grounded since they derived their meaning from interaction with an environment (Harnad, 1990). In this case, symbols were grounded, not to the real world, but to a simulated one. Nevertheless, this kind of model allows the researcher to avoid dictating the interpretation of linguistic signals which is enough to observe certain kinds of self-organization in an emerging communication system.

---

correspond to the other by illuminating cases where segments of chromosome act as a unit of selection, but are longer than an individual cistron.

Ecological approaches have been used to model both biological and cultural aspects of the evolution of language usually by using a genetic algorithm for the evolution aspect and learning via back-propagation for cultural transmission during which parents teach offspring labels for different perceptual inputs. It should be clear from the previous section that there is nothing necessary about this distinction. Genetic algorithms and back-propagation are both optimizing algorithms and so either could be used for learning.

Robot Communication

As in ecological simulations, a major motivation of robotic approaches is in the modeling of communication systems that have symbols grounded in an environment with which agents interact. However, what makes the robotic approach different is that robots have a material embodiment with their communication systems grounded in a physical environment (albeit a controlled laboratory setting) to which agents have access via cameras and other sensors.

The robotic approach has been applied to various aspects of language. Steels and Vogt (1997) have used it to simulate the emergence of grounded symbols, Steels (1998; this volume) has used it to simulate the beginnings of syntax and De Boer (1997; this volume) has used it to simulate the evolution of sound systems for communication.

Comparison on Issues

This section is about how computational models are being used to inform the debate on some of the major issues within the field. For each of these issues, it is instructive to ask how producing a computational model can inform the debate. This question will come up again and again.

Innateness

A central concern in the language evolution literature is to understand the extent to which processes governing linguistic competence and performance are innately specified and computational models have been used to inform this debate in a number of ways. In some cases, they are being used to demonstrate that certain aspects of syntax are learnable without requiring a model to have prior (i.e. innately specified) knowledge of them (e.g. Batali, 1994, 1998; Elman, 1990). This won't always be a profitable approach because the biological evolution that debatably gave rise to innate knowledge can be viewed as a kind of learning too. So to say that something is learnable does not in itself help to distinguish between the relevant hypotheses. After all, if an agent in a computational model could learn to build a nest, it would not demonstrate that birds in the natural world are not

hatched with this ability. This has consequences for researchers on both sides of the issue. For nativists, it is only possible to form a learnability argument about language if constraints such as *on the basis of the limited data available to the child* and *without negative feedback* are placed on the learning procedure. By the same token, a model that seeks to refute this claim must demonstrate learning with these conditions in place since biological evolution has operated without them. Negative feedback, for instance, is provided in the form of the environment culling genes for certain kinds of traits from the gene-pool. So while demonstrating that a language can be learnt using an unconstrained learning procedure does not help to inform the debate, demonstrating learnability under the conditions that a child does would.

Evolution is more likely to converge on good design features if related designs have sufficient plasticity to enable them to emulate these features (this is the Baldwin effect). Phenotypic plasticity can result from the ability to learn, but is not restricted to learning. Muscle development, callus formation and skin tanning are also examples where exposure to certain kinds of environmental stimuli trigger changes in the phenotype. Phenotypic plasticity in a population means that the fitness of closely-related individuals will tend to be more correlated. This effectively means that a peak in the fitness landscape will have a broader base making it easier to find by sampling the space of variations. The perils of trial and error and of missed opportunities will mean that individuals with innate knowledge of some useful skill will always have an advantage over those that can only acquire the skill through learning. Given this, one should expect adaptive abilities that are learnable in one generation to become easier to learn in subsequent generations (see Batali (1994) for a model incorporating this idea). Following this argument to its logical conclusion, we might expect natural selection to continue to shape neural structures or biases to facilitate the learning of adaptive abilities until such a point that these neural dispositions effectively embody the ability or until the cost involved in terms of growth and maintenance of these neural structures exactly offsets the gain in fitness that they confer. There appears to be a continuum here over which evolution might have traversed from the ability to learn, to biased learning, to innate knowledge.

Bates, Elman, Johnson, Karmiloff-Smith, Parisi and Plunkett (1998) have suggested one reason why this kind of adaptive pathway might be implausible. They have expressed their incredulity about the possibility that linguistic and other knowledge could be genetically encoded by saying "it is difficult to understand how 10^{14} synaptic connections in the human brain could be controlled by a genome with approximately 10^6 genes."[3] This appears to be a surprising failure of imagination for these authors given the frequent use of expressions such as 'infinite use from finite means' in the linguistics literature. To understand *how* information about 10^{14} connections could be compressed to this extent one could further appeal to the general finding from chaos theory that complexity (even apparent randomness) can arise from the interplay of a few simple components (e.g.

[3] Recent news from the human genome project indicates that the true number of genes is much less than even this number. The current estimate is in the order of 3.5×10^5 which is around a third of the Bates et al. (1998) estimate.

Kauffman, 1995). So why should we expect the complexity of the human genetic endowment to be of the same order as the structures they code for? As Dawkins (1991) has stressed, DNA is not a blueprint for the mature organism. A more appropriate analogy, he says, is that of a recipe.

> A recipe in a cookery book is not, in any sense, a blueprint for the cake that will finally emerge from the oven ... a recipe is not a scale-model, not a description of a finished cake, not in any sense a point-for-point representation. It is a set of *instructions* which, if obeyed in the right order, will result in a cake. (Dawkins, 1991:295)

If DNA is not a blueprint, we should not expect the complexity of its instructions to be of the same order as the complexity of the resulting form, but it isn't obvious that the neural structures embodying linguistic knowledge are complex anyway. While the complexity of language *data* has been cited as evidence against its learnability – the generalizations required being too deep to uncover by general-purpose learning mechanisms and the data available to the child too unreliable, linguists who offer such arguments do not generally believe that the *mental structures* constituting linguistic knowledge are themselves complex. The trend in generative linguistics especially over the past twenty years has been to unify the features of universal grammar under fewer and fewer principles. Perhaps as a consequence of this success, the computational system of human language is assumed to be extremely elegant. For instance, Uriagereka (1998) compares its elegance to the growth functions that give rise to patterns in peacock feathers and flower corollas.

Deacon (1997: 329) has also argued that the Baldwin effect would be unlikely to operate in the linguistic context because languages change too rapidly. A skill has to confer an advantage with enough regularity for selection to act and if that advantage is only a consequence of conforming to arbitrary conventions then as languages vary so would the selection pressure and with a pace that would be far too rapid with respect to the biological time-scales required for genetic assimilation. Yet we do observe stability in cross-linguistic universals over at least historical time-scales (Uriagereka, 1998:46f) which is at least more than we should expect if languages can vary freely. This suggests that variation is constrained either by an innate endowment or by properties of universals that make convergence toward them highly likely in the process of language change. Even if innate knowledge of language is not a cause of this stability, if the knowledge is adaptive and the stability is enough to allow the Baldwin effect to take hold, it is likely to be a result of it.

Adaptive benefit

The extent to which properties of language are adaptive is another major area of contention. Uriagereka (1998)[4], argues that grammar may have evolved not as an

[4] This view is typically attributed to Chomsky although he has never put a clear version of it in print.

adaptation, but as a particular kind of exaptation[5] or perhaps as a spandrel[6]. Part of the reason for this view is that the 'language faculty' seems to be too elegant to have been crafted like this (principles don't appear to follow from multiple determinants and they don't appear to exhibit the messiness that is characteristic of adapted structures). Another reason involves personal incredulity about functional explanations of any linguistic constraint that disallows semantically interpretable sentences. Take the following example from Uriagererka (1998:65):

(3) *What have you discovered the fact that English is?[7]

Uriagereka (1998:50) expresses his incredulity thus: "What's the evolutionary advantage of having [a constraint], if it *disallows* the communication (in those terms) of a perfectly fine thought?"

Firstly, 'in those terms' is an important qualification here because this thought certainly *can* be expressed by other means. Secondly, even if this constraint is maladaptive in some ways then this would not be unusual for an adaptation. Cziko (1995) makes this point in relation to the anatomy of the vocal tract:

> ...unlike mammals that maintain separate pathways for breathing and feeding, thus enabling them to breathe and drink at the same time, adult humans are at a much higher risk for having food enter their respiratory systems; indeed, many thousands die each year from choking ... The risk of choking to which we are exposed results from our larynx being located quite low in the throat. This low position permits us to use the large cavity above the larynx formed by the throat and mouth (supralaryngal tract) as a sound filter ... We thus see an interesting trade-off in the evolution of the throat and mouth, with safety and efficiency in eating and breathing sacrificed to a significant extent for the sake of speaking. (Cziko, 1995)

Some universals may be difficult to explain in terms of gradual adaptations, but others seem quite adaptive by contrast. Pinker and Bloom (1990) suggest some adaptive explanations for case systems, subject-verb agreement and many other universals. Genes that facilitate their possessor making use of a given linguistic feature might replicate more successfully as a consequence of something as arbitrary as the conventionality of the feature within the linguistic community. The linguistic system is part of the environment that selects the genes of language speakers.

Deacon (1997) points out that the reverse may also be true – a linguistic system could be regarded as evolving to fit a niche with the language learner being part of

---

[5] Gould's (1991) label for what Darwin called *preadaptation*. In Gould's terms an exaptation is "a feature, now useful to an organism, that did not arise as an adaptation for its present role, but was subsequently coopted for its current function." (p.43).

[6] Gould and Lewontin's (1979) label for a feature that exists simply because something like it must be present. This is in contrast to adaptations which exist because they are functional in terms of reproductive success. Pinker and Bloom (1990) offer the redness of blood as an example.

[7] Chomsky (1965:228 fn.5) provides another example of an ungrammatical sentence with an unambiguous meaning.

the environment that selects its features. "Language operations that can be learned quickly and easily by children will tend to get passed on to the next generation more effectively and more intact than those that are difficult to learn" (Deacon, 1997:110). Deacon also argues that the limitation on the amount of data that a language learner can use to reconstruct the language of its community provides another kind of selection pressure which may make certain features of language look maladaptive or 'quirky' if viewed in terms of the benefit to language speakers. Chomsky and colleagues reject adaptationist explanations precisely because they are incredulous about functional explanations of linguistic properties for which the language-learner is benefactor, but "if, as linguists often point out, grammars appear illogical and quirky in their design, it may only be because we are comparing them to inappropriate models and judging their design according to functional criteria that are less critical than we think" (Deacon, 1997:110f). The perspective afforded by Deacon's language-adaptive view could explain some of this quirkiness while remaining an adaptive explanation of sorts.

If we accept this view, we should expect the linguistic system to exhibit the appearance of design for minimizing errors in transmission from generation to generation. Kirby's (in press) model of regular compositionality appears to lend itself well to this kind of analysis since a non-compositional language (i.e. one in which each meaning is expressed via a unique symbol) would be less likely to be transmitted intact than a compositional one in which the same number of meanings can be expressed by combining in various ways, symbols from a smaller set. Independent evidence for this view has been provided in a computational model by Batali (2000) and a purely mathematical model by Nowak, Komarova & Niyogi (2001).

The learning bottleneck produces a selection pressure against irregularity that is particularly strong for low-frequency forms. Kirby demonstrated that when a pressure is introduced for shorter signals (which places limits on the extent to which they can be regular) the incidence of irregularity becomes stable and this strongly agrees with what we see in natural languages where highly irregular forms have a very strong tendency to be also highly frequent (forms of *to be* for instance) and regular morphology is associated with forms appearing less frequently.

The idea that languages themselves evolve rather than (or concurrently with) their speakers is not a new one[8], but many have ruled out the possibility. Among them Uriagereka (1998:33ff), but he does so as a result of a misunderstanding of the unit of selection in linguistic evolution:

> Is there any meaning to the claim that English is fitter for survival on the American plains than the language of the Navajo? If English dominates the continent, it does so because of the strongest army in the world, whose finest attribute isn't precisely its verbal brilliance. (Uriagereka, 1998:34)

He argues that even when language change occurs in the absence of such sociological factors, there is no sense in which languages that disappear are less well-adapted than those that remain. That much language change appears to be

[8] Deacon (1997) reviews the history of this idea.

cyclical gives weight to this view. At the very least, it suggests that not all language change that occurs in modern times is adaptive.

It is easy to reach premature conclusions about adaptive explanations if one isn't adhering to the logic of gene-centrism and Uriagereka is probably right to reject an account of language evolution in which the unit of selection is a language for the same reason that modern biologists reject the idea of selection at the level of the individual or the species. The paradox goes away when considering individual linguistic features as competing rather than whole languages.

Another apparent problem has been highlighted at a more fundamental level concerning the adaptive advantage of speaking in the first place (e.g. Ackley & Littmann, 1994; Batali, unpublished; Cangelosi & Parisi, 1998). The benefit to the hearer is taken to be obvious if communication is about the exchange of information, but "[w]hat is the advantage of producing the signal to the individual that produces it? Why should an individual that produces the appropriate signals live longer and have more offspring than other individuals that fail to do so?" (Cangelosi and Parisi, 1998:85). Again, this paradox gets its potency from a confusion over the unit of selection and framing the question this way suggests a paradox that simply does not exist. Under a modern, gene-centric view of evolution, it is the reproductive success of genes and not individuals that is relevant, and the former does not always entail the latter. There may be genuine questions that need to be answered about the reproductive advantage to a speaker's genes, but the wealth of literature on kin selection and evolutionarily stable strategies suggest likely explanations that make this seem much less paradoxical.[9]

Briefly, if an individual's behavior increases the likelihood of close kin being replicated, and if that behavior has a genetic basis, then it might still be selected in the gene pool even if it reduces the reproductive success of the specific individual in question. This is because the kin that benefited are likely to possess these genes as well. For example, genes that promote parental care (such as attending to a baby's cries) increase the likelihood of their own replication even though they are at the expense of the parent as an individual, because the parental care genes are likely to be inherited in the crying baby. Incidentally, for an infant, crying appears to be an instance where communication does benefit the 'speaker' (hunger and other unpleasantness go away with some reliability), so at least in this case, explaining the reproductive advantage to the signaler's *genes* is trivial.

Some aspects of language clearly don't demand an adaptive explanation. For instance, conventionality can arise without any selection pressure at all. A feature that is neutral with respect to fitness can become conventionalised across a population simply because, given time, every individual will come to have a common ancestor.

Some linguistic universals might not be adaptive while others are. Of those that are, some may be adaptive in terms of the benefit to language users while others may be adaptive to the linguistic systems themselves. Chomsky, Pinker and Deacon have adopted mutually exclusive versions of what are essentially compatible hypotheses.

---

[9] See Dawkins (1989) for an accessible introduction that is also a primary text in this field.

We should expect that the interactions between biological and cultural evolution to be complex. A biologically evolved language acquisition device would be a central feature of the environment to which a culturally evolving language is exposed and as such would be the primary determinant in the selection of linguistic features. At the same time, stable features of a linguistic system may be assimilated into the genome of its speakers. Computational models of these processes will help us to understand the dynamics of adaptive interactions, but if no adaptive process is ever simulated for the emergence of a given feature then this will be one for Chomsky.

Of course, demonstrating the emergence of syntax in a broad sense, is not the same as demonstrating the emergence of the particular kind of syntax that linguists call universal grammar, and while the properties of universal grammar are still being debated, it will remain something of a moving target. Nevertheless, computational models serve to demonstrate the consequences of a model's assumptions and what is possible in principle.

Conclusion

Although we can't directly observe the historical emergence of language, we can observe the emergence of modern languages such as creoles that appear to spring from impoverished forms known as pidgins. Pidgins are formed when speakers of a mix of languages are forced to communicate with one another and they lack many of the properties we usually associate with natural languages such as recursion, movement and overt morphology. While pidgins lack these features, creoles are full-fledged languages with structures characteristic of English or any other natural language. Given this, it is surprising that the transition from pidgin to creole usually occurs in only a single generation – despite children never being exposed to structures that are as rich as those they actually acquire. Also surprising is that creoles formed independently in distant regions of the world seem to share many fundamental properties such as a basic SVO word-order.

Computational models of the evolution of language must be reconciled with evidence like this as well as that obtained from a variety of other means. From the paleontological record, it is possible to learn about the environmental conditions in which our ancestors lived including the kinds of social requirements that might have necessitated language and, by indirect means (like looking at skull shape), it is possible to make at least some inferences about brain developments that occurred in parallel. Comparisons between this data and our present knowledge of the neural anatomy of humans and our primate relatives allow us to make inferences about which parts of the brain are necessary for language. Such evidence is problematic for theories that are difficult to relate to such changes. Linguistic data concerning language universals and psycholinguistic data about the way in which language is acquired in childhood also constrain the form that any theory of language origins can take.

As in any scientific discipline, a theory about the origin of language is stronger if it is falsifiable by virtue of the predictions it makes and, in this case, predictions can be made about what should be found in the fossil record, the nature of

language acquisition, the features of modern languages and the differences that should be evident in the neural anatomy of human and non-human species. In addition, the evolutionary narrative that a theory proposes should be plausible and simulating the process computationally can provide an essential test of this. Subtleties arising from co-evolutionary arms-races, kin selection, the Baldwin effect and other processes can be counterintuitive and this is where simulation is most useful. Of particular value are the unexpected features that emerge from a model during simulation that match evidence obtained by other means.

Arguments from personal incredulity are an ideal target for computational modeling because a model need only demonstrate a single possibility to destroy them whether the possibility presented is *the* correct explanation for the given feature or not. Fortunately, for computational modelers, this field is filled with arguments of this kind. Bates et al. (1998) are incredulous about the possibility that linguistic knowledge is encoded in the genome. Chomsky (1972) is incredulous about the possibility that natural selection can provide a meaningful explanation of the emergence of universal grammar. Uriagereka (1998) is incredulous about the possibility that language change is adaptive. There are probably others. Of course, in each case, a failure to demonstrate a possibility will give weight to the given position.

References

Ackley DH, Littman ML (1994) Altruism in the evolution of communication. In: Brooks R, Maes P (eds) *Proceedings of the Fourth Artificial Life Workshop*. MIT Press, Cambridge MA

Batali J (1994) Innate biases and critical periods: combining evolution and learning in the acquisition of syntax. In: Brooks R, Maes P (eds) *Proceedings of the Fourth Artificial Life Workshop*. MIT Press, Cambridge MA

Batali J (1998) Computational simulations of the emergence of grammar. In: Hurford J, Knight C, Studdert-Kennedy M (eds) *Approaches to the evolution of human language: Social and cognitive basis*. Cambridge University Press, Cambridge UK

Batali J (2000) The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe EJ (ed.) *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press, Cambridge UK.

Bates E, Elman J, Johnson M, Karmiloff-Smith A, Parisi D, Plunkett K (1998) Innateness and emergentism. In: Bechtel W, Graham G (eds) *A companion to cognitive science*. Basil Blackwell, Oxford, pp 590-601

Berwick RC (1996) Art imitates life? Simulating evolution to solve engineering problems. *Boston Review*

Cangelosi A, Harnad S (in press) The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*

Cangelosi A, Parisi D (1998) The emergence of a 'language' in an evolving population of neural networks. *Connection Science*, 10: 83-97

Cangelosi A (1999) Modeling the evolution of communication: From stimulus associations to grounded symbolic associations. In: Floreano D, Nicoud JD, Mondada F (eds)

Proceedings of ECAL99 the Fifth European Conference on Artificial Life(Lecture Notes in Artificial Intelligence) Springer-Verlag, Berlin

Chomsky N (1972) *Language and mind: enlarged edition*. Harcourt Brace Jovanovich, New York.

Christiansen MH, Devlin JT (1997) Recursive inconsistencies are hard to learn: A connectionist perspective on universal word order correlations. In: *Proceedings of the 19th annual Cognitive Science Society conference*. Lawrence Erlbaum Associates, Mahwah, NJ, pp 113-118

Cziko G (1995) Without miracles: Universal selection theory and the second Darwinian revolution. MIT Press, Cambridge MA

Dawkins R (1989) *The selfish gene (2nd ed)*. Oxford University Press, Oxford

Dawkins R (1991) *The blind watchmaker*. Penguin Books, London

Deacon TW (1997) *The symbolic species: The coevolution of language and human brain*. Penguin, London

de Boer B (1997) Generating vowel systems in a population of agents. Presented at the *Fourth European Conference on Artificial Life*, ECAL 97, Brighton, UK

Edelman GM (1987) *Neural Darwinism: the theory of neuronal group selection*. Basic Books, New York

Elman JL (1990) Finding structure in time. *Cognitive Science*, 14: 179-211

Fogel LJ, Owens AJ, Walsh MJ (1966) *Artificial intelligence through simulated evolution*. Wiley, New York

Gould SJ (1991) Exaptation: a crucial tool for an evolutionary psychology. *Journal of Social Issues*, 47: 43-65

Gould SJ, Lewontin R (1979) The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proceedings of the Royal Society of London*, 205: 581-598

Harnad S (1990) The symbol grounding problem. *Physica D*, 42: 335-346

Holland JJ (1975) *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor MI

Kauffman SA (1995) *At home in the universe: The search for the laws of self-organization and complexity*. Oxford University Press, Oxford

Kirby S in press. Spontaneous evolution of linguistic structure: an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* Special issue on Evolutionary Computation and Cognitive Science

Kirby S (1999) Syntax out of learning: the cultural evolution of structured communication in a population of induction algorithms In: Floreano D, Nicoud JD, Mondada F (eds) *Proceedings of ECAL99 the Fifth European Conference on Artificial Life(Lecture Notes in Artificial Intelligence)* Springer-Verlag, Berlin

Kirby S, Hurford JR (1997) Learning, culture and evolution in the origin of linguistic constraints In: Husband P, Harvey I (eds) *Proceedings of the Fourth European Conference on Artificial Life*. MIT Press, Cambridge MA

Koza JR (1990) Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. *Technical Report STAN-CS-90-1314*, Stanford University

Livingstone D, Fyfe C (1999) Dialect in learned communication. In: Dautenhahn K, Nehaniv (eds) *AISB '99 Symposium on Imitation in Animals and Artifacts*, Edinburgh, The Society for the Study of Artificial Intelligence and Simulation of Behaviour

MacLennan BJ, Burghardt GM (1994) Synthetic Ethology and the evolution of cooperative communication. *Adaptive Behavior*, 2: 151-188

Montana DJ, Davies LD (1989) Training feedforward networks using genetic algorithms In: *Proceedings of the International Conference on Genetic Algorithms* Morgan Kaufmann

Nowak MA, Komarova NL, Niyogi P (2001). Evolution of universal grammar. *Science*, 291: 114-118

Parisi D, Cecconi F, Nolfi S (1990) Econets: Neural networks that learn in an environment *Network*, 1: 149-168

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

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation In: Rumelhart DE, McClelland JL, the PDP Research Group (eds) (1986) *Parallel distributed processing: Explorations in the microstructure of cognition (Vol 1: Foundations)*. MIT Press, Cambridge MA

Schwefel H-P, Rudolph G (1995) Contemporary evolution strategies In: Mor'an F, Moreno A, Merelo JJ, and Chac'on P (eds) *Advances in Artificial Life: Proceedings of the third international conference on artificial life*. Springer-Verlag, Berlin

Skinner BF (1966) The phylogeny and ontogeny of behavior. *Science*, 153: 1205- 1213

Skinner BF (1981) Selection by consequences. *Science*, 213: 501-504

Steels L (1998) The origin of syntax in visually grounded robotic agents. *Artificial Intelligence*, 103: 133-156

Steels L, Vogt P (1997) Grounding adaptive language games in robotic agents. In: Husband P, Harvey I (eds) *Proceedings of the Fourth European Conference on Artificial Life*. MIT Press, Cambridge MA, pp 474-482

Uriagereka J (1998) *Rhyme and reason: An introduction to minimalist syntax*. MIT Press, Cambridge, MA

situata nel contesto e gli studi sull'interazione sociale nei luoghi di lavoro come esempi di cognizione distribuita.

Morana Alac sostiene che non si può studiare la cultura senza considerare la mente e il cervello. Illustra diversi approcci che studiano il rapporto tra evoluzione e cultura. Individua nella creatività il tratto specifico della cognizione umana che le nuove teorie evoluzionistiche della cultura dovrebbero spiegare.

Angelo Cangelosi e Huck Turner

L'emergere del linguaggio

1. Il linguaggio come sistema complesso

In diverse parti di questo volume è stato più volte discusso come la mente sia un sistema complesso, gli organismi biologici siano sistemi complessi e il cervello sia a sua volta complesso. Il linguaggio dipende da tutti questi sistemi, e di conseguenza ne condivide i vari livelli di complessità. Questa complessità è spiegata dal fatto che il sistema linguistico è caratterizzato da una serie di elementi che interagiscono tra loro in maniera distribuita, autonoma e non lineare. Il processo di interazione e auto-organizzazione di queste componenti porta all'emergenza di strutture linguistiche e comportamenti complessi, come la sintassi e la comunicazione linguistica tra gruppi di individui.

Le componenti di base del sistema linguistico appartengono a una varietà di domini. Alcune sono abilità biologiche e adattive, altre dipendono da fattori individuali, altre ancora da fattori sociali. Le *componenti biologiche* che contribuiscono allo sviluppo del sistema linguistico dipendono prevalentemente dal nostro cervello. Ad esempio, è il nostro sistema neurale che gestisce i meccanismi di apprendimento, come ad esempio un'elevata plasticità neurale nel periodo critico di acquisizione del linguaggio. Diversi *fattori adattivi* ed evoluzionisti hanno una diretta relazione con il linguaggio, e possono spiegare ad esempio il valore adattivo di abilità linguistiche e l'origine delle lingue. Tra le *abilità individuali* che contribuiscono allo sviluppo di comportamenti linguistici vi sono la capacità di percepire il mondo esterno, di produrre e percepire suoni, di formare categorie e di creare collegamenti astratti e logici

Il lavoro dei due autori è stato reso possibile grazie al supporto di UK Engineering and Physical Science Research Council (Grant: GR/NO1118).

tra queste categorie e conservarli in memoria [Harnad 1987]. Per esempio, questi fattori individuali contribuiranno prevalentemente allo sviluppo di abilità linguistiche come un sistema di comunicazione vocale, la capacità di creare significati, e di organizzare una propria base di conoscenza. Tra i *fattori sociali* che influenzano il linguaggio vi sono la tendenza alla formazione di legami familiari e di gruppo, e la necessità di comunicare con i componenti del gruppo e all'esterno del gruppo. Questo porta, ad esempio, al processo di apprendimento e trasmissione del linguaggio, e alla differenziazione tra linguaggi e dialetti in diversi gruppi sociali.

Tutte queste componenti di base del sistema linguistico sono organizzate in maniera distribuita, interagiscono tra loro in maniera non lineare a gerarchica, e sono soggette a un processo di auto-organizzazione. Il concetto di organizzazione *distribuita* significa che non vi è alcuna componente che ha un ruolo di supervisore centrale con diretto controllo su ogni altro elemento. Ciascuna abilità (percettiva, cognitiva, sociale, adattiva) fornisce un contributo parziale al sistema linguistico globale. Il linguaggio è un sistema basato su processi di interazione *non lineare*. In un sistema lineare, le dinamiche del sistema sono additive, cioè il suo comportamento globale è il semplice risultato della somma dei contributi di ciascun componente. Invece, in un sistema non lineare, il risultato finale ha sempre qualcosa in più della semplice somma degli elementi. Questo rende impossibile predire il contributo specifico di un solo componente in isolamento. Quando in un esperimento di psicolinguistica si studia una specifica facoltà linguistica in isolamento, si rischia di perdere di vista il contributo che altri componenti hanno sull'abilità oggetto di studio. Per questo è importante che i risultati sperimentali siano interpretati tenendo conto delle interazioni non lineari tra l'abilità osservata in laboratorio e il ruolo di altri comportamenti inerenti al linguaggio.

Il linguaggio è un sistema *gerarchico* con differenti livelli che dipendono uno dall'altro. Infatti, le abilità linguistico-comunicative sono organizzate dal basso verso l'alto, dove i comportamenti del livello inferiore, come ad esempio le abilità fonetiche, avranno influenza su quelle di livelli superiori, come il livello semantico-lessicale. Per via di quest'organizzazione gerarchica, il sistema funziona secondo principi di *auto-organizzazione*. Esso è in grado di trovare, in maniera autonoma, stati di equilibrio stabile in cui l'interazione dei diversi elementi individuali, sociali, neurali e adattivi produce comportamenti funzionalmente ottimali, come è appunto la nostra facoltà linguistico-comunicativa.

Il risultato del processo di auto-organizzazione è chiamato *emergente*, perché imprevedibile e non facilmente spiegabile. Nello studio del linguaggio, è possibile osservare l'emergere di diverse proprietà e abilità. A livello individuale, durante il processo di acquisizione del linguaggio nel bambino/a emergono gradualmente una serie di complesse abilità linguistico-comunicative, come l'acquisizione del lessico, l'apprendimento di conoscenze sintattiche e lo sviluppo di competenze comunicative. A livello sociale ed evoluzionista, il processo di origine ed evoluzione del linguaggio può essere considerato un processo emergente. Infatti, l'auto-organizzazione di processi adattivi, sociali e neurali deve aver portato alla graduale emergenza di facoltà sociocomunicative, basata su abilità vocali o gestuali, e su abilità cognitivo-linguistiche sempre più complesse.

2. La simulazione del linguaggio

Abbiamo appena definito il linguaggio come un sistema complesso. In particolare, abbiamo spiegato che il comportamento linguistico è basato su diversi livelli di conoscenza (fonetico, lessicale-semantico, sintattico, pragmatico) che interagiscono in maniera non lineare tra loro, e che le abilità linguistiche sono in completa interdipendenza con altre capacità cognitive e sensomotorie, come percezione, categorizzazione, problem solving. Tutto questo ha implicazioni per il tipo di metodo che uno può usare per studiare l'apprendimento e l'evoluzione del linguaggio. Il classico *metodo analitico* delle scienze naturali, come quello basato su esperimenti di laboratorio in biologia o sulle deduzioni logiche in matematica, è ottimale per l'analisi (cioè decomposizione) di un sistema nei suoi componenti di base. Ma non è possibile utilizzare tale metodo con i sistemi complessi non lineari, come è appunto il linguaggio, perché questi sono per definizione non linearmente decomponibili. È possibile applicare metodi sperimentali analitici quando si studia una caratteristica limitata del linguaggio (ad esempio, esperimenti di laboratorio sulla morfologia della forma passata dei verbi), ma non quando si vogliono studiare gli aspetti di interazione tra livelli e tra comportamenti (ad esempio, evoluzione di varie abilità linguistico-comunicative). Per studiare il linguaggio è necessario ricorrere all'alternativa dei *métodi sintetici* basati sulla simulazione al calcolatore [Cangelosi 1998; Steels 1997]. I metodi sintetici, come l'uso di reti neurali artificiali, algoritmi genetici, e le tecniche di Vita Artificiale, permettono di studiare fenomeni

linguistici complessi perché usano un approccio costruttivo. In genere, il modello computazionale simula le componenti di base del sistema, come le parole, significati, sintassi, pragmatica, e le regole di interazione tra questi elementi, come ad esempio i processi percettivi uditivi, e la fondazione di simboli (*symbol grounding*) che verrà discussa in dettaglio più avanti, per arrivare a studiare l'emergenza di comportamenti e abilità linguistiche. Nel caso dello studio dell'acquisizione del linguaggio ci sono diversi metodi computazionali, come le reti neurali artificiali che simulano il processo di apprendimento di diverse abilità linguistiche (cfr. Di Ferdinando in questo volume). Nello studio dell'evoluzione del linguaggio, metodi come gli algoritmi genetici permettono di simulare il processo darwiniano di selezione naturale.

La simulazione e l'uso di metodi sintetici non sono qui proposti come in alternativa ai classici metodi sperimentali. Al contrario, le simulazioni sono un nuovo strumento scientifico che si aggiunge ai metodi di ricerca tradizionali e permette di superarne alcuni limiti. Infatti, lo scopo principale delle simulazioni al calcolatore è quello di esprimere una teoria in termini operativi, cioè di istruzioni per un programma al calcolatore. Questo permette di creare una specie di laboratorio sperimentale virtuale [Parisi 2001] con il quale è possibile investigare la coerenza e validità interna di una teoria, la plausibilità delle sue assunzioni teoriche, ecc. L'esecuzione di esperimenti simulativi servirà alla generazione di nuove predizioni empiriche, che possono successivamente essere verificate anche tramite metodologie sperimentali.

Nei successivi paragrafi presenteremo esempi di modelli simulativi per l'acquisizione del linguaggio e per lo studio dell'origine e dell'evoluzione del linguaggio. Per i modelli che saranno brevemente descritti, verrà posto l'accento sul tipo di ipotesi generate dal modello e sulla corrispondenza con dati empirici. Il lettore interessato a una più dettagliata rassegna critica di tali modelli può leggere il lavoro di Christiansen e Chater [1999] sui modelli connessionisti dell'apprendimento linguistico e di Cangelosi e Parisi [in stampa] per l'analisi dei modelli simulativi di evoluzione del linguaggio.

2.1. Modelli simulativi di acquisizione del linguaggio. Gran parte dei modelli simulativi che studiano l'apprendimento del linguaggio è basato sul connessionismo, cioè sulle reti neurali artificiali (cfr. Di Ferdinando in questo volume). Questo perché l'approccio connessionista permette di simulare l'organizzazione e processi subsimbolici e con modelli ispirati al funzionamento del cervello [Parisi 1989; Rumelhart e McClelland 1986a].

I primi modelli connessionisti del linguaggio si sono concentrati sullo studio della lettura. Per esempio, uno tra i primi e più conosciuti modelli è quello della lettura di Seidenberg e McClelland [1989]. Essi hanno addestrato una rete neurale per la lettura fonetica di parole inglesi. Nel modello è simulata la situazione in cui una sequenza di lettere (rappresentazione grafemica in input della parola) è riconosciuta come parola che viene letta (rappresentazione fonetica in uscita). Nella fase di test alla fine dell'apprendimento, l'errore a livello delle unità ortografiche è stato considerato come una misura della prova di decisione lessicale, mentre l'errore a livello delle unità fonologiche è stato considerato come una misura della prova di lettura e denominazione. Lavorando su questi dati, Seidenberg e McClelland [*ibidem*] hanno mostrato che il modello è in accordo con molti dati psicologici sull'elaborazione dei diversi tipi di stimoli linguistici. Per esempio, le parole regolari come «gave» erano pronunciate più velocemente rispetto a quelle con eccezioni fonetiche come «have». Gli autori hanno analizzato la loro simulazione anche per le implicazioni che essa ha su modelli teorici generali della lettura. In contrapposizione con il modello della lettura a due vie [Coltherart 1986], con una prima via lessicale diretta di analisi fonetica, usata per le parole a pronuncia regolare, e una seconda via grafemico-fonetica, usata per le parole irregolari e le non-parole, Seidenberg e McClelland [1989] usano un modello con un singolo meccanismo. Infatti, la rete usa un solo strato comune di unità nascoste per le parole regolari, irregolari e per le non-parole. Le reti usano una rappresentazione distribuita che impedisce la formazione di unità nascoste lessicali. Un'estensione di questo modello è stata di recente proposta da Zorzi e collaboratori [1998]. In questa nuova simulazione viene mostrato come la rete neurale usa le due vie funzionali per la lettura attraverso diverse connessioni tra l'input grafemico, le unità nascoste e quelle di output fonetico.

Buona parte dei modelli connessionisti per il linguaggio riguarda la simulazione di vari aspetti morfosintattici e lessicali. Per esempio, alcuni studiano l'apprendimento di strutture sintattiche [Elman 1990] e gli effetti dell'età di acquisizione lessicale [Ellis e Lambon-Ralph 2000]. Vari altri lavori si sono focalizzati sulla morfologia. In particolare, uno degli aspetti più estensivamente studiato è quello della formazione del tempo passato dei verbi (*past tense* in inglese). Il modello connessionista originario sulla

formazione del passato è quello di Rumelhart, McClelland [1986b]. Attraverso l'esperienza di apprendimento la rete impara un meccanismo generale per la formazione del passato dei verbi. Rumelhart e McClelland [*ibidem*] hanno osservato che la prestazione del modello durante l'apprendimento riflette alcuni dei fenomeni evolutivi osservati nei bambini durante l'acquisizione della morfologia del tempo passato dei verbi. Per esempio, vi è una fase intermedia nella quale vi è una temporanea inversione della curva di apprendimento. Per alcuni periodi la rete compie temporaneamente un numero maggiore di errori, in particolare producendo errori di iper-regolarizzazione, cioè di attribuzione della forma regolare del passato (aggiunta del suffisso -ed) ai verbi con forma irregolare. Questo fenomeno, corrispondente a una curva di apprendimento a U rovesciata, era stato precedentemente osservato nei bambini inglesi. Gli autori hanno usato questa corrispondenza tra modello computazionale e dati empirici per sottolineare la bontà del loro modello, e per usarlo in contrapposizione ai modelli simbolici della mente basati sull'uso di regole.

Questo lavoro, e in generale l'approccio connessionista allo studio del linguaggio, sono stati aspramente criticati da Pinker e Prince [1988] i quali hanno evidenziato alcuni limiti del modello. Per esempio, Pinker e Prince hanno fatto notare che alcune regole normalmente usate per la trasformazione del passato non possono essere rappresentate dal modello. Inoltre, anche nell'interpretazione evolutiva degli errori, il modello non è capace di usare alcune forme di iper-regolarizzazione di verbi irregolari, e l'uso temporaneo di passato con tutte e due le forme regolari e irregolari. Usando queste argomentazioni, Pinker e Prince hanno concluso che il ricorso a sistemi di regole per la spiegazione del linguaggio umano e del suo sviluppo è essenziale e non sostituibile da modelli connessionisti. In seguito, alcune di queste critiche al modello sono state superate da successivi studi più sistematici del modello del tempo passato [Plunkett e Marchman 1993].

2.2. Modelli simulativi di evoluzione del linguaggio. Nell'ultimo decennio vi è stato un crescente interesse verso lo sviluppo di modelli computazionali di evoluzione del linguaggio. Questo perché si sono sviluppate una serie di metodologie simulative, come gli algoritmi genetici, che permettono di studiare processi adattivi ed evoluzionisti. Inoltre, data la difficoltà di ricorrere a prove dirette dell'evoluzione del linguaggio, l'uso di modelli computazionali permette di supplire a tali limiti attraverso metodologie

sintetiche che simulano le condizioni evolutive, sociali, cognitive che hanno portato all'emergere del linguaggio. Queste metodologie possono essere di notevole utilità se affiancate allo studio delle altre prove indirette di evoluzione del linguaggio, come gli studi comparati sui sistemi di comunicazione animale [Hauser 1996].

Tra i primi lavori che hanno simulato l'evoluzione di abilità linguistiche, vi è il modello di Hurford [1991] che simula l'evoluzione della capacità di comunicazione linguistica con durata variabile del periodo di acquisizione del linguaggio. Confrontando condizioni variabili di selezione selettiva dipendenti dal comportamento linguistico e dai diversi tipi di relazione genitore-figlio, Hurford [*ibidem*] mostra che in tutte le condizioni si sviluppa un periodo critico di acquisizione del linguaggio che finisce verso la pubertà. Tale periodo, una volta evoluto, rimane presente in maniera stabile. I risultati della simulazione mostrano anche che il periodo critico non ha un vantaggio adattivo assoluto, ma il suo valore adattivo è il risultato dell'interazione dei fattori genetici che influenzano le caratteristiche dell'organizzazione delle fasi temporali della vita.

A partire da queste simulazioni iniziali di Hurford, è stata sviluppata una varietà di modelli evoluzionisti che hanno investigato diverse problematiche dell'emergenza del linguaggio. Per esempio, alcuni modelli hanno investigato il processo di emergenza e auto-organizzazione del lessico, mentre altri ancora hanno studiato l'evoluzione della sintassi.

Tra i modelli che simulano l'emergenza di sistemi lessicali, vi sono i lavori di Hutchins e Hazelhurst [1995] e Cangelosi e Parisi [1998]. In alcuni di questi modelli gli autori si sono focalizzati sui processi di auto-organizzazione della comunicazione in gruppi di organismi artificiali [ad esempio Steels 1997]. Altri lavori sono basati sull'auto-organizzazione di un lessico di comunicazione tra robot [Steels e Kaplan 1999]. Alcuni modelli di evoluzione lessicale studiano il ruolo di reti neurali nell'evoluzione di lessici condivisi, sia in situazione di segnali di comunicazione geneticamente determinati [Cangelosi e Parisi 1998], sia con lessici ontogeneticamente appresi [Hutchins e Hazelhurst 1995]. Nel modello di Cangelosi e Parisi [1998] le reti neurali controllano il comportamento di organismi che devono comunicare riguardo a fonti di cibo. Poiché i pesi delle connessioni delle reti non cambiano durante la vita degli organismi, i segnali di comunicazione prodotti dagli individui sono considerati geneticamente innati. Solo durante il processo di selezione naturale e riproduzione i pesi vengono modificati per via di mutazioni casuali, e ciò determina variazioni nel sistema di comu-

nicazione. La stessa rete neurale gestisce non solo il comportamento comunicativo (generazione e comprensione di segnali) ma anche il comportamento di classificazione dei tipi di cibo. Gli organismi devono riconoscere ed evitare i funghi velenosi, ma raccogliere quelli commestibili. Tale modello permette di studiare l'interazione tra abilità cognitive e facoltà linguistiche nella stessa rete neurale. Per esempio, l'analisi dei risultati della simulazione mostra la stretta interdipendenza tra categorizzazione e comunicazione. Solo dopo che gli organismi evolvono l'abilità di discriminazione tra funghi velenosi e commestibili è possibile che evolvano un lessico condiviso, e questa osservazione concorda con l'ipotesi di Burling [1993] sulla interdipendenza evoluzionista tra cognizione e linguaggio. Inoltre, la simulazione permette l'analisi dei principi di organizzazione del lessico, come quello del contrasto, in base al quale differenze tra segnali corrispondono sempre a differenze tra significati, e quello della convenzionalità, cioè che per alcuni significati vi è una forma che ci si aspetta sia usata nella comunità linguistica.

Un diverso gruppo di modelli simulativi si è focalizzato sull'emergenza della sintassi. Alcuni lavori [Kirby 2001] hanno studiato l'emergenza della composizionalità e delle irregolarità sintattiche anche in assenza di processi selettivi, altri [Batali 1994] si sono focalizzati sul periodo critico di acquisizione della sintassi e sull'emergenza della sintassi attraverso processi interattivi di negoziazione, e infine i modelli di Cangelosi e Parisi [2001; Cangelosi 2001] hanno investigato l'evoluzione delle classi di verbi e nomi. Tra tutti questi modelli, solo quelli di Batali e Cangelosi e Parisi utilizzano reti neurali, e permettono di studiare l'interazione tra fattori neurali e l'evoluzione della sintassi. Per esempio, nel lavoro di Batali [1994] reti neurali ricorrenti sono utilizzate per studiare l'emergenza di un periodo critico per l'acquisizione della sintassi. Nella simulazione viene mostrato che se una rete viene inizialmente sottoposta all'apprendimento di una grammatica, e poi è sottoposta a un linguaggio con diversa sintassi, questo provoca una difficoltà a imparare il nuovo linguaggio. Cioè le reti passano attraverso un periodo critico di acquisizione oltre il quale sono incapaci di apprendere facilmente nuovi linguaggi. L'autore ha interpretato il risultato rifacendosi al concetto della Marchmann [1993] secondo cui il periodo critico dell'apprendimento ha l'effetto di intrappolare la rete nella prima soluzione che ha dovuto apprendere. Questa spiegazione è in contrasto con le teorie che interpretano il periodo critico come il risultato di processi di maturazione di uno specifico meccanismo di apprendimento del

linguaggio, come la limitazione iniziale di abilità cognitive sensoriali o di memoria che favorisce l'apprendimento seguente di forme complesse del linguaggio [Elman 1993].

Nel modello dell'evoluzione delle classi sintattiche di verbo e nome, Cangelosi e Parisi [2001; Cangelosi 2001] usano un sistema di reti neurali e Vita Artificiale per mostrare come vi sia una tendenza evoluzionista a evolvere linguaggi composizionali che utilizzano regole grammaticali tipo «verbo + nome». In una prima simulazione [Cangelosi 2001] il linguaggio sintattico emerge per auto-organizzazione. All'inizio dell'evoluzione, tutti gli individui partono da una situazione di assenza totale di conoscenza linguistica e devono imparare a comunicare attraverso la combinazione di una o due parole. Essi devono comunicare riguardo a diverse fonti di cibo. Sebbene vi sia pressione selettiva solo per il compito di foraggiamento, e non per le abilità linguistiche, dopo alcune centinaia di generazioni nella popolazione emerge un linguaggio condiviso da tutti gli organismi. L'analisi della struttura sintattica del linguaggio mostra come vi sia una forte tendenza a far emergere linguaggi basati sull'uso di verbi (ad es., «raggiungere», «evitare») e nomi (ad es., nome del colore dei funghi), invece che usare parole singole, o regole combinatorie non composizionali.

In successivi sviluppi di questo modello [Cangelosi e Parisi 2001], è stata fatta una analisi dettagliata delle fasi di evoluzione delle classi sintattiche di verbi e nomi. L'evoluzione di questo linguaggio sintattico passa attraverso due stadi, il primo nel quale gli organismi imparano a usare solo i nomi, e il secondo nel quale l'uso dei verbi migliora fino a superare la prestazione dei nomi. Questo fenomeno riflette le osservazioni sull'acquisizione della grammatica nei bambini, dove la classe dei nomi precede sempre quella dei verbi [Tomasello e Brook 1999]. L'analisi delle rappresentazioni neurali interne mostra che i verbi vengono rappresentati in una maniera ottimale rispetto ai nomi (cfr. par. 3.2). Questa migliore organizzazione delle rappresentazioni linguistiche provoca diversi vantaggi adattivi, non solo nei comportamenti linguistici, ma anche in altre abilità puramente cognitive.

3. Problemi e questioni nello studio dell'emergere del linguaggio

Nel paragrafo precedente abbiamo visto che la simulazione permette di studiare una varietà di comportamenti linguistici. Con le metodologie computazionali, come le reti neurali e la Vita Arti-

ficiale, è possibile studiare l'interazione tra i diversi fattori individuali, neurali, sociali, adattivi che contribuiscono al funzionamento del sistema complesso del linguaggio e all'emergere di abilità linguistiche. Questo permette di investigare alcuni dei problemi e questioni di ricerca ancora aperti nello studio dei processi di acquisizione linguistica nei bambini, e in quelli di origine ed evoluzione della comunicazione. Nei successivi paragrafi discuteremo in dettaglio il contributo di modelli simulativi su due specifiche questioni di ricerca, e cioè il problema del *symbol grounding*, e quello del controllo neurale del linguaggio.

3.1. Il problema del «symbol grounding». Il linguaggio è un sistema basato sull'uso di una serie di simboli (parole). Ciascuno di questi simboli ha uno o più significati (semantica) e più simboli possono essere combinati tra loro con regole grammaticali (sintassi) per esprimere ulteriori e più complessi significati. Gli aspetti lessicali, semantici e sintattici sono strettamente intercorrelati tra loro nel processo di formazione del significato, e cioè nella creazione di un legame tra il sistema cognitivo-linguistico dell'individuo, le sue abilità sensomotorie, e il mondo nel quale questi vive e interagisce. Questo processo è normalmente chiamato *symbol grounding*. Esso è facilmente risolvibile in sistemi cognitivi reali come il nostro, ma in modelli cognitivi computazionali è spesso trascurato e non risolto. Harnad [1990] ha per primo definito il problema del *symbol grounding* nei modelli computazionali. Un modello plausibile della cognizione deve necessariamente includere il collegamento autonomo e intrinseco tra simboli e loro referenti. Secondo alcuni scienziati cognitivi, come i cognitivisti, la manipolazione dei simboli è un processo autonomo da quello della formazione di significati. Un generico e ipotetico collegamento tra simboli e referenti del mondo reale può essere ipotizzato nella mente del ricercatore e ciò è sufficiente perché il sistema sia di per sé esplicativo della cognizione. Per altri ricercatori, invece, un modello deve simulare sia i simboli sia i significati, e i due sistemi devono essere intrinsecamente e autonomamente collegati tra loro. Per affrontare il problema in maniera più sistematica e non banale, Harnad [*ibidem*] propone un sistema ibrido subsimbolico/simbolico con una configurazione dal basso verso l'alto, nel quale ogni componente di base subsimbolica sia funzionalmente collegato al livello superiore simbolico. Per integrare i due livelli nel sistema cognitivo umano Harnad [1987] fa riferimento alla teoria della percezione categoriale, che studia l'organizzazione delle nostre

abilità di categorizzazione. Per esempio, quando dobbiamo classificare un colore, il *continuum* sensoriale dello spettro di luce è suddiviso in categorie soggettive distinte di colori (blu, giallo, rosso) per via dei due processi della percezione categoriale di *discriminazione* e *identificazione*. La teoria della percezione categoriale, in contrasto con la Legge di Weber, sostiene che non vi è corrispondenza tra (il logaritmo del) la dimensione fisica di uno stimolo e la percezione psicologica di esso. Piuttosto, gli individui riducono la differenza percettiva psicologica di oggetti all'interno di una categoria, e accentuano le differenze tra gli elementi di categorie diverse. Per esempio, una tonalità di colore giallo-verde, la cui distanza dai due colori giallo e verde può essere equivalente da un punto di vista ottico, sarà vissuta come percettivamente più simile a un altro colore di tonalità gialla una volta che sia stata identificata come appartenente alla categoria giallo.

Nella sua teoria della percezione categoriale, Harnad [*ibidem*] propone un modello delle abilità cognitive umane organizzato secondo tre processi gerarchici: 1) *discriminazione*, cioè la capacità di separare gli oggetti e di deciderne la somiglianza/differenza tra gli oggetti ed eventi, 2) *identificazione*, cioè la capacità di assegnare un nome agli oggetti, e 3) descrizione *proposizionale* degli oggetti, degli eventi e degli stati del mondo. Per ciascuno di questi tre processi esiste una tipologia di rappresentazione mentale: *a*) rappresentazioni *iconiche*, cioè trasformazioni mentali analogiche delle proiezioni dell'immagine degli oggetti sulla nostra superficie sensoriale; *b*) rappresentazioni *categoriali*, che permettono il processo di identificazione degli oggetti attraverso la formazione di categorie; *c*) rappresentazioni *simboliche*, basate sulla combinazione dei nomi (simboli) delle categorie.

L'origine dei primi due tipi di rappresentazioni è probabilmente innata e si è evoluta nella nostra specie, come nel caso della categorizzazione di colori. Comunque, poiché sarebbe impossibile evolvere una capacità innata di identificazione di tutte le categorie percettive umane, alcune di queste rappresentazioni vengono apprese. I due tipi di rappresentazioni iconiche e categoriali sono ancora sensoriali e non simboliche. Il nome dell'oggetto ottenuto dal processo di identificazione ha solo un valore tassonomico, non ha nulla di simbolico e non può essere manipolato in base a regole logico-sintattiche. Perché una rappresentazione categoriale diventi il significato cui un simbolo si riferisce, è necessario che ne sia formata una rappresentazione simbolica. Su questa rappresentazione sarà poi possibile applicare le regole di manipolazione di simboli che permettano la formazione di proposizioni (frasi) ri-

guardo a ulteriori relazioni di appartenenza a categorie. Per esempio, a partire dal nome «cavallo» che è intrinsecamente collegato alle rappresentazioni iconiche e categoriali del cavallo (il nome è cioè fondato, *grounded*, sulla realtà), e dal nome «strisce» che è ugualmente collegato alla realtà, è possibile ottenere una nuova descrizione simbolica della categoria zebra, attraverso la proposizione «zebra = cavallo + strisce» [Harnad 1990]. I nuovi simboli (ad esempio, «zebra») sono creati a partire dalla combinazione linguistica di simboli di base («cavallo, strisce»), e hanno la proprietà di acquisire indirettamente il *grounding* dalle due categorie di base. Grazie a questo principio è possibile identificare una zebra (o un unicorno) pur non avendola mai vista prima, semplicemente da una descrizione verbale di essa. A partire dai simboli elementari di una tassonomia di nomi è, in principio, possibile generare il resto dei simboli di un linguaggio naturale, inclusi i simboli per relazioni logiche come le parole: «no», «e», «tutti», «qualche», ecc. Infatti, alcuni dizionari usano un limitato lessico di base per definire tutte le parole in esso contenute. Grazie a questo linguaggio di base, le nuove parole possono ereditare il loro collegamento intrinseco con gli oggetti e gli eventi del mondo reale. Con ciò è anche possibile esprimere concetti astratti senza diretti referenti nella realtà percettiva.

Harnad [*ibidem*] ha più volte proposto i modelli connessionisti delle reti neurali come possibili candidati per un sistema ibrido connessionista/simbolico nel quale la rete neurale svolga il compito di percezione categoriale e di *grounding* delle rappresentazioni. I simboli estratti dal modello connessionista sarebbero poi manipolabili da sistemi simbolici o da altri sistemi connessionisti addestrati a elaborazioni simboliche. Le reti neurali sono un naturale candidato del processo di elaborazione cognitiva di base che è la percezione categoriale per via della loro capacità generale di apprendimento e classificazione di pattern sensoriali. Per esempio, l'uso di un robot controllato da una rete neurale è un buon esempio di come sia possibile avere un modello cognitivo per l'estrazione di rappresentazioni categoriali e simboliche del mondo reale [Harnad 1995]. Questa soluzione connessionista è stata studiata con diversi modelli di percezione categoriale e *symbol grounding*.

Il primo modello connessionista della percezione categoriale è stato presentato da Harnad e collaboratori [1991] per un compito semplice di classificazione e denominazione di linee. Gli autori hanno utilizzato una rete neurale per la classificazione di 8 linee in due categorie di linee lunghe o corte. Il criterio in base al quale i ricercatori hanno deciso l'appartenenza di ciascuna linea alle due classi si basa sulla scelta arbitraria della misura di lunghezza corta/lunga. Una rete neurale a tre strati è stata prima addestrata al compito di auto-associazione in maniera da ottenere una rappresentazione compressa e distribuita a livello delle unità nascoste. A questo punto la rete è stata addestrata a un nuovo compito che, oltre alla auto-associazione, prevedeva l'apprendimento tramite *backpropagation* dell'appropriata etichetta di classificazione della linea «lunga» o «corta». Dall'analisi delle rappresentazioni interne della rete con compito di denominazione è stato trovato un chiaro effetto di percezione categoriale, e cioè una diminuzione della distanza tra le attivazioni di stimoli di una stessa categoria, e un aumento della distanza tra le due categorie rispetto alla rappresentazione ottenuta nella rete precedentemente addestrata al solo compito di auto-associazione. Cioè, il compito di classificazione ha modificato la corrispondenza tra le distanze fisiche dello stimolo e le distanze (psicologiche e neurali) delle rappresentazioni.

Altri modelli connessionisti hanno studiato in dettaglio il processo del *symbol grounding* e del trasferimento di *grounding* da parole di base a nuove parole. L'esempio dell'apprendimento di categorie e nomi di animali (ad esempio, zebra) è stato simulato in un modello connessionista [Cangelosi *et al.* 2000; Riga *et al.* 2001]. La rete neurale è stata prima addestrata a categorizzare e denominare immagini di cavalli, di strisce, e di altri animali e pattern visivi. Poi la rete è stata addestrata a imparare nuovi concetti attraverso la combinazione dei nomi delle precedenti categorie, come ad esempio «zebra» = «cavallo» + «strisce». L'immagine della zebra non viene mai presentata durante questa fase di apprendimento linguistico. Nella fase di test, la rete vede un'immagine di zebra e deve denominarla. Essa è in grado di denominare questa figura come «zebra», mostrando che il *grounding* dei due nomi di base è stato trasferito al nuovo nome. Inoltre, l'analisi delle rappresentazioni interne di questa rete mostra come i meccanismi di percezione categoriale siano alla base del *symbol grounding* e del suo trasferimento a nuove parole. Questi modelli simulativi confermano l'ipotesi di Harnad [1987; 1990] sul ruolo della percezione categoriale nell'apprendimento di simboli. Inoltre essi propongono un sistema di *symbol grounding* totalmente basato su sistemi connessionisti, dove la stessa rete neurale gestisce sia le operazioni subsimboliche di *grounding* sia quelle di manipolazione simbolica.

3.2. Linguaggio e cervello. Il cervello, insieme agli altri fattori di apprendimento individuali, comportamenti sociali, e fenomeni adattivi, contribuisce al processo di auto-organizzazione del linguaggio e di emergenza di complesse facoltà linguistiche. Lo studio del controllo neurale del linguaggio è stato, finora, prevalentemente basato sullo studio di patologie neuropsicolinguistiche in pazienti con lesioni cerebrali in aree coinvolte nella funzione linguistica. Solo di recente, con lo sviluppo di tecniche di visualizzazione cerebrale (*brain imaging*), come PET e fMRI (cfr. Chieffi in questo volume), sono stati effettuati studi sperimentali per l'identificazione fine delle strutture neurali funzionalmente coinvolte nel controllo del comportamento linguistico normale e patologico. Per esempio, Martin e collaboratori [1995] hanno usato tali metodologie per l'identificazione dei correlati corticali di abilità sintattiche. In particolare, questi autori hanno guardato al coinvolgimento di strutture corticali motorie per l'elaborazione della classe dei verbi (ad esempio, nomi di azioni), e al coinvolgimento di aree sensoriali per l'uso di nomi (ad esempio, nomi di colori).

In parallelo con lo sviluppo di nuove metodologie sperimentali di visualizzazione cerebrale, la diffusione dei modelli di reti neurali ha contribuito allo studio dei correlati neurali del linguaggio. Sebbene la maggior parte dei classici lavori connessionisti del linguaggio [Christiansen e Chater 1999] non intendano proporre modelli plausibili del cervello, una parte vuole intenzionalmente simulare le strutture neurali coinvolte in abilità linguistiche. Per esempio, il modello ACTION di Taylor e Taylor [2000] è basato sulla diretta simulazione delle diverse regioni cerebrali coinvolte nell'elaborazione di conoscenze sintattiche.

Un diverso gruppo di modelli simulativi si propone di studiare il comportamento linguistico, e il suo substrato neurale, in una cornice teorica e metodologica più integrata, in cui il linguaggio è una tra le tante abilità sotto il controllo dello stesso sistema neurale. Tale approccio, che usa metodologie di Vita Artificiale, considera anche il contesto sociale e adattivo nel quale l'organismo interagisce, e che porta all'emergenza di capacità linguistiche [Cangelosi 1998]. Nei paragrafi precedenti sono stati descritti alcuni modelli di Vita Artificiale, come quello dell'emergere per auto-organizzazione delle classi dei verbi e nomi [Cangelosi 2001]. Nell'estensione di tale modello [Cangelosi e Parisi 2001], viene simulato un linguaggio, basato sull'uso di verbi e nomi, che è direttamente imposto dal ricercatore. Questo permette di focalizzarsi su nuovi aspetti dell'evoluzione della sintassi, e in particolare sulla differenziazione

neurale tra i verbi e i nomi. Gli organismi hanno il compito di imparare a manipolare due oggetti A e B, e cioè rispettivamente una barra verticale o una orizzontale. La rete neurale di ogni organismo ha una retina per vedere la posizione dell'oggetto e la sua forma, riceve in input segnali propriocettivi sull'estensione del braccio, e può ascoltare un comando verbale (fig. 1). Gli organismi ricevono dei comandi verbali tipo «allontanare A», «allontanare B», «avvicinare A» e «avvicinare B». Essi devono imparare ad associare i verbi allontanare/avvicinare all'azione di allontanare/avvicinare l'oggetto indicato. Durante la loro vita, gli organismi sono esposti a diverse condizioni sperimentali, per esempio situazioni in cui ascoltano il comando verbale (che può essere o solo il verbo, o solo il nome, o il verbo e nome insieme) senza vedere l'oggetto, o in situazioni dove la rete riceve informazioni sia visive che linguistiche. Poiché vi è una pressione selettiva a evolvere gli organismi che si comportano meglio nelle diverse situazioni sperimentali (ad esempio, allontanare qualunque oggetto quando viene detto il verbo «allontanare»), dopo alcune centinaia di generazioni la popolazione di individui avrà reti neurali che imparano l'elaborazione di informazioni linguistiche, visive e motorie.

L'analisi delle rappresentazioni interne delle reti neurali durante l'esecuzione dei diversi compiti linguistici e motori permette di costruire un modello del controllo neurale di tali compiti. Per

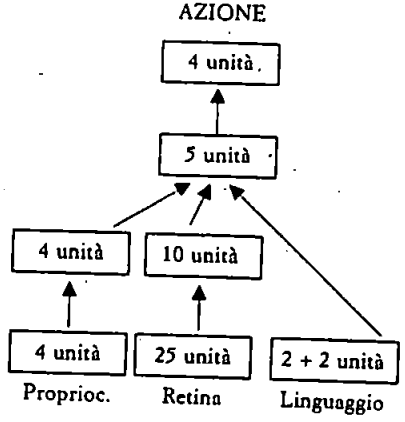


FIG. 1. Architettura della rete neurale nel modello di Cangelosi e Parisi [2001] sull'evoluzione dei verbi e nomi.

tra ricercatori che usano diversi metodi. Gli esempi usati in questo saggio, e in genere l'approccio interdisciplinare e comparativo allo studio della mente proposto nel presente volume, rendono auspicabile l'integrazione tra metodologie sperimentali e simulative per la comprensione dei sistemi complessi in psicologia.

Roberta Lorenzetti

La coscienza incarnata

Il tema della definizione della coscienza e della possibilità di ricerca scientifica relativamente a essa costituisce oggi uno dei punti di intersezione interdisciplinare più complessi e dibattuti. Filosofia della mente e del linguaggio, biologia, matematica, fisica, psicologia e neuroscienze riconoscono nella coscienza un tema di rilevanza sostanziale. Il dibattito si articola, tra altri vari temi, nella discussione comparata dei risultati ottenuti negli specifici ambiti di ricerca e nella ridefinizione della validità e applicabilità dei metodi di indagine al complesso rapporto tra mente e cervello, tra processi cognitivi ed esperienza fenomenica, tra l'attività di comprendere, il soggetto che comprende nel suo sentire individuale e la sua capacità di agire nell'ambiente naturale e sociale.

Solo tracciare le linee essenziali del dibattito contemporaneo (trascurando le linee di evoluzione storica di questo tema) appare un compito di vastità pressoché inaffrontabile e necessiterebbe, comunque, di possibilità di espressione che certamente esulano dai limiti convenzionali di un capitolo (per una descrizione panoramica si consigliano i seguenti volumi: Benzoni e Coppola [2000]; Di Francesco [2000]). Pertanto ci limiteremo a offrire alcuni approcci, o punti di vista, che possono costituire spunti di riflessione sul tema della *coscienza incarnata* (parr. 3, 4, 5).

Le seguenti parole di Edelman e Tononi [2000; trad. it. 2000, 250] propongono una considerazione della coscienza condivisa dall'autrice: «La coscienza è un processo fisico radicato nel corpo di ogni individuo, che è unico; tale radicamento nel corpo non può mai essere sostituito da una mera descrizione». Cercare «il corpo della mente», il corpo della coscienza, costituisce il tentativo di rendere conto di esperienze comuni, usuali, della nostra vita quotidiana che trovano nell'ovvia unitarietà del nostro essere soggetti la risposta a domande dotate di significato filosofico e psicologico complesso e profondo. Eccone alcune:

- esperire emozioni, attitudini, pensieri, ovvero la complessa

esempio, è possibile calcolare le distanze euclidee tra le rappresentazioni delle unità nascoste, per capire come i diversi verbi e nomi sono rappresentati nella rete neurale. Maggiori distanze intercategoriali (cioè tra le diverse categorie di verbi, e/o di nomi) corrispondono a una migliore rappresentazione neurale delle classi sintattiche in oggetto. La figura 2 confronta queste distanze intercategoriali per i nomi e per i verbi nei due strati nascosti della rete neurale. Il primo strato corrisponde alle strutture neurali sensoriali più vicine all'input retinico, mentre il secondo strato corrisponde alle strutture neurali più vicine allo strato di uscita motorio per il movimento del braccio. Mentre nel primo strato nascosto vi è una maggiore differenza tra le classi dei nomi, nel secondo strato più vicino all'output motorio i verbi sono meglio differenziati. Tali risultati sono in accordo con le osservazioni sperimentali di visualizzazione cerebrale descritti da Martin *et al.* [1995], in cui le aree corticali motorie sono più attive con i nomi di azioni, e quelle sensoriali con i nomi di colori.

La similarità tra i dati sperimentali e i dati della simulazione permette di estendere tali modelli computazionali per studiare il ruolo di fattori adattivo-evoluzionisti nel controllo neurale del linguaggio. Per esempio, come già descritto precedentemente, nel

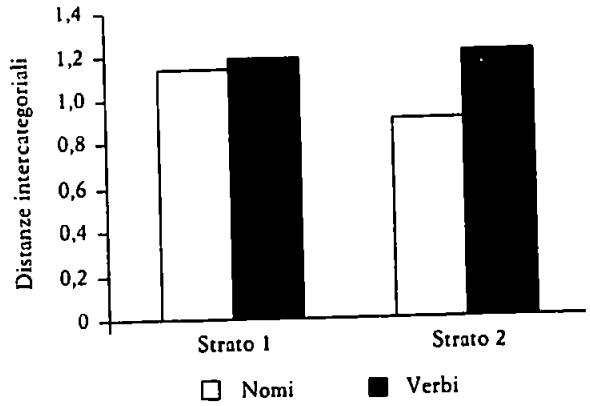


FIG. 2. Distanze intercategoriali tra nomi (colonna bianca) e tra verbi (colonna grigia) nella simulazione di Cangelosi e Parisi [2001]. Si noti l'incremento della differenza tra verbi e nomi nel secondo strato nascosto, quello più vicino alle unità motorie di uscita. Questo incremento indica che il secondo strato nascosto si è specializzato per ottimizzare la differenza tra verbi diversi.

modello è stato osservato che le diverse classi di parole emergono in fasi diverse. Durante l'evoluzione, i nomi vengono appresi prima dei verbi, e ciò è correlato al tipo di rappresentazioni interne che le reti sviluppano. Inoltre, lo stesso tipo di approccio può essere usato per simulare l'emergenza di strutture funzionalmente distinte (moduli) per il controllo di diverse abilità linguistico-sintattiche. Infatti, simili modelli di Vita Artificiale sono stati usati per lo studio dell'emergere di modularità in reti neurali [Calabretta e Parisi 2001].

4. Conclusione

In questo capitolo è stato mostrato come il linguaggio sia un sistema complesso, perché diversi fattori individuali, neurali, sociali e adattivi interagiscono in maniera non lineare e difficilmente prevedibile. Questa complessa interazione porta all'emergenza di particolari strutture linguistiche, come la sintassi, e di uniche abilità linguistiche umane, come l'apprendimento del lessico e il *symbol grounding*.

Il metodo della simulazione al calcolatore è di particolare aiuto nello studio dei fenomeni emergenti del linguaggio. In particolare, le metodologie sintetiche, come le reti neurali e la Vita Artificiale, permettono di simulare il processo di emergenza di abilità linguistiche e quindi di investigare il contributo di ciascun fattore e della interazione tra elementi. Per esempio, in alcune simulazioni è stato possibile analizzare in dettaglio l'interazione tra il processo di evoluzione e quello di apprendimento, permettendo la comprensione del funzionamento dell'effetto Baldwin di assimilazione genetica. Inoltre, altri modelli hanno esplorato il meccanismo del *symbol grounding* e del controllo neurale di abilità linguistiche.

Il metodo simulativo non è qui proposto in alternativa ai metodi scientifici tradizionali di psicologia e delle scienze naturali. Esso è uno strumento complementare di ricerca che insieme agli studi empirici permette di generare nuove ipotesi e teorie esplicative dei processi di emergenza di complesse strutture e facoltà linguistiche. A tutt'oggi, l'integrazione tra i due approcci metodologici non è stata sempre facile o possibile, perché i modelli simulativi sono talvolta troppo astratti, o le troppe semplificazioni rendono difficili i confronti diretti tra dati sperimentali e dati della simulazione. Inoltre, vi sono separazioni storiche tra discipline e tra approcci metodologici che rendono difficile la comunicazione