A Perceptual Basis for Noun Phrase
Syntax

Hayden Walles

a thesis submitted for the degree of
Doctor of Philosophy
at the University of Otago, Dunedin,
New Zealand.

March 2010
Abstract

Human language is the result of both biological and cultural evolution. To have the best chance of understanding language we must seek all the constraints of that evolution. In the first part of my thesis I propose the general hypothesis that the sensorimotor system is one of those constraints and argue that regardless of whether language is the result of biological evolution, cultural evolution, or both, we should expect idiosyncrasies of the sensorimotor system to be reflected in linguistic structure.

The bulk of the thesis explores a particular version of this hypothesis – namely that visual attention and classification of objects are reflected in noun phrase syntax. Within the noun phrase the noun stem (e.g. “dog”) and number morphology (e.g. “-s”) are contributed by separate syntactic elements; I argue that this reflects a separation of functionality in the sensorimotor system.

To begin an exploration of this hypothesis I draw upon existing models of visual attention by Itti and Koch (2000) and object classification by Mozer and Sitton (1998), adapting and combining them into a new computational model. The key new idea in the model is that object classification is cardinality blind which means its output is the same whether presented with one token of a class or many tokens of the class. This allows groups of similar objects to be handled at once. I implement a model of classification which, like primate object classification, is location invariant. In my model cardinality blindness emerges naturally from location invariance. I argue the same thing happens in primates, reviewing neurophysiological evidence for this. To cater for a cardinality-blind classifier I also implement extensions to a standard model of visual attention. The combined classifier and attentional models elegantly reproduce a number of human results, including Gestalt grouping by similarity, global precedence (Navon, 1977) and the role of stimulus similarity in visual search (Duncan and Humphreys, 1989). These results show that the model does useful work in an account of the visual system.

With the visual foundation established I propose a simple model of the interface between visual cognition and noun phrase syntax. Within my model the information corresponding to the noun stem is produced by the classifier and is cardinality blind so carries no number information. The information corresponding to singular or plural number morphology is produced separately by the attentional system. The decomposition of information in my
model corresponds to the same decomposition of information in noun phrases. I conclude that cardinality blindness in the visual system can explain this aspect of noun phrase syntax, supporting the general hypothesis that natural language syntax reflects properties of the sensorimotor system and inviting further theories of this nature.
Acknowledgements

I could not have completed this project without the advice and assistance of many people. Alistair Knott, my principal supervisor, buoyed me up with his enthusiasm while my co-supervisor Anthony Robins ensured my feet remained firmly on the ground. Both inspired me with their depth of knowledge and attention to detail. I doubt I could have found better supervision anywhere.

The university provided financial assistance along with a fertile source of expertise in the many disciplines the thesis touches on. The Department of Computer Science was supportive throughout and I would like to particularly acknowledge Mike Atkinson both as postgraduate coordinator and as my examination convener. I would also like to thank my examiners, Brendan McCane, Michael Corballis and Srinivas Narayanan.

My thesis forms just a small part of an ambitious project headed by Alistair Knott and has benefited from the advice and criticism of others within the Artificial Intelligence Research Group. I relied especially on Andrew Webb, Michael Liddle, Greg Caza and Lech Szymanski, fellow students and good friends who were always available for intellectual and moral support.

Last but not least I want to thank my family for their support while I grappled with this challenge, and express my boundless gratitude to my partner Erica, without whose encouragement I might be grappling with it yet.
For my mother and father
## Contents

1 Introduction .............................................................................. 1
   1.1 Generalities ................................................................. 1
   1.2 Sensorimotor Cognition and Language ............................... 1
   1.3 Visual Object Cognition and Noun Phrases ....................... 2
   1.4 Outline of the Thesis ...................................................... 3

2 A Survey of Perspectives on Language and the Sensorimotor System 5
   2.1 Introduction .................................................................... 5
   2.2 Explaining How Language Works ...................................... 6
      2.2.1 The Chomskyan Approach ......................................... 6
      2.2.2 Cognitive and Functional Linguistics ........................... 9
      2.2.3 Conclusion ............................................................ 11
   2.3 Explaining the Origins and Evolution of Language .......... 11
   2.4 Principles of Biological Evolution and their Applicability to Language Evolution 12
      2.4.1 Natural Selection: The Modern Synthesis .................. 12
      2.4.2 Alternatives to Natural Selection ............................... 15
   2.5 Comparisons Between the Communication Systems of Humans and Other Species 16
      2.5.1 Homologies or Homoplasies? ...................................... 17
      2.5.2 Hockett’s Design Features ........................................ 17
      2.5.3 Monkeys and Gibbons ............................................. 17
      2.5.4 Great Apes .......................................................... 21
      2.5.5 Other Analogues: Dolphins, parrots and bees ............. 25
      2.5.6 Conclusion .......................................................... 25
   2.6 Neuroscience Perspectives on the Evolution of Language in Humans 26
      2.6.1 Introduction .......................................................... 26
      2.6.2 Mirror Neurons ..................................................... 26
      2.6.3 Gestures .............................................................. 26
      2.6.4 Sequencing ......................................................... 27
      2.6.5 Predicate-Argument Structure ................................. 27
   2.7 Archaeological Perspectives on the Evolution of Language in Humans 27
      2.7.1 Physical Anthropology ............................................. 27
      2.7.2 Culture as a Proxy of Language ................................. 28
   2.8 Cultural evolution .......................................................... 29
      2.8.1 Introduction .......................................................... 29
      2.8.2 Iterated Learning Models ......................................... 29
      2.8.3 Historical Linguistics .............................................. 30
   2.9 Conclusion .................................................................... 30
      2.9.1 Language is the Result of Natural Selection .............. 32
      2.9.2 Sensorimotor Processing Almost Certainly Constrains Language 33
      2.9.3 Noun Phrases and Objects ....................................... 34
3 Visual Object Perception and Attention

3.1 Introduction ................................................. 37

3.2 Primary Visual Cortex ........................................ 38
  3.2.1 Early Vision ............................................ 38
  3.2.2 Cortex .................................................. 39
  3.2.3 Features, Tuning and Cell Types ......................... 39
  3.2.4 Retinotopy ............................................. 40
  3.2.5 Columns ............................................... 41
  3.2.6 Channels and Pathways ................................ 41

3.3 Secondary Visual Cortex and Beyond ......................... 42
  3.3.1 Hierarchy and Specialisation ......................... 42
  3.3.2 Interconnectivity ..................................... 42
  3.3.3 V2 .................................................... 43
  3.3.4 What and Where - Ventral and Dorsal Streams ....... 43

3.4 Object Detection ............................................. 44

3.5 Object Classification ....................................... 44
  3.5.1 Location and Scale Invariance ....................... 44
  3.5.2 Cardinality Blindness ................................. 45
  3.5.3 A Survey of Approaches to Visual Classification ... 46
  3.5.4 Five Models of Visual Classification .................. 47
  3.5.5 Jacobs's (2001) Model ............................... 47
  3.5.6 Marr's (1982) Model ................................ 47
  3.5.7 Ullman's (1996) Model ............................... 49
  3.5.8 Riesenhuber and Poggio's (1999) Model .............. 49
  3.5.9 Mozer and Sitton's (1998) Model .................... 49

3.6 Visual Attention ............................................ 51
  3.6.1 Theories of Attention ................................ 51
  3.6.2 The Medium of Attention ............................. 52
  3.6.3 Modality of Attention ................................ 54
  3.6.4 Models of Attention .................................. 55

3.7 Conclusion ................................................. 64

4 A Cardinality Blind Classification Model .................. 65

4.1 Introduction ................................................. 65

4.2 Network Design ............................................. 65

4.3 Training .................................................... 70
  4.3.1 Training Algorithm .................................. 70
  4.3.2 Training Regime ..................................... 71

4.4 Location and Scale Invariance ............................ 72

4.5 Cardinality Invariance ..................................... 73
  4.5.1 Cardinality Invariance with Three Shapes ............ 73
  4.5.2 Cardinality Invariance with Four Shapes ............. 76
  4.5.3 Comparing Cardinality Invariance across Group Sizes 77

4.6 Repeatability ............................................... 78

4.7 The Generality of Cardinality Blindness .................. 78
  4.7.1 Other Cardinality Blind Models ..................... 78
  4.7.2 Biological Plausibility .............................. 79

4.8 Conclusion ................................................. 80

viii
5 A Model of Serial Visual Attention and Group Classification

5.1 Introduction ....................................................... 81
5.2 Parallel Attention Component: Saliency Analysis ................. 82
  5.3.1 Local Contrast ............................................. 84
  5.3.2 Homogeneity ............................................... 84
  5.3.3 Combination of Partial Saliency Maps ......................... 84
  5.3.4 An Intermediate Result: Performance of the Saliency Analysis Component ... 87
5.4 Serial Attention Component: The Selection Mechanism ......... 87
  5.4.1 The Default Classification Scale .......................... 89
  5.4.2 Inhibition of Return ....................................... 90
5.5 Performance of the Combined Attention and Classification Systems .......... 90
  5.5.1 Classifying Groups by their Form ........................ 90
  5.5.2 Global Precedence ......................................... 90
  5.5.3 Simple Visual Search ..................................... 91
5.6 Selection Mechanism Refined .................................. 96
  5.6.1 Performance of the Refined Selection Mechanism .......... 97
5.7 Predictions of the Model and Further Extensions ................ 98

6 Perceptual Correlates in the Noun Phrase .......................... 101

6.1 The Syntactic Marking of Number in Noun Phrases ............... 101
6.2 The DP Hypothesis .............................................. 102
  6.2.1 A Recap of X-bar theory .................................. 102
  6.2.2 Introducing DP ............................................ 102
6.3 D Contributes a Referent, N Contributes a Type ................. 104
6.4 NumP ............................................................ 105
6.5 Computing Reference and Type With the Visual Model ........... 107
6.6 Computing Cardinality with the Visual Model .................... 107
6.7 Sequencing of Visual Operations and Syntactic Hierarchy ...... 108
6.8 Two Loose Ends ................................................ 108
  6.8.1 Other Numbers ............................................. 108
  6.8.2 Other Modalities .......................................... 110
6.9 Conclusion ....................................................... 110

7 Summary and Conclusions .......................................... 111
## List of Tables

2.1 Design features of language. .................................................. 18

4.1 Results of testing scale and location invariance. .......................... 73
4.2 Results of testing cardinality invariance with three items. ............... 76
4.3 Results of testing cardinality invariance with four items. ............... 77
4.4 Results for comparison of cardinality invariance with different display sizes. .. 78
4.5 Translation and cardinality invariance results for five different networks. .... 79

5.1 The simple search results. .................................................. 96
5.2 The refined search results. .................................................. 97

6.1 Māori pronouns showing singular, dual and plural number marking. ....... 102
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>A schematic representation of the proposed structure of the noun phrase.</td>
</tr>
<tr>
<td>2.1</td>
<td>The family tree showing the evolutionary relationship between the anthropoid primates.</td>
</tr>
<tr>
<td>2.2</td>
<td>A hypothetical model of vervet monkey processing in response to perceiving a leopard.</td>
</tr>
<tr>
<td>2.3</td>
<td>The evolutionary relationship between humans and the great apes.</td>
</tr>
<tr>
<td>2.4</td>
<td>The Saussurean sign, a simple mapping between meaning and sound.</td>
</tr>
<tr>
<td>2.5</td>
<td>The apparent unit of selection.</td>
</tr>
<tr>
<td>2.6</td>
<td>The actual unit of selection.</td>
</tr>
<tr>
<td>2.7</td>
<td>The linguistic unit of selection.</td>
</tr>
<tr>
<td>3.1</td>
<td>The pathways of early vision.</td>
</tr>
<tr>
<td>3.2</td>
<td>Layers of V1, the primary visual cortex.</td>
</tr>
<tr>
<td>3.3</td>
<td>A diagram of a centre-surround receptive field.</td>
</tr>
<tr>
<td>3.4</td>
<td>The receptive fields of cortical simple cells.</td>
</tr>
<tr>
<td>3.5</td>
<td>Illustration of the log-polar representation used in cortical area V1.</td>
</tr>
<tr>
<td>3.6</td>
<td>Gestalt grouping principles.</td>
</tr>
<tr>
<td>3.7</td>
<td>Group classification.</td>
</tr>
<tr>
<td>3.8</td>
<td>A selection of different object identification and classification models.</td>
</tr>
<tr>
<td>3.9</td>
<td>Convolution.</td>
</tr>
<tr>
<td>3.10</td>
<td>Two Navon figures.</td>
</tr>
<tr>
<td>3.11</td>
<td>A diagram of the crossmodal attentional interactions thought to exist.</td>
</tr>
<tr>
<td>3.12</td>
<td>A diagram of the FTT model according to Treisman (1998).</td>
</tr>
<tr>
<td>3.15</td>
<td>A diagram of the model by Mozer and Sitton (1998).</td>
</tr>
<tr>
<td>3.16</td>
<td>A diagram of the biased competition model by Deco et al. (2002).</td>
</tr>
<tr>
<td>4.1</td>
<td>The general structure of my convolutional neural network.</td>
</tr>
<tr>
<td>4.2</td>
<td>Convolving layers of the network.</td>
</tr>
<tr>
<td>4.3</td>
<td>Abstracting layers of the network.</td>
</tr>
<tr>
<td>4.4</td>
<td>The shapes used in the training of the network.</td>
</tr>
<tr>
<td>4.5</td>
<td>Maps of the classifier’s accuracy by location.</td>
</tr>
<tr>
<td>4.6</td>
<td>Examples of the kinds of patterns used in the cardinality invariance experiments.</td>
</tr>
<tr>
<td>5.1</td>
<td>The broad structure of the overall computational model.</td>
</tr>
<tr>
<td>5.2</td>
<td>Examples of computing the saliency maps.</td>
</tr>
<tr>
<td>5.3</td>
<td>A proposal for how saliency measures at several neighbouring spatial frequencies (scales) could be computed.</td>
</tr>
<tr>
<td>5.4</td>
<td>Examples of computing the saliency maps when the grouping principles conflict.</td>
</tr>
<tr>
<td>5.5</td>
<td>Results of similarity-proximity conflict tests.</td>
</tr>
<tr>
<td>5.6</td>
<td>Performance of the model when classifying the form of a perceptual group.</td>
</tr>
<tr>
<td>5.7</td>
<td>Illustrations of the model’s global precedence.</td>
</tr>
<tr>
<td>5.8</td>
<td>Example search displays.</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Generalities

Human language is the result of both biological and cultural evolution. This is not hard to see. In the wild our nearest relatives, chimpanzees, communicate with only a limited number of simple signals. Even when instructed by humans they seem incapable of the rich, productive system that we think of as language. Something about our biological evolution in the four to seven million years since we diverged from chimpanzees is crucial to language as we know it.

Cultural evolution of language is evident all around us. We can document this evolution over the past two thousand years or so from the time when written language first appeared. Over the course of this time languages have evolved, splitting apart just like biological species. Unlike biological evolution we are witness to these changes taking place before our eyes as new words and new constructions appear. Just as species adapt to new conditions, so do languages and their speakers.

Evolution, when directed by the process of natural selection, can be thought of as a way of satisfying constraints. Entities (organisms or languages) reproduce, and those that are more successful at coping with prevailing constraints on reproduction will flourish.

To give a simple example in terms of a familiar metaphor: if there is an advantage to squeezing through round holes, then entities reproducing in that environment will tend to be round, not square. The environment constrains evolutionary development. Entities that have evolved to meet constraints bear the mark of those constraints. The roundness of the entities betrays a world of round holes, and the roundness of the holes allows us to predict the roundness of the evolved entities.

Language is a product of evolution. By understanding the constraints on that evolution – all the constraints – we can get a clearer picture of how language works. And conversely, by finding the linguistic equivalents of features like roundness we can say something about the constraints that led to their emergence and development.

1.2 Sensorimotor Cognition and Language

One of those constraints is sensorimotor processing. Language and sensorimotor cognition are not often studied together because they seem so far apart. Perception and motor control are paragons of concreteness, while language is the paragon of abstract thought. The methods of study of each show this same bias.

But sensorimotor processing is a constraint on language. An elegant example of how the sensory system can influence language this way is the cross-linguistic research conducted by Berlin and Kay (1969) on colour names. They found that languages with only two basic colour names used them to label black and white (or, more correctly, dark and light). Languages with three colour names labelled these plus red. In languages with more colour names those with the same number of basic names used them to label the same colours, with little variation. What’s more, the actual colours (measured objectively based on the frequency of light) were the same when labelled in different languages. Kay
and McDaniel (1978) later found that these key frequencies corresponded to the maximum sensitivity of the cells in the retina that detect colour. Nobody contends that the eye is part of an autonomous language module. But visual cortex perceives colour as interpreted by the retina, which constrains how other parts of the brain represent colour. Thus in a very basic way language can be influenced by our physiology – even aspects of physiology usually considered very remote from language.

The broad goal of this thesis is to argue for the hypothesis that studying sensorimotor processing can help the study of language. I argue for this in two ways. First, by discussing language as an evolved entity and why its evolution is partly constrained by (and therefore likely to bear the stamp of) sensorimotor processing. Second, by considering a specific hypothesis. This second argument makes up the bulk of the thesis.

1.3 Visual Object Cognition and Noun Phrases

In particular, I am interested in the hypothesis that the visual and cognitive mechanisms which humans use to attend to and classify objects in the world are encoded in language in the structure of noun phrases. Noun phrases are phrases built around nouns that can serve as grammatical subjects and objects. For example, “a dog”, “some trees” and “John” are simple noun phrases. One of their linguistic functions is to represent concrete objects in the world, though they can have other functions as well, particularly in expressing quantification and reference to abstract entities.

My interest is focussed on three aspects of the structure of noun phrases: the grammatical expression of type, reference and number. Number is reflected in several places in language: in noun phrases it is particularly reflected in the marking of singular or plural number. Though there are variations, the distinction between one and many is common among natural languages. Linguistic evidence from a variety of languages suggests that within the noun phrase the noun proper denotes a type only, while other syntactic elements contribute information about the referent or referents on which the type is predicated as well as the number of referents. Figure 1.1 illustrates the proposed structure. From a purely linguistic point of view, these distinctions seem fairly arbitrary: it is not obvious why it has to be this way.

From a perceptual point of view, though, things are different. I am certainly not the first to note that there is a perceptual precedent for the type-reference distinction. This point has been well made by Hurford (2003b). But I have created a perceptual model of visual object attention and classification that, in addition to the type-reference distinction, quite naturally produces the distinction between singular and plural and, what is more, conforms neatly with other linguistic evidence and theory relating to grammatical number.

The visual model and its results hinge on cardinality blindness or cardinality invariance. There are several variables for which the primate object classification system in inferotemporal cortex (IT) is invariant – that is, when that variable changes in the input, the classification is unaffected. One such variable is location: to some extent the output of the primate visual classification system abstracts away from location. It doesn’t matter where on the retina an object’s image appears, the object can still be classified. Another example is scale: it doesn’t matter what the apparent size of an object is, it can be classified.

I define a classifier as cardinality invariant if it is largely unaffected by the difference between one
object and several objects of the same type presented simultaneously. Such a system reacts similarly if presented with a single dog or with a group of dogs. This means a cardinality invariant classifier can respond to homogeneous groups of objects by performing group classification.

I have discovered that a class of artificial neural networks called convolutional neural networks (CNNs) are cardinality blind. And I argue that the primate visual classifier in inferotemporal cortex is also cardinality blind.

Because it provides type alone, without number information, the output of a cardinality-blind visual classifier corresponds to the representation of type in the noun phrase.

Attention is usually understood as limiting the input to the classifier in some way. I have created a computational model of selective attention designed on the assumption that the classifier is cardinality blind. It models attention to different locations (spatial attention) and spatial frequencies (scale attention) in its input and in conjunction with a CNN classifier reproduces several known psychological phenomena: Gestalt grouping principles, global precedence and similarity effects in visual search. Selective attention corresponds to the representation of reference in the noun phrase.

Finally, having established the psychological credibility of the whole model I show how the distinction between singular and plural number emerges quite naturally as a side-effect of selective attention. This corresponds to the representation of grammatical number in the noun phrase.

The visual-attention model thus produces reference information in one place, type information in another place and number information in another, consistent with the decomposition of this information linguistically inside the noun phrase. It stands as a possible sensorimotor explanation for an aspect of linguistic structure.

By reaching this goal, I also provide another argument for my broader goal, demonstrating the usefulness of the sensorimotor approach to the study of language. Examining the constraints that sensorimotor processing places on language is a fruitful way to go forward and I hope that my result, even if it is falsified in future, invites others to propose and test linguistic correlates of sensorimotor processing.

1.4 Outline of the Thesis

The structure of this thesis is as follows. Chapters 2 and 3 together make up a review of the literature relevant to my topic. I have found it convenient to split the review into two chapters. My research attempts, to some degree, to synthesise two relatively disparate disciplines, the study of language and the study of visual perception and attention. It makes sense to review them separately as the kind of material reviewed, and the reason for reviewing it, is different in each case.

Chapter 2 reviews the approaches to the study of language and the principles of evolution as they apply to language to justify the general hypothesis that language can reflect sensorimotor correlates. Along the way it provides the theoretical framework in which the linguistic side of my specific hypothesis is couched.

In Chapter 3 I examine human object perception and attention, summarising current knowledge at a concrete and detailed level, preparing the ground for the description of my own computational model.

That presentation begins in Chapter 4 where I present my CNN model of visual classification and show that it is cardinality blind.

It continues in Chapter 5 where I present my model of selective attention that makes use of the CNN’s cardinality blindness for group classification. It is here that I show my complete model reproduces existing psychophysical results relating to Gestalt grouping principles, Navon’s (1977) global precedence and visual search.

Finally, in Chapter 6 I tie the perceptual model to the linguistic data by proposing how my model can produce the information conveyed by the noun phrase and help explain its structure.

Chapter 7 summarises the thesis and points to possible future work.
Chapter 2

A Survey of Perspectives on Language and the Sensorimotor System

2.1 Introduction

In some ways this chapter is a deviation from the main body of the thesis: most of the thesis is about a computational model of sensory processing. In other ways this chapter is what the thesis is all about, explaining language.

The purpose of this chapter is to survey what is known about the relationship between language and sensorimotor cognition. It’s about embodied approaches to language. This includes, but is by no means limited to, the field of linguistics, and even within linguistics there is considerable division between theoretical and experimental approaches. I am guided in my travels through the literature by three principles.

The first principle relates directly to my research question. I seek a direct role for the sensorimotor system in the structure of language. Therefore I am interested to find theories or experimental results that suggest a link between linguistic cognition and general cognition – or better yet with sensorimotor cognition. There are a number of such theories already, some of which I look at in Section §2.6. In particular Hurford’s (2003b) proposal that the predicate-argument distinction in language corresponds to a division of visual processing in the brain foreshadows my own theory.

The second principle is to survey methodological approaches to the study of language. These range from traditional linguistic theorising which relies on mathematical formalism abstracted from actual language use to neuroscience approaches that hunt for the way language is implemented in cortex. This is important because I need to select an appropriate approach for my own consideration of language and presumably adopt some of the relevant analytical tools in my research.

The third principle is the need to be satisfied on a point that has often been neglected or skimmed over in the past with regard to human language, evolutionary plausibility. By taking evolution into account from the beginning I need not fear creating another incompatible theory and I can also restrict the scope of study to just those theories that are compatible.

These three principles are mutually reinforcing in many ways. For instance, evolution works by adapting old features to new tasks so the connection between language and general cognition is relevant to that investigation as well.

There are many ways this chapter could have been divided up, none of them ideal, so I have chosen to begin by following a traditional dichotomy in linguistics between Saussure’s synchronic (contemporary and descriptive) and diachronic (historical and dynamic) linguistics. Section §2.2 deals with synchronic explanations of language. The remaining sections deal with historical explanations of language, with an emphasis on the role of evolution. This part of the chapter begins by examin-
ing biological evolution and moves towards topics more relevant to cultural evolution. Section §2.3 introduces and lays out this investigation in more detail, which is presented in Sections §2.4 to §2.8.

In Section §2.9 I draw together common threads and use them to justify my general hypothesis that the sensorimotor system is a fertile source of linguistic explanations. I then set up the linguistic plausibility of my hypothesis that noun phrase structure represents object perception and attention.

2.2 Explaining How Language Works

2.2.1 The Chomskyan Approach

Much of modern linguistics owes its existence to Noam Chomsky, either by adoption of his theoretical system or in rebellion against it. Chomsky’s theory of generative grammar has undergone several revisions, the two most recent going under the names of Government and Binding Theory (Chomsky, 1981; Haegeman, 1991) and the Minimalist Program (Chomsky, 1995). The important contributions are outlined below.

Many linguistic theories share some of these components with the Chomskyan approach. For example, Head-driven Phrase Structure Grammar (HPSG, see Pollard and Sag, 1994) uses X-bar theory but diverges from Chomsky on other points.

2.2.1.1 Universal Grammar

Central to Chomsky’s vision of language is Universal Grammar (UG). Chomsky believes in a largely innate language faculty whose design is the same in all humans. This faculty embodies the rules of UG, which is not a grammar as such, but a set of constraints that restrict the kinds of grammars that a child can learn.

There are a few arguments in favour of UG. One is the undeniable fact that languages seem to be built around the same pattern. Another, known as the poverty of the stimulus, claims that the examples of language that a child observes are insufficient to allow a general-purpose learning system to pick up the language. Chomsky thus theorises that the domain of learnable languages is restricted by UG (but see section §2.8.2 for a counterargument).

UG is laid out more explicitly as principles and parameters: it consists of universal principles which all languages obey, as well as parameters which are alternative ways of doing things adopted by different languages. Parameters are set by a child from linguistic examples they acquire during infancy.

Chomsky’s goal seems to be a minimal set of principles and parameters within UG (hence adoption of the term minimalism for the latest effort).

2.2.1.2 Competence and Performance

Chomsky stresses a division in language between competence, which he sees as the domain of the linguist, and performance. Competence is the individual’s knowledge of the language, while performance is everything involved in actually generating or processing language. Competence is therefore an abstraction of the language system and it conveniently allows linguists to formally describe languages and language structure without claiming anything about processing.

For example, grammar is usually described as a directed process beginning with items from the lexicon which are combined via phrase structure rules and transformed before being fed to either phonology (the subsystem concerned with production and comprehension of linguistic sounds and signs) or semantics (the subsystem concerned with meaning). This is the so-called syntactocentric view (Jackendoff, 2002). Linguists allow this because competence theories say nothing about actual implementation by users, which is a performance issue. Presumably performance actually requires information to flow from phonology to semantics (or vice versa) mediated by a syntactic module.
2.2.1.3 Universal Grammar as a Module

Of particular note is that although Chomsky champions language as a part of the brain, he does not believe it is useful to study other cognitive functions as a means to explaining language. Others are happy to adopt various parts of Chomsky’s theory, including even the idea of language universals, but do not always maintain his isolationist approach.

Fodor (1983) takes an even more extreme view than Chomsky, stating that not only are other cognitive functions unnecessary for the study of language but that language is a strictly independent cognitive module.

2.2.1.4 Phrase Structure

*Generative syntax* or *generative grammar* refers to the system consisting of phrase structure rules (the grammar) and the lexicon of words or morphemes whose combination the grammar dictates. Generative syntax is the source of combinatorial variations in Chomskyan linguistics. Chomsky’s formalisation of grammar as phrase structure rules allows linguistic structure to be specifically laid down. The system is generative because the repeated application of grammar rules generates the valid sentences.

Modern versions of Chomsky’s theories are lexically driven, that is the phrase structure rules and lexicon are combined and phrases grow from the lexicon by satisfying constraints on what can be grouped with what.

*Phrases* – sentences and their components – have structure indicated visually with trees. For example, this structure

```
S
  NP  VP
    "Jim"  V  NP
      "hit"  "Fred"
```

shows a phrase labelled S (for sentence) composed of an NP (for noun phrase) and VP (for verb phrase). The VP in turn combines a V (verb) with another NP. The verb is an item from the lexicon. Sometimes phrases show themselves as single words, as in the NPs above, but NPs can have their own phrase structure:

```
S
  NP  VP
    "The ball"  V  NP
      "smashed"  "a window"
```

Here the triangles indicate implied structure that is not shown between the phrasal nodes of the tree and the lexical items at its leaves.

Why propose phrase structure at all, and why this particular structure? It is not arbitrary. Evidence is available to test these structures. A hypothetical phrase must pass the substitution test - that is, it must be possible to substitute a pithy syntactic equivalent (though not semantic equivalent)
for the phrase. For example, in the phrase structure above we can substitute “did so” for the entire VP and the sentence is still grammatical.

\[
\text{S} \quad \begin{array}{c}
\text{NP} \\
\text{VP} \\
\text{“The ball”} \\
\text{“did so”}
\end{array}
\]

### 2.2.1.5 Generalising Phrase Structure: X-bar Theory

Jackendoff (1977) proposed the influential and widely accepted X-bar theory which posits a schematic structure for all natural language phrases. X-bar is a *lingua franca* of linguistics and is crucial to understanding the interface between my perceptual model and language. The brief presentation here is based on those of Carnie (2007) and Radford (1988).

The phrase structure rules discovered by applying the substitution rule show many similarities. For example, noun phrases can take the form

\[
\text{NP} \quad \begin{array}{c}
\text{D} \\
\text{“the”} \\
\text{N} \\
\text{“destruction”} \\
\text{PP} \\
\text{“of the window”}
\end{array}
\]

which parallels that of the sentence. The X-bar schema attempts to capture this similarity by proposing that all phrase structure is based on this shape:

\[
\text{XP} \quad \begin{array}{c}
\text{Spec} \\
\text{X’} \\
\text{X} \\
\text{Comp}
\end{array}
\]

Each lexical item (represented schematically by X) is supposed to carry such a piece of structure around it. It is thus called the *head* of the phrase. The *specifier* (Spec) and *complement* (Comp) are positions that are filled by other XPs, the exact species depending on constraints placed by the head. These positions are not always obligatory, sometimes they are optional.

The X-bar schema applies to all syntactic categories, and systematising the names to coincide with the structure we see that what I have been calling S is really a VP, VP is a V’ and the unknown phrasal element in the NP above is an N’. (In fact this is not the complete X-bar schema structure, but is sufficient for this discussion)
2.2.1.6 Head Movement

Phrase structure rules are only part (if a large part) of Chomsky’s theory. Chomsky believed some constructions are not the result of orderly combination of words into larger and larger phrases, so he introduced the concept of movement and created a transformational grammar.

(1) Fred can hit Jim.

(2) Can Fred hit Jim?

Here the statement and question have all the same words but in a slightly different order. But these two sentences are supposed to share underlying structure and differ only because movement has taken place.

(3) Can, Fred t₁ hit Jim?

In this case head-movement has taken place, where the item at one head has moved to another head position higher up the tree. Here the modal “can” has moved from the head of the inflectional phrase (IP), leaving a trace t₁, and landed in the head position of the complementizer phrase (CP). Sometimes transformations of this kind are required. When head movement occurs the heads can combine with each other to form a complex head. This will become significant in Chapter 6.

Until the advent of minimalism phrase structure rules were supposed to generate a deep structure (later D-structure) to which transformations were applied to produce surface structure (later S-structure). In the latest revision this separation has been discarded and transformation rules are applied alongside phrase structure rules.

2.2.1.7 Formal Semantics

Syntax receives the most attention, but there are two other major components, phonology and semantics. These are generally treated as independent but communicating layers. While phonology is important, I am not concerned with it here, but semantics is very important.

The Chomskyan approach treats semantics as something that can be read off the syntax and, in line with the strict mathematical formalism for syntax, there is a corresponding formalism for the semantics. This is usually predicate calculus, a symbolic logic system that is well understood and unambiguous. In particular it abstracts semantics entirely away from the brain.

2.2.2 Cognitive and Functional Linguistics

Cognitive linguists and functional linguists purport to fundamentally challenge the assumptions of Chomskyan linguistics. While Chomsky believes linguistic structure is arbitrary, these linguists think it directly reflects the psychological or environmental pressures on language.

Cognitive linguistics (Lakoff, 1987; Langacker, 1991; Lee, 2001) is based on the assumption that language is just an aspect of human cognition. Language is embedded in and interwoven with other faculties, so any account of language must be in terms of the associated cognition (Lee, 2001).
**Functional linguistics** encompasses the requirements of cognitive linguistics, but also demands that environmental and social factors be taken into account. Functional linguistics therefore also considers language change, acquisition and evolution to provide valuable data points for explaining language as it is now (Givón, 1995; Halliday, 1985).

In cognitive linguistics all linguistic structure is seen as the result of cognitive processing. Units of language have only two components, phonology and meaning (Langacker, 1986). Units are combined by tying together parts of their meanings, which also produces a combined phonology. Because there is no separate syntactic aspect to this language model (as in the traditional, Chomskyan approach) any change in appearance of a sentence represents a change in meaning.

Cognitive linguists also emphasise that the semantics of language are *embodied*. Because the mind interprets the world in a particular way, they argue, semantics must be considered relative to that embodied interpretation, in contrast with the objectivist semantic models employed in formal logic, as well as traditional linguistics. Embodied meaning could be viewed as an imperfect image of the world around us, but it can also carry information that is not apparent in the world. Cognitive linguists believe that these purely psychological interpretations are represented in language by sentences with similar meanings which *construe* the same situation in different ways. For example a road can run up a hill or down it, and the physical situation under discussion is identical. However “up” and “down” construe the situation differently. The same argument can apply to many other linguistic phenomena such as passives and question formation.

Traditional, Chomskyan linguists confront the problem of different sentences with the same (or very closely related) objective meaning by postulating that the sentences really are the same in some deep form but are transformed by rules of the language to reach different surface forms. Cognitive linguists instead believe the differences stem from genuine differences in meaning. This might be viewed as a central “dogma” of cognitive linguistics: that different linguistic structure *always* indicates different semantics.

Although cognitive and functional linguists disagree with traditional formalist accounts of language, such as Chomsky’s, this does not mean their own theories lack rigour. The current model of syntax (the form or phonological side of language) is *construction grammar* (Goldberg, 1995, 2003).

Constructions, which are form-meaning pairs, fit naturally into the cognitive linguistics framework. Constructions include words (the only level at which generative linguists believe atomic meaning enters language) as well as structural patterns that can be combined into more complex form-meaning pairs. So “rose” and “bomb” are constructions, but so are idioms like “raining cats and dogs”, underspecified idiomatic patterns (“The more you laugh the less it hurts”, “The greater the current the greater the danger”) and general rules such as sentence formation from a noun phrase and verb phrase.

Construction grammar allows for the existence of many constructions whose meanings are not compositional (determined by the meanings of their parts) which other linguistic schools dismiss as peripheral. At the same time it allows for general grammatical rules.

Construction grammar may be accepted as a syntactic theory by cognitive linguists, but semantics is more difficult to describe because cognitive linguists believe that language is a product of general cognition, not a specific language module, so a comprehensive explanation of semantics would effectively be a complete account of human psychology as well. For this reason cognitive linguistics accounts of meaning deal in psychological models, eschewing traditional formalism. For example, Lakoff (1987) makes use of Rosch’s psychological theory of categories and a metaphor principle to explain meaning.

Lakoff also embraces the use of metaphor to explain the meaning of language, and this technique has been widely adopted within the field (Lee, 2001). The use of metaphorical language is supposed to indicate something about human cognition. An example is the use of spatial metaphors to describe abstract notions like relationships (friends can be close, or drift apart, for example) and politics (where opinion is categorised on a spectrum ranging from far right to far left).

---

1 The idea that structural differences represent semantic differences does not follow from the premise that linguistic structure is an epiphenomenon of general cognition. One can imagine, in principle, just such a system where identical meaning can be expressed in two distinct ways (perhaps depending on a single neuron somewhere whose activity is random). Unless one takes an extremist view that construal encompasses the entire state of the brain (including random neurons) it is difficult to see how the dogma follows.
Since they believe language is a product of general cognition, cognitive linguists believe the only way to explain cross-linguistic similarity is via shared traits of human cognition, or shared function of the language. Functional linguists (Givón, 1995, for example) reach the same conclusion, but because they take a functionalist approach are not necessarily of the belief that language is entirely the result of general cognition.

Finally, cognitive linguists also disagree with Chomskyans about the learnability of language. The conventional claim that new learners of a language are confronted with too difficult a task is rejected. Learners acquire constructions using their general cognitive abilities to extract the necessary patterns from the input. Interestingly, there seems to be mounting evidence that children learn language in this kind of way, rather than the way proposed by Chomskyans (where language ability improves in large jumps as various grammatical switches are set in the child’s mind) (Tomasello and Brooks, 1999).

Cognitive and functional linguists are especially sceptical of the dependence on logical formalism by traditional linguists, viewing this as an abstract symbolic game with its own rules that does not necessarily coincide with linguistic reality. This is just part of the single biggest difference between traditional (especially Chomskyan) linguistics and either of these approaches: Chomskyans concern themselves with syntax, semantics and occasionally pragmatics of language, but cognitive and functional linguistics are also concerned about the teleology of language. Chomsky argues that the form of universal grammar is arbitrary and therefore needs no explanation, while Lakoff, Langacker, Givón and others argue that the best way to uncover language structure is by considering its purpose.

As a side-effect of this, cognitive and functional linguists cannot take the escape of claiming they are only trying to describe linguistic competence. Language is a product of the human brain and its environment so there must be a real linguistic structure that can be described if enough time is devoted to the study, and it is this language (of both performance and competence under Chomsky’s dichotomy) that cognitive and functional linguists are seeking to explain. It is of course quite possible that traditional linguistics might arrive at the same description of language as the newer approaches, but since Chomsky and his followers only demand a description of human language, for which there may conceivably be many, it would only be through coincidence that they arrived at the way that humans actually use language – as they are quite willing to accept.

2.2.3 Conclusion

There are three key aspects to the cognitive and functional approaches which particularly differentiate them from the more conventional, Chomskyan approach.

First, the theory of syntax is different. This is the least insurmountable difference, however. For example, Jackendoff (2002) proposes a theory of syntax which is clearly based on his X-bar theory (Jackendoff, 1977) but also includes less general rules which make it seem very like construction grammar.

Second, cognitive and functional linguistics use embodied semantics where the semantics are the representations used by the brain directly, contrasting with Chomskyan linguistics where semantics are disembodied formal systems.

Lastly, where Chomskyan linguistics proposes universal rules which come built-in to the language module, cognitive and functional linguistics explain language universals as resulting from cognitive constraints. Because semantics constrains language it is elevated to a higher position than in the Chomskyan school where it tends to be in the background.

That these three differences are resolved one particular way for the Chomskyan and another way for the cognitive and functional linguists is more accident than design. As discussed further in section §2.9 I choose to take embodied semantics and non-linguistic cognition as the source of universals from the cognitive school, and X-bar syntax from the Chomskyan school.

2.3 Explaining the Origins and Evolution of Language

Linguists have, in the past, given little consideration to the origins of language. Partly this is because the analytical tools have not been available to deal with the problem, but many linguists have deliber-
ately chosen a research framework that renders research into origin of language irrelevant (Christiansen and Kirby, 2003).

Recently this attitude has changed, with some linguists expressing interest in discovering how language came about. From outside the discipline anthropologists, neuroscientists, psychologists and others have found language an irresistible target for research – and speculation.

New fields inevitably take some time to settle into a coherent research framework, and evolutionary linguistics is still working at this. Both linguists and non-linguists have ignored evidence from other disciplines. Gradually the realisation is emerging that to study the evolution of language, for which any one discipline provides little information, will require a multi-disciplinary effort (Bickerton, 2003; Newmeyer, 2003).

The change in attitude towards language evolution indicates that language investigators have realised the enormous value that evolutionary thinking can bring to bear. When evolution is considered in a rigorous and principled manner it provides a valuable constraint on the explanations of both language and cognition.

The sections that follow review the current approaches to language evolution research, of which there are two main forks. Sections §2.4 to §2.7 deal with biological evolution which has presumably had at least some role in human use of language, since other animals appear not to use it. Theories of evolution, empirical data on communications and language evolution in our own and other species are discussed. Some researchers (such as Pinker, 2003) believe biological evolution is entirely responsible for language, but others believe “cultural” evolution has had a major part to play in language development. Section §2.8.2 examines this.

There is little coherence among theories of language origins, a fact also noted by Carstairs-McCarthy (2004). Different researchers create different stories about language evolution, and while there are some common threads there are seldom two authors who agree on more than one or two points.

In evolutionary biology such stories are sometimes called just so stories, after Kipling’s fanciful tales of how leopards got their spots and so forth. While this term is sometimes used pejoratively, I believe in this case it is justified simply by the plethora of stories spun by serious researchers to explain just one phenomenon, the evolution of human language. Because they are so disparate I introduce many examples throughout the chapter where they seem appropriate. Section §2.6 collects four particularly relevant stories which propose a sensorimotor evolutionary origin of language.

### 2.4 Principles of Biological Evolution and their Applicability to Language Evolution

#### 2.4.1 Natural Selection: The Modern Synthesis

The principal, and most widely accepted explanation for the evolution of organisms today is a combination of random variation and natural selection, the principle discovered independently by Darwin and Wallace (Darwin, 1859; Darwin and Wallace, 1858). Natural selection is the name for the process whereby traits which confer reproductive advantages on organisms tend to become more prolific in a population. Darwin and Wallace both noted that any population has constraints placed on its growth, and that organisms better adapted to their environment will survive longer and reproduce more than less well-adapted competitors. These constraints are often referred to as selection pressures.

Natural selection is the principal mechanism biologists use to explain evolutionary change of organisms. After the rediscovery of Mendel’s (1866) genetics work, natural selection was given a rigorous foundation and an evolutionary theory emerged which is known as the modern synthesis (of Darwinian evolution and Mendelian genetics), or neo-Darwinism. (Fisher, 1930; Mayr, 1942)

Modern evolutionary theory has built on this foundation, adopting novel mathematical tools in a very successful programme which is capable of explaining much of the interesting variation of life on Earth in a reasoned manner. These tools come from fields such as economics, complexity theory and computer science, among others.
In particular, modern Darwinism considers natural selection as a process that works at the level of the *gene*, rather than the individual, group, or species. This is the selfish gene interpretation made popular by Dawkins (1976). Genes that produce bodies and behaviours that propagate more of the same genes will come to dominate a genetic population, that is they will be effectively “selected”.

### 2.4.1.1 Adaptations

Modern evolutionary theory is sometimes described as adaptationist. This is because evolutionary biologists assume natural selection will cause organisms to evolve in a way that adapts them well to their environment. Traits that appear to have evolved this way are called adaptations.

Though this principle may appear obvious, there is sometimes confusion over the term. It is important to note that biologists are not demanding that *every* trait be an adaptation. One can claim all adaptations are the result of natural selection without requiring natural selection to be the cause of every trait (Cronin, 1991).

### 2.4.1.2 Language and Altruism

The value of language is not restricted to the individual, it is a social phenomenon requiring competition between members of a group. At first glance altruistic behaviour in nature poses a problem for natural selection, which is generally supposed to be a self-interested mechanism. For example a predator alarm call seems difficult to explain because it aids others by alerting them to danger but tells the predator where you are. In fact there are several mechanisms available which allow natural selection to choose altruistic behaviour (Cronin, 1991).

**Kin selection** (Fisher, 1930; Haldane, 1932; Hamilton, 1963, 1964) Natural selection operates on genes, not individuals, so if an altruistic gene can locate other copies of itself in other individuals, and help those, it will be favoured. The obvious way this can happen is for individuals to help their close relations.

**Game theory** (Axelrod, 1984; Axelrod and Hamilton, 1981; Trivers, 1971) Sometimes behaviour is not entirely altruistic; there is an assumption by the giver that the receiver will return the favour later. At first it seems such a system would be ripe for exploitation by cheaters who accept altruism but do not repay. However analysis using game theory shows that under the right circumstances stable strategies will evolve that are altruistic and cooperative.

**Manipulation** (Dawkins, 1982) Dawkins’s idea of an extended phenotype, that is that effects beyond the traditional phenotype (body) such as birds nests are under genetic control, allows an interesting interpretation of apparently altruistic behaviour. Perhaps individuals observed selflessly aiding others are actually being manipulated, their muscle power and resources being used by someone else’s genes. This explains the behaviour of birds whose nests are parasitised by cuckoos, cuckoo genes having adapted to precisely exploit and manipulate the host bird.

**Language via Social Cognition** Dunbar (1996) is a leading proponent of social cognition as the key to language evolution. In section §2.5.4.2 I note that many non-human primates use grooming as a way to reinforce social structure and maintain social order. Dunbar believes that as hominin group sizes increased grooming became impracticable as a way to maintain such large social networks and that language, for the purpose of gossip, evolved to take its place.

Others (Pinker, 2003, for example) believe language shows too much function to have evolved as a grooming replacement, but that social factors place an important obstacle in the way of language evolution that cannot be ignored. There is a paradox involved in the evolution of human language, illustrated by Krebs and Dawkins (1984), that as soon as language evolves there is nothing to stop people lying for their own advantage and manipulating others. Honesty is not evolutionarily stable: an honest population cannot survive if invaded by a deceitful mutant. Even a population of sceptical individuals who assess the honesty of others using experience cannot survive against liars. For this reason we expect a selection pressure against language.
Not everybody seems to agree that social factors were so important in driving language evolution. Bickerton (2002) backs off somewhat from a previous emphasis on social cognition to claim that the utility of language was more important near the beginning of its evolution than its social contribution. He avoids the problem of deception by considering language at a stage when symbols were being combined in very crude ways (possibly using different modalities). Such simple language could be used by small foraging groups to indicate food sources to the larger group so that the whole group can make good foraging choices. The language is too simple to express social concepts, and there is no advantage to be gained in deception.

Most theorists who favour cultural evolution (see Section §2.8) to explain the emergence of part or all of language favour a scheme where language adapts to meet its function. However this does not necessarily mean they favour utility over social uses, since social interaction is a perfectly valid function that language might adapt to fulfill.

2.4.1.3 Language and Sexual Selection

Kin selection and other altruistic mechanisms can explain many apparently non-adaptive traits, but there are other strikingly non-adaptive examples they cannot account for. The classic example is the peacock’s tail, which is extremely non-adaptive and apparently benefits nobody.

Darwin (Cronin, 1991; Darwin, 1871; Fisher, 1915, 1930; Trivers, 1972; Zahavi, 1975, 2003) had an answer for this in a special kind of natural selection called sexual selection. There are several variants of sexual selection, but all rely on traits in one sex being selected for simply because the other sex prefers it. Since increased sexual attraction increases reproductive potential these traits come to dominate.

In the case of the peacock’s tail, all that is required is for peahens to prefer long, gaudy tails and they will evolve. The same effect is observed in many species. Sexual selection only recently enjoyed a resurgence after a century of being largely ignored.

Okanoya (2002) has studied Bengalese finches and concluded that the syntax in their song could have evolved by sexual selection. Building on Marler’s (1970) favourable comparison of bird song development and human language development he argues that human syntax could also have evolved this way. Female Bengalese finches show signs of greater attraction to syntactically complex songs, and comparisons with their wild ancestors show that the domesticated males use more complex songs, perhaps because inhibitory selection pressures are reduced in domesticity. Okanoya believes that if syntax evolved this way then only later became involved in intentional communication then the problems of how language evolved despite dishonesty are sidestepped. Pinker (2003) has argued that syntax could not evolve by sexual selection because it is too complicated and functional, however other authors (such as Carstairs-McCarthy, 1999) have noted just the opposite, that syntax is not well matched to its function, so although it seems unlikely, Okanoya’s theory is not impossible.

Fitch (2002) provides a more compelling argument for sexual selection in the evolution of speech, if not syntax. He has studied laryngeal anatomy in several species using new techniques and found that the previously accepted wisdom that humans are the only animals with a lowered larynx is wrong. A lowered larynx allows a wider variety of sounds to be produced, presumably an advantage to language users. However even dogs can lower their larynx to a similar position on demand, and red deer also have a permanently lowered larynx which is probably the result of sexual selection for large-sounding calls. Fitch argues that human laryngeal lowering could easily have come about by the same means and only later have been coopted for its versatility in speech production.

2.4.1.4 Language and the Baldwin Effect: Pseudo-Lamarckism

Lamarckism is the term applied to a loose coalition of obsolete evolutionary theories which all have two factors in common:

1. features of organisms develop and atrophy in response to use and disuse respectively, and
2. These acquired characteristics are passed on to offspring.

The classic example is the claim that giraffes obtained their long necks by striving to reach high leaves which stretched their necks. Once acquired the long neck would then be passed on to offspring.

Lamarckism cannot work in practice for a variety of reasons (Cronin, 1991; Dawkins, 1982). However, under certain conditions, natural selection can mimic the inheritance of acquired characteristics, a process known as the Baldwin effect (Baldwin, 1896; Ku and Mak, 1998).

The Baldwin effect relies on plasticity of an organism. In principle, if a particular trait is beneficial to an organism’s fitness then natural selection could favour an adaptation which facilitates acquisition of the trait from interaction with the environment, rather than the trait itself. Examples are the formation of a callus on irritated skin, and tanning after exposure to ultraviolet light. In each case the change is brought about as a response to an environmental influence.

The Baldwin effect is pseudo-Lamarckian because it assumes the plasticity of use and disuse in an organism, but rather than having offspring inherit the phenotypic changes from their parents, they inherit the ability to respond to environmental stimuli in the same way. Exposed to a similar environment, as is likely in nature, the offspring then develop similar phenotypes to their parents. Natural selection would tend to adjust the range of plasticity to best allow organisms to fit the environment.

The Baldwin effect is often spoken of in language evolution. Pinker and Bloom (1990) argue the Baldwin effect could explain a purely biological evolutionary root for language. At the other extreme, proponents of cultural language evolution (see section §2.8.2) cover themselves by arguing that once stable language has developed it could become integrated into the genome by the Baldwin effect, given enough time.

### 2.4.2 Alternatives to Natural Selection

Although natural selection is the overwhelmingly dominant explanation for evolutionary change in biology, several alternatives crop up in the literature on language evolution. I address these here mainly to note that the alternatives provide little, if any, explanatory power. Despite this they have been adopted by many competent researchers outside evolutionary biology largely through ignorance (as admitted by Bickerton (2003) for example).

#### 2.4.2.1 Genetic Drift

One alternative to natural selection is endorsed by biologists, and this is genetic drift. Genetic drift is the inevitable result of chance in the workings of evolution. Ideally, natural selection would be able to choose the best and brightest candidates to survive, but in reality chance plays a part. Cronin (1991, p. 88) explains that natural selection acts by allowing a non-random sample of a population to survive, but notes that sampling error will always occur. Genetic drift is the result of this error, and its effects are expected to be strongest in selectively neutral situations.

However genetic drift is not useful for explaining adaptive design, since the changes it leads to are genuinely random. For this reason it provides little explanatory power, except as a last resort.

#### 2.4.2.2 Saltationism

Saltationism (from the Latin saltus, jump) is an evolutionary principle that expects evolutionary change to take place in bursts that occur from one generation to another. Although jumps of this kind, known as macromutations, are not unknown (for example the Attennapaedia mutations in Drosophila melanogaster that cause legs to grow from the antennae sockets) (Ridley, 1996, p. 600), this is not considered an important factor in evolutionary change (though if such mutations are beneficial, natural selection will take advantage of them).

An argument against saltationist theories originally proposed by R.A. Fisher and elaborated by Dawkins (1986) uses an analogy with a microscope. Imagine a microscope that is almost, but not quite in focus. Say the lens is a millimetre too high. Now consider changes that could be made to the focus. If we adjust the mechanism by a small amount, say half a millimetre (in either direction), then
there is an even chance that the focus will improve, albeit slightly. On the other hand, if we make a centimetre adjustment the focus will be much worse, no matter which direction we move.

Fisher's point is this: we assume that any organism that is breeding is already fairly well adapted to its environment. Statistically then, the smaller the change in its offspring, the more likely they are to be well adapted too. Macromutations are unlikely to improve fitness; in fact they are more likely to reduce fitness. Most biologists embrace this principle of gradualism: evolution proceeds in small steps.

2.4.2.3 Punctuated Equilibrium

Punctuated equilibrium is the name given to an amorphous idea posed by Gould and Eldredge (1977). The authors claim that evolutionary change is not a slow and “gradual” process as usually depicted, but a combination of long periods of stasis during which a species stays relatively stable punctuated by very short bursts of evolutionary change driven by natural selection.

The idea is widely misrepresented and misunderstood. Dawkins (1986) gives an elegant rebuttal of punctuated equilibrium, of which I give a scant summary.

Dawkins first blows away the smoke surrounding punctuated equilibrium. There is no claim that change occurs in a single generation, so Gould and Eldredge are not saltationists. Dawkins observes that they do set up a straw man in the form of “gradualism” which does not exist. He eventually concludes that the only radical point made by the authors is that the rate of evolutionary change is not free to vary continuously, but has only two “gears”: stop and very fast. In particular they claim this is enforced by active resistance by genomes to evolve, and thus remain in stasis. However Dawkins notes there is no evidence for such resistance, after all animal breeders find no “snapping” point of variation.

The conclusion is that punctuated equilibrium is not a useful term in evolutionary biology.

2.4.2.4 Exaptation

Gould and Lewontin (1979) rail against what they imply is adaptationism gone mad. Evolutionary biologists, they say, concentrate far too much on finding adaptive reasons for traits when some, perhaps many traits are actually spandrels, side-effects of other changes. Gould and Vrba (1982) later named these traits exaptations. An exaptation is any trait which did not originally evolve for the purpose it now serves.

Exaptation has received a patchy welcome in the expert literature, despite the revolutionary impetus of its creators. Dennett (Cronin, 1991; Dennett, 1995) provides a critical examination of the exaptation principle, and finds it wanting in substance. All traits must trivially be exaptations, he reasons, because all complexities of life evolved from simpler forms. For example, the human jaw is an exaptation because it originally evolved as a gill, which was presumably itself an exaptation.

Proponents of exaptation also imply that side-effects of evolutionary change can exist and somehow be outside the realm of natural selection. This is patently untrue. If a naturally selected trait brings with it a side-effect, the side-effect is also subject to selection. Indeed unless the “side-effect” is selectively neutral it will help influence the evolutionary fate of the original trait.

Thus it seems that the term exaptation is not a particularly useful one to employ. The conventional term preadaptation, referring to a trait which is adapted to a new function, is all that seems necessary.

2.5 Comparisons Between the Communication Systems of Humans and Other Species

The previous section explained the parts of modern evolutionary theory that are applicable to the question of language evolution. This section looks at what science has found by using these theories to help assess communication and cognition in other species.

Sections §2.5.1 and §2.5.2 introduce useful concepts for categorising evolutionary communication and cognition, while sections §2.5.3 to §2.5.5 consider actual examples.
2.5.1 Homologies or Homoplasies?

In biology one trait is **homologous**\(^2\) to another similar trait if they are both derived from a common source. For example a bat’s wing is homologous to the human arm because both evolved from the foreleg of a common ancestor. However a bat’s wing is not homologous to a fly’s wing because, although they perform similar functions, they are derived from different structures. In this case the traits are said to be **homoplastic**. In an evolutionary context homoplastic traits are usually the result of convergent evolution, where similar selection pressures have resulted in similar solutions to similar problems.

One of the considerations in the sections below is whether the communicative and cognitive traits found in non-human animals are homologous or homoplastic to aspects of human language. Most animals are so distantly related to humans that we would not expect behavioural or cognitive homologies: these examples are considered in section §2.5.5. Primates are more difficult. Many researchers believe human language is just a highly developed version of whatever communication system was employed by distant human ancestors. A distinction is usually made between the great apes and other primates, a distinction I maintain here. Monkeys and gibbons are considered in section §2.5.3, the great apes in section §2.5.4.

However, before discussing specific examples, we need a working definition of the traits we are looking for. For this I turn to the design features of language proposed by Hockett in section §2.5.2.

2.5.2 Hockett’s Design Features

Hockett (1960a,b) was one of the first to consider the evolution of language in a rigorous framework. He did so by stipulating what he considered to be thirteen design features of language (see table 2.1) which can then be attributed to various species (both extant and extinct).

Many modern communications researchers, such as Hauser (1996) are critical of Hockett’s system because it does not address the function or relative importance of the features, which are important to the evolutionary and ecological consideration of communications. For example, we could claim that a particular species uses semantic communications, but that says nothing about why semantic communications might be useful.

However Hauser is speaking from the point of view of ethology and behavioural ecology where the aim is to understand communications of all animals. When comparing the use of humanistic language features by other animals I feel that at least some of Hockett’s features are a good way to characterise and ultimately compare these groups. Also, many of these features occur commonly in the discussion of language evolution (of all kinds), and other attempts to specify design features of language (such as Pinker, 2003) incorporate these.

Therefore in the discussion below I will find it useful to use some of these features to categorise different groups.

2.5.3 Monkeys and Gibbons

Primatologists traditionally make several divisions within the order Primata, though some of these have fuzzy edges due to disputes over organisation. The *cladograms* (evolutionary family trees) in Figures 2.1 and 2.3 illustrate the accepted evolutionary relationships between the Anthropoids, consisting of monkeys, apes and humans as described by Byrne (2000) and Noble and Davidson (1996). The dates themselves are somewhat problematic and fluid (the uncertainties are discussed in Kumar et al., 2005) however the relationships between different species and groups are more solidly established. In this section I explore the behaviour of monkeys and gibbons (lesser apes), while the great apes are considered in section §2.5.4.

---

\(^2\)Homology is a fundamental concept in biology, entire books are written about it (like Hall, 1994), yet there is still some dispute, as well as misunderstanding, over its use. Analogy and homology are used in the rigorous fashion laid down by Owen (1843, 1848) (though the term and concept predate him), homology being a *kind* of analogy, not its *opposite*. The term for traits that are analogous but not homologous is homoplastic, after Lankester (1870), who also wanted to replace the term homologous with the term homogenous.
1. **Vocal-auditory channel**
   Signals are carried as sound from the mouth to the ear.

2. **Broadcast-transmission and directional reception**
   Signals propagate in all directions, and recipients can locate the source.

3. **Rapid fading (transitoriness)**
   The signal is closely bounded temporally (unlike writing or spoor, for example).

4. **Interchangeability**
   Organisms can say anything they hear.

5. **Total feedback**
   Organisms can reflect on everything they say.

6. **Specialisation**
   The signal serves only the purpose of communication.

7. **Semanticity**
   Signals have meaning, that is they are consistently associated with objects or events in the environment.

8. **Arbitrariness**
   Semantic signals are arbitrary (the form of the signal is not predicted by the form of its meaning).

9. **Discreteness**
   A discrete set of signals is used.

10. **Displacement**
    Organisms can refer to objects or events displaced in time or space.

11. **Productivity**
    Organisms can express things that have never been expressed before by combining signals. Productive systems are *open*, non-productive systems are closed.

12. **Traditional transmission**
    Although the communication apparatus may be innate, the conventions of the system are learnt from other users.

13. **Duality of patterning**
    Meaningful units are created by combining several meaningless units in particular ways.

---

Since Hockett’s system was published other modalities have been properly recognised for language, such as sign languages (Stokoe et al., 1965). Although a strong case can be made for speech as a design feature, it is not essential, however this only affects the first and second features.

**Table 2.1: Design features of language, as defined by Hockett (1960a,b).**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Vocal-auditory channel</td>
<td>Signals are carried as sound from the mouth to the ear.</td>
</tr>
<tr>
<td>2. Broadcast-transmission and directional reception</td>
<td>Signals propagate in all directions, and recipients can locate the source.</td>
</tr>
<tr>
<td>3. Rapid fading (transitoriness)</td>
<td>The signal is closely bounded temporally (unlike writing or spoor, for example).</td>
</tr>
<tr>
<td>4. Interchangeability</td>
<td>Organisms can say anything they hear.</td>
</tr>
<tr>
<td>5. Total feedback</td>
<td>Organisms can reflect on everything they say.</td>
</tr>
<tr>
<td>6. Specialisation</td>
<td>The signal serves only the purpose of communication.</td>
</tr>
<tr>
<td>7. Semanticity</td>
<td>Signals have meaning, that is they are consistently associated with objects or events in the environment.</td>
</tr>
<tr>
<td>8. Arbitrariness</td>
<td>Semantic signals are arbitrary (the form of the signal is not predicted by the form of its meaning).</td>
</tr>
<tr>
<td>9. Discreteness</td>
<td>A discrete set of signals is used.</td>
</tr>
<tr>
<td>10. Displacement</td>
<td>Organisms can refer to objects or events displaced in time or space.</td>
</tr>
<tr>
<td>11. Productivity</td>
<td>Organisms can express things that have never been expressed before by combining signals. Productive systems are <em>open</em>, non-productive systems are closed.</td>
</tr>
<tr>
<td>12. Traditional transmission</td>
<td>Although the communication apparatus may be innate, the conventions of the system are learnt from other users.</td>
</tr>
<tr>
<td>13. Duality of patterning</td>
<td>Meaningful units are created by combining several meaningless units in particular ways.</td>
</tr>
</tbody>
</table>

---

Figure 2.1: The family tree showing the evolutionary relationship between the anthropoid primates. The approximate dates of division from common ancestry are indicated at the right in millions of years ago (Ma). The family tree of humans and great apes is exploded in Figure 2.3.
There is some dispute over whether some monkeys or lesser apes possess any of the characteristics that distinguish human language from nonhuman animal communication, depending on the tradition in which the commentator has been trained, and the definitions used. As explained above, I will analyse the communications used by these creatures using Hockett’s features.

The classic monkey communication example is that of vervet monkeys (*Cercopithecus aethiops*), a widespread species of old world monkey. Vervet calls were first analysed by Struhsaker (1967) who identified, among other calls, three alarm calls that were apparently predator-specific, and which elicited predator-specific responses in call recipients.

### 2.5.3.1 Ethological Conceptions of Meaning in Vervet Monkey Calls

Further work by Seyfarth et al. (Cheyne and Seyfarth, 1990; Seyfarth et al., 1980) using playback experiments revealed several interesting points. An individual’s response was determined solely by the signal - its acoustic properties - rather than context. In response to the “leopard” signal, vervets run up trees or climb higher if they are already in a tree. In response to the “snake” call they stand on two legs and search the ground for snakes. In response to the “eagle” call they hide under bushes.

The second point is that signals which vary in intensity and duration do not change the response. Response may be more sluggish if a “weak” signal is given, but the same action is taken.

The researchers, and others who follow them (Hauser, 1996) use these points as evidence that vervet calls are referential, that is they refer to the external world rather than to affective state (emotion and motivation), the traditional explanation for non-human calls. A signal refers to an event, in this sense (Hauser, 1996, p. 507), if it is reliably associated with the event, a much weaker relationship than evoked by the term in related fields like linguistics. The reason for this was pointed out by Quine (1970), and is usually referred to as referential opacity.

That vervet calls are referential in this sense seems inescapable in light of the evidence outlined above. However all this means is that the “leopard” signal “stands for” the leopard in the part of vervet cognition that deals with predator evasion: hearing the alarm is just as good a reason to climb a tree as seeing a cat lurking in the undergrowth.

To illustrate this, we can embark on a brief thought experiment. Suppose that we define meaning such that a signal $A$ means eventuality $B$ to a receiving organism if wherever the organism relies on $B$ to produce a particular output, $A$ produces the same output. For a signalling organism $A$ means $B$ if whenever $B$ arises $A$ is produced. Then, referring to the hypothetical model of vervet cognition in Figure 2.2 we see that the vervet’s leopard call might mean several different things, and may mean something different to signaller and receiver.

The box at the top represents the cognition of a monkey that spots a leopard, with the solid arrows showing the progression through perception, affect and action modules, each of which produces intermediate output stored in a “working memory”. The result is that the monkey makes for the nearest tree and climbs, and also (implicitly) emits a signal. The box at the bottom of the figure shows the cognition of a second monkey that does not see the leopard, but hears the first monkey’s signal. The dashed arrows indicate the possible route taken by the signal.

For example, a traditional ethologist would have claimed that the call only carries affective information, so the signal would travel something like the middle path of Figure 2.2. Modern ethologists of the Marler school believe the call refers to the leopard in the environment, and the signal in this case would travel something like the left path. One could imagine other possible routes too, such as a signal link between the tree-climbing motor programmes of signaller and receiver. Finally, one could imagine combinations of these, for example a monkey squawking because it is climbing a tree, but receivers changing their affective state as a result of hearing this.

I noted above that vervet alarm call intensity and duration does not affect response much, which is taken as evidence that more than just affect is involved. One can also imagine methods for eliminating various signalling pathways of the kind shown in Figure 2.2 by observation or experiment, and this is the path apparently taken by modern research in this tradition, on several species.
Figure 2.2: A hypothetical model of vervet monkey processing in response to perceiving a leopard, and three possible meanings of the vervet monkey alarm call.
2.5.3.2 Commonalities Between Monkey Calls and Human Language

The issue of reference is an important one in the ethological tradition in which vervet research is conducted, and for evaluating some claims about the uniqueness of human language. For the student of human language there are other important points.

The ethological work shows that vervet calls are semantic, arbitrary (it would seem) and discrete. However the calls are innate, the repertoire apparently fixed by the genes, clearly an adaptive solution to the problem of multiple predators requiring different escape routes. The selection pressures are obvious, since monkeys that do the wrong thing (run into the bushes when a leopard is about, say) get eaten. There is a slight possibility that some of the predator-call association is learned, but this may be a side-effect of infants’ developing perception. So vervets do not transmit their calls traditionally. Their calls also make no use of productivity, duality of patterning or displacement.

Vervet monkeys are the most extensively studied in this area, but there are other examples of primate communication that differ from vervet calls. Hauser (1996, pp. 37–38) describes some other interesting examples from primate communications.

In some primate species individuals vocalise before moving to another area, the vocalisation apparently carrying information to coordinate the group. These vocalisations have not yet been analysed in detail, but Hauser notes this is screaming out for grammar (and is the same kind of scenario claimed by Pinker, 2003, as a selection pressure for language).

Mitani and Marler (1989) analysed the calls of male agile gibbons (Hylobates agilis) and found strong evidence for some kind of ordering rules in their calls. These calls consist of strings of several discrete signals, and in certain cases semantics depends critically on ordering. Mitani and Marler suggest several other primate species do this, too. Robinson (1984) concluded that wedge-capped capuchin monkeys (Cebus olivaceus) use syntactic complex calls.

This evidence suggests that some primate species make use of duality of patterning, and possibly productivity.

2.5.3.3 Conclusion

The conclusion I reach is that few conclusions can be drawn about the relationship between lesser ape and monkey communication and human language. Although the linguistic line between humans and animals clearly cannot be drawn as boldly as it once was (human language conveys meaning, animal communication conveys affective state), there is so far scant evidence that these animals utilise more than a handful of the features attributed to language. Most importantly, no primate groups seem to stand out as possessing these traits, suggesting that any sophisticated communication methods encountered are the product of parallel evolution.

2.5.4 Great Apes

The great apes (gorillas, chimpanzees, bonobos and orangutans – see Figure 2.3) have, understandably, received more attention by both cognitive and language researchers. I examine the three broadly different approaches below: ethology, cognitive ethology and ape acculturation and language teaching.

2.5.4.1 Teaching Language to Apes

Researchers have been investigating the possibility that great apes could learn language for about a century. Steinberg et al. (2001) give a good historical review.

Furness (1916) spent four months trying to teach an orangutan to speak. The subject learnt only two words, “papa” and “cup” and the experiment ended when the ape died of a fever, during which it repeated its vocabulary.

Kellogg and Kellogg (Kellogg and Kellogg, 1933; Kellogg, 1968), husband and wife psychologists, tried raising a female chimpanzee called Gua alongside their son Donald. The experiment ended after nine months when the Kelloggs feared Donald was aping the ape’s behaviour too much. The Kelloggs gave Gua no additional speech training, and she did not learn to speak, but they allege that by the end of the experiment she could understand about 95 words and phrases. Such spontaneous
Figure 2.3: The evolutionary relationship between humans and the great apes. The approximate dates of division from common ancestry are indicated at the right in millions of years ago (Ma).

language learning has not been reported since - possible explanations include Clever Hans tricks, and the effect of Donald on Gua’s development. A similar experiment by Hayes (1951) produced less impressive results.

Gardner and Gardner (1969, 1975) decided that apes were physiologically incapable of speech, an approach which has been generally adopted since. The Gardners raised a female chimpanzee, Washoe, in an institutional environment, teaching her a simplified form of ASL (American Sign Language). After four years Washoe had supposedly learned about 130 signs and some 2–3 word utterances, such as “go sweet”, at roughly a 2–3 year old human level. The validity of the signs and the combinations has been questioned (Wallman, 1992).

Other signing ape experiments have been undertaken with chimpanzees (Fouts, 1973; Fouts et al., 1989; Terrace, 1979; Terrace et al., 1979), Gorillas (Patterson, 1978a,b, 1980), and an orangutan (Miles, 1983, 1990) and artificial language teaching experiments (using tokens or symbols instead of signs) have been done with Chimpanzees (Premack, 1970, 1971, 1976; Rumbaugh, 1977; Savage-Rumbaugh and Rumbaugh, 1978).

Savage-Rumbaugh et al. has continued this technique (Savage-Rumbaugh et al., 1986, 1993) with the bonobo (Pan paniscus) Kanzi, who at 5 years supposedly had the grammar of a two year old and about 250 “words”. The researchers also claim Kanzi uses fully productive syntax, though this is highly disputed because Kanzi lives in a very social environment with looser environmental controls. Despite this, other primate researchers, such as Tomasello (2000) believe Kanzi provides the most rigorous ape-language research to date.

Steinberg et al. (2001) criticise work in this area for concentrating too much on production by apes, neglecting comprehension (researchers often believe their subjects do understand a symbol, but only get excited when the ape uses it). Although there may be little quantitative information on ape comprehension, the research suggests a learning pattern similar to humans, where production lags comprehension (see, for example Gardner and Gardner, 1969).

Wallman (1992) and Tomasello (2000) criticise ape-teaching studies methodologically, claiming they are almost all flawed and can provide only anecdotal evidence. Indeed Wallman reports that one of the chimpanzee researchers, Terrace, started an optimist but changed his mind as the result of his

---

3Clever Hans was a nineteenth century horse who could apparently solve arithmetic problems that were presented on a blackboard, giving the answer by stamping a foot the appropriate number of times. A committee of enquiry discovered that Hans only gave the right answer if the questioner knew the solution, and psychologist Oskar Pfungst showed that Hans was actually acting on unconscious cues given by his questioners.
It would seem that whatever capacity apes have to use language is extremely limited. It seems they lack something unique to humans that facilitates language. Next we turn to the similarities between apes and humans, by looking at primate cognition.

### 2.5.4.2 Comparing Cognition of Apes and Humans

Apes are very similar to humans but do not have language. When trying to understand the key cognitive developments in human evolution that made language possible it is useful to understand how ape cognition differs from our own. Note that this does not rule out shared aspects of cognition as explanations for parts of language. At the same time any cognitive differences between apes and humans are good candidates for critical developments that allowed language as a whole to emerge.

Although there was some early work on ape cognition (notably Köhler, 1925; Yerkes, 1916), dominance of the behaviourist school of psychology in the middle of the twentieth century suppressed this kind of research. During this time the ape language-teaching studies described in Section §2.5.4.1 were the main method of cognitive analysis (whatever the researchers’ motivations for such studies, they could fit into either a cognitive or behaviourist theoretical framework since they were purportedly studying the behaviour of language, not the cognition behind it).

More recently, following the cognitive revolution in psychology, ape cognition has again become a respectable research area. Primate comparative psychology draws upon both the ethological and enculturation studies already looked at, but also makes use of a different experimental paradigm, that of the experimental psychologist. In this section I summarise relevant discoveries regarding ape cognition, drawn mostly from the Tomasello and Call’s (1997) comprehensive review.

#### Space and Objects

An important part of spatial cognition (when compared to human models) is object permanence. This was initially investigated by Piaget (1954), who created several stages of object permanence. Two are relevant here: stage 4 object permanence meant the subject could retrieve an object after they saw it hidden under a cover, stage 5 object permanence meant the subject could correctly retrieve an object after they saw it hidden under one cover, then moved and hidden under a second cover. According to Tomasello and Call (1997), stage 4 and 5 permanence is well established in chimpanzees and gorillas, as well as other non-apes.

Object permanence is presumably required for a further feat of spatial memory, cognitive mapping. Early observations by Tinklepaugh (1932) with chimpanzees have been confirmed, and apes have good memory for food locations and changes to the non-immediate environment. Chimpanzees and orangutans also show ability for sophisticated reasoning about locations, following energy-efficient routes to pick up hidden food for example. Gorillas do not show this behaviour.

A final relevant point is the flexibility of spatial skills, that is can apes use spatial reasoning outside its “natural” domain? The answer seems to be yes, shown in tasks such as computer-maze solving and novel spatial problem solving.

#### Features and Categories

Tomasello and Call (1997) describe some similarities and some disparities between ape and human feature recognition and categorisation.

So-called learning to learn is one aspect of feature-based cognition. This describes an experimental method where subjects are taught a number of different discrimination tasks, and the learning to learn effect is observed if they begin to generalise the kind of learning required and thus require less training on subsequent discriminations. Humans, apes and other mammals have this ability, which is taken by many researchers to indicate conceptual mediation of the learning task, with some kind of conceptual representation of the “rules” of the task.

Similar evidence for cognition comes from cross-modal perception studies, which indicate apes have some mental representation of objects that allows them to be recognised through a different modality than that of first encounter.

One important difference between humans and apes in this area is the natural kinds of categories employed. Apes apparently make use of specific categories (e.g. crow) and generic categories (e.g. predator), but not basic categories (e.g. bird) as humans apparently universally do.
Apes share with other primates the ability to form relational categories, an ability that is perhaps unique to the order. Categorisation and sorting by constraints such as “same as” and “larger than” have been documented experimentally.

**Social Learning** Social learning is of particular interest to evolutionary biologists since it is a method of information transmission outside biological evolution (see also sections §2.9.1 and §2.8.2). It is also of importance for language since language is inherently a social phenomenon and requires social learning in its acquisition. Boyd and Richerson (1985) argue that social learning is the most useful adaptation for species in rapidly changing environments since it allows behaviour to adjust far more rapidly than natural selection can provide.

Tomasello and Call (1997) list several degrees of animal learning from goal directed behaviour.

1. Animals learn things about the results of others’ behaviour, but only with reference to the physical environment, rather than behaviour itself.
2. Animals learn the possibility of state-changing actions by observing the manipulations of others (emulation learning).
3. In communication, signals are learnt by ritualisation during development. This is strongly implied in chimpanzee communication, but has not been studied elsewhere.
4. Imitation. The evidence for imitation is patchy, with only human-raised apes showing strong imitative skills (in the wild apes apparently do not imitate). Possibly monkeys are imitators too, but there is insufficient research to back this claim.

In particular the proposition that chimpanzee communication is learnt by ritualisation while immature is interesting, indicating cultural transmission.

**Social Strategies, Communication and Theory of Mind** Animals living in social groups are able to exploit an entirely new set of environmental factors, namely their conspecifics, through social strategies. The social strategies which have received most attention are deception and communications, with a link sometimes posited between them.

There is good evidence for deception in apes (and other primates). Deception comes in degrees, just as it does in humans. There are reports of concealment; for example Goodall (1986) reports that when chimpanzee groups patrol for other chimpanzee groups they are quieter than usual. In another example, this time of active misleading, Goodall describes a young chimpanzee who was not getting enough food in a group feeding situation. This chimpanzee would suddenly move off, an action which is usually followed by a group no matter who initiates it. He then led the group in a circle back to the food where he was now in a better position. Tomasello and Call (1997) also describe a few examples of counter-deception (most of which were observed in chimpanzees, see Menzel, 1973). Byrne and Whiten (1990) summarise reported cases of primate deception.

According to the data reviewed by Tomasello and Call (1997), ape communication is of a similar level to that seen in other primates, though perhaps slightly more flexible, especially in light of the enculturated apes described above. One kind of communication that is important for many primates, including apes, is grooming or huddling, used to reinforce social relationships like kinship, “friendship” and dominance.

One final and contentious point about cognition in apes is theory of mind (Premack and Woodruff, 1978), tied to both deception and communication. Apes with theory of mind are supposed to understand the intentions and beliefs of others, but there is no strong evidence for theory of mind in any ape, according to Tomasello and Call (1997). Indeed, building on evidence that chimpanzees have no coherent knowledge of others’ mental states regarding seeing, propose escaping the “argument by analogy”. By this they mean abandoning the assumption that similar behaviour in humans and non-humans has similar psychological causes (essentially this is a claim for a homoplastic relationship between complex behaviour in apes and humans).
Deception and theory of mind are important for some theories of language evolution (see section \(\S\) 2.4.1.2) which claim development was driven by social cognition. In particular, without a theory of mind to understand the beliefs of others, how could human language develop with its apparent function of manipulating and adding to others beliefs?

**Conclusion** In summary, apes possess good spatial skills. Their classification of features seems similar to ours but they use different natural kinds. They seem capable of some cultural transmission, but are not adept imitators. They also seem capable of deception, though they perhaps lack theory of mind.

### 2.5.5 Other Analogues: Dolphins, parrots and bees

Primates are not the only focus of studies of communication, nor always the most interesting. The paradigms above have also been applied to other branches of the animal kingdom.

Indeed Frisch (1967a,b), one of the founders of ethology, found that honeybees use a very complex communication system, using dance to inform conspecifics of the location of food. Bees communicate information about distance, direction and quality of food using a semantic communication system which is at least partly arbitrary and partly discrete - and of course makes use of displacement. Birds are also heavily researched. Marler (1970) noted many parallels between bird song and human language, including traditional transmission, interchangeability and total feedback, as well as more specific parallels, such as left-hemisphere dominance for production, babbling in infants and immature birds and a critical period of learning.

Hailman and Ficken (1987) analysed the calls of black-capped chickadees (*Parus atricapillus*) and showed that their calls, made up of four elements \(A-D\) conformed to the regular language

\[
\text{suggesting chickadees use some productivity (but not, apparently, duality of patterning).}^4
\]

Pepperberg (1993) tried teaching a simplified form of English to an African grey parrot (*Psittacus erithacus*), apparently with great success. Her subject, Alex could discuss the colour, shape and composition of over 100 objects as well as abstractions about them such as number and colour. Combining these results with analysis of Alex’s phonological errors suggest he could use an arbitrary semantic communication system involving duality of patterning, production and traditional transmission.

However Alex’s reported syntactic ability is not as good as that reported for dolphins. Herman and Wolz (Herman et al., 1984; Herman and Wolz, 1984) trained two bottlenose dolphins (*Tursiops truncatus*), Akeakamai and Phoenix, to use a simple language, each using a different modality. Akea learnt a gesture-based language and Phoenix a sound-based one, both consisting of about 30 words organised into simple phrases (up to five words long). Of great interest was the fact that not only could the dolphins learn the languages, but they also showed productive ability (could understand sentences they had not heard before) and apparently used syntax to decode meanings. In particular they could distinguish semantically reversible sentences appropriately (sentences where roles must be determined from their position).

### 2.5.6 Conclusion

This is a wide field. Despite being taught, it seems that even our nearest relatives cannot grasp language. Yet in many other ways our cognition seems similar to theirs. There really is *something* that humans have but other apes lack which allows us to use language with great facility. Exactly what this is we cannot yet say.

---

4The notation \(X^*\) means the symbol \(X\) repeated any number of times, including zero.
2.6 Neuroscience Perspectives on the Evolution of Language in Humans

2.6.1 Introduction

Our closest living relatives, the great apes, do not have language and so it seems there must be something special about the human brain that makes language possible. There are a number of theories that posit such links. I am particularly interested in the theories that propose a link between sensorimotor cognition and language. In this section I review four such theories including one, in Section §2.6.5, which is closely related to the theory I will develop about noun phrase structure.

2.6.2 Mirror Neurons

Rizzolatti and Arbib (1998) suggest that language might have evolved out of mirror neurons, which are found in the F5 region of the monkey brain. Mirror neurons are so named because they fire both when the monkey performs an action (such as grasping) and when the monkey observes another individual performing the same action. Monkey F5 is partly homologous to Broca’s area in humans, which is important for language. Rizzolatti and Arbib (1998) believe a shared system of communication could evolve from this mirror system because signals stimulate the same neurons whether made by oneself or someone else (Arbib, 2003, calls this feature parity, and it is approximately Hockett’s interchangeability – see Table 2.1). Furthermore, since these cells occur in the cortical motor system Arbib (2003) believes this allows language to be under intentional control. This is known as the mirror system hypothesis (MSH).

Hurford (2004) argues against the MSH as an important driver of language evolution. He believes MSH cannot explain the Saussurean sign, the mapping from sound to meaning (Figure 2.4). MSH could explain the parity of sounds or concepts, the components of the sign, but not the mapping between them. The mapping is arbitrary, while mirror neurons associate actions which have the same form independent of who performs them. Hurford also notes that mirror behaviour is probably very common in nature, from behaviour like flocking (imitating neighbours movements) to neural control over pigmentation in cephalopods to blend into the background.

However Arbib continues to favour the MSH, proposing a series of steps from a simple mirror system which probably evolved as a way to correct motor errors to a system which allows imitation and then symbolic communication. Arbib considers the development of arbitrary, intentional symbolic communication to be main problem of language evolution. After this syntax would develop culturally. Arbib’s argument receives some support from the fact that humans seem to be far better at imitation than other primates, something that a developed mirror system could provide.

2.6.3 Gestures

Corballis (2003) believes that early language was based entirely on gestures, much as modern sign
languages are. In his story the arms and hands were originally required for human language because in primates these are under intentional control, unlike the non-intentional primate communications controlled by the limbic system. However the left-brain communication centres were not wasted by human language, and this led to left-brain dominance effects which are most visible in the right-handedness of the vast majority of humans (other primates do not show this asymmetry).

2.6.4 Sequencing

Lieberman (2003) puts forth two views relating to language development. The first, relating to laryngeal lowering, is discussed below. His other argument is that language coopted the basal ganglia, a part of the brain responsible for executing and monitoring motor commands at a low level. The basal ganglia are good sequencing, and Lieberman cites evidence that damage to this area can cause aphasias. He believes that homologues of ordinary motor sequencing are responsible for the sequences and structure of phonology, syntax and even semantics by ordering cognitive states. The basal ganglia are an important part of Lieberman’s functional language system (FLS) which is supposed to be distributed over several parts of the brain.

2.6.5 Predicate-Argument Structure

A related argument made by Hurford (2003b) is that predicate-argument \((P(x))\) structure in language is simply a reflection of the same structure in the sensorimotor system. His interpretation of the distinction between dorsal and ventral processing (see Section §3.3.4) in human vision attributes predicates to the ventral stream and arguments to the dorsal stream. This distinction is then reflected in the rest of reasoning and language as a matter of course.

Hurford tries very hard to make the link, but others are sceptical. Why, for instance, if predicate argument structure is so basic to language is only one argument (the subject noun phrase) set off from any other arguments of a sentence’s verb phrase. Also, since the predicates detected by the ventral stream include such things as colour and category, why are these things expressed using adjectives and nouns? If there is a causative connection between vision processing and predicate-argument structure in reasoning, this is apparently not so directly reflected in language.

As I show in this thesis, these problems are not insurmountable. My own hypothesis owes a lot to Hurford’s.

2.7 Archaeological Perspectives on the Evolution of Language in Humans

Comparative biology is one of two mainstays of empirical evolutionary biology. The other is palaeontology, which in the case of recent human development means physical anthropology and archaeology. Language development has been a focus of speculation in physical anthropology ever since Darwin (1871), and today has regained importance as the (dubiously important) marker of when humans became really human (other possibilities having fallen out of fashion: man the toolmaker, man the hunter and so forth). Archaeology treats language more circumspectly, but acknowledges the importance of contributions from linguistics (Blench and Spriggs, 1997).

2.7.1 Physical Anthropology

Davidson (2003) is critical of stories about language evolution that rely heavily on assumptions about fossils. The anatomical features that are important for language are largely soft (larynx, brain, nerves etc.) and so leave only indirect traces in the fossil record. Davidson (2003, p. 145) provides a brief survey of this evidence, summarised below.

Brain Cranial casts are supposed to allow inferences about how big the brain was overall as well as the size and shapes of particular regions (such as Broca’s area). However these conclusions rely
on assumptions that cerebral regions are fixed, as well as the uncertainty of determining brain shape from cranium shape.

**Throat and Larynx** Flexion, or curvature, of the base of the skull supposedly indicates a lowered larynx. Lieberman (2003) uses this to argue *Homo Neandertalensis* lacked the vocal range of *Homo Sapiens*. However others (Lieberman and McCarthy, 1999) show the shape of the skull depends on development of other parts of the skull, so is unlikely to be purely the result of linguistic pressures. Also, Fitch (2002) shows that laryngeal lowering is more widespread in animals than previously thought, and in particular that anatomy of dead animals, even the soft parts, is inadequate to determine the extent of laryngeal lowering.

**Hyoid bone** The earliest surviving hominin hyoid (the u-shaped bone that supports the tongue) is within the range of modern humans, however it is also within the range of many modern non-human primates.

**Hypoglossal canal** The hypoglossal canal (the gap in the skull through which the nerve that controls the tongue runs) has only been within the range of modern humans for about 400,000 years, according to Kay et al. (1998). However this is disputed by DeGusta et al. (1999) who claims all fossil hominin canals are within the range of modern humans.

**Spinal cord** According to MacLarnon and Hewitt (1999) the thoracic region is serviced by fewer nerves in early hominins than either modern humans or Neanderthals.

The only undisputed fossil evidence is then the spinal cord data, and even this may only be due to its recency. Because the fossil data is so equivocal Davidson and others (Carstairs-McCarthy, 1999; Richards, 1987) believe that little information regarding language evolution can be gleaned from the hominin fossil record.

### 2.7.2 Culture as a Proxy of Language

The alternative approach to using skeletal remains as pointers to linguistic ability is to look for signs of language in artifacts left behind by fossil hominins. Archaeologists and anthropologists generally assume that human culture is closely tied to human language and attempt to divine something about linguistic progress from remains of tools, campsites and other effects. However, just as with fossil remains, artifacts that have survived thousands or millions of years for us to find them provide only a skeletal description of the minds that created them. Archaeologists advise caution when artifacts are interpreted.

Davidson (2003) examines current ideas regarding stone tools, one of the most common kinds of artifacts, and finds them wanting with regard to language evolution. There are three recognised modes of tool production found in the fossil record (also known as *industries*). Davidson presents the prevailing view as a close association between different hominin species and the progression of tool modes.

However now that more evidence has accumulated, this simple model is no longer adequate. Foley (1987) argued for a close species-mode link, but more recently (Foley and Lahr, 2003) backs away from that interpretation. It turns out that Mode 1 tools (Oldowan industry, simple flakes) were used by *Homo habilis*, *Homo ergaster* and *Homo erectus*, Mode 2 tools (Acheulean industry, large shaped hand-axes) by *H. erectus*, *H. ergaster* and *Homo heidelbergensis*, while Mode 3 (Levallois and Mousterian industries, variety of shaped tools) made by *Homo neandertalensis*, *Homo helembi* and *Homo sapiens*.

Foley and Lahr (2003) continue to interpret the complex distribution of tool modes around time, space and species as an indication that tool manufacture has a phylogenetic base, assuming that the cognitive processes behind tool making are conserved between species.

Davidson (2003) argues a different line, saying that culture implies cultural transmission (whether by language, imitation or whatever). Cultural transmission should only occur *within* species, though, so he finds it difficult to see how tool modes could have been culturally transmitted since this would
require interspecies transmission. In short, he believes that stone tools have little to say about the presence or absence of language (a - perhaps the method of cultural transmission).

Davidson (2003, see also Noble and Davidson, 1996) prefers an approach to the archaeological record that focuses on imposition of form, rather than technical sophistication as judged by our own standards. With respect to language this means looking for concomitants of a linguistic mind in artifacts, and for Davidson and Noble the key is symbolic thinking. A concomitant of symbolic thinking, they believe, are the design features of productivity and displacement (see section §2.5.2 for description of these) derived from Hockett and Ascher (1964).

2.8 Cultural evolution

2.8.1 Introduction

In this section I examine the question of how one can explain the universality among humans of some language features without recourse to biological evolution.

Of course no cultural evolution model can stand alone. Some theorists believe a language faculty evolved entirely through natural selection of the brain; cultural evolutionary thinkers prefer to think of language as an emergent property of a network of human brains, but they still require some biological preadaptations before this can begin. After all, humans are still the only species with language, so the question of necessary and sufficient biological features for language remains.

In Section §2.8.2 I discuss a promising simulation approach for studying the effects of constraints on language structure when language evolves culturally. In Section §2.8.3 I discuss the more conventional linguistic study of cultural evolution within language, historical linguistics.

2.8.2 Iterated Learning Models

The paradigm known as the Iterated Learning Model (ILM) (Hurford, 2000, 2003a; Kirby, 2000; Kirby and Christiansen, 2003; Kirby and Hurford, 2002), and related models (Batali, 1998; Steels, 1998; Steels et al., 2002), show that under the right conditions culturally transmitted communication systems can evolve similar features to those found in human languages. Many of these universal features can thus be explained by ILM practitioners as manifestations of more general constraints on learning.

The main innovation of the ILM approach is that it provides a rigorous way to test theories of cultural evolution. Kirby and Christiansen (2003) explain that the iteration of learning repeatedly moves a language from the minds of speakers (I-language) to utterances (E-language) and back. Constraints on the transmission channel and learning properties of the speakers constrain I-language and E-language respectively, and lead to evolutionary change that better suits these constraints.

Specific questions in the ILM paradigm are examined mainly by computer simulation, in a manner strikingly similar to that used by biologists testing their own evolutionary theories. The simulations take the form of language games between agents which are abstracted as far as possible to allow examination of just the question at hand. The actual implementation of the agents is not so important, indeed this method easily transcends the traditional break between symbolic and subsymbolic approaches.

The ILM is a relatively new approach, but several such assumptions are apparent. Here are some examples.

Observational learning Agents in ILM simulations are assumed to be capable of learning meaning/utterance mappings by observing their use by other agents. As noted in section §2.5.4.2 this kind of learning is not widespread among primates, so this is not a trivial assumption.

Perfect information ILM models seem to rely on agents always having perfect information about which utterance refers to which meaning. The difficulty of this discrimination of good learning examples from bad learning examples seems to receive little consideration.
Rule learning The ILM requires agents to generalise from specific learning examples to rules of the language. This property is not difficult for a brain to meet, but many species use only stereotyped communication.

Predispositions Importantly the predictions that ILM makes about languages arise from constraints posed artificially on the system. For example there is extensive evidence (Christiansen and Chater, 2001) that aspects of human preferences for parsing (such as avoiding multiple centre embeddings) could be the result of the non-linguistic constraint of serial learning. The implication is that serial learning is a biological predisposition of human learning that also constrains the form of language.

2.8.3 Historical Linguistics

If language is considered an aspect of evolving culture, then historical linguistics is the discipline that documents the nature of this evolution. Historical linguists use recognised trends of language to predict how earlier protolanguages would have been in sound, grammar and meaning, but this process only works on the scale of hundreds or thousands of years. (Crowley, 1997; Newmeyer, 2003)

Of particular interest to those seeking a cultural explanation of the origins of language is whether language evolution shows directionality - do the identified trends show a consistent change from one style to another? The existence of unidirectional trends would challenge what Newmeyer (2002) calls the uniformitarian assumption, that all languages are in some sense equal and have been for a long time, and give strength to the view that the languages of today evolved from simpler languages of the distant past.

Most historical linguists answer that there are no unidirectional changes. Trends in one direction are countered by similar numbers of opposing trends, as with strengthening and weakening of sounds, and some even run in well-defined cycles like the path from isolating languages through agglutinating languages to inflecting languages and back again.

Newmeyer believes that the uniformitarian assumption is wrong, however, and cites some examples that seem to show directional change. Perkins (1992) supposes that complex deixis (reference) systems are associated more with simpler cultures, perhaps as a way to isolate themselves. Webb (1977) considers the possibility that transitive have and equivalents are the result of property rights in Western cultures. These examples are even more interesting in that they suggest languages adapt to their cultural environment, a view consistent with cultural language modellers as discussed in section 2.8.2.

In fact historical linguistics shares a great deal with evolutionary biology in the way of methods and even terminology. The comparative method is essential to both, and by examining historical written records a kind of fossil history can be traced for the relevant languages. One recently revived subfield, known as glottochronology, involves applying the techniques of molecular phylogeny discussed in section 2.4.2.1 to words in an attempt to date and locate changes in a family of languages (Balter, 2004; Gray and Atkinson, 2003).

However languages evolve much faster than organisms, and with such a limited window into the past there is little hope that historical linguistics can reveal any specific facts about the origin of language. What the field does reveal is the trends of language change occurring now, from which evolutionary thinkers might be able to glean information about the extent of cultural language evolution.

2.9 Conclusion

As might be expected for a field that is so young, and draws upon so many different disciplines, theories to account for language are many and varied. This makes summarising them very hard. The approach I take here is to list the main conceptual dimensions in which these theories sit. Some apply better to some theories than others. By examining these features I can build an idea of those on which I need to take a stand and those where I can afford to remain uncommitted.
Symbolism versus subsymbolism

One dimension, which extends into wider cognitive science, is the difference between symbolic and subsymbolic (or connectionist) methodologies. More than most areas of cognition language lends itself to the symbolic approach because it is itself a symbolic system. This is the approach taken by some. When modelling the cognition behind this symbolic system others prefer to take a subsymbolic approach, especially those interested in grounding language directly in the brain.

Symbols versus structure

Another dimension is the evolutionary primacy accorded to different aspects of language. Some theorists believe symbolic representation to be the key development necessary for language (Hockett’s semanticity and arbitrariness). Others believe structure – the combination of symbols – is the key development (Hockett’s productivity feature).

Syntax

Perhaps related to this is the kind of syntactic theory involved. The most elaborate syntactic approaches belong to those theories, like the Chomskyan ones, which focus heavily on syntax. Those where syntax is not so pre-eminent, like cognitive and functional linguistics, adopt less formal syntactic theories.

Semantics

The type of semantics is another dimension. Chomskyan linguistics proposes disembodied semantics based on formal logic. Cognitive and functional linguistics propose embodied semantics based in cognitive processing.

Sharing of resources

Another dimension ranges from the use of discrete and dedicated mental resources for language on one hand to the use of entirely shared and general mental resources on the other. This is a continuum. At the discrete end is the conventional position that there is an innate language module. At the other end are statistical learning theories that propose language acquisition is just a special case of the kind of general learning used for everything else. Other theories sit along this continuum.

The extreme modularist view (Fodor, 1983) is not tenable. All the models I have reviewed build on existing structures. And except as the result of an evolutionary jump the cognitive bases of language must have been subject to selection pressures from outside language. But as discussed in section §2.4.2.2 these sudden evolutionary jumps are extremely unlikely. As a last hope for the strict modularist it is possible that the evolutionary basis of language is separated from sensorimotor processing by a chasm that evolution cannot bridge. But this seems unlikely, especially in light of known peripheral connections like the colour names discussed in Chapter 1.

The medium of evolution and language universals

The origin of cross-linguistic universals forms another dimension. Are these effectively arbitrary and hard coded into the brains of speakers, either as hard rules or tendencies, as proposed by Chomskyan? Or are they the necessary or at least probable result of other constraints imposed by some part of the linguistic environment as shown by the ILM school?

Another way to look at this is through the medium of language evolution. Some theories concentrate on biological evolution, others on cultural evolution. In some cases this is because the theorists believe only one or the other is necessary. More usually theorists assume both are involved while taking an interest in either the biological or cultural aspect.

With all these differences (and there are others, too) it may seem pointless to attempt any kind of summary. But let us survey the field again, keeping in mind this overarching question: does sensorimotor processing impact the structure of language? In answering this question I find that most of the differences melt away. Some are irrelevant, some are subordinated to others and some cease to exist.

In Section §2.9.2 I will show how consideration of the sharing of resources, allows this question to be answered. First, though, I will show that the medium of evolution, isn’t really a difference at all when answering this question.
Figure 2.5: The apparent unit of selection. At first glance, biological evolution apparently relies on the replication of organisms.

Figure 2.6: The actual unit of selection. In the modern view it is genes, not organisms, that are subject to natural selection. The effective path of replication is shown by the dashed line. The actual path of replication is much more complicated, though, requiring the intermediate stage of an organism. This path defines the constraints on the evolution of the gene.

2.9.1 Language is the Result of Natural Selection

Although it is often convenient to treat biological evolution and cultural evolution separately, they are really just two instances of the same abstract process. Natural selection is the driving force behind the biological and cultural aspects of language evolution. Figure 2.5 shows a naive idea of the biological “unit of selection”, the actual entities that replicate and compete via natural selection. This used to be thought to be the organism or even the entire species.

In Figure 2.6 we see the real unit of selection. It is genes, not organisms, which compete (Dawkins, 1976). An organism is just a gene’s way of making another gene, and the organism and its environment are the path to which genes must adapt to replicate successfully and flourish. Genes do this by producing organisms that best fit the constraints placed by the environment and other genes sharing the same organism. In Figure 2.7 we see the same approach applied to cultural evolution of linguistic entities. Now the entities are lexical and syntactic items and rules. These rules live in minds, but reproduce via linguistic utterances from which the rules are inferred by other minds. The path of replication now takes in the medium of communication along with the cognitive mechanisms for uttering sentences and inferring rules from them, and cognitive and social constraints. For example,
Figure 2.7: By analogy with Figure 2.6 we arrive at the linguistic unit of selection. These are the rules and words that make up the grammar and lexicon of a language. Again the dotted line shows the effective path of replication of rules. The solid lines show the true path, which takes in the mechanisms required to utter sentences, transmission through the medium and inference of rules by the learner. Again, this path defines the constraints on the evolution of the rules.

rules that the brain cannot learn from examples will not flourish, nor will words that are not useful for communication.

This approach to linguistic evolution is essentially the same as that of the *meme* proposed by Dawkins (1976), with the scope limited to linguistic knowledge.

When one realises that natural selection applies to language evolving either biologically or culturally, the medium of evolution suddenly becomes less important. Natural selection fits entities to constraints, and a great many of the constraints are shared, for the most part, regardless of the medium of evolution. In particular, the sensorimotor system constrains both kinds of evolution.

The impact of sensorimotor constraints on language has been explored before. Of special interest is Regier (1992) which uses a connectionist perceptual model to provide constraints on acquisition of spatial terms. The approach taken is similar in many respects to that taken in this work. Regier also discusses other research in the same vein.

### 2.9.2 Sensorimotor Processing Almost Certainly Constrains Language

Given the evolutionary concordance presented in the previous section I believe I can address the question of sensorimotor effects on language just by examining the way resources are shared between the two faculties.

I can divide models of language into three types as follows.

**The first type of model** These are the narrow biological models that propose overlap between the neural circuits of the sensorimotor system (or their homologues) and the neural circuits responsible for language. In this case we would expect sensorimotor idiosyncrasies to manifest themselves directly in language simply because language coopts sensorimotor circuitry. These are purely biologically inspired models.

**The second type of model** These are the embodied models that make language a part of general cognition with no direct dependency on the sensorimotor system. Here linguistic rules either emerge through cultural evolution, driven by constraints on general learning, or evolve biologically as innate tendencies in general learning (e.g. via the Baldwin effect, Section §2.4.1.4). Section §2.9.1 shows us that these rules would be subject to the same kinds of selection pressures either way.
As well as pressures on transmission and perception of speech and signs there is pressure on learning associations between language elements and their semantics, which, in the case of the world around us, are filtered through the sensorimotor system. There is a fairly clear evolutionary pressure for the interface between language and the sensorimotor system to be as simple as possible. We should therefore expect to discover features of language that correspond to idiosyncrasies of the sensorimotor system that partly constrains it.

**The third type of model** These are models that propose language is an entirely independent module with its own rigid and arbitrary rules (at least relative to sensorimotor processing). The paradigm is Fodor’s (see Section §2.2.1.3). In this case any correspondence between linguistic and sensorimotor phenomena would be coincidental. Models where syntax or semantics are treated as abstract formal systems are liable to fall into this category.

As I showed in the start of this section the third type is also the least likely to correspond with reality. Both the first two types are far more likely.

I do not propose that every model of language evolution can be pigeonholed precisely into one of these types. Most models seem to fit into two or more and it is this that makes me especially hopeful that the sensorimotor system affects language structure: even if the third type corresponds to reality it is likely to do so only in concert with one or both of the other categories.

Based on this assumption I conclude that, contrary to first impressions, the sensorimotor system is very likely to have an impact on the structure of language. Indeed the question isn’t really whether sensorimotor processing affects linguistic structure, but how it does so.

To make a start considering this question I turn at last to my own specific hypothesis relating sensorimotor processing to linguistic structure. This is the subject of the next section.

### 2.9.3 Noun Phrases and Objects

I am interested in linking aspects of the noun phrase to aspects of object perception and attention. The idea that noun phrases correspond to objects may seem an obvious one, and the relationship is often asserted. Langacker (1986, 1987) actually took the care to tie this relationship down. When the assertion that nouns refer to objects is made, it is with nouns like *concrete* in the following sentence.

(1) Fred poured the concrete.

But nouns can just as easily refer to things that are not objects in the conventional sense, as in *construction* here.

(2) The construction was going well.

This is often thought to mean that nouns therefore don’t represent objects, or any semantic class, at all.

Langacker claims, as part of his cognitive grammar, that nouns do have a semantic counterpart, a thing. A thing is defined to be a region in some domain, where the domain itself may be abstract.

Even though the domains into which nouns refer under this scheme may be abstract, Langacker’s (1987) approach relates things back to the spatial domain semantically using metaphor as described in Section §2.2.2. This unification of the semantics of the noun phrase is hopeful, but it still relies on the semantics of references to concrete objects at root.

I intend to model the semantics of the noun phrase via sensorimotor processing. I won’t commit to a particular way that non-perceptual objects fit into my theory – perhaps it is by some metaphorical system on the semantic side like Langacker’s. Perhaps constraints on referring to perceptual objects during language acquisition set the pattern for reference to all objects, perceptual or abstract. Perhaps the cognitive processing corresponding to abstract objects is actually separate from that for perceptual objects, but homologous to it. There are many possibilities, all I need to know is that they exist, and I can move on to model perceptual object processing.

To do that I should tie down exactly what it is I am modelling. Strictly speaking, I am not modelling the semantics of noun phrases as objects. Instead I am modelling their semantics as the
processing that goes on in perceiving and attending to objects. This will be explained more fully in Chapter 6.

My methodology will be a mixture of those that I reviewed in Section §2.2. From the Chomskyans I will be using X-bar theory as my syntax model. But my semantics will be far closer to cognitive linguistics, depending on embodiment and cognitive constraints. Finally I will have a foot in both camps with regard to the source of some universals – I am happy to explain the behaviour of noun phrases referring to perceptual objects in terms of my model, but as I stated above I won’t speculate too much on why all noun phrases behave this way.

My approach is very close to that taken by Hurford (2003b, see Section §2.6.5 for details). Hurford proposes that the distinction between predicate and argument in the clause corresponds to a division of labour in the visual system between the ventral and dorsal streams. This division will also play a key role in my theory of noun phrase structure. In Chapter 6 I will show how my visual model explains three elements of the noun phrase and their relationships: reference (Hurford’s argument), type (Hurford’s predicate) and linguistic number, which Hurford does not touch on.

The remainder of the thesis, apart from Chapter 6, is devoted to developing a computational model of object processing embodying this hypothesis. Next I turn to review the literature relevant to object perception and attention.
Chapter 3

Visual Object Perception and Attention

3.1 Introduction

This chapter is a survey of models of visual perception of objects, and of the attentional mechanisms which are associated with this process. Two main approaches stand out in the study of any kind of cognition, the psychological approach and the neuroscience approach. Psychologists tackle the problem by studying human behaviour in controlled situations and postulating cognitive structures and processes to explain the observed behaviour. Neuroscientists observe or manipulate the actual neural activity associated with different behaviours using evoked response potentials, brain lesion studies, drug studies and advanced scanning methods such as functional magnetic resonance imaging.

In fact the parts of the perceptual system which are most worthwhile to model computationally are those where the two disciplines overlap, since they represent psychological processes specified well enough for implementation on a computer and neuroscientific knowledge abstract enough to test in another medium.

Since the two methodologies are complementary I have chosen not to keep their accounts separate below, but instead try to present a coherent story of the visual pathways, with particular regard to the aspects relevant to my own investigation into visual cognition.

When discussing object perception and related processes there can be a confusion with the terminology used. Everyday English is not particularly well equipped to distinguish between the various aspects of perception, with the result that useful terms such as recognition and identification risk losing their utility. For that reason I begin by defining some of these terms as I intend to use them in the hope of avoiding such confusion. My definitions borrow from conventions, though note that they are, at least to some degree, arbitrary.

Object detection is the process of determining that there is an object and where it is located.

Object classification places a detected object or its features into a class (such as chair, face, red or smiling).

Object identification matches a detected object with a previously encountered object (your chair, John).

Selective attention is the selection of parts of the visual field for further processing. It relies on detection and restricts the field of classification. Throughout this thesis the term attention is used interchangeably with selective attention.

These are all difficult problems for which there are currently only partial solutions.
Figure 3.1: The pathways of early vision. Light falls on the retina where photo-sensitive cone and rod cells produce an image which undergoes some processing before travelling to the superior colliculus in the brain-stem and the lateral geniculate nucleus. The signal then travels from the LGN to primary visual cortex in the occipital lobe. Processing in the superior colliculus is not considered of great importance in humans.

I am interested in object classification because I believe it correlates linguistically with noun phrases. I am also interested in detection and attention because I believe that these correlate with the reference and number components of the noun phrase.

I cannot hope to cover everything here, but in the rest of this chapter I review relevant aspects of visual cognition, beginning with early vision from the retina to visual cortex in Sections §3.2 and §3.3. Then I examine each of detection, classification and attention in Sections §3.4 to §3.6. Finally, in Section §3.7 I present a firmer idea of my own model.

### 3.2 Primary Visual Cortex

My description of these early stages of vision is drawn largely from Kolb and Whishaw (1990), a neuropsychology text, Grill-Spector and Malach (2004), a review of functional areas in visual cortex, Lee (1996), a review of very early vision and Gegenfurtner (2003), a review of colour vision. Other sources are noted in the text.

#### 3.2.1 Early Vision

Visual processing begins in the retina, which is not just a sensor but also a processor. It continues in the lateral geniculate nuclei before the visual data reaches V1, or primary visual cortex, as illustrated
in Figure 3.1. I am not concerned with the details of this early processing but it is worth knowing that input to visual cortex is already a filtered version of the information available at the retina.

By the time information reaches visual cortex it has been divided into separate channels for the luminance and colour information (the magnocellular layer for luminance, parvocellular layer for red-green colour and the koniocellular layer for blue-yellow). The information from each eye is also maintained in separate channels at this early stage.

### 3.2.2 Cortex

The visual cortex, like all cortex, is composed of six layers. The surface of the visual cortex is also divided into several functional regions according to conventions based on anatomical and physiological differences. The axons of LGN cells enter primary visual cortex (V1, also known as the striate cortex) in layer IV. This layer is also subdivided into several sublayers according to anatomical and physiological considerations. Figure 3.2 illustrates the layers of V1 and their assumed specialisations.

### 3.2.3 Features, Tuning and Cell Types

Neurons in V1 are sensitive to many different aspects of the visual input they receive in their receptive field. Cells which are sensitive to a particular value or range of values of a feature are also said to be tuned to that feature. For example, cells can be sensitive to position of a stimulus within the receptive field, orientation of the stimulus, direction of its motion, its size (also known as spatial frequency), its colour, its depth or which eye the stimulus is perceived in. The tuning of cells is not just randomly distributed through V1, however. Many tuned features are organised as discussed below.
The function of cells within V1 is reasonably understood (Hubel and Wiesel, 1968). There are at least four broad divisions of cells.

Centre-surround cells are similar to centre-surround cells in the LGN and retina. This type is illustrated in Figure 3.3.

Simple cells are also centre-surround, but the shape of their receptive field is elongated so that the centre is a bar with a particular orientation (see Figure 3.4).

Complex cells are sensitive to the orientation of bars or edges, but not to their position within the receptive field. Many complex cells are also sensitive to the direction in which the stimulus moves, and also to the size (or spatial frequency) of the stimulus.

Hypercomplex cells are similar to complex cells but are most sensitive to moving stimuli that are bounded at one or both ends. This allows them to detect features such as corners.

Complex cells are the most common within V1.

3.2.4 Retinotopy

Cells in V1 are arranged in an orderly way according to several criteria (Grill-Spector and Malach, 2004). The coarsest of these is the retinotopy of the representation in V1. This just means that...
neighbouring regions of the retinal image are represented in V1 by neighbouring cells, that is the representation is retinotopic. It should be noted, though, that while the topology of the retinal image and cortical representations are the roughly the same, they are not the same shape. The cortical representation is usually described as log-polar. The two axes of the representation are eccentricity, or distance from the centre of the retina, and angular position on the retina. Because the fovea occupies far more space in the representation than the periphery this distorts the eccentricity axis much as a logarithmic transformation does. The cortical retinotopic organisation is illustrated in Figure 3.5.

3.2.5 Columns

At a finer level of detail a different arrangement is apparent in V1. This is the well-known columnar organisation discovered by Hubel and Wiesel (1968). They found that as one travels through V1 perpendicular to the surface, cells that are sensitive to orientation are sensitive to the same orientation. Thus there appear to be columns of cells with the same orientation sensitivity. Moving across the surface the orientation sensitivity of columns varies, though not always smoothly. Cells also seem to group into independent columns according to ocular dominance.

Although at first it may seem difficult to accommodate this columnar organisation with the previous discussion of retinotopy, the problem is resolved when one considers the scales involved. V1 covers a large part of the surface of the rear of the brain; and a macroscopic examination of cortical activity reveals retinotopic organisation. At small (i.e. sub-millimetre) scales, however, each tiny patch of cortex reveals the richer orientation-column structure. Cells within orientation columns have receptive fields whose positions vary randomly over several receptive-field widths, but moving across the cortex the average receptive field position varies smoothly according to the retinotopic mapping.

3.2.6 Channels and Pathways

Retinotopy and columns are not the only levels of organisation in V1 (Grill-Spector and Malach, 2004) Although the brain is marked by its level of interconnectivity, there are some reasonable clearly defined pathways within V1. Broadly speaking, information enters at layer IV of V1 and propagates towards
more superficial (lower numbered) layers and thence to other parts of the brain. This propagation is modulated, however, by neurons in the deeper layers which receive their input from the superficial layers. Thus there are feedback loops within primary visual cortex.

The flow of visual information is not homogeneous, either. The distinction between the magnocellular and parvocellular connections made in the LGN (and originating in the retina) is preserved through processing in V1. As Figure 3.2 shows, these pathways follow different routes and leave from different layers.

Processing of the magnocellular pathway seems homogeneous; cells are orientation tuned for both form and motion and also have some tuning for depth information. Within the parvocellular pathway processing layers II and III there are differences. The cells in these layers are divided into blobs and interblobs (named because of their appearance under staining) and they appear to have functional as well as anatomical differences. Blobs prefer specific colours at low spatial frequencies and are not orientation selective, interblobs are not tuned for colour but sensitive to orientation and high spatial frequencies.

These combinations of features are usually associated with particular visual functions: the magnocellular pathway with motion, depth and dynamic form, the blobs with colour and the interblobs with static form (Shipp and Zeki, 2002). So within the pathways there are channels for different aspects of the visual signal (depth, colour, and so forth).

3.3 Secondary Visual Cortex and Beyond

V1 is the nexus between the early visual system and the rest of the brain; the processing subsequent to V1 is less well understood largely because by now the visual signal does not directly encode the image presented to the retina. The methods employed in studies of the earlier visual system (tracing projections, recording the responses of single cells or imaging whole regions while stimuli are presented, for example) still provide useful information for the study of later areas but the results must be interpreted with caution.

Having said this it should be noted that later “modules” of the visual system appear to work on similar lines to V1, and the principles discussed in section §3.2 apply equally well here. Feature detection, channels, columns and retinotopy are all found in these parts of the visual system (indeed analogous arrangements occur in other parts of the sensory system).

3.3.1 Hierarchy and Specialisation

In the early visual cortex at least, processing seems largely concerned with picking out features from the visual cortex. One can extrapolate this to a claim that visual processing is more or less a series (or hierarchy) of increasingly complicated feature detectors, each feeding on the output of the module below. Feature detectors become more specialised as the hierarchy proceeds, eventually dividing into their own subchannels.

This conception is popular, and is consistent with general theories of brain function (Grill-Spector and Malach, 2004). However it should also be kept in mind that there are alternatives and additions to this cascading feature-detection model (Ullman, 1996). For example, several models of attention propose that even early visual processing involves competition between neurons at each level of the hierarchy or modulating feedback from higher levels of the hierarchy to lower ones. Section §3.6.4 reviews some of these. Some physiologically plausible object recognition theories require some simple feature extraction, but do not require the computation of high-level features.

3.3.2 Interconnectivity

One new aspect that was not discussed in section §3.2 is the proliferation of projections linking one part of the visual cortex to another or carrying signals to and from subcortical structures (see Callaway, 2004, for a review). The visual signal does not simply propagate through a sequence of visual modules,
after which it reaches some final representation in the brain. Instead the modules are linked into a (probably densely connected) network, mostly reciprocally.

The purpose of all these interconnections is not well understood. Some no doubt coordinate the activity of various specialised subsystems and allow efficient parallel processing of the visual signal. Some models, to be discussed below in section §3.6 require projections running from “later” to “earlier” processing modules to carry gating or priming information for attention. Very probably many connections from late processing modules to early modules enhance or suppress “interesting” signals to better assist the late processing.

Despite this network of activity in the visual cortex there is also the widely assumed view that visual processing represents a stream of information flowing from V1 through secondary visual cortex to later (association) cortical areas where the information is used in the planning of actions. In fact, as will be seen below, there are two distinct pathways heading from the occipital lobe where V1 resides to the parietal and temporal lobes respectively. The magnocellular and parvocellular pathways that originate in the retina seem to continue as separate channels through the visual system, but before they physically separate undergo further processing in close proximity.

3.3.3 V2

According to Shipp and Zeki (2002), V2, like V1, is also arranged retinotopically at the macroscopic level. Microscopically, its most striking level of organisation is stripes that run across the surface of the cortex. There are effectively three types of stripes: thin stripes, interstripes and thick stripes. Stripes are an anatomical feature, revealed by staining, but again they appear to be functionally different too. Thick stripes receive input from layer IV\textsubscript{B} and so continue the magnocellular pathway. Thin stripes receive input from blobs in V1, interstripes from interblob regions in V1, continuing the separate channels of the parvocellular pathway.

Superficially, V2 keeps the pathways and channels separate, but in fact V2 is about integration of signals just as much as specialisation of processing. The stripes are distinct in the middle layers of V2, but not as much in the deeper and more superficial layers. Information flow follows a similar pattern through the layers in V2 as in V1 (and indeed, throughout the visual cortex) so the input channels have an opportunity for cooperative signalling and feedback within V2 before heading to other brain areas.

V2 performs both specialisation and integration, but the magnocellular and parvocellular pathway separation is still maintained. The relationship at this stage appears to be in the nature of sharing. This makes some sense given that the pathways originally contain signals from distinct types of retinal cells. The division which holds at the retina may not be suitable when the pathways later become specialised for more abstract functions which require the combination of information from both pathways.

3.3.4 What and Where - Ventral and Dorsal Streams

Subsequent to V2 our knowledge of the visual system becomes much less sure, and there is no longer a single stream of information. The magnocellular and parvocellular pathways, which have already become specialised for extracting different features from the visual signal, become physically separated.

The magnocellular pathway projects from V2 through V3, a further retinotopically arranged cortical area, to the middle temporal area, MT, also known as V5. MT also receives connections from V2 and V1 more directly, but again from the magnocellular pathway. MT contains columns sensitive to motion direction from a variety of abstract stimuli (not just luminance and colour but also texture variation) and projects on to regions in the parietal lobe. For this reason it is known as the dorsal stream (Ungerleider and Mishkin, 1982).

The parvocellular pathway connects from V2 to V3 and to another region V4. V4 is a retinotopically arranged region that contains many colour sensitive cells, and for this reason is sometimes considered part of a “colour centre” in visual cortex. There is some disagreement about the importance of V4 for colour (Gegenfurtner, 2003), however, and V4 contains cells sensitive to complex spatial features as well. Connections from V4 lead into the temporal lobe creating another stream
known as the ventral stream. I am particularly interested in inferotemporal cortex (IT) which lies within the ventral stream.

The dorsal stream has become associated with determination of spatial relations, the ventral stream with the shape of objects (commonly referred to as “where” and “what”, respectively). Evidence for the division of labour into “what” and “where” includes cells in the ventral stream that are sensitive to particular objects or kinds of objects such as faces, body parts, animals, tools, buildings and so on (Grill-Spector and Malach, 2004). The dorsal stream, by contrast, appears to be involved in processes such as motion, directing spatial attention, and planning motor activity.

The dissociation of the two streams is illustrated starkly in various neurological disorders. Patients with optic ataxia can recognise visual targets but not reach for them. Visual agnosics can orient to and even grasp visually presented objects but not recognise or describe them. In so-called blindsight the patient responds at some level to visual cues (pointing, orienting, even grasping) but is not consciously aware of seeing anything (Goodale and Milner, 1992; Weiskrantz, 1986).

As emphasised above, however, the streams should not be considered entirely separate. Feedback and other interconnections ensure that within the visual cortex the streams interact. While the “what” and “where” labels have been criticised for oversimplifying the situation, the hypothesis that these two pathways perform different jobs has become widely accepted (Goodale and Milner, 1992).

3.4 Object Detection

Before a perceived object can be placed into a class or identified with a previously encountered object it has to be detected in the image. This means that the parts of the image that correspond to an object (or objects) need to be found. This implicitly includes deciding that there is an object to process in the image.

Detection is related to the processes of attention that find salient information in the visual field, and the reader is directed to section §3.6 for a discussion of these. Detection is also supposed to serve the classification and identification processes, however, and some points seem more appropriate to discuss here.

This aspect of detection is usually seen as the problem of segmenting the input image into the figure and ground. Theories of how such segmentation works vary. The process might work using only information in the image by using “general principles” that derive likely coherent parts of the image. Jacobs (2001) takes this view, describing ways to find the shape of an object (see Figure 3.8a).

We already know of principles that fit this theory, the Gestalt grouping principles, and these are not controversial. They are illustrated in Figure 3.6. The Gestalt principles are usually illustrated in this abstract, geometrical sense, but they are generally thought to represent heuristics for object detection. The proximity principle represents the tendency for parts of an object to be near one another, similarity reflects their tendency to be made of the same stuff, and so forth.

Gillam (2001) proposes two new principles (common region and connectedness). These continue the interpretation of Gestalt principles as rules for creating proto-objects (Scholl, 2001), entities which the visual system suspects are objects. Proto-objects represent a “rough guess” about objecthood which may be revised later when more intensive processing is brought to bear.

3.5 Object Classification

3.5.1 Location and Scale Invariance

The IT pathway (or “what” pathway, see Section §3.3.4) seems to perform only a classification of the input, though the units of classification, their granularity, and other details yet elude us.

One of the characteristics of the IT pathway that points to this conclusion is the invariance of responses under a wide range of conditions (Logothetis and Sheinberg, 1996; Riesenhuber and Poggio, 2002). For example, cells that respond to a particular visual stimulus respond similarly over a range of retinal positions (translation invariance) and scales (scale or size invariance). At the cellular level this invariance is limited (DiCarlo and Maunsell, 2003, report sensitivity to position changes of as little
as 1.5°) but significant. Invariance is exactly what one would expect from a system whose principal purpose was classification, and IT could conceivably contain an invariant distributed representation.

### 3.5.2 Cardinality Blindness

More recent evidence suggests that IT cells also exhibit a different kind of invariance. This is invariance to number, or cardinality in the visual input. By referring to a visual classifier as cardinality invariant, or cardinality blind, I just mean that it is largely unaffected by the difference between one object and several objects of the same type. Such a system would react similarly if presented with a single dog or with a group of dogs. This means a cardinality invariant classifier can respond to homogeneous groups by performing group classification. Figure 3.7 illustrates this.

More precisely, a classifier that detects a single type in a cardinality invariant manner should respond in (approximately) the same way to any number of tokens of that type. It should also be able to distinguish between a homogeneous group and a group made up of different types. When presented with a heterogeneous group, its response should be a failure to classify, since the objects do not share the same type.

Cardinality invariance is a property of the classifier alone. I am not calling on other subsystems to help achieve the effect by restricting input to a single instance at a time. To make use of the effect, however, other subsystems are needed to restrict input to homogeneous groups.

There is evidence that IT is blind to cardinality. Nieder and Miller (2004) showed monkeys two displays containing one to five items, controlled for confounding cues such as shape, area, circumference, density and location. The monkeys judged whether the displays contained the same number of items. They found that, during this task, only between five and ten percent of anterior IT (AIT) neurons responded to cardinality in the visual input. And almost half of those showed a significant effect of type on their response, indicating that they were not detecting only number, but number in conjunction with type. As Nieder and Miller state, “Thus, it seems that aITC neurons are primarily sensitive to the physical appearance of the displays and do not extract numerosity information per se” (p. 7460).

Further evidence from Zoccolan et al. (2005) is also consistent with cardinality invariance. They examined AIT neurons of monkeys looking at displays containing multiple objects. They found the
response of a cell to multiple items was roughly the average of its responses to the items on their own. The cell’s response to its preferred stimulus is then cardinality invariant, while the response is strongly curtailed when preferred and non-preferred stimuli are presented together.

There is also evidence from psychological experiments that classification of homogeneous groups can proceed in parallel, like van der Heijden (1975) who presented subjects with displays containing one, two or three items of which some, all or none were defined as search targets. For stimuli containing only targets, response times decreased as the number of targets increased, suggesting not only that potential targets could be processed in parallel, but that multiple targets facilitated processing.

I believe that, while evidence exists for cardinality invariance in IT, this is the first time it has been identified as such and given a name. For instance, Nieder and Miller (2004) demonstrate the cardinality blindness of AIT neurons only as a side-effect of demonstrating the cardinality sensitivity of neurons in parietal cortex.

3.5.3 A Survey of Approaches to Visual Classification

According to Ullman (1996) object classification (what is often referred to as recognition) can be approached in three main ways.

1. The invariant properties or feature space methods. These methods assume that different classes are defined by properties that are invariant no matter how the object appears to a viewer. These include methods such as Gibson’s (1966) invariants which might include calculations involving simple information like perimeter length and area.

Such simple methods were quickly discredited for general vision problems (though they are of use in some technical fields), but the modern technique of feature extraction and feature spaces is really just a more sophisticated version of invariant properties. Features are extracted from the input image. Each feature can be considered as a dimension along which the input varies, and classes are then assumed to correspond to compact regions within this feature space.

The difficulty then lies in finding features that are invariant, or designing a matching system flexible enough to deal with missing (and spurious) features. The definition of classes in the feature space is also problematic.
2. Parts and structure methods. In these methods objects are classified by detecting primitive units in the input image and using their configuration and shape to match an object template for the class. This can be done in two or three dimensions. The best known example of the structure method is Marr’s (1982) 3D model which was built of “generalised cones” from the input image. Biederman’s (1985) was a similar model with a greater number of primitive three dimensional shapes.

The problem with classification by structure is that the image often includes only some of the parts that define the relevant class, and that the primitives themselves are often difficult to isolate.

3. Alignment methods. This is the approach favoured by Ullman (1996), though it also has problems. The alignment approach consists in directly matching views of an object with views representing classes. Though this sounds impractical at first, Ullman shows how he and other researchers have built a method for representing all the views of an object using just a few canonical views and mathematical transformations which compensate for changes in orientation, illumination and so forth.

The difficulties lie in finding ways to determine if two views are of the same object using the allowed transformations in a timely (and plausible) manner.

These methods are not mutually exclusive, but could conceivably be combined in a single system.

3.5.4 Five Models of Visual Classification

In this section I discuss five different models proposed for human vision, a selection intended to illustrate the kind of theories available, including the model in Section §3.5.9 which I adopt and develop in Chapter 4.

3.5.5 Jacobs’s (2001) Model

Jacobs (Figure 3.8a) describes a model which uses grouping in the image array as a form of generic object recognition. Short oriented line segments (of the sort perhaps delivered by V1, see section §3.2) are grouped into two dimensional shapes (in this case using Markov chains), which Jacobs believes can be used as an index to select a group of object templates for more specific classification to work on. Although the model does not spell out the more specific methods (though it might work as a preprocessing stage to something like Ullman’s alignment scheme) this generic classification is clearly a structural method.

3.5.6 Marr’s (1982) Model

Marr (Figure 3.8b) hypothesised a comprehensive system for classification. He proposed a roughly three stage model. First, the image is processed to find local discontinuities. These are used to find edges, lines, terminations and other primitives that make up the primal sketch. Information about contours from this and further image processing (based on stereopsis, shape from shading, colour differences and motion for instance) is then used to create the $2\frac{1}{2}$D sketch, a retinotopic map in which each location is associated with distance and orientation of the surface visible there. Finally, the $2\frac{1}{2}$D sketch is used in combination with constraints on object form to compute the 3D model, a representation built of generalised cones as described above in the discussion of structural methods. Marr believes in the structural approach and explicitly eschews the other approaches, proposing interesting explanations for apparent problems for the model. For example, neurological patients who are incapable of classifying objects from non-typical views are often seen as supporting some kind of view-based alignment method. Marr (Marr, 1982, p. 316) instead sees this as inability to use any but the simplest methods of primitive extraction in the 3D model. Deprived of primitives in difficult views, classification fails.
(a) An outline of the model described by Jacobs (2001), an example of shape-based identification.

(b) Marr’s (1982) influential model, devised using a heavily analytical methodology with a structural approach.

(c) Ullman’s (1996) model, in a similar tradition to Marr’s (1982), but with a view alignment approach.

(d) Riesenhuber and Poggio’s (1999) model. A combination of the feature space and view-matching approaches.

(e) The object classification component of Mozer and Sitton’s (1998) model, a feature space approach in the form of a convolutional neural network (CNN).

Figure 3.8: A selection of different object identification and classification models.
Marr’s approach to vision was rigorous, analytical and blunt; the early parts of his model include processes as they must be, in his view, determined by analysing the problem vision is to solve and breaking it into components. The influence of this model on vision research was profound. However the details of the later stages (which in many ways were speculation) survives as only one of many theories of object perception. In particular, Marr’s insistence that most of vision could be explained as stimulus driven (which even he was less strict about at the late stages of classification) is challenged by methods that make extensive use of top down constraints to ease the difficulties of the problem.

3.5.7 Ullman’s (1996) Model

This model (see Figure 3.8c), which was partly described above, is in the same analytical tradition as Marr’s work but with a radically different result. Ullman envisages a process, as yet undetermined, that works out how to transform the input image in such a way that it matches a view of a known object, or decides that no such transformation exists. This process must search all available views as well as determining the transformation, clearly a difficult task. The transformation itself is tightly constrained by rules that describe the way two dimensional views change in different circumstances (the key result for the alignment approach) but Ullman also believes the transformation is constrained by dynamic high level hypotheses about the object. In his final model classification is a “hypothesise and test” system that eventually settles on an interpretation of the input image. Ullman also proposes additional methods to ease the transformation finder’s job, such as a preprocessing stage that uses simpler cues to narrow down the choice of objects.

3.5.8 Riesenhuber and Poggio’s (1999) Model

View matching is also the strategy used by Riesenhuber and Poggio (1999) and the neurologically inspired model of Serre et al. (2005) which draws on it, shown in Figure 3.8d. Here primitive feature extraction is followed by a hierarchy of abstraction and feature combination operations that produce location and scale invariance, matching the input to one of a set of template views (represented by the view-tuned cells). The winning view provides object classification and orientation.

The chief technical innovation in this approach is the use of the MAX operator to achieve invariance. The output of a MAX unit depends only on the maximum of its inputs. This implies that MAX units are blind to cardinality in their input, which surely helps explain why Serre et al. find cardinality blindness (see Section §3.5.2) in their overall system.

3.5.9 Mozer and Sitton’s (1998) Model

This final model, shown in Figure 3.8e, is the convolutional neural network (CNN). This is a method that has found some success as both a psychological model (Mozer and Sitton, 1998) and as a computer vision technique (LeCun and Bengio, 1995). It is a connectionist model, an artificial neural network (ANN) trained using the back propagation algorithm. It takes as input either the raw image or a set of parallel retinotopic feature maps and produces as output activation of units, each representing an object class.

The convolutional neural network differs from the conventional ANN in that the weights connecting units in successive layers are not independent across the array. Within each layer of the network, units are organised into several retinotopic feature maps. In Mozer and Sitton’s (1998) network, for example, the input layer consists of five feature maps, representing four orientations and terminators. Successive layers have a smaller spatial extent but more features, culminating in the output layer with only a single unit in each feature map, corresponding to object classifications. The weights that connect units in successive layers are similarly arranged, so that a unit in one layer will typically receive input from several units in each feature map of the previous layer. But the weights only extend in a small radius, and the same weights are used across the entire network by convolving them.

Convolution is a useful mathematical process, illustrated for a simple image array in Figure 3.9. A small window (in the CNN case, the weights array) is slid over a relatively large image and the result at each position is a representation of the image as “seen” through the weights array. The
Figure 3.9: The inputs to convolution (shading represents magnitude of each value). The weight matrix is convolved with the image by sliding it around to every possible position, and storing the sum of the products of corresponding values at each position. The result is a new “image”.

Weight matrix

Image or feature map

Resulting image
effect of convolving the weights with the image is that the network gains a large amount of spatial
independence: stimuli presented in one part of the input during training can still be classified even if
they appear in a different place.

The CNN is a feature extraction approach. Mozer and Sitton (1998) apply it to stylised letters,
LeCun and Bengio (1995) apply it to real written character recognition, both with some success.
Unfortunately, this performance with single letters does not generalise well to multiple letters. The
network performs adequately with two simultaneous letters, but with three or more its accuracy suffers
greatly. This is the motivation for the introduction of attention in these models: as a way to extract
only the relevant part of the stimulus needed for object classification without performance degrading.
(see section §3.6.4)

However, as I will show in Chapter 4, there is a loophole here, because although performance
suffers when a CNN is presented with several heterogeneous items, it works well when presented with
several homogeneous items: it is cardinality blind.

3.6 Visual Attention

In this section I examine visual attention, an important aspect of cognition which is clearly related to
perception of the world, though whether its role is dominant, subservient, as a counterpart to other
perceptual processes or a mixture of these is unclear. To explore this complex area I first turn to
some general theoretical principles of attention in section §3.6.1, examine two of these, the medium
of attention and the modality of attention, more closely in Sections §3.6.2 and §3.6.3 and with these
in mind look more specifically at current models in section §3.6.4.

3.6.1 Theories of Attention

There are, and have been in the past, several different theories of attention. The differences are not
always because of differences in explanation, but also because different theories usually concentrate
on particular aspects of attention. Such aspects include engaging, directing, shifting and maintaining
attention as well as actually suppressing unattended stimuli.

Broadbent (1958) is usually credited with the first definite theory of attention, usually referred
to as the filtering or bottleneck approach. The idea was that attention was a device that acted
early on impinging stimuli, limiting what would reach the later stages, which were assumed to have
a limited processing capacity. Broadbent’s theory was modified in the light of evidence to relative
filtering, allowing non-attended stimuli to pass through but receive less processing. The filter in
Broadbent’s model is imposed because of the bottleneck of limited resources, but the exact location
of the bottleneck during processing is a matter of dispute. Broadbent favoured early selection, that is
attention selected stimuli after very little processing had occurred. In late-selection versions stimuli
are processed more, to the point where they are identified, before attention selects between them.

In fact it is now believed that more than one type of attention is at work, which has lessened
hostility between early-selectionists and late-selectionists, though not of course the importance of the
distinction for individual components. Although the details of Broadbent’s original approach have
not weathered so well, the principles of limited processing capacity and the need to select only some
stimuli for processing remains an important assumption behind other attention theories.

Another related theory holds that the phenomenon of attention is due to the sharing of limited
processing resources across stimuli (Kahneman, 1973, for example). Because capacity is still limited,
the more stimuli attended to the slower the processing of each will be. Resource sharing also elegantly
explains the loss of efficiency when attention is divided.

As mentioned above, there seems to be more than one type of attention. An important distinction is
made between transient and sustained attention (Nakayama and Mackeben, 1989). Transient attention
(also known as reflexive attention, exogenous attention and the orienting response) occurs swiftly
when a cue appears but only lasts a few hundred milliseconds before fading. Sustained attention (also
known as voluntary attention or endogenous attention), on the other hand, takes several hundred
milliseconds to reach maximum effectiveness, after which it can be maintained apparently indefinitely.
These observations have led to a theoretical interpretation that transient attention is a bottom-up process that attracts attention to perceived changes in the world, while sustained attention is a top-down process that directs processing towards interesting stimuli. Thus there is also some interaction with the issues of early and late selection.

A distinction is also often made between attention acting for different functions, for example perceptual attention is distinguished from motor attention. Even within the visual system there might be different attentional processes active in the dorsal and ventral streams. Some kinds of attention are apparently cross-modal (limiting both visual and auditory processing, for example).

3.6.2 The Medium of Attention

The previous section dealt with some general theoretical issues of attention. In this section I explore the medium of attention, the domain in which information is divided and selected. There are four commonly acknowledged media which are not mutually exclusive – there seems to be evidence for each – but it is unclear how the different versions interact. I also discuss a fifth, scale-based attention, which is not so common in the literature.

3.6.2.1 Space-based Attention

Spatial attention is usually characterised in terms of a spotlight (Posner, 1980) or zoom-lens (Eriksen and Yeh, 1985). In the spotlight model stimuli that fall within the radius of some position in the visual field receive more processing than those outside, thus attention works much like a spotlight. This is modified in the zoom-lens model so that the size of the spotlight varies in size to allow varying levels of detail to be picked out. The spotlight metaphor is useful; observation has revealed that it must have a centre-surround shape similar in size to the receptive fields of cells in late areas of visual cortex (such as MT). The zoom-lens model fares less well, however, and at least in this simple metaphorical form is not borne out by the evidence (Steinman and Steinman, 2002).

Spatial attention is uncontroversial; attention can be directed and maintained at a particular location of the visual field even if there is nothing to attend to (e.g. in expectation that something will appear), but it is only the most obvious medium of attention.

3.6.2.2 Feature-based Attention

The existence of feature-based attention emerged from research in paradigms like visual search and priming. In a typical model (such as the feature integration theory of Treisman, 1998, described in detail in Section §3.6.4.1) certain “features” in the visual field receive more attention than others. Features can be things like orientation, colour or the like, the important point being that they are not defined spatially, instead feature-based attention is supposed to cover the entire visual field with instances of the attended feature that occur anywhere receiving more attention than other features. In visual search this allows the target item to pop out from a large field of distractors because of its defining features.

Feature-based attention of some form exists, as these visual search experiments show (though its extent and details remain hotly contested).

3.6.2.3 Object-based Attention

Feature-based attention is quite disjoint from spatial attention. The third medium, object-based attention, is in some ways a mix of the two. Supporters of object-based attention hold that attention is dished out at the level of perceptual objects, so that one or more objects in the world (for example a chair) can be picked out for careful processing while surrounding objects and the background are suppressed (even if they overlap or obscure the attended object in the visual field).

There is good evidence that objects can act as units of attention in this way, summarised in Scholl (2001). For example, Neisser (1967, 1979) found that when presented with two superimposed movies, subjects could attend only to one. While this does not isolate individual objects as units of attention it shows that something other than spatial attention is at work. Egly et al. (1994) found when part of
an object is attended to attention somehow spreads to the entire object, giving a kind of priming effect which facilitates tasks related to that object. Pathological studies showing that unilateral neglect is at least partially tied to an object-based reference frame (Rafal, 1998; Robertson and Marshall, 1993), patients with Balint’s syndrome can perceive only one object at a time. The evidence all suggests that perceptual objects are treated as units by attention.

Within the general category of object-based attention, Tipper and Weaver (1998) note a distinction between object-centred attention and scene-based attention. Object-centred attention is directed at single objects, or parts of single objects. Scene-based attention is directed at assemblies of objects that are dealt with together. Tipper and Weaver believe that scene-based attention may operate mainly in familiar environments, where delivering negative selection to the familiar surroundings enhances unusual objects.

### 3.6.2.4 Action-based Attention

Tipper et al. (1998) have concluded that a different kind of attention is used to select targets for actions. Through experiment they have found that during tasks that require the hand to be moved towards targets placed among distractors, attention seems to operate in a coordinate system based around the hand. For example the effect of distractors on the task is affected by their distance from the hand. Tipper et al. believe this is evidence for a distinct attention mechanism used to select targets of actions.

A somewhat different view is expressed by Schneider (1999, and also Schneider and Deubel, 2002) who also consider the difference between perceptual (visual) attention and action attention. That there are different kinds of attention is undisputed, but the mechanism for this is debatable. Schneider (1999), for example, believes that a single selection mechanism operates early in the visual cortex. But while the same stimuli are passed to the ventral and dorsal streams (see section §3.3.4) the differences in processing lead to different information and reference frames being used for the perceptual attention (in an object-based frame, processed in the inferior temporal cortex) and attention for action (in an action-based frame in the posterior parietal cortex). Tipper et al. (1998) interpret the same evidence as indicating at least two different kinds of selection occuring much later in the processing of visual information.

### 3.6.2.5 Scale-based Attention

Navon (1977) describes the phenomenon of global precedence (or global dominance). In humans a global figure interferes with tasks relating to the local figures which make it up but not the other way around (see Kimchi, 1992, for a review and Navon, 2003, for recent reflections on the phenomenon). Figure 3.10 shows two figures that demonstrate global precedence.

The crucial experiment in Navon (1977) (experiment 3) demonstrates the “inevitability” of global processing. Subjects were shown figures like those in Figure 3.10. On each trial the stimuli were shown for 40ms in a random quadrant of the display and masked immediately afterwards.

For a series of blocks of trials the subjects were told to concentrate on the shapes at the global or local level (subjects were given both conditions in a session). The task given was to classify the shape at the relevant level by pressing one of two keys.

The results showed two things. First, that subjects were slower to respond when directed to the local level compared to the global level for all conditions. Second, that when the global shape was different from the local shape this interfered strongly with classification of the local shape (slowing reactions) but not with classification of the global shape.

Global precedence can be thought of as a phenomenon of scale-based attention. In many ways it seems to behave analogously to space- or object-based attention but in the scale (or spatial frequency) domain, with its own predispositions, like the precedence of global figures over their constituents (Filoteo et al., 2001).
3.6.3 Modality of Attention

I am chiefly concerned with visual attention (as is a lot of the literature), however there are some aspects of attention in other modalities that are of interest. In particular, relatively recent research is beginning to show that attention interacts between the modalities, sometimes in unexpected ways.

Attention has long been appreciated in vision, audition and somatosensation. Each modality possesses some different media of attention. Audition, for example, involves tone-based attention which is unavailable to the other modalities. On the other hand, evidence for common media of attention between the modalities is emerging. For example Scholl (2001) presents evidence that object-based attention is at work in audition, though based on pitch rather than location.

Even more interesting is the approach of Driver and Spence (1998) who found that attention within one modality can interact with attention in other modalities. In experiments they found that cues given visually could prime targets presented by touch and vice-versa. The mechanism behind their results is unclear. It is difficult to tell whether a single attentional representation is integrating various sources or if separate representations are in limited communication with each other. Also, although the phenomenon occurs for both exogenous and endogenous attention, the latter seems to be less potent at spreading across modalities. Driver and Spence (1998) also found that tasks can sometimes be facilitated in a location just because a saccade is planned to direct the eyes at that location. This is not limited to vision, either, and planned saccades seem to be able to prime audition. Figure 3.11 shows possible interactions between the three modalities and saccade generation.

It is also unclear how these attention phenomena relate to other crossmodal interactions such as the McGurk effect (McGurk and MacDonald, 1976) where seen and heard cues of lip movements interfere, or the ventriloquist effect, the familiar illusion that voices emanate from where we see a moving mouth rather than the actual sound source. This latter occurs not just with a ventriloquist’s dummy but in more mundane situations like watching a movie or television.

Combined with the other evidence this led Driver and Spence to speculate that neurons known to have multimodal sensitivity are responsible for at least some of these effects. A popular location for these responsible neurons is the superior colliculus which receives multimodal input and directs movements of the head and eyes to exogenously salient stimuli. The neurons of this subcortical structure seem to dynamically remap stimuli as the body’s position and posture change ensuring that
information from various modalities is represented in a uniform spatial manner.

This theory could also explain the discrepancy between exogenous and endogenous crossmodal attention as other authors (e.g., Tipper and Weaver, 1998) have proposed that endogenous attention is driven by cortical structures that do not share the representation in the superior colliculus. Of course, this leaves wide open the question of which cortical structures are responsible for endogenous crossmodal attention, especially since crossmodally sensitive cells are spread through the cortex.

### 3.6.4 Models of Attention

There are many models of attention available. Some of these are related refinements of one another, and so fit into groups. Where this is the case I have tried to present progenitor models before their derivatives.

#### 3.6.4.1 Feature Integration Theory

Treisman’s (1998) feature integration theory (FIT) is the result of a long evolution of the original version presented by Treisman and Gelade (1980). Figure 3.12 illustrates this theory. It is a psychological rather than computational model, though some computationally compatible versions are described below (see section §3.6.4.3). The model’s premise, an elaboration of the spotlight model, is that early perceptual processes operate in a massively parallel way to extract features across the visual field, such as orientation, colour and intensity. However only a small region of the feature maps that result from this processing are usable by later processes such as object recognition. Tasks which rely on simple feature discrimination can be computed using the parallel feature maps and so occur swiftly and in constant time. However tasks that require information from different feature maps to be integrated must be mediated by attention. This is best illustrated in the visual search paradigm. A visual search task in which the targets differ from the distractors in a single primitive feature (for
Figure 3.12: A diagram of the FIT model according to Treisman (1998). Solid lines indicate information flow, dashed lines indicate attentional control.
example red targets against blue distractors) are completed in the parallel preattentive stage because
the target “pops out” of the search field. But if the targets share a feature with the distractors
(red horizontal lines against red vertical lines and blue horizontal lines, say) then slow serial search
is required to consider each item in turn, integrate its features and decide whether it matches the
template.

Treisman codifies this by introducing a master map of locations, or location map, a retinotopic
map whose activity records the position and extent of the attentional spotlight. The location map
is controlled by several other processes. It receives bottom up activation via the “where” pathway
(see section §3.3.4) (the details are left unspecified). It also receives input from the feature maps,
which draw attention to conspicuous features. Top down commands can direct the spotlight for
the purposes of endogenous attention or serial search. Object files, Treisman’s posited temporary
object representation (see Kahneman et al., 1992) can draw attention to their associated objects. The
map can also be controlled indirectly via the feature maps if the latter are inhibited by a top down
command, perhaps as the result of a cue (e.g. to make red targets pop out, suppress all the other
colour feature maps). The inhibition from the feature maps will then tend to suppress unwanted
exogenously salient items.

Group scanning  FIT has proven a remarkably resilient model, though it has had to undergo
extensive revision to cope with criticisms over 25 years (Quinlan, 2003). One that I particularly want
to mention is group scanning.

After the original FIT theory was proposed, experiments uncovered more complex patterns of
visual search performance. Treisman (1982) found that perceptual grouping affects search because
subjects serially scan groups of items where possible, not just individual items.

Treisman tested both feature and conjunction search; I describe only the conjunction results here
because that is where the effect emerges and I am more interested in the conjunction condition.

Subjects were shown stimuli in which red Hs and green Xs made up the distractors, and the target
was a green H (which differs from the distractors only by the conjunction of colour and shape). Both
target-present and target-absent displays were used. In one set of stimuli the display size was varied
from one letter up to thirty six letters. In the other set there were always thirty six letters but the
number of homogeneous distractor groups was varied.

The letters were arranged in regular square matrices. In the display size condition these measured
either 2x2, 3x3 or 6x6. In the grouped condition distractors were either arranged randomly (36 groups),
in alternating horizontal pairs (18 groups), in alternating 2x2 squares (9 groups), in alternating 3x3
squares (4 groups) or all distractors were the same (1 group).

The task given was to decide as quickly as possible whether a display contained the target and
press one key if it did, another key if it didn’t.

Plotting reaction time against display size showed the expected linear slope as the number of
distractors increased. In the target-absent condition the slope was steeper. But plotting the reaction
time versus number groups in the grouped condition also showed a positive relationship, not quite
linear, between the number of homogeneous distractor groups and reaction time (and again the target-
absent condition had a steeper slope).

Treisman and Gormican (1988) drew this finding into the FIT model with the group scanning
theory. In that model, when parallel search (i.e. pop-out) fails because the target cannot be discrimi-
nated from the distractors, attention is used to limit the spatial scope of the parallel search to a region
where parallel search by feature discrimination can work. Parallel search then continues inside the
attended area.

3.6.4.2  Selective Routing

Koch and Ullman (1985) and Itti and Koch (2000) have presented a simple model aimed at explaining
direction and shifting of attention. Koch and Ullman (1985) introduced the important concept of the
saliency map, a spatial map of units that are excited in a bottom-up manner by “salient” features in
the input. In the later model by Itti and Koch (2000) and Walther and Koch (2006) (Figure 3.13) the
saliency map is computed from the retinal input by first filtering the input by colour, orientation and
Figure 3.13: A diagram of the model of Itti and Koch (2000) and Walther and Koch (2006).
intensity at several spatial scales. Centre surround filtering (section §3.2) of the raw filtered output produces a feature map in the style of Treisman and Gelade for each simple visual feature at each of several spatial scales. Each of these maps records local contrast in the given feature at the given scale. The activation in these maps is then summed into a single conspicuity map for each feature, which records local contrast for that feature at all spatial scales. Finally, the conspicuity maps are summed into a single saliency map. This part simulates bottom-up attention using mechanisms known to exist in the visual cortex (see section §3.2). The salient locations are then fed to another map where winner-take-all competition (WTA) ensues. WTA (Feldman and Ballard, 1982) is a neural network approach that makes use of self feedback to select only one of the salient locations. Attention is then directed to this location. Finally, winners in the WTA inhibit the corresponding location in the saliency map, a kind of inhibition of return (IOR, Posner, 1980) that allows the next most salient location to emerge and so on, a process which simulates serial search of noteworthy locations in the retinal image.

This approach models bottom up attention, with the addition of an attentional mechanism (IOR) that simulates shifts in attention. The other models discussed here place emphasis on different aspects of attention, but many share the structural elements of this model, such as the saliency map, massively parallel early feature extraction and WTA competition. The model is interesting because it shows how attention could be driven solely by the stimulus even during apparently complex tasks like visual search.
3.6.4.3 Guided Search 2.0

Wolfe’s (1994) Guided Search 2.0 (GS2) model (see Figure 3.14), a revision of an earlier model (Cave and Wolfe, 1990; Wolfe and Cave, 1989; Wolfe et al., 1989), is based on feature integration theory. The important addition, or elaboration, is a subsystem that uses the information available in the parallel feature maps to guide the attention spotlight in a bottom-up manner. This corresponds to the part of Figure 3.12 labelled “location pathway”. GS2 is also an implemented computational model.

The model has three main components. Linear filters extract raw feature information from retinal input, which in GS2 is coarse coded, meaning that each feature type is represented as a linear combination of several broad features (for example, orientation could be represented in terms of vertical, horizontal and both diagonal orientations). The output of each filter is then passed to the bottom-up map in the corresponding feature map. The bottom-up map detects features with spatial contrasts and records these. The other half of each feature map is a top-down map, which records the attention required by top-down commands for each feature. So each feature map really consists of two maps, one for bottom-up processes and one for top-down. A weighted sum of all these maps is recorded in a single activation map, the weights determined by top-down commands that emphasise or suppress the maps according to the task at hand. Wolfe notes that in future versions of GS these weights could also be set by a component that monitors the model’s performance during a task. Finally, the activation map itself records the current location of spatial attention and is used to decide which parts of visual input are passed through to higher visual processes, just as in FIT.

As mentioned above, the key additions are the endogenous contrast and competition that allow the stimulus to direct attention in visual search. In traditional FIT serial visual search is commanded from above if required, but in GS2 the serial behaviour of the attention spotlight can be controlled at least partly from below. In particular GS2 provides a solution to a puzzle of FIT - how do the controlling attentional processes know where to put the spotlight during serial search? In simulation the model’s behaviour compares well to that of a human in many respects.

3.6.4.4 MORSEL

This computational model of attention has been under development for some time as part of the MORSEL model. In Mozer and Behrmann (1990) it is combined with BLIRNET, a recognition network with similar properties to (and limitations of) the CNN described in Section §3.5.9. That CNN supplants BLIRNET in later versions (Mozer, 2002; Mozer and Sitton, 1998).

Though taking a different approach to the need for attention, Mozer and Sitton construct a similar attentional model to those described above. Figure 3.15 shows the object recognition system proceeding up the page from retinal input to activated object templates. Onto this is bolted the attention mechanism, which takes input from the feature maps (as well as endogenous commands), computes a saliency map using WTA competition, and gates the feature map input to the convolutional network.

If taken too literally, this is a circular problem. After all, how can one extract only the details necessary to identify object $x$ if one has no idea what $x$ is yet? What the authors really propose is a good enough solution to the problem, a mechanism which filters out enough of the noise and leaves enough of the signal for good performance under a reasonable range of conditions.

It has an advantage over other models, like that of Itti and Koch (2000), because it tests the effectiveness of the attention model on the perceptual processes that are supposed to require it (in this case classification). Mozer and Sitton report that the model matches human data well for several tasks such as visual search and attention shifts.

3.6.4.5 Stimulus Similarity Theory

Duncan and Humphreys (1989), in a rebuttal of FIT, concluded that the motivation for the difference between parallel and serial search was not the primitive features that targets shared with distractors, but rather a more general measure of similarity between the two.

In one of the experiments they describe (experiment one), for example, displays consisted of two, four or six items arranged on the perimeter of a circle. Distractors were either upright Ts or Ts tilted 90° (either all the same or a random mixture). The target was either an upright L (corresponding to
Figure 3.15: A diagram of the model by Mozer and Sitton (1998).
FIT conjunction search) or a T tilted 45° (corresponding to FIT feature search). Displays were either made up of large shapes or small shapes. Both target-present and target-absent displays were used.

Displays were briefly presented to the subjects and the task given was to indicate whether the target was present or absent by pressing one of two keys.

The reaction time and error results showed several features inconsistent with FIT. In both target conditions the search slope was very shallow, suggesting parallel search. There was no difference between the two target conditions. The homogeneity of the distractors had a slight effect, but the biggest effect was size of the items, with bigger items having much shallower (occasionally even negative) slopes.

The conclusion drawn by Duncan and Humphreys based on this experiment, as well as the others they performed and some other similar results, was that the simple binary distinction between parallel and serial search proposed by FIT doesn’t fit the data. More factors are involved that produce continuous variation of search slopes in different conditions depending on a number of variables.

Duncan and Humphreys proposed that during search (and other tasks requiring attention) stimuli were grouped according to similarity, forming groups of objects that could be accepted or rejected en masse. Highly similar distractors contrasted against distinct targets would result in fast searches because distractors could easily be grouped together, leaving targets to pop out. But if the distractors were themselves of several distinct types, or any distractors were similar to the targets, then the process would take much longer. This stimulus similarity theory (SST) was proposed to deal with some experimental results that early formulations of FIT did not deal with; FIT has since been modified to accommodate these (see Section §3.6.4.1, especially group scanning), but stimulus similarity remains a viable alternative.

### 3.6.4.6 Biased Competition

These models, which owe much to Duncan and Humphreys (1989) and Desimone and Duncan (1995), do not make use of a saliency map and have a different interpretation of features.

Desimone and Duncan (1995) introduce biased competition as a general principle to explain attention. Attention, in their view, is not something that occurs at a single stage in processing. Instead it can occur at many levels and works by stimuli competing against each other for saliency. The stimuli can compete using just bottom up signals, but this competition can also be biased by top down signals. Such biases might include emphasising particular features to match a template, or emphasising locations. A model based on this has been implemented by Deco et al. (2002) as shown in Figure 3.16, and simulations show behaviour similar to experimental findings for serial and parallel search.

This is an important point about biased competition models. There is nowhere in the system (no saliency map) that indicates where an attentional spotlight is directed; instead the activity of stimulus representations themselves can be interpreted as saliency information, from the visual cortex to the motor cortex. However the implicit biases of early visual processing (feature detection and local contrast for example) combined with the explicit biases of top down commands give the illusion of an attentional spotlight sweeping over the visual field. Note that this kind of model deals naturally with feature and space-based attention.

The biased competition models are driven more strongly by neuroscientific findings than models in the FIT tradition, while both make use of psychophysical data. As more experimental data has come to light, both approaches have had to adapt and now contain superficial resemblances (for example, GS2 in Figure 3.14 and the model in Figure 3.16 both allow for top down cues accentuating feature maps). The main theoretical difference remains that FIT posits attention as a filter lying between early and late processing which selects only useful stimuli to pass through. Biased competition, however, sees attention as more amorphous. Selective attention still serves the purpose of filtering out unwanted information, but this is done as part of the processing itself and so can operate at several places on the way from perception to action. The particular way this is accomplished – biased competition – predicts behaviour that happens to be close in many cases to that of the traditional spotlight model.
Figure 3.16: A diagram of the biased competition model by Deco et al. (2002). Competition occurs directly in the feature maps and the posterior parietal spatial map, biased by top-down preferences. The feature maps are supposed to be located in the ventral stream and the PP spatial map in the dorsal stream. The mutual interconnection is used for binding features into objects.
3.7 Conclusion

My hypothesis is that the processing of visual objects corresponds to the structure of the noun phrase. This means that components of object processing correspond to components of the noun phrase and occur in the right order.

I am going to concentrate on three “data points” here: reference, number, and type. As Chapter 6 reveals in detail, these correspond to separate syntactic components of the noun phrase.

Type corresponds directly to the result of object classification. I wanted a model of classification that produces type and type alone – in particular I didn’t want it to produce number, so it needed to be cardinality blind. I also wanted it to be a neurally plausible model of IT. For this I chose to develop Mozer and Sitton’s (1998) CNN model. In Chapter 4 I will present a location and scale invariant CNN that works with images of geometrical shapes and demonstrate that it is cardinality blind.

Selective attention seems a natural fit for the reference component of the noun phrase. In particular any model of attention that represents the focus of attention in a unified, rather than distributed manner directly maps to the representation of reference in the noun phrase. FIT, selective routing, MORSEL and guided search all meet this criterion, while biased competition does not. I also wanted a model of detection and attention (generally lumped together as the “attention model”) that takes advantage of the classifier’s cardinality blindness. I chose to borrow and develop the selective routing model of Itti and Koch (2000) and Walther and Koch (2006) for my own purposes because it is well developed, proven and robust.

This attribution of reference to attention (“where/how” pathway, dorsal stream) and type to classification (“what” pathway, ventral stream) is similar to Hurford’s (2003b) approach, as discussed in Sections §2.6.5 and §2.9. In Chapter 6 I will show that combining the classification and attention models of Chapters 4 and 5 also allows extraction of my third data point from the noun phrase, number information.
Chapter 4

A Cardinality Blind Classification Model\(^1\)

4.1 Introduction

This chapter provides a computational model in which cardinality blindness emerges naturally from spatial abstraction mechanisms commonly thought to exist in inferotemporal cortex (IT). The main evidence I present comes from modelling of the IT classifier using the convolutional neural network (CNN) architecture I outlined in Section §3.5.9. A CNN maps two dimensional “feature maps” onto a set of outputs corresponding to recognised classes. To recap briefly: it does this using a series of retinotopic maps that alternately compute novel features from previous layers and abstract over space. This spatial abstraction leads to location invariance (or translation invariance) quite naturally, and is one of the interesting properties of the CNN. This also makes it attractive as a model of visual classification which is also location invariant. Surprisingly, though, the CNN also turns out to be cardinality invariant, probably as a side-effect of its location invariance.

My implementation is based on that of Mozer and Sitton (1998). They used the CNN as a prelude to their model of visual attention, to show that a classifier similar to IT could not reliably classify more than one object at a time, and must therefore require a limiting mechanism to select only one object. This turns out to be true for heterogeneous groups. However a CNN’s response can be quite robust when presented with groups of homogeneous objects, as we will see.

I present the CNN model in detail in Sections §4.2 and §4.3. Sections §4.4 to §4.6 put the network to the test and show that it is indeed cardinality blind. Finally, Section §4.7 examines the plausibility of the CNN as a model of primate IT and Section §4.8 concludes by placing the classifier in the context of the overall model.

4.2 Network Design

A convolutional neural network is a partially connected feed-forward neural network that uses shared weights and a hierarchical configuration (LeCun and Bengio, 1995; LeCun et al., 1992; Mozer and Sitton, 1998; Rumelhart et al., 1986). The particular network I used, though artificial, has parameters that make it roughly comparable to IT. This point is addressed more fully in Section §4.7.

The units of the network are arranged in a series of planes, with units in each plane connected to units in the one above by a layer of weights. The network I used had the design shown in Figure 4.1. There were nine planes and eight layers. Units within each plane are clustered into cells, which are arranged retinotopically. Every cell in a particular plane contains the same number of units, one for each feature that the plane represents. Each unit in a cell represents the strength of its associated feature at the

\(^1\)This chapter is based on Walle et al. (2008). The results, except where noted, have been updated to reflect the new network design adopted since that paper appeared.
Figure 4.1: The general structure of my convolutional neural network. Plies of cells are connected by layers of weights. Ply dimensions are given in cells, each cell contains one unit for every feature represented by that ply.
The network learns to abstract over space. Training in a variety of locations, very good translation invariance (close to 100%) can be achieved. That is, the network recognises shapes even in locations they were never trained in. Together with local connectivity of the input ply without interaction between different features, little change in the number of cells between plies, while abstracting layers reduce the number of cells.

Successive plies divide the visual field more and more coarsely, so contain fewer cells than their predecessors, each of which has a wider receptive field than those in earlier plies. However later plies inside a cell’s window will be the same irrespective of where in a ply the cell is located. In addition, the weights for corresponding units in different cells of a ply are constrained to be the same, effectively sharing the weights. This means that the response to activity inside a cell’s window will be the same irrespective of where in a ply the cell is located.

Successive plies divide the visual field more and more coarsely, so contain fewer cells than their predecessors, each of which has a wider receptive field than those in earlier plies. However later plies generally represent more features than earlier plies, and therefore contain more units per cell.

The function and structure of the weight layers alternates throughout the network between convolution and abstraction. Convolving layers compute combinations of features in the previous ply with little change in the number of cells between plies, while abstracting layers reduce the number of cells of the input ply without interaction between different features.

Local connectivity and weight sharing give the network some degree of translation invariance—that is, the network recognises shapes even in locations they were never trained in. Together with training in a variety of locations, very good translation invariance (close to 100%) can be achieved. The network learns to abstract over space.

The sharing of weights is not biologically plausible, but should be considered a simplification to hasten learning. Biologically, we might expect weights linking cells in different locations to converge due to presentation of stimulus types during learning across most of the visual field. The weight sharing just codifies this assumption into the network.

Many different shapes and sizes of CNNs are possible. I chose, for simplicity, to use one with a uniform structure throughout. This decision, along with the choice of integration window size, determined the number of plies and their sizes. The only other free parameter was the number of cells receiving input from a small square region of the ply beneath, meaning they are connected locally and can only make use of local features.
features in each layer, which was fairly arbitrary (work performed with other feature profiles suggests better results than those presented here may be possible).

In convolving layers, an output unit receives input from every unit within its $2 \times 2$ integration window. A unit receiving input from a ply representing $n$ features will have $4n + 1$ inputs (including a bias). Figure 4.2 illustrates this. Weights are shared spatially, so that a unit in the output ply representing a particular feature will have its inputs modulated by the same weights as all the units in the output ply corresponding to that feature. Weights are not shared between different features in the output ply. The window of a cell in the output ply is offset one cell from those of neighbouring output cells, as illustrated in Figure 4.2.

The structure of a convolving layer basically implements a generalisation of the convolution operation (see Section §3.5.9) commonly used in machine vision. In the special case of input and output representing only one feature each, and biases being set to zero, the process really would be just convolution. The convolution layer allows spatially local combinations of input features to be computed.

Weights in abstracting layers are simpler. Input and output plies contain the same number of features and there is no interaction between features. A unit receiving input from its $2 \times 2$ window will have 5 inputs (including a bias). The window of a cell in the output ply precisely abuts but does
Figure 4.3: Abstracting layers. (a) An example of the complete integration window of an output unit for an abstracting weight layer. The unit has a square integration window measuring two cells on each side. The input and output plies here both represent two features (the number of features represented never changes over an abstracting layer). Each cell therefore contains two units, one for each feature represented. An output unit representing a particular feature receives input from all units representing the same feature inside the window in the ply below, plus a bias which can vary in the same way as the weights. In abstracting layers, each unit’s weights are constrained to be identical (though the bias may differ). Different units within a cell, however, have different weights. Corresponding weights are shared between cells. (b) The relationship between integration windows for abstracting weight layers. Squares demarcate cells, circles represent units and broken lines illustrate integration windows. Units in each output cell only receive input from units inside the cell’s integration window. The integration windows of units in neighbouring cells do not overlap, but exactly abut.
not overlap with the windows of neighbouring cells. This is illustrated in Figure 4.3. The effect is that
the integration windows of cells in the output ply tile the input ply. Weights are shared even further
within abstracting layers, with all weights for a feature constrained to be identical. This means that
each abstracting layer really has only two variable parameters per feature: one weight shared among
all the inputs units, and the bias.

Apart from the varying structure of the layers, unit activation is computed in the same way
throughout the network. For a unit with \( n \) inputs \( p_1 \ldots p_n \) (excluding the bias) and \( n + 1 \) weights
(including the bias) \( w_1 \ldots w_{n+1} \) the unit’s activation, a weighted sum, \( \sigma \) is computed:

\[
\sigma = \sum_{i=1}^{n} p_i w_i + w_{n+1}
\]

which for an abstracting unit can be simplified further to:

\[
\sigma = w_{\text{ply}} \sum_{i=1}^{n} p_i + w_{\text{bias}}
\]

because of weight sharing.

The output of the unit is then computed via the logistic function:

\[
f = \frac{1}{1 + e^{-\sigma}}
\]

This is conventional for feed-forward networks.

4.3 Training

There are two components to the training – the algorithm used and the details of the training examples
used. I will treat these separately.

4.3.1 Training Algorithm

The network was trained using the \textsc{rprop} algorithm (Riedmiller, 1994). This is a variation of the \textsc{backprop}
algorithm (Rumelhart et al., 1986). This algorithm works to minimise the squared error
of each output unit:

\[
\epsilon = (d - f)^2
\]

where \( d \) is the desired output and \( f \) the actual output. These errors on each output unit are propagated
back through the network where they determine weight changes in intermediate layers.

Although this algorithm is not biologically plausible, CNNs are typically trained using some version
of it and this does not make them less useful as models of the brain. I am intending the trained network
to model the mature human classification system – I am not intending to model the learning process
which produces this system.

A number of training examples, matched with desired output, are used during training. During a
single cycle (or batch or epoch) of training, each example is presented to the network, the discrepancy
between the actual and desired output is registered in the units of the final ply and this is propagated
backwards through the network. Once all of the examples have been presented, the error information
stored throughout the network is used to update the weights.

Each time a pattern is presented, the value \( \delta_{N,i} \) for unit \( i \) in the last ply of an \( N \) ply network is
computed using

\[
\delta_{N,i} = f_{N,i}(1 - f_{N,i})(d_{N,i} - f_{N,i})
\]

where \( d_{N,i} \) is the desired output of the unit and \( f_{N,i} \) is the actual output of the unit. For units
in earlier layers, the value \( \delta_{l,i} \) for unit \( i \) in the \( l \)th ply which connects to \( n \) units in the next ply is
computed recursively by

\[
\delta_{l,i} = f_{l,i}(1 - f_{l,i}) \sum_{k=1}^{n} \delta_{l+1,k} w_{l+1,k}
\]
where \( w_{l,i}^{l+1,k} \) is the weight connecting the \( r \)th unit of ply \( l \) and the \( k \)th unit of ply \( l + 1 \). The units summed over are precisely those that the unit in question connects to in the subsequent ply. The size and layout of this set differs depending on the location of the unit and what kind of layer (convolving or abstracting) connects the plies.

As the \( \delta \) values are computed, error derivatives are also computed. After the presentation of a single pattern the partial derivative of each unit’s error, \( \epsilon \) with respect to each of its incoming weights is

\[
\frac{\partial \epsilon}{\partial w_{l,i}^{l+1,k}} = -2\delta_{l+1,k}f_{l,i}
\]

which is an error gradient of the unit. It is associated with the weight \( w_{l,i}^{l+1,k} \) and where weights are shared the different derivatives are summed.

The partial derivatives are accumulated over all of the training examples in the batch. At the end of the batch weight changes can be made.

RPROP requires that apart from its actual weight value, each weight has two values associated with it between batches. The first is the \( \Delta \) value, which determines the size of weight changes. The other is the immediately previous value of the partial error derivative. The old partial derivative allows the algorithm to detect when the sign of the partial derivative has changed, indicating it has jumped over a local minimum. The \( \Delta \) value is adjusted up or down to accelerate learning when big changes are possible, and slow it down when fine control is required.

For a given weight let \( m(t) \) indicate the value \( m \) at batch \( t \). Then

\[
\Delta(t) = \begin{cases} 
\eta^+ \Delta(t-1), & \text{if } \frac{\partial \epsilon}{\partial w}(t) \frac{\partial \epsilon}{\partial w}(t-1) > 0 \\
\eta^- \Delta(t-1), & \text{if } \frac{\partial \epsilon}{\partial w}(t) \frac{\partial \epsilon}{\partial w}(t-1) < 0 \\
\Delta(t-1), & \text{otherwise}
\end{cases}
\]

where \( 0 < \eta^- < 1 < \eta^+ \). Now the weight update \( \Delta w(t) \) is

\[
\Delta w(t) = \begin{cases} 
-\Delta(t), & \text{if } \frac{\partial \epsilon}{\partial w}(t) > 0 \\
+\Delta(t), & \text{if } \frac{\partial \epsilon}{\partial w}(t) < 0 \\
0, & \text{otherwise}
\end{cases}
\]

based on the \( \Delta \) values computed earlier. There are two additional points to note. The \( \Delta \) values are constrained to never exceed a constant value \( \Delta_{\text{max}} \) and never go below a constant value \( \Delta_{\text{min}} \). This is performed by simple thresholding during the computation of the values. And in the event that the sign of \( \frac{\partial \epsilon}{\partial w} \) changes, there is no adaptation of the weight on the next batch, which is achieved by setting the recorded value of \( \frac{\partial \epsilon}{\partial w} \) to zero for the next batch. The partial derivatives are initialised to zero before training begins, and the \( \Delta \) values to a value \( \Delta_0 \).

The values of the parameters \( \Delta_{\text{min}}, \eta^-, \eta^+ \) and \( \Delta_0 \) in my implementation were \( 1 \times 10^{-6}, 0.5, 1.2 \) and 0.1 respectively. \( \Delta_{\text{max}} \) had the value 0.08 except in the first two layers (at the input end) where it had the value 0.5.

### 4.3.2 Training Regime

Prior to training, all weights in the network were initialised to small random numbers.

I trained my network with greyscale images equal in size to the \( 31 \times 31 \) retina. Depending on the scale these were either filtered or presented directly, as described below. There was only one cell in the final output ply, with a receptive field spanning the entire retina, and that cell contained seven units. Six of these units were associated with the six shapes shown in Figure 4.4 during training. The seventh output was associated with random noise during training.

Small shapes were each trained at a randomly chosen third of possible retinal locations (independently and without replacement). Large shapes were each trained at a random third of all possible retinal locations for each of four densities. These included solid shapes as shown in Figure 4.4 as well
as the same shapes with pixels randomly ablated to the background colour with probabilities \( \frac{1}{6}, \frac{1}{3}, \frac{1}{2} \). Thus for the low-frequency training total spatial coverage was likely. The small shapes were presented to the high-frequency inputs only and the large shapes to the low-frequency inputs only. During operation only one of the sets of inputs was used at a time, the other being suppressed entirely (At this stage the decision about which set of inputs is used to process a given shape is made by hand. Later it will be automated – see Section §5.4). Noise examples were each presented to both sets of inputs one after the other.

There were 1566 positive high-frequency training examples and 2152 positive low-frequency training examples. There were also 371 negative (or noise) examples which were each fed to the low- and high-frequency inputs in turn. New noise examples were generated on each cycle of training. This number is approximately 10% of the number of non-noise examples. The noise examples were divided among eleven groups, and each group was allocated a different density between 0.05 and 0.5. The probability of a pixel being set in a noise example was equal to the density of its group. If set, a pixel took a random value between zero and one, otherwise the pixel took the value zero. This provided noise of varying density. New noise examples were used for each batch. Thus the training set consisted of 4089 examples: 1566 positive high-frequency examples, 2152 positive low-frequency examples, 33 noise examples for each of nine density groups with densities at intervals of 0.045 starting at 0.05, and 8 noise examples with a density of 0.5.

**4.4 Location and Scale Invariance**

First I tested the network’s translation and scale invariance. A translation invariant classifier successfully classifies stimuli regardless of where in the visual field they are presented. The obvious way to test this was to present every shape in every one of the possible locations on the retina and see how the network performed.

The classifier is blind to which of the two sizes of shapes are presented. This is not due to generalisation during learning, but arises because both sizes are trained.

Table 4.4 and Figure 4.5 show the results. For the small shapes, a successful classification was made at more than 90% of locations for four cases, and no shape did worse than 80% of locations. This indicates a high level of translation invariance. Note that during training each shape is presented in only approximately one third of the possible locations.

For the low-frequency shapes performance is close to perfect with single items, though because of the training regime these probably were trained at all possible locations.
<table>
<thead>
<tr>
<th>Scale</th>
<th>Shape</th>
<th>Total trials</th>
<th>Hits</th>
<th>Misses</th>
<th>Correct winners</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>Arm</td>
<td>272</td>
<td>271 (99.6%)</td>
<td>0 (0.0%)</td>
<td>271 (99.6%)</td>
</tr>
<tr>
<td></td>
<td>Arrow</td>
<td>256</td>
<td>256 (100.0%)</td>
<td>1 (0.4%)</td>
<td>256 (100.0%)</td>
</tr>
<tr>
<td></td>
<td>Cross</td>
<td>324</td>
<td>323 (99.7%)</td>
<td>0 (0.0%)</td>
<td>324 (100.0%)</td>
</tr>
<tr>
<td></td>
<td>Ell</td>
<td>256</td>
<td>256 (100.0%)</td>
<td>4 (1.6%)</td>
<td>256 (100.0%)</td>
</tr>
<tr>
<td></td>
<td>Square</td>
<td>256</td>
<td>249 (97.3%)</td>
<td>0 (0.0%)</td>
<td>251 (98.0%)</td>
</tr>
<tr>
<td></td>
<td>Triangle</td>
<td>256</td>
<td>256 (100.0%)</td>
<td>1 (0.4%)</td>
<td>256 (100.0%)</td>
</tr>
<tr>
<td>Small</td>
<td>Arm</td>
<td>784</td>
<td>635 (81.0%)</td>
<td>25 (3.2%)</td>
<td>693 (88.4%)</td>
</tr>
<tr>
<td></td>
<td>Arrow</td>
<td>784</td>
<td>708 (90.3%)</td>
<td>25 (3.2%)</td>
<td>742 (94.6%)</td>
</tr>
<tr>
<td></td>
<td>Cross</td>
<td>784</td>
<td>726 (92.6%)</td>
<td>40 (5.1%)</td>
<td>753 (96.0%)</td>
</tr>
<tr>
<td></td>
<td>Ell</td>
<td>784</td>
<td>784 (100.0%)</td>
<td>16 (2.0%)</td>
<td>784 (100.0%)</td>
</tr>
<tr>
<td></td>
<td>Square</td>
<td>784</td>
<td>683 (87.1%)</td>
<td>43 (5.5%)</td>
<td>726 (92.6%)</td>
</tr>
<tr>
<td></td>
<td>Triangle</td>
<td>784</td>
<td>769 (98.1%)</td>
<td>37 (4.7%)</td>
<td>770 (98.2%)</td>
</tr>
</tbody>
</table>

Table 4.1: Results of testing scale and location invariance. Each shape was placed at every possible position on the 31 \times 31 retina and classification was attempted. The total number of trials thus performed for each shape is shown. The table shows three results for each shape. The hits column is the number of examples that produced a correct classification. The misses column is the number of examples that produced at least one incorrect classification. If the network activates more than one output unit for a given pattern, both conditions can hold, therefore these two need not add up to the total number of trials. For these columns a classification is defined as an output unit with output greater than 0.5. The correct winners column is the number of examples for which the correct output unit had the highest output, regardless of whether it crossed the threshold.

Figure 4.5 shows that performance varies from shape to shape but is close to perfect in the centre of the classifier’s retina with most errors occurring in borders about four pixels wide around the edges. In the experiments presented later the attended stimuli are automatically centred in the classifier’s retina and so tend to avoid these border effects.

### 4.5 Cardinality Invariance

Having established the performance of the network on single stimuli, I could now explore performance at group classification. To do this I explored the CNN’s performance in detail when classifying groups of varying homogeneity made up of three items (Section §4.5.1) and four items (Section §4.5.2). I also examined performance with homogeneous groups at a range of cardinalities (Section §4.5.2). All the cardinality invariance experiments were performed with small shapes only.

#### 4.5.1 Cardinality Invariance with Three Shapes

Because each member of a group could appear in many possible retinal locations, an exhaustive survey was impractical. Sampling was therefore employed. One thousand blocks of 216 trials were performed. Each block began with the selection of three random locations on the retina, the choice constrained only so that shapes placed at any two of those locations would always be separated by at least two pixels. Now a test input was created for each combination of the six shapes at those three locations, \(6^3 = 216\) inputs in all. See Figure 4.6 for example stimuli. Finally the test examples were fed to the network and the classification results noted. Altogether 216,000 inputs were presented to the network.
Figure 4.5: Maps of the classifier’s accuracy by location. For each trained shape there are two maps: the left is for the large version of the shape, the right is the small version. The shape was placed at each possible position on the retina and the map shows the responses of the corresponding output unit arranged topographically. Dark areas show locations where the correct shape was the strongest output and also above 0.5. White areas show misclassifications.
Figure 4.6: Examples of the kinds of patterns used in the cardinality invariance experiments. Examples are labelled XXX if they are homogeneous, that is the three shapes are identical. Examples are labelled XXY if they contain a pair of identical shapes and a third different shape. Heterogeneous examples, those with all three shapes different, are labelled XYZ. The ordering in the labels is not significant: every distinct combination of the shapes at the different locations was included in the experiment.
Table 4.2: Results of testing cardinality invariance with three items. One thousand blocks of 216 trials were performed. The figures in the table are for all trials, grouped by the homogeneity of the input. The groups are labelled with the forms XXX, XXY and XYZ, representing inputs containing three identical shapes (homogeneous case), exactly two identical shapes and three distinct shapes (heterogeneous case) respectively. Figure 4.6 shows examples of the inputs that fall into each group.

An example was considered classified if an output other than the error output was greater than 0.5. If more than one output was greater than 0.5 then the output with the greatest value was the result of classification. All other cases were considered unclassified. For each group, the proportion of that group that produced a classification and the proportion that went unclassified are given. These are the most important values. For those that were classified, there is a further breakdown according to how well the classification matched the shapes actually present in the input. Note that as the heterogeneity of the inputs increases, the classifier becomes more reluctant to assign a class.

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
<th>Examples</th>
<th>Classified</th>
<th>Matched X</th>
<th>Matched Y</th>
<th>Matched None</th>
<th>Unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXX</td>
<td>Three identical shapes</td>
<td>6000</td>
<td>98.8%</td>
<td>98.7%</td>
<td>0.1%</td>
<td>1.2%</td>
<td></td>
</tr>
<tr>
<td>XXY</td>
<td>Two identical shapes</td>
<td>90000</td>
<td>79.4%</td>
<td>58.8%</td>
<td>20.1%</td>
<td>0.5%</td>
<td>20.6%</td>
</tr>
<tr>
<td>XYZ</td>
<td>Three distinct shapes</td>
<td>120000</td>
<td>66.9%</td>
<td>66.5%</td>
<td>0.4%</td>
<td>33.1%</td>
<td></td>
</tr>
</tbody>
</table>

The results are shown in Table 4.2, grouped by the homogeneity of the input. The key points are the proportions of examples that were classified and unclassified for each group. The homogeneous groups were mostly classified correctly, remaining close to the performance with single shapes, and easily meeting the first part of the definition of cardinality blindness given in Section §3.5.2.

The table also clearly shows that the proportion that went unclassified increases with heterogeneity of the group. For completely heterogeneous patterns the network refuses to classify about a third of the time. This result is not as robust (relative to the definition of cardinality blindness) as that for homogeneous groups, but I consider that it at least partially meets the second part of the definition.

4.5.2 Cardinality Invariance with Four Shapes

This was a repeat of the previous experiment, but using four shapes in each example, rather than three, yielding $6^4 = 1296$ trials per block. Table 4.3 shows the results, grouped similarly to those in Table 4.2. For this experiment there are more levels of homogeneity, but the key results remain the increase in the proportion of unclassified examples heading down the table, and the difference between these for the totally homogeneous and totally heterogeneous groups. Homogeneous groups are still classified with performance close to or even better than that with single shapes. Completely heterogeneous groups fail to classify about half of the time, better than the result with three shapes. This also suggests that (as Mozer and Sitton, 1998, showed) as the number of heterogeneous items
<table>
<thead>
<tr>
<th>Group</th>
<th>Examples</th>
<th>Classified</th>
<th>Matched X</th>
<th>Matched Y</th>
<th>Matched None</th>
<th>Unclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td>XXXX</td>
<td>6000</td>
<td>98.1%</td>
<td>97.9%</td>
<td>0.2%</td>
<td>1.9%</td>
<td></td>
</tr>
<tr>
<td>XXXY</td>
<td>120000</td>
<td>78.6%</td>
<td>67.2%</td>
<td>11.0%</td>
<td>0.4%</td>
<td>21.4%</td>
</tr>
<tr>
<td>XXXY</td>
<td>90000</td>
<td>77.2%</td>
<td></td>
<td></td>
<td></td>
<td>22.8%</td>
</tr>
<tr>
<td>XXXZ</td>
<td>720000</td>
<td>65.2%</td>
<td></td>
<td></td>
<td></td>
<td>34.8%</td>
</tr>
<tr>
<td>WXYZ</td>
<td>360000</td>
<td>52.2%</td>
<td></td>
<td></td>
<td></td>
<td>47.8%</td>
</tr>
</tbody>
</table>

Table 4.3: Results of testing cardinality invariance with four items. One thousand blocks of 1296 trials were performed. The groups are labelled and divided into groups similarly to those in Table 4.2. Note that as the heterogeneity of the inputs increases, the classifier becomes more reluctant to assign a class.

increases classification performance decreases steeply. That in turn means that as the number of items increases performance more closely matches my definition of cardinality blindness (see Section §3.5.2).

4.5.3 Comparing Cardinality Invariance across Group Sizes

Having demonstrated cardinality blindness at two particular cardinalities I now examined the performance of the network at different cardinalities. For every single shape in every position, and for 3000 randomly arranged homogeneous groups of size two, three, four and five, classification was attempted. The arrangement of groups was restricted in the same way as for the other cardinality invariance experiments.

For each test example the units that passed the 0.5 threshold were tallied. For homogeneous groups of input type X, the number of times output X was over the threshold (correct group classifications) were counted up and appear in the top part of Table 4.4. The number of times any output other than X was over the threshold (false positives) were also counted up and appear in the bottom part of Table 4.4. The type is correctly identified most of the time. Homogeneous groups are classified similarly at different cardinalities, with only a very slight drop in performance with increasing cardinality. The results also show that the network seldom mistakes a group of one type for that of another.
Table 4.4: Results for comparison of cardinality invariance with different display sizes. The first five rows describe the response of each output unit when presented with a homogeneous group made up of its corresponding type. A response in this situation is a correct classification. The last five rows describe the responses of all the other output units when presented with a homogeneous group made up of the remaining output’s type. A response in this situation is a false positive.

We might expect, in fact, that classification performance improves as the number of items presented increases. Perhaps the presence of more items would make the group classification more resilient to noise. Perhaps more examples would make the system more certain of its classification. I am not aware of any studies examining this effect in humans. Note, though, that performance for groups is better than singly presented items. The very slight decrease in performance shown in Table 4.4 may be an artifact. The more items present the more likely one will hit one of the blind spots shown in Figure 4.5. Maybe this decreases performance and masks a facilitatory effect with increasing display size. This is currently just speculation, but the effect of the number of items on accuracy (both in the model and experimentally) is worthy of closer examination.

4.6 Repeatability

All the experiments described above were performed with a single network. Do the invariances noted arise in this kind of CNN generally, or is this a special case? My model is very slow and somewhat temperamental to train (though this has shown some improvement). Occasionally networks refuse to converge, and often they do not generalise some of the shapes well. However the results always reflect at least some translation invariance, and similarly at least some cardinality invariance. Table 4.5 summarises invariance performance for four consecutively trained networks, plus the one (in the final column) used in the experiments presented above. In most cases training was stopped because the training error was no longer decreasing. It should be noted that the other networks were not identical to the one presented here: though very similar they have fewer features at the input and intermediate layers and were trained with small shapes alone (as described in Walles et al. (2008)). It was not thought necessary to repeat this test with the new network configuration especially as it is very time intensive.

4.7 The Generality of Cardinality Blindness

4.7.1 Other Cardinality Blind Models

This is not the only model to exhibit some kind of cardinality blindness. For instance the models of Riesenhuber and Poggio (1999) and Serre et al. (2005) (see Section §3.5.8) are cardinality blind. Their model shares some features with mine. It is hierarchical, for instance, and is scale and location
Table 4.5: Translation and cardinality invariance results for five different networks, including the one used in the experiments (final column). To measure translation invariance each shape is classified in every position as in experiment one. The minimum, maximum and mean of the correct classifications is displayed. The proportion of XXX patterns correctly classified corresponds to the second value in Table 4.2 and the proportion of XYZ patterns unclassified corresponds to the final value of that table.

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs trained</td>
<td>34953</td>
<td>7393</td>
<td>3657</td>
<td>8839</td>
<td>4353</td>
</tr>
<tr>
<td>Min. translation invariance</td>
<td>67.2%</td>
<td>34.3%</td>
<td>62.9%</td>
<td>89.3%</td>
<td>81.0%</td>
</tr>
<tr>
<td>Max. translation invariance</td>
<td>95.9%</td>
<td>92.1%</td>
<td>97.8%</td>
<td>99.0%</td>
<td>100.0%</td>
</tr>
<tr>
<td>Mean translation invariance</td>
<td>85.0%</td>
<td>77.0%</td>
<td>85.9%</td>
<td>95.3%</td>
<td>91.5%</td>
</tr>
<tr>
<td>XXX correctly classified</td>
<td>70.8%</td>
<td>66.8%</td>
<td>73.1%</td>
<td>91.4%</td>
<td>98.7%</td>
</tr>
<tr>
<td>XYZ unclassified</td>
<td>72.1%</td>
<td>69.1%</td>
<td>60.1%</td>
<td>52.6%</td>
<td>33.1%</td>
</tr>
</tbody>
</table>

invariant, too. Importantly, my results show that cardinality blindness is not tied to a very specific classifier design. Though of course there must be necessary conditions under which it will appear, they may not be too numerous.

Perhaps, indeed, cardinality blindness is a feature of any classification device that is also translation invariant. If so, then it neatly follows that IT must be blind to cardinality because of its translation invariance. But even if the link between these two invariances is not universal, the structural and functional similarity of the network to IT makes its behaviour relevant.

4.7.2 Biological Plausibility

Like IT, the network contains a series of stages, with the elements of later stages representing larger portions of the visual field than those of earlier stages. Though my CNN is a crude device compared to the human visual system, their gross structures are comparable. My CNN uses eight layers, corresponding to synaptic connections, between the retina and the classification output, compared with at least five between the retina and similarly class-selective regions of IT (Mesulam, 1998). The low resolution of my retina means that units in the input ply have receptive fields covering 3% of the retina, but those of the seventh ply cover a modest 45%, comparable to the largest found in TE (Logothetis and Sheinberg, 1996). The receptive fields of the eighth ply nearly cover the entire retina, and those of the ninth ply actually do. But since a biological system would likely represent the output in a distributed fashion, rather than with one neuron per class as I have (for convenience), it’s reasonable to assume TE cells are modelled by the seventh or eighth plies.

The early structure of the visual system (V1 and V2) with its familiar feature-detecting cells (Hubel and Wiesel, 1968) is consistent with the CNN model. By modelling the ventral pathway with a CNN I am effectively hypothesising that this structure continues in later processing stages. This hypothesis leads to a model which is translation invariant, just like IT, and also cardinality invariant. I suspect, however, that it is translation invariance rather the particular structure of the CNN that is necessary for cardinality invariance. It would be interesting to know which other translation invariant classification models are also cardinality invariant, especially those which more closely match biological reality.

Observations of cells in IT indicate they are not as translation invariant as was once thought. DiCarlo and Maunsell (2003) show that cells are somewhat selective for location. Yet IT specialises in recognition, which does require translation invariance, and this is strongly supported by other evidence (Goodale and Milner, 1992; Ungerleider and Mishkin, 1982). These observations are only superficially in conflict. Even my CNN model consists, mostly, of units that are selective for location. But the output units, and hence the classifier as a whole, are translation invariant. An invariant classifier need not be made of invariant parts. If TE cells are modelled by the seventh or eighth layers then no unit
is fully translation invariant, yet the distributed representation is ready to be cleansed of translation information with minimal processing (via one or two layers).

4.8 Conclusion

It is interesting to consider how my model of IT relates to visual perception as a whole. The model is solely a classifier: it does not concern itself with determining the locations of objects (a job that is supposed to fall to the dorsal pathway). This also means that it does not involve any of the processes that fall under the general heading of visual attention.

As Mozer and Sitton (1998) argued when they presented their own CNN model of IT, attention is necessary for the IT classifier to operate in a real environment. They presented different sized groups of stimuli to their network, and found that as the size of the group increased, classification performance dropped sharply. They concluded from this that a classifier like that in IT needed a filtering mechanism – attention – to make sure only one stimulus (or perhaps two) reached the classifier at a time. Attention, then, restricts the flow of information from the real world to the input of the classifier so that it can do its job. My results lead to a similar conclusion, but with the restriction slightly relaxed. A CNN makes an accurate classifier, but its performance drops sharply if asked to classify more than one type at a time. I could conclude, then, that a filtering mechanism is needed to make sure that only homogeneous groups reach the classifier.

In the next chapter I describe an implementation of just such a mechanism.
Chapter 5

A Model of Serial Visual Attention and Group Classification \(^1\)

5.1 Introduction

In this chapter I describe the attentional component of my overall model, incorporating the detection and attention aspects of object perception (see Section ³.1 for definitions of these). I also test the combined classification-attention model in a purely perceptual context.

The properties of visual attention should be partly determined by the input requirements of the IT classifier. Attention needs to know where the currently attended stimulus is, how big it is and so on so that it can direct the classifier appropriately. As we saw in Chapter 3 visual attention is a complex process. It can be directed by factors like common fate that transcend mere location, taking into account structure, motion and other aspects of the visual field that are of use in dividing the world into parts that are manageable or important.

My model so far suggests that the classifier can manage inputs as long as they contain a single type, no matter how many tokens of that type appear. This would constrain visual attention to provide such manageable chunks. And so it seems to be. One of the well known Gestalt principles groups stimuli by homogeneity (see Section ³.4). The human visual system does tend to cluster like with like, just as one would expect of an attention system that grouped homogeneous items for presentation to a classifier that was capable of dealing with them because it was blind to cardinality.

It should be noted, by the way, that grouping by homogeneity can be explained without recourse to a limitation of the classifier. The Gestalt factors also come into play when dividing visual input into objects. For instance, objects made of a particular material often have a characteristic texture, which helps us to distinguish them from their background. Grouping by homogeneity acknowledges that objects are often made of the same stuff throughout.

The nature of the original need for grouping by similarity is not critical anyway. The ability of the IT classifier to do group classification would be valuable even if it was a phylogenetic afterthought. Parallel classification of homogeneous groups saves time and one of the things I am going to demonstrate in this chapter is how it can speed up ecologically useful processes like visual search.

To sum up so far: if the classifier in IT is cardinality blind, this may be expected to have consequences for the design of the attentional system. Selective attention is commonly believed to play a filtering role for visual classification by restricting the visual field so that later visual processing is not overwhelmed (Moran and Desimone, 1985; Treisman and Gelade, 1980). However there are many ways to slice up visual stimuli, depending on the constraints of later processing. If IT can perform group classification then the attentional system could make use of this ability and select homogeneous groups of objects for classification as well as single objects.

In Section ⁵.2 I describe the broad structure of the attentional model. In Sections ⁵.3 to ⁵.6 I describe in detail how it models the preattentive grouping principles of proximity and similarity, the

\(^1\)This chapter is based on Walles et al. (2010).
combination of this information to generate salient proto-objects, and the attentive (serial) selection of those proto-objects. Finally, in Section §5.7 I identify the value of this model and describe some testable predictions it makes.

5.2 Attention Model Overview

Figure 5.1 shows the outline of the overall model, including the attentional components.

The attention model can be divided into two interacting stages: a preattentive, or parallel, stage and an attentive, or serial, stage (again, see Figure 5.1). This reflects the well known idea that attention is a mechanism which manages the interface between parallel and serial processing in the brain (Broadbent, 1958).

The job of saliency analysis is to analyze the local contrast and texture homogeneity of the input in parallel. These are used to implement the Gestalt grouping properties of proximity and similarity respectively. The result of this is a saliency representation, or saliency map – a notion introduced by Koch and Ullman (1985) and extended by Itti and Koch (2000) and Walther and Koch (2006) (see Section §3.6.4.2). This representation is the point of communication between saliency analysis and the selection mechanism.

The selection mechanism uses the saliency representation to decide how best to deploy attention and, once processing of attended stimuli is complete, the representation is updated and then used to redeploy attention.

The saliency representation is also used to gate the input to the classifier. Input is gated in two different ways. It is gated by location, which is a well-known idea. And it is also gated by scale, which is a new idea I have introduced. Given the region and scale corresponding to the currently attended proto-object, only primitive features corresponding to that region and scale will reach the classifier from early vision.

Although inputs to the system as a whole measure 128x128 pixels, inputs to the classifier always measure 31x31 pixels. This is a practical limitation of the classifier and the disparity is resolved by always centring the attended region in the classifier's input for classification purposes.

5.3 Parallel Attention Component: Saliency Analysis

The saliency analysis module parses the visual field and produces a saliency representation – basically a saliency map. This map identifies a number of proto-objects, salient regions tagged with corresponding salient scales. Each proto-object contains either a single item or a set of items grouped according to the Gestalt principles of proximity and similarity. Salient regions are identified by computing partial saliency maps from the input. Two local contrast maps are computed, using Laplacian of Gaussian filters tuned to two different spatial frequencies (see Section §5.3.1). A single texture homogeneity map at the higher spatial frequency is computed using the statistical histogram-based system of Liu and Wang (2000) (see Section §5.3.2). The two local contrast maps and the homogeneity map are combined to produce the saliency representation (see Section §5.3.3).

Prima facie, there may seem to be a conflict between local contrast and homogeneity as indicators of saliency. Saliency computed from contrast, and saliency computed from homogeneity (which implies no contrast) seem to fight each other. However, while the principles may be in conflict at a single spatial scale, I suggest that they are complementary at different spatial scales. The kind of things which are salient for us are objects which contrast from their background and are made up of uniform texture elements. Thus uniformity is required at a higher spatial frequency than contrast. The appropriate way of combining local contrast and homogeneity is reminiscent of Marr's (1982) idea that regularities found in natural images can be exploited by combining information across spatial frequencies.

Figure 5.2 illustrates the concept. The low-frequency saliency map (tuned to large items) is the weighted sum of the low-frequency contrast map and the high-frequency homogeneity map. The high-frequency saliency map (tuned to small items) is the collection of regions of high contrast alone.
Figure 5.1: The broad structure of the overall computational model. The classifier (ventral pathway) and the components on the left making up the attention system (dorsal pathway) both receive input from the retina. The attention mechanism can also gate the input to the classifier ensuring that only information from desired locations and spatial frequencies reaches it.
weighted sum of these produces a scale-independent saliency map, not shown in the figure.) Textures
which are optimally analysed at the higher spatial frequency provide a measure of similarity which
complements the measure of proximity provided by local contrast at the lower spatial frequency.

My model makes use of only two primitive spatial frequencies and so only produces one saliency
map by combining texture and contrast information (by combining the low-frequency contrast with
high-frequency homogeneity). However I also use the high-frequency contrast alone to provide a
saliency map at the high frequency, allowing my model to operate at two scales. The lack of texture
information for the high-frequency saliency measure does not affect any of my experiments but I
expect that a more realistic system would incorporate more spatial frequencies. By chaining together
a range of spatial frequencies, a corresponding range of saliency measures can be produced, each
corresponding to a different spatial frequency, or scale, as shown in Figure 5.3.

5.3.1 Local Contrast

Saliency analysis is based on the model presented by Itti and Koch (2000) and Walther and Koch
(2006). Local contrast computation begins by taking the luminance of the input image and scaling the
values into the range -128 (black) to 127 (white). Local contrast is then computed by convolving with
two Laplacian of Gaussian filters, one for each spatial frequency (\(\sigma = 1\) and \(\sigma = 15\)). The absolute
value of each resulting value is then taken.

5.3.2 Homogeneity

The similarity measure is computed by the procedure described in Liu and Wang (2000). This pro-
cedure samples a small 7x7 pixel region around each pixel in the input image, computing its spectral
histogram which can be thought of as a high-dimensional feature vector, and finally finds the closest
match to this histogram among those belonging to a set of texture templates derived from images of
the small shapes in Figure 4.4 both closely packed and sparsely scattered. Once each pixel is assigned a
category (square, ell, etc. or background), boundaries are determined by comparing four-neighbouring
pixels. Whenever a pair differs in category, the pixel that was least certainly classified (determined
by the distance between it’s histogram and that of its category’s template, measured as a \(\chi^2\) value) is
marked as a texture boundary. In the resulting boundary map homogeneous regions are marked with
zero, boundaries with one.

For the experiments presented here I wanted some stimuli to be considered similar enough for
saliency analysis to group them even though they were distinct. To this end I defined that boundaries
between ells and squares, crosses and arrows, arrows and arms, arrows and triangles and triangles and
arms would not be marked in the boundary map.

The resulting boundary map is then combined with the low-frequency local contrast map by a
weighted sum, 15.5 times the contrast value minus 1.45 times the boundary map value. The result
of this sum is the low-frequency saliency map. The high-frequency saliency map is just the same as
high-frequency local contrast map.

5.3.3 Combination of Partial Saliency Maps

The saliency maps are then thresholded, the high-frequency map using the threshold value 0.4, the
low-frequency map the threshold value 0.060. Regions which are 4-neighbour contiguous are then
identified and labelled. Any labelled region in the low-frequency map containing fewer than 55 pixels
is discarded. The point-wise sum of these maps yields the master saliency map in which contiguous
regions are also identified and labelled.

The final stage of saliency analysis is the extraction of proto-objects. A proto-object is a region
with an associated scale. In my model, for convenience of implementation, they are represented in a
series of maps, one for each proto-object – though a single map could be used because proto-objects
do not overlap. First, the left- and top-most proto-object with the highest master saliency is found
by scanning from left to right and top to bottom. If there is a low-frequency proto-object at that
point, the low frequency is selected as the associated scale, otherwise the high frequency is selected.
Figure 5.2: A saliency map at a particular spatial frequency is computed by combining the local contrast map at that spatial frequency with the homogeneity map at a higher spatial frequency. The visual stimulus input to the system is locally homogeneous, but also stands out as a global figure from the white background. The low-frequency contrast map captures the latter property, while the high-frequency contrast map captures the former property. These two maps are combined to produce the low-frequency saliency map.
Figure 5.3: A proposal for how saliency measures at several neighbouring spatial frequencies (scales) could be computed. At each scale two kinds of analysis take place. Local contrast is activated strongly by isolated, contrasting regions corresponding to the scale in question. Homogeneity is activated strongly by regions whose elements at the scale above this are locally similar. The section enclosed by dotted lines corresponds to the set of maps depicted in Figure 5.2. In my model only two spatial frequencies are used, as in the examples, but this illustration shows how the principle can be extended to generate a series of scale-specific saliency maps.
Standard morphological dilation (Gonzalez and Woods, 1992) is then applied to the corresponding region (radius 2 pixels for high frequency, 4 pixels for low frequency). Finally, any pixels in the result that overlap existing proto-objects are removed as well as any pixels overlapping any active regions in the master saliency map and any pixels that have been inhibited by previous attention operations (see Section §5.4).

The resulting region is added to the set of proto-objects, labelled with its associated scale. The region as well as any overlapping regions in the non-selected scale are also suppressed in the saliency maps, and the above process is repeated until the master saliency map is empty. Once the set of proto-objects is computed, one is chosen at random and the associated region and scale become the subjects of attention. I do not select proto-objects by decreasing order of saliency, as is typically done, because my stimuli are very simple and the standard measure of “degree of saliency” doesn’t really apply. The randomisation of selection can be viewed as the addition of noise to simulate the variation of saliency found in real-world stimuli. After the winner is selected, suppression of the saliency maps introduced during proto-object computation is then removed.

5.3.4 An Intermediate Result: Performance of the Saliency Analysis Component

Before I discuss the serial component of the attentional model, I will review the performance of the saliency analysis module just described. The most novel feature of the module is the incorporation of a homogeneity term in its computation of saliency. This means we can look at how the system trades off local contrast and homogeneity when these two cues to objecthood conflict. Figure 5.4 illustrates situations in which this can happen, and the problem that it causes for saliency analysis.

Figure 5.5 shows examples of the model’s performance when these grouping principles interact. Each case was generated by presenting the given input to the system and recording all of the proto-objects produced.

If items are close enough (Figure 5.5a) then grouping can occur even among heterogeneous items. At an intermediate separation (Figures 5.5b and 5.5c), grouping is determined by homogeneity. Grouping always requires some proximity, though, and if items are separated widely enough they cannot group even if similar (Figure 5.5d).

Quinlan and Wilton (1998) explored the interaction of the Gestalt properties of similarity and proximity in humans. They found that proximity is a strong grouping factor, that similarity sometimes overpowers proximity, and that, when the two are in direct conflict, subjects resolved the conflict one way or the other depending on a threshold that varied among subjects.

In my model proximity is the stronger grouping factor. If two stimuli are close enough, proximity will always trump similarity. When the two are in direct conflict the conflict is always resolved one way or the other. By altering the weights in the sum of the local contrast map and texture homogeneity map I can simulate the individual differences in grouping threshold found by Quinlan and Wilton.

5.4 Serial Attention Component: The Selection Mechanism

The behaviour of the serial attention components can be understood as a combination of different attentional shift operations. There are two kinds of attentional resources in play: spatial attention and scale attention. The classifier can be restricted to attend only to certain locations in the visual field, and can be restricted to attend only to features at one of two spatial frequencies. Therefore there are two different attention shift operators.

1. A change in the location processed by the classifier.

2. A change in the classification scale, the scale of the primitive visual features used as input to the classifier.

Each operation has lasting side-effects on the saliency representation. Operations are constrained to occur one at a time. I model a sequence of attentional operations as a sequence of time steps.
Figure 5.4: It is possible to devise stimuli where the principles of proximity and similarity conflict. (a) shows a region of dense but heterogeneous texture; (b) shows a region of sparse but homogeneous texture. In both cases the outcome in the conflicting saliency map is shown as indeterminate, but in practice these conflicts must be resolved somehow. Their resolution, and the resulting effects, are explored in Section 5.3.4.
At each step the selection mechanism selects a proto-object at random and allows the classifier to process the associated region and scale. The selection mechanism then alters either the location or scale. The location is changed by inhibiting the current location in the saliency map and selecting a new most salient region. The scale can be changed in two ways. One is as a side-effect of a location change. Whenever a new proto-object is attended to, the classifier is constrained to use the spatial scale associated with the object, its proto-object scale, which is a function of its size. Another is by switching the classification scale without changing location, so that the currently attended proto-object is reclassified at a higher frequency. Once any changes are made the cycle begins again.

5.4.1 The Default Classification Scale

When a proto-object is first attended to, the classification scale used is determined by the proto-object scale. If the proto-object scale is large, a large classification scale will be selected. If the proto-object scale is small, a small classification scale will be selected. Specifically, the classification scale which is selected is the one which allows through the “primitive visual features” from which the form of the proto-object can be constructed. Every proto-object can be presumed to have a form. The default classification scale selected by the proto-object is one which allows this form to be recognised.

I suggest that the default classification scale is the largest scale for which the classifier can generate a meaningful representation of the proto-object. If a coarser scale is used, the form of the proto-object will not be recognised, because the primitive visual features used will not contain enough detail. However, it is important to note that meaningful information about the proto-object can be obtained by selecting a classification scale which is finer than the default scale. In this case, the information will be about the texture of the proto-object, not its form. Thus the definition of “default classification scale” reflects a preference for form analysis over texture analysis. There is also some evidence that low-frequency stimuli are processed more quickly than high-frequency stimuli (Hughes et al., 1996). Once a region is attended to the system can be certain there is some form at that frequency on which to attempt classification. After processing at the low frequency the model tries processing at the high frequency. The reason for trying the high frequency afterwards is to take advantage of the usefulness

Figure 5.5: Results of similarity-proximity conflict tests. In (a) the elements are so close together that proximity wins and a single group is formed. In (b) and (c) the elements all have the same spacing, but in (b) the similarity of the adjacent pairs allows them to form two groups, while in (c) there is no similarity and no groups form. In (d) the elements are so far apart that, even though they are all similar, they cannot group.
Figure 5.6: Performance of the model when classifying the form of a perceptual group. The thick border around the attended region indicates scale-based attention is directed to the low spatial frequency. Note that the system has classified the form of the attended group (as an arm) rather than the type of its constituents (ells).

5.4.2 Inhibition of Return

Once a proto-object is chosen and processing of the attended item is complete, any overlapping regions in the non-attended scale are inhibited in the non-attended scale saliency map. The region corresponding to the attended item is also inhibited in the master saliency map. This implements the well-known inhibition of return (IOR) operation (Posner, 1980, discussed in Section §3.6).

5.5 Performance of the Combined Attention and Classification Systems

5.5.1 Classifying Groups by their Form

An immediately apparent result of this model is its ability to classify the configuration of a group of items. When attending to a group of objects detected as a single proto-object, as in Figure 5.6, the model classifies the form of the group, not the type of its constituents. This is because the proto-object scale determines the classification scale. The classifier can operate at either spatial frequency, but only one at a time, and the attentional model acknowledges this with its implementation of scale-based attention (see Section §3.6.2.5).

5.5.2 Global Precedence

Section §3.6.2.5 describes the phenomenon of global precedence, roughly speaking the preference of attention for global figures over their local constituents.
My system also exhibits global precedence. Presented with a global figure made of smaller figures it will attend first to the region and scale appropriate to the large figure. There are two components in my system that coincide to produce the entire effect. First, because of the attentional grouping principles, regions corresponding to global figures are preferred to the individual regions corresponding to their local components. Second, when a region is selected, the low-frequency scale is initially preferred for classification. Combining these means that the global figure will be attended and classified before its components. Figure 5.7 illustrates the system’s global precedence.

There is dissent (Kimchi, 1992; Navon, 2003) over whether global precedence is a perceptual or attentional phenomenon. In my system the preference for low-frequency regions occurs at the end of saliency analysis when proto-objects are generated, and the preference for low-frequency classification occurs in the selection mechanism. Perhaps global precedence is the result of perception and attention.

In his original description of global precedence, Navon (1977) presented both empirical support for the hypothesis and the theoretical foundations for it. He argued that global precedence is an ecologically sensible approach because it is better to begin with globally salient stimuli and proceed to the details rather than analyse every scene exhaustively using time and energy to extract unnecessary detail.

My model is designed to incorporate grouping as a way to take advantage of group classification. Though not inspired by global precedence, this is consistent with global precedence. It serves the same purpose as Navon intended: group classification can save time extracting information, but only if stimuli are processed globally first. Though this may not be the kind of advantage envisioned by Navon, group classification is a specific example of a motivation for global precedence.

My model also suggests a reason for the failure of Navon (2003) to show an effect of local heterogeneity on the precedence of global processing. In my model the decision about whether to treat stimuli as a global or local involves interaction between proximity and similarity – it is possible to have heterogeneous global figures which are classified without difficulty at the low frequency because the local figures are densely packed.

5.5.3 Simple Visual Search

Visual search is a well-studied task in humans. Several theories have been advanced to account for the results found, some of which are described in Section §3.6.4. I am particularly interested in the contrast between two of these, Treisman’s (1998) feature integration theory (FIT) and the stimulus similarity theory (SST) of Duncan and Humphreys (1989).

FIT, in its original formulation at least, distinguished clearly between slow serial search and fast parallel “pop out”.

Duncan and Humphreys (1989) suggested that the slope of reaction time versus display size varies more than these two conditions suggest. Sometimes the search slope is steep, other times it is shallow and it seems to vary continuously, not in the binary fashion predicted by FIT.

SST allows for continuously varying search difficulty controlled by two interacting variables. The first is the similarity of the target to non-targets. Higher similarity slows search. The second is the similarity of non-targets to one another. Higher similarity here speeds search. Pop-out then becomes an extreme case in ideal circumstances and item-by-item serial search is an extreme case where distractors are heterogeneous and the target is similar to the distractors.

Duncan and Humphreys speculated that items are grouped for consideration during search and the search time therefore depends on the “groupiness” of the display. Distractors are easy to group with one another and discard if they are all similar. Targets are easy to group with distractors and therefore be missed if they are too similar.

Duncan and Humphreys propose a model different from mine. Where my model has a serial attentional component as well as a group classification mechanism, their model has no serial component: all attentional mechanisms operate in parallel. Their model has a strong preattentional flavour while mine is attentional as well as preattentional.

However my model does fit neatly into the paradigm they created and I will use it to test my model. I used four parallel conditions to test search performance of my model, based on those used by Duncan and Humphreys (1989). The conditions were:
Figure 5.7: Illustrations of the model’s global precedence. In cases where there is a salient low-frequency region that will be attended first at the low frequency. Thick borders around a region indicate attention to the low spatial frequency, thin borders attention to the high spatial frequency. (a) The system’s performance on a homogeneous low-frequency shape. At \( t_1 \) the shape is presented. At \( t_2 \) the shape is recognised as a salient region and classified at the low spatial frequency as an ARM. At \( t_3 \) attention remains on the same region but at the high spatial frequency. Now the classifier reports ELL – an occurrence of group classification. (b) The system’s performance on a heterogeneous low-frequency shape. The interpretation is similar to (a) but this time group classification proves impossible.
Figure 5.8: Example displays for each of the cases in the search experiments. Similarity is arbitrarily defined in the saliency analysis phase. Els are similar to squares, crosses to arrows, arrows to arms, arrows to triangles and triangles to arms.

- target similar to distractors, distractors heterogeneous (t-d similar, d-d different)
- target similar to distractors, distractors identical to each other (t-d similar, d-d identical)
- target dissimilar to distractors, distractors heterogeneous (t-d different, d-d different)
- target dissimilar to distractors, distractors identical to each other (t-d different, d-d identical)

The four conditions represent variation along two dimensions, target-distractor similarity and distractor-distractor similarity.

The system was required to find a single target in displays of varying size (the target is always present). The displays are carefully set up to produce the desired condition, while controlling for other variables, similar to the strategy used by Duncan and Humphreys. My displays are less controlled than those used with humans because some of the variables that must be controlled in human subjects need not be in my model. I do not have to guard against inter-trial priming, for example. Figure 5.8 shows example displays for the different conditions.

Each trial display was passed to the system, which was configured for search as follows. The attention system directs the sequence of operations on the proto-objects. At each step attention is directed at the selected object. Classification at the high spatial frequency is attempted regardless of the attended frequency. (This is only to save time. When the low frequency is salient an extra
Figure 5.9: An example sequence of operations during simple search. At $t_1$ the input is presented and at subsequent time steps attention is directed as shown until the target (ell) is found. Thick borders around a region indicate attention to the low spatial frequency, thin borders attention to the high spatial frequency. In the visual search configuration, only high-frequency targets were considered. But low-frequency shapes had to be attended before their elements, conforming to Navon’s (1977) global precedence principle. Note that several distractors are classified at once at the high frequency at times $t_3$ and $t_5$. 
Figure 5.10: Search results. For each of the four conditions the mean number of steps to find the target is plotted against the number of distractor items present. The regression line for each condition is also plotted.

step is added for the presumed classification attempt at the low frequency.) If the result is a positive classification and this corresponds to the search target, search is complete and the attention system halts. Otherwise the object is inhibited and selection begins again. If the system runs out of candidates for selection then all inhibition is removed, the low-frequency saliency map is entirely suppressed and selection begins anew (this time selecting only high-frequency objects). In the unlikely case that the target is still not found, the system halts with a failed search result.

Search displays were all based on the same spatial configuration. Twenty items were arranged in five groups of four each in the locations shown in Figure 5.8.

Search trials were arranged into blocks. In each block a basic distractor arrangement was created. One group was chosen randomly and the target placed at the left side of it. Next fifteen search trials were run. First the original display was run. Then one item was removed randomly with the constraints that at least one item remained in each group (the leftmost item) and the items in each group remained contiguous. Now this case was run and the process repeated until only one item remained in each group.

Each block was organised to provide the conditions above. For example, in the target-distractor similar, distractor-distractor different condition no adjacent distractors are similar to one another, but each distractor that is adjacent to the target is similar to it.

For each trial the number of selection steps needed to locate the target was recorded. Figure 5.9 shows the steps taken by the system during one trial. These results (interpreted as a reaction time) were then plotted against display size to allow for comparisons.

The results are summarised in Figure 5.10 and Table 5.1. The key result is that the different conditions provide different slopes. There is no dichotomy between flat pop-out search (indeed my simulation does not model pop-out at all) and slow serial search. Instead, there are a range of gradients for different search conditions. My simulation reproduces Duncan and Humphreys’ experimental results
Table 5.1: Results of simple search. For each condition a linear regression was performed with the results in (a). A two-tailed \( t \)-test was performed to test significance of the regression for each slope (null hypothesis that slopes is zero) and the \( t \) value and corresponding \( p \) value are shown for each line. The slopes of each of the lines were tested to determine if they were the same (null hypothesis that slopes are equal) and the results of those significance tests are shown in (b) with the number of degrees of freedom, the \( t \) value and the corresponding two-tailed \( p \) value.

(a)

<table>
<thead>
<tr>
<th>condition</th>
<th>slope</th>
<th>intercept</th>
<th>( r^2 )</th>
<th>( n )</th>
<th>( t )</th>
<th>( p &lt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-d similar, d-d identical</td>
<td>0.99</td>
<td>-0.87</td>
<td>0.33</td>
<td>2400</td>
<td>34.33</td>
<td>3.2 ( \times 10^{-12} )</td>
</tr>
<tr>
<td>t-d similar, d-d different</td>
<td>0.87</td>
<td>0.14</td>
<td>0.25</td>
<td>4480</td>
<td>38.25</td>
<td>3.2 ( \times 10^{-12} )</td>
</tr>
<tr>
<td>t-d different, d-d identical</td>
<td>0.26</td>
<td>2.46</td>
<td>0.09</td>
<td>4800</td>
<td>22.28</td>
<td>3.2 ( \times 10^{-12} )</td>
</tr>
<tr>
<td>t-d different, d-d different</td>
<td>0.62</td>
<td>0.18</td>
<td>0.27</td>
<td>3840</td>
<td>37.21</td>
<td>3.2 ( \times 10^{-12} )</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>conditions</th>
<th>d.f.</th>
<th>( t )</th>
<th>( p &lt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-d similar, d-d identical v. t-d different, d-d identical</td>
<td>7196</td>
<td>24.21</td>
<td>1.1 ( \times 10^{-16} )</td>
</tr>
<tr>
<td>t-d similar, d-d different v. t-d similar, d-d identical</td>
<td>6876</td>
<td>2.89</td>
<td>0.0039</td>
</tr>
<tr>
<td>t-d similar, d-d different v. t-d different, d-d identical</td>
<td>9276</td>
<td>21.56</td>
<td>0</td>
</tr>
<tr>
<td>t-d different, d-d different v. t-d similar, d-d identical</td>
<td>6236</td>
<td>10.12</td>
<td>2.2 ( \times 10^{-16} )</td>
</tr>
<tr>
<td>t-d different, d-d different v. t-d similar, d-d different</td>
<td>8316</td>
<td>7.45</td>
<td>1 ( \times 10^{-13} )</td>
</tr>
<tr>
<td>t-d different, d-d different v. t-d different, d-d identical</td>
<td>8636</td>
<td>16.29</td>
<td>0</td>
</tr>
</tbody>
</table>

in the following cases. In the ideal case where targets are dissimilar to distractors but distractors are similar to one another the search slope is close to flat. In the cases where targets are similar to distractors the slopes are highest. I obtain these results even though my model differs from that of Duncan and Humphreys. There are some results which I do not reproduce. In their experiments they found no effect of distractor-distractor similarity when targets are very different from distractors and only a small effect of target-distractor similarity when distractors are all very similar. I found some effect of distractor-distractor similarity when targets are very different from distractors. I also found a large effect of target-distractor similarity when distractors are all identical.

In summary, my results are broadly consistent with those found by Duncan and Humphreys (1989), the main difference being the magnitude of the effect in the t-d similar, d-d identical case, and the moderate effect I found in the t-d different, d-d different case. These differences quite possibly arise from the differences in the stimuli used. In their experiments (see Section §3.6.4.5 for a precis) they found that results varied quite considerably just with the size of the stimuli used. Duncan and Humphreys' results are particular to their stimuli and indeed one of the points of their experiments was to show that search is a complicated and multifactorial phenomenon.

5.6 Selection Mechanism Refined

As it stands, when the selection mechanism visits a heterogeneous group which cannot be group classified the group is assumed not to contain the target and the next salient region is visited. If all salient regions are visited without finding the target, every item is revisited in turn at the high spatial frequency. This is an expensive policy if the target ever appears in a heterogeneous group of items. I think it makes more sense to examine the contents of unclassifiable groups before moving on.

With this in mind I introduce for the final experiment a new kind of attention shift, a change in the scope of the attention system itself. If attention is directed to a heterogeneous group of objects,
the system has the choice of establishing the currently attended region as a saliency map in its own right, at a higher spatial frequency, so that attention can be directed serially to individual stimuli one by one.

This allows the system to “zoom in” on a region that it is interested in for closer analysis. To do this the scope of the other two kinds of attention switching are limited to the currently attended stimulus. For example, the system may have come across a group of heterogeneous items. While the group is still selected, it can adopt the group as the new scope and then attend to the items that make up the group one by one. When the group is exhausted the scope returns to the entire visual field. This introduces a kind of search within a search.

With recursion added the selection mechanism has a new option in the case where a low-frequency object is currently attended but no positive classification is available. In this situation the system begins a new inhibition map, suppressing the old one. The new inhibition map initially inhibits all but the region corresponding to the previously attended object. Low frequency saliency is entirely suppressed and the system continues with selection. If the system runs out of objects then low frequency saliency is reactivated, the system switches to the original inhibition map, the original (low frequency) object is inhibited and selection continues. The effect of this is to make the system rely on group classification if it can, but if it cannot to take a closer look at the group before moving on. It is still possible if a group is misclassified that the system will have to visit every item individually to find the target.

### 5.6.1 Performance of the Refined Selection Mechanism

I repeated the search task with the refined model. The model was the same as before except that now, whenever the system encountered a low frequency region that cannot be classified at the high frequency the region is established as the new scope (one extra step), the high-frequency contents of the region are attended serially and the scope is returned to the whole display (one more extra step).

The results are summarised in Figure 5.11 and Table 5.2. The t-d different conditions have not changed, while the t-d similar conditions both have shallower slopes (and their slopes are no longer significantly different). There remain three significantly different search slopes, however.
5.7 Predictions of the Model and Further Extensions

Beginning with the assumption that human visual classification is cardinality blind I constructed an attention model to take advantage of this.

To cater for a cardinality-blind classifier, attention must group homogeneous items together for the purposes of classification when it can. This is what the attention component of my model does. It combines saliency analysis based on proximity and homogeneity with three kinds of attentional shift that allow the gains of group classification to be achieved where possible while falling back on conventional serial processing in other situations. These three shifts are: shifting the attended region, shifting the classification scale (Section §5.4) and shifting the attentional scope to zoom in on the components of a composite object (Section §5.6). The resulting model is consistent with experimental results about Gestalt properties, global precedence and visual search.

Along with the comparisons I have already made, my model makes some other testable predictions. The easiest test, assuming that I am modelling overt attention, is that saccades should follow the pattern I have established. The model of Duncan and Humphreys (1989) is a parallel one and does not require attention or eye movements to work. It may, therefore, be possible to distinguish my model from that of Duncan and Humphreys with an experiment that combines a visual search task with eye tracking and asks whether people do move their eyes during search. A parallel theory would say no, a serial theory (like mine) would say yes. If people do move their eyes then the way they move their eyes can test a further prediction of my model. I predict that eyes would move among homogeneous groups, which are each processed in parallel.

There are strong similarities between my approach to this problem and the group scanning model introduced by Treisman and Gormican (1988) (see Section §3.6.4.1 for detail on approach). MacQuistan (1994) even examined the effect of t-d similarity on group scanning. The hypothesis was that when t-d similarity is high groups will be small and so search will be slowed. Search was indeed slowed in this case, but I hypothesise instead that this is because groups are large in this case and likely to
contain the target, requiring a lot of serial search inside the groups. Unfortunately it is not obvious how these explanations can be picked apart experimentally in this particular case.

In both Treisman and Gormican (1988) and Duncan and Humphreys (1989) pop-out is just a special case of the more general search mechanism. I have not modelled pop-out, but I believe that pop-out should be modelled by the function that delivers the saliency representation, rather than as something to be modelled by group classification. In other words, I conceive attentive (serial) and preattentive (parallel, pop-out) search being performed by distinct mechanisms sharing the saliency representation. In my model apparent parallel effects during serial search are accounted for by group classification, not feature discrimination. This may provide a point of difference from the group scanning theory if it can be shown that the constraints on pop-out (feature discriminability) are different from those on group scanning (stimulus similarity), thus indicating different systems are involved.

The model can still be improved in a number of ways. Simulating more spatial frequencies with more realistic bandwidths (1-2 octaves according to Hughes et al., 1996), more grouping factors and pop-out would all make the model more robust and realistic. An overarching goal, drawing on these improvements, would be the handling of naturalistic images.

My interest in developing this model is, of course, for its potential linguistic payoff. With the visual foundation established I am ready to propose a simple model connecting visual cognition and with noun phrase syntax. That is the subject of the next chapter.
Chapter 6

Perceptual Correlates in the Noun Phrase

The perceptual model described in Chapters 4 and 5 is plausible on its own as an explanation for certain perceptual results. But it also provides an explanation for a linguistic phenomenon, that of number, and its place in the noun phrase. Number in this sense refers to the marking of nouns as singular or plural (in English; in some other languages, as described below, there are finer distinctions as well).

I begin in Section §6.1 by describing the syntactic marking of number in noun phrases in more detail. Then I will present the linguistic framework showing the distinction between reference, number and type in the noun phrase in Sections §6.2 and §6.4. In Section §6.6 I lay out a scheme that takes information in the model and produces the distinction between singular and plural for an attended region. In Section §6.7 I show that not only does the separation of perceptual information correspond to the separation of syntactic entities in the noun phrase but that the sequence in which they emerge is also consistent with noun phrase structure. In Section §6.8 I address two points for further work and in Section §6.9 I review what I have found.

6.1 The Syntactic Marking of Number in Noun Phrases

Number is signalled in a few places in language – obviously in the case of words like “one”, “two”, “three”, and more complicated phrases like “four million, two hundred and six”. It also comes up in quantifiers like “some”, “lots” and “many”. But the manifestation that I am interested in is the syntactic number feature that attaches itself to nouns. In English, for example, there is a difference between singular and plural noun phrases: “the dog” versus “the dogs”. Furthermore, the subject of a sentence and its verb must agree in the number feature: “The dog runs away” versus “⋆The dogs runs away”.

This distinction suggests that there are at least two separate components to nouns: the noun itself indicating a class, and the number feature indicating whether one or more members of the class are intended. In an introductory text on the subject, Haspelmath (2002) notes that the regularity of number marking is best represented by hypothesising the existence of a rule for producing fully inflected nouns by combining an uninflected noun stem (e.g. “dog”) with a morpheme indicating number (e.g. “-s” for plural or the null morpheme for singular). Also, the linking of the number feature with verbs also suggests it has a life independent of its noun, as it were.

Other languages exhibit even more variations of the number feature. For instance Māori includes personal pronouns that refer to exactly two people (dual number) or to three or more people (plural), as shown in Table 6.1. According to Corbett (2000) the Larike language includes marking for exactly three things (trial number) and other languages may possibly, though controversially, include marking

\[\text{An example marked by a star indicates ungrammaticality. An example marked by a question mark indicates dubious grammaticality.}\]
Table 6.1: Māori pronouns showing singular, dual and plural number marking (Biggs, 1973).

<table>
<thead>
<tr>
<th>Number</th>
<th>Singular</th>
<th>Second person</th>
<th>Third person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singular</td>
<td>au, ahau I, me</td>
<td>koe you</td>
<td>ia he/she</td>
</tr>
<tr>
<td>Dual</td>
<td>tāua you and I</td>
<td>kōrua you two</td>
<td>rāua they two</td>
</tr>
<tr>
<td>Plural</td>
<td>tātou we all (and you)</td>
<td>koutou you all</td>
<td>rātou they all</td>
</tr>
<tr>
<td></td>
<td>mātou we all (but not you)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

for three things, four things and so on (Rijkhoff, 2002). But across languages there are some generalisations that can be made which allow English to serve as a reasonable model for language generally. In those languages which do divide noun phrases up by number, the distinction between singular and plural seems to be most basic (Greenberg, 1963). The underlying cause of the singular versus plural distinction is, perhaps, to be found in the components of perception I have modelled.

6.2 The DP Hypothesis

6.2.1 A Recap of X-bar theory

In section §2.2.1.5 I described X-bar theory, which proposes that heads project phrases with the following schema.

```
XP
   Spec
     X'
       X Comp
```

X stands for the head, a lexical item, XP for the maximal projection of X, and the Spec and Comp positions are filled by other XPs. The constraints that determine which XPs can appear in the Spec and Comp positions are part of the information carried by each head.

At the end of Section §2.2.1.5 I motivated the following structures for noun phrases and clauses.

```
NP
  D
    “the”
  N
    “destruction”
  PP
    “of the window”

VP
  NP
    “The ball”
  V
    “smashed”
  NP
    “a window”
```

6.2.2 Introducing DP

There is an important inconsistency with the X-bar treatment of noun phrases just described. I stated that the specifier position in an XP is to be filled by another XP, however in the example NP given the specifier is filled by a determiner, “the”, a lexical item rather than a phrase. This inconsistency can be removed by adopting the DP (determiner phrase) hypothesis. This was proposed by Abney
(1987) for a number of reasons, the chief of which was to bring the out similarities between noun phrases and sentences.

Prior to Abney (1987) the structure of the noun phrase was roughly as follows:

```
NP
  D  N'
  "the"  N
  "cats"
```

Abney revised this by proposing that D heads its own phrase, the DP or *determiner phrase*. This creates a deeper structure:

```
DP
  D'
  D  NP
  "the"  N'
  "cats"
```

Making noun phrases conform rigorously to X-bar theory was not the only factor behind the DP hypothesis. One piece of evidence (borrowed from Carnie, 2007) concerns the "s"-genitive.

(1) the window we broke’s glass

The possessive marker "s" is not attached to any particular word (or category of word) but rather seems associated with an entire phrase as if it were a little word in its own right.

(2) [the window we broke]’s glass

And in addition it seems to behave like a determiner – we can’t have another determiner in there, for instance.

(3) * the window we broke’s the glass

But taking this into account means we run into problems with the old noun phrase structure.
There is nowhere in the structure to the left of the determiner to attach the possessor phrase. However by adopting the DP hypothesis, “’s” becomes the head of the DP, and the possessor (itself a DP) can fill the specifier position of the DP.

Another argument for the DP comes from the semantic importance of the job determiners do. To see this we only need look at quantified sentences and their semantic interpretations.

(4) Many players hid.

In a compositional model of semantics the meaning of a determiner like “many” is a structure which specifies the outer form of the logical formulation which captures the meaning of the sentence, just leaving “slots” for the predicates. Thus in Montague semantics (Dowty et al., 1981) the meaning of “many” is the following lambda expression:

$$\lambda P \lambda Q\text{many}(x, P(x), Q(x))$$

This semantic contribution suggests that the determiner position is not a specifier but a head. The semantic contribution of the determiner seems consonant with that provided by other functional heads in the clause (Alexiadou et al., 2007).

Abney (1987) and Alexiadou et al. (2007) contain further justifications, theoretical and empirical, for the DP hypothesis. Bernstein (2001) is a useful review of newer empirical justification for the hypothesis.

### 6.3 D Contributes a Referent, N Contributes a Type

It is generally agreed, and has been since Abney (1987) proposed the DP hypothesis that determiners contribute referents semantically. This is even commonly believed in other linguistic schools. For instance Rijkhoff (2002) claims that the semantic correlate of the determiner is location. The referent is the thing that the DP is “talking about”. Nouns do not contribute referents, they only contribute types. We can see this by considering the difference between the noun phrases in the following sentences.

(1) A dog barked.
(2) The dog barked.

Both are talking about something, and that something is a dog. The determiner tells us how to “produce” the referent of which “dog” is predicated. In the first sentence, the indefinite determiner “a” tells us that the referent is brand new: we have to create it. In the second sentence, the definite determiner “the” tells us that the referent is presupposed: we can find a unique dog in the recent context. To summarise: D contributes a referent; N contributes the type which predicates of it.

6.4 NumP

What contributes singular or plural number? Is it D? Is it N? Or is it another syntactic projection? Zamparelli (1995) presents evidence for additional layers in the DP, so that there are heads for type (the noun, projecting NP), number (the so-called weak determiners, projecting what I will call NumP) and reference (the so-called strong determiners, projecting DP).

As evidence that there is a third layer, consider these sentences.

(1) Mount Taranaki is Mount Egmont.
(2) Mount Taranaki is a mountain.

The first expresses a simple identity between two objects. The second predicates the mountainhood of an object. Note that “a mountain” in Example 2 is not referential. But it still has a determiner. Also, it still has number, which we can see clearly by contrasting it with the following example.

(3) Taranaki and Cook are sm mountains.²

The structures for the two kinds of sentences seem to be parallel, but consider the following sentence:

(4) ? Mount Taranaki is Mount Egmont and a mountain.

It is of (at best) dubious grammaticality. According to Zamparelli (1995) this is because the two phrases being coordinated, “Mount Egmont” and “a mountain”, are not of the same category. “Mount Egmont” is a full DP while “a mountain” is a different kind of XP.

Because of the inclusion of number information this projection is often called NumP and this is the name I give it in the remainder of the text.

We now have the tripartite syntactic division of labour shown in Figure 6.1.

Note that D contributes reference, Num contributes number and N contributes type.

So we find now that the structure of our example “noun phrase” is really something like the following.

²Here “sm” stands for unstressed “some”.

105
Figure 6.1: The final structure of the DP. Note the three projections contributed by the determiner (D), number (Num) and noun (N).
At first there may seem to be a difficulty with the morpheme indicating number (“-s”) preceding the noun in the tree, when in English it comes afterwards, attached to the noun. This is accounted for by head movement (see section §2.2.1.6). The head of the NP (“cat”) moves and merges with the head of NumP (“-s”) to form the complex head “cats”.

The existence of NumP has also been argued for by Ritter (1991, 1995), which is the definitive argument for NumP, but is based on Hebrew examples and thus harder to demonstrate for a general audience. For a useful summary of that argument see Bernstein (2001). Rijkhoff (2002) makes an equivalent argument for separation using functional grammar rather than X-bar theory.

### 6.5 Computing Reference and Type With the Visual Model

Having established that the referent, number and type of the noun phrase are separate entities, I want to show that they are in my visual model, too, by showing how my model produces the corresponding information. In this section I will show how my model produces the reference and type information. The separation of these and their association with noun phrases is not so new, being the point made by Hurford (2003b).

Just as Hurford proposes, in my model the type is produced by the classifier, presumably in the ventral (“what”) stream, while the referent is produced by the dorsal (“where/how”) stream (see Section §3.3.4 for more on this distinction). They are distinct pieces of information produced in distinct places, just as the noun and determiner are distinct in the noun phrase.

Once the system attends to a region and successfully classifies it, the class is output directly from the classifier. This is the type. The referent is provided by whatever object is currently attended in the saliency representation. These correspond directly to the noun and the determiner respectively in the syntax of the noun phrase.

### 6.6 Computing Cardinality with the Visual Model

Now I need to show how the cardinality, which corresponds to the number, can be found.

Once the system attends to a region and successfully classifies it, the selection mechanism can determine the cardinality (singular or plural). This emerges quite directly from information about
spatial scale which the attentional system computes. To explain how, let me refer back to the two
terms defined in Section §5.4. The classification scale is the scale at which classification has just been
performed. The proto-object scale is the original scale at which the currently attended proto-object was
attended, before any changes in scale attention took place. At any point after a successful classification
the selection mechanism can produce cardinality information by comparing the classification scale to
the proto-object scale. If they are the same, singular number is generated. If they are different then
plural number is generated. That is all there is to number computation. This piece of information

corresponds directly to the Num of the noun phrase syntax.

Figure 6.2 illustrates number computation via this interpretation of the model.

6.7 Sequencing of Visual Operations and Syntactic Hierarchy

Note that the visual operations involved in classifying an object must occur in a particular sequence
in my model. The first operation is the selection of a proto-object; that is, of a referent. The next
operation is the selection of a classification scale (The classification scale is a function of the proto-
object scale, so it can only be determined after the proto-object is selected). Finally, the proto-object
can be classified and a category can be evoked.

It is interesting that the sequential order of these operations mirrors the right-branching hierar-
chical structure of the DP. The projection which dominates is DP, which contributes a referent. DP
introduces NumP, which contributes singular/plural number. NumP introduces NP, which contributes
an object type.

This may, of course, be a coincidence. But Knott (2009) argues that LF structures (the strictly-
right branching structures closest to semantic representations in the Chomskyan paradigm) can in
fact be interpreted as descriptions of sequences of sensorimotor operations, at least for sentences
that describe concrete objects and events. Exactly how sensorimotor sequences might be employed
linguistically — for both production and comprehension — is beyond the scope of this thesis, but is
not neglected by Knott and remains an interesting issue for further work.

6.8 Two Loose Ends

6.8.1 Other Numbers

In addition to singular and plural number, some languages have dual or even trial number for referring
to two or three things, respectively (According to Rijkhoff, 2002, there are some controversial claims
for even higher numbers). My model does not produce anything corresponding to these numbers. I
can think of three possibilities for additional numbers. There may be more.

First, it is possible that they could be incorporated into the scale-based scheme devised above for
computing plurality. The current system relies on disparity between the scales at which things are
salient and classifiable. The singular-plural distinction is a very coarse one, relying on any disparity.
If the system were more sensitive to the disparity (and perhaps the shape of a group) then judgements
about two or three objects in a group could be made. For example, two similar items side by side have
a proto-object scale approximately twice that of their classification scale (at least in one direction,

to be determined after the proto-object is selected). Finally, the proto-object
can be classified and a category can be evoked.

This idea also nicely coincides with a known cross-linguistic universal (Greenberg, 1963) that
languages with dual number also have plural number and those with trial also have dual. Number
computation based on sensitivity to spatial frequency also has this property. Determining that a group
contains two components is a more sensitive operation than determining that it contains more than
one, so whenever dual cardinality is available, plural cardinality must also be available. A system
relying on this approach would be limited to low numbers, as the linguistic evidence suggests, because
the more items present the harder it becomes to distinguish exactly between \(n\) items and \(n+1\) items.
Figure 6.2: Attention examples annotated with both the type and the number computed as described in the text. The referent corresponds to the selected proto-object in each case. Thick borders around a region indicate attention to the low spatial frequency, thin borders attention to the high spatial frequency. (a) is the input presented to the system. In (b) a new region with low-frequency proto-object scale is selected and classified first at the low-frequency classification scale ($t_1$) and then at high-frequency classification scale ($t_2$). At each step the classification scale is compared with the proto-object scale: if they are the same singular cardinality is output, otherwise plural cardinality is output. In (c) the same thing happens but this time the high frequency is the proto-object scale.
A second approach to fine number distinctions is to suggest that some other number system is in play. Humans are able to “count” items in parallel up to around four, a phenomenon called subitization (Nieder and Miller, 2004). If subitization is not just a more sensitive version of the approach I have proposed (which does not seem very parsimonious, admittedly) then maybe subitization is responsible for providing the additional numbers.

Finally, it is possible that dual, trial and so forth are manifestations of the same kind of numerical information that lies behind cardinal numerals like two and three, information which may come from some innate subitization mechanism or counting but which is qualitatively different from the one-many distinction of the singular versus plural.

6.8.2 Other Modalities

I have examined only vision as a basis for noun phrase structure. That congenitally blind people have no linguistic problems may seem at first to undermine my argument but I do not think so.

First, the relevance of vision depends on the precise nature of the proposed link between visual attention and noun phrase structure. If this link represents evolution to fit semantic constraints posed by the visual system (the second type of language evolution model in Section §2.9.2) then there is probably no difficulty. In general humans depend on vision the most and so vision will have a great impact on the kinds of linguistic rules that proliferate. Since the blind are exposed to just as much language and have the same general learning biases a link of this type provides no difficulty. It does raise the question of what would happen to language if the dominance of vision were removed, though. Though in the land of the blind the one-eyed man may be king it’s quite possible the king is predisposed to learn a different kind of language from his subjects.

If the linguistic machinery shares neural structures with the perceptual machinery or is homologous to the perceptual machinery (the first type of model from Section §2.9.2) then there may be some explaining to do. If the link is the result of language relying on homologous neural structures to those used by vision then, again, there is no problem as that region is not necessarily tied to current visual function. If language and visual perception actually overlap neurally, though, then the kind of link I have proposed would only be feasible if it relied on cross-modal rather than vision-specific processing. This is not incredible, as Section §3.6.3 shows.

I wouldn’t want to speculate too far in this area without more careful consideration but I don’t believe that modelling one particular modality (especially vision) makes my claims any less plausible.

6.9 Conclusion

In this chapter I have reviewed some syntactic arguments suggesting that the determiner, noun and number marking are contributed by separate syntactic projections within the noun phrase. I have presented a simple method by which my model can compute reference (which corresponds to the determiner), cardinality (which corresponds to linguistic number) and type (which corresponds to the noun).

I have also shown that, in the influential X-bar theoretical approach to syntax, noun phrase structure ordering (determiner, number, noun) suggests a sequence of operations (reference/location, cardinality, class) which corresponds neatly with the order of operations of my model.

Finally I discussed a way in which my model could plausibly be extended to account for other kinds of number (such as dual and trial) and perhaps even subitization. This would be a very worthwhile extension to pursue.

I conclude that the perceptual model of object classification and attention described in Chapters 4 and 5 provides a plausible explanation for several aspects of the basic syntax of noun phrases.
Chapter 7

Summary and Conclusions

I set out two goals in the introduction to this thesis. The broad goal was to argue for the hypothesis that studying sensorimotor processing can help the study of language. The narrower goal was to consider a specific hypothesis of this kind, namely that the mechanisms of object classification and attention are encoded in the syntax of noun phrases.

The bulk of the thesis, Chapters 3 to 6, is concerned with this second goal and I will discuss it first before moving on to my broader conclusions.

In Chapter 3 I reviewed the literature relating to object perception and attention. There I showed that there is a well-known separation between visual classification and attention in the brain, the ventral (via inferotemporal cortex, IT) and dorsal (via parietal cortex) pathways, respectively. I found that the idea of a pathway dedicated to determining the type and another dedicated to determining location and objecthood fitted well with the separation between the noun (type) and determiner (reference) in the noun phrase, a distinction already made by Hurford (2003b).

With an eye to developing my own combined model of visual classification and attention I examined existing models of both.

I found that an important feature of any successful classifier is its invariance to changes in certain variables. For human visual classification these appear to include translation (or location) invariance and scale invariance. These invariances are usually empirically based or argued from purely visual precepts. But they also make sense if one thinks of classification as the mechanism behind the noun, because nouns don't carry location or scale information either. I also presented some evidence for cardinality invariance in primate IT, invariance to the number of items of a single type presented simultaneously. This is information that is represented elsewhere in the noun phrase and again is not carried by nouns and so is consistent with the idea that nouns represent the output of the visual classifier.

From among a range of computational models of visual classification I highlighted the one I would go on to develop as part of my own model, Mozer and Sitton’s (1998) convolutional neural network, chosen because of its general structural and computational similarity to IT.

I also reviewed many models of attention and concentrated on those where the representation of attention is not distributed, finally choosing that of Itti and Koch (2000) and Walther and Koch (2006) to develop as part of my own model because of its maturity and proven record.

In Chapter 4 I developed the classifier model, adapting the original structure of Mozer and Sitton to cope with images instead of abstract feature maps and to handle two different spatial frequencies. Thus I designed my CNN to be scale invariant, and its structure ensures a strong tendency to translation invariance. As most of Chapter 4 is devoted to proving, the CNN is also cardinality blind. I claim, based on this model and the evidence presented in Chapter 3 that primate IT is also cardinality blind.

In Chapter 5 I went on to develop a computational model of attention, taking into account the cardinality blindness of the classifier which allows group classification, the simultaneous classification of groups of homogeneous items. The computational model I presented invokes the Gestalt principle of grouping by similarity to ensure that homogeneous groups are attended as wholes when possible. I also modelled the Gestalt principle of grouping by proximity and showed that the conflict between
these in the model is consistent with that observed in humans. The model also naturally prefers global figures over their component local figures and so demonstrates global precedence, another phenomenon already observed in humans. In my model global precedence is a characteristic of the system’s scale-based attention, a companion to the more conventional location-based attention. Finally, I showed that group classification in conjunction with serial attention leads to a variety of reaction time versus display size functions in visual search. Group classification allows serial search to proceed faster than a naive implementation, sometimes much faster. My model is consistent with the spirit and results of Duncan and Humphreys’s (1989) stimulus similarity theory (SST), though my model differs from the one that they propose. My model also shows some superficial similarity to the group scanning theory proposed as a component of Treisman and Gelade’s (1980) evolving feature integration theory. Group scanning was proposed to model the same kinds of discrepancies in the data (a continuum of search functions rather than the two extremes of pure serial and parallel search) as with SST and my own model, but achieves it through somewhat different mechanisms. Where my model groups things that are similar and so can be easily processed together, group scanning groups things which need extra processing to be distinguished. These results encourage me to claim that the perceptual model developed in this thesis is a plausible one. On purely perceptual grounds it is a reasonable model of human performance. I also noted in this chapter a few future directions for development and improvement of the model.

Armed with this model, in Chapter 6 I presented an idea about the interface from visual/attentional mechanisms to natural language representations: specifically, to the syntactic structure of the noun phrase. Given the representations which are made available in my visual model, the interface can be stated quite simply. The output of the attention mechanism (reference) corresponds to the determiner. Another output of the attention mechanism (proto-object scale in relation to classification scale) corresponds to the number projection. And the output of the classifier (type) corresponds to the noun. Lastly I showed how the right-branching hierarchical structure of the noun phrase, interpreted as a trace of temporal events as proposed by Knott (2009), is consistent with the order in which my model makes the corresponding information available.

My conclusions from this aspect of the thesis are twofold. First: cardinality invariance is probably a property of IT and can usefully be employed to help explain a range of behavioural results in the field of visual attention. My model embodies just such an explanation. Second: with a cardinality blindness inspired model it is possible to produce a very plausible and neat mapping between three components of the noun phrase (determiners, number and nouns) and aspects of visual processing (spatial/object attention, scale attention and classification).

This brings me to the broader conclusion. In Chapter 2 I surveyed what is known about the relationship between language and sensorimotor cognition. I found a perplexing variety of theories, but also reached an important understanding of how sensorimotor cognition could affect linguistic structure. I found that, rather than being surprised that there could be such links, we should expect them. Linguistic structure very likely does reflect sensorimotor processing (the converse is also true). Sensorimotor processing is one of the constraints on the evolution of language and so language will reflect the peculiarities of sensorimotor processing.

This is already acknowledged for peripheral aspects such as colour names. I have sought to show, and believe I have shown in this thesis, that sensorimotor cognition is a fruitful source of inspiration for explanations of deeper linguistic knowledge, too. We interact with the world through the sensorimotor system and it seems wasteful to ignore this in theories of how we talk about the world.

My conclusion, then, is that cardinality blindness is useful for understanding visual object cognition, that visual object cognition, based on cardinality blindness, is reflected very directly in the structure of the noun phrase, and that sensorimotor cognition in general is potentially a very fruitful source of linguistic explanations.
References


122


