Sensorimotor cognition and natural language syntax[1]

Alistair Knott, Dept of Computer Science, University of Otago
alik@cs.otago.ac.nz

March 1, 2010

[1]This manuscript is still in draft form—comments welcome!

Abstract

This book is about the interface between natural language and the sensorimotor system. It is obvious that there *is* an interface between language and sensorimotor cognition, because we can talk about what we see and do. The main proposal in the book is that the interface is more direct than is commonly assumed. To argue for this proposal I focus on a simple concrete episode—a man grabbing a cup—which can be reported in a simple transitive sentence (e.g. the English sentence *The man grabbed a cup*). In the first part of the book I present a detailed model of the sensorimotor processes involved in experiencing this episode, both as the agent bringing it about and as an observer watching it happen. The model draws on a large body of research in neuroscience and psychology. I also present a model of the syntactic structure of the associated transitive sentence, developed within the entirely separate discipline of theoretical linguistics. This latter model is a version of Chomsky's 'Minimalist' syntactic theory, which assumes that a sentence reporting the episode has the same underlying syntactic structure (called 'logical form') regardless of which language it is in. My main proposal is that these two independently motivated models are in fact closely linked: specifically, that the logical form of the sentence reporting the cup-grabbing episode can be understood as a *description* of the sequence of sensorimotor processes involved in experiencing the episode. I argue that the logical form of the sentence can be given a detailed sensorimotor characterisation, and, more generally, that many of the syntactic principles which are understood in Minimalism as encoding innate linguistic knowledge are actually sensorimotor in origin. This reinterpretation of Chomskyan syntax places it squarely within mainstream cognitive science. In fact, I suggest it offers a way of reconciling Chomsykan syntax with the empiricist models of language which currently dominate the field. In particular, it permits the development of a model of language *processing* which is compatible with Minimalism (which in its original conception just a model of 'syntactic competence'). I conclude the first part of the book by presenting a neural network model of sentence production, whose basic recurrent architecture should be familiar to empiricist linguists, but which can also be understood by Minimalist linguists as a model of the mechanism which learns a mapping between the logical form of sentences and their surface form in a particular language. The network can learn the kind of syntactic parameter settings which Minimalists assume are responsible for the differences between languages; but it can also learn the statistically-defined surface patterns which play a prominent role in empiricist models of language, and which are problematic for traditional Minimalism.

In the second part of the book, I move beyond the simple cup-grabbing scenario, and extend the sensorimotor interpretation of Minimalist logical form to a number of other syntactic topics, including the internal syntactic structure of noun phrases, the noun-phrase/clause interface, predication, quantification and relative clauses.

I would like to thank several colleagues and students who contributed in several places to the work I describe in this book: Lubica Benuskova, Greg Caza, Mike Liddle, Michael MacAskill, Anthony Robins, Martin Takac, Joost van Oijen, Hayden Walles and Andrew Webb.

I would also like to thank the following people for helpful comments and discussions: Michael Arbib, Bob Berwick, Noam Chomsky, Liz Franz, Caroline Heycock, Giuseppe Longobardi, Liana Machado, Liz Pearce, Raffaella Rumiati, Tim Shallice, Mark Steedman, Martin Takac, Alessandro Treves and Jean-Roger Vergnaud. Needless to say, the book does not necessarily reflect their views, and any errors in the manuscript are mine.

Thank you Mele, Mia and Helen xxx

Contents

Chapter 1

Introduction

Human language is a powerful tool. Much of its power—and presumably much of its benefit to humans—comes from its ability to represent situations other than the perceptual here-and-now. We can use language to describe events which are distant in space or time, which are counterfactual, or imaginary, or desired. In fact, we can use language to talk about anything we like. The meaning of a linguistic utterance comes from its component words, and the way they are put together, rather than from the situation in which it is uttered. The mechanism which allows words to combine together to convey meanings is called **syntax**. Syntacticians develop models of how words can be combined into sentences, and of how the meanings of sentences are conveyed by the words of which they are made. We have learned a lot about these processes, but we are still far from devising a model which captures the subtlety and astonishing representational power of human language. How syntax works is still to a large extent an open, and controversial, question.

While it is remarkable that we can use language to express arbitrary meanings, it is in a way equally remarkable that we can use language to describe our immediate sensorimotor experience. We can talk about what we see and do, or about objects and events which are close at hand. When we do this, we are not making use of the full representational power of language. But presumably it is through associations with experience that linguistic expressions initially *acquire* much of their representational power. The remarkable thing about our ability to express sensorimotor experience in language is the degree of compression which is involved. Surface language is a simple medium: a sequential stream of words. When we describe a concrete event we witness, we convert a rich multimodal complex of sensory and motor stimuli evoked through contact with the world into a short expression in this simple medium. How this conversion is achieved is also very much an open question.

In this book, I will be pursuing both of the open questions just mentioned. My main

aim is to investigate the syntax of natural language. But I also want to investigate the interface between language and the sensorimotor system. This is a secondary interest: I want to study this interface because I think this might reveal something interesting about syntax. Of course, much can be learned about syntax by studying language as a self-contained mechanism. This is the way it is traditionally studied. But there are several recent traditions in cognitive science exploring the hunch that the interface between language and sensorimotor cognition has something interesting to tell us about syntax. The most venerable of these is 'cognitive linguistics' (see e.g. Lakoff and Johnson, 1980; Lakoff, 1987; Langacker, 1987; 2008). A more recent tradition is 'embodied cognitive science' (see e.g. Harnad, 1990; Brooks, 1991; Ballard *et al.*, 1997; Barsalou, 2008), which extends beyond language to consider the sensorimotor grounding of cognition generally. More recently still, the hunch has been pursued within neuroscience, in research into the neural substrates of language (see e.g. Rizzolatti and Arbib, 1998; Feldman and Narayanan, 2004; Arbib, 2005; Dominey *et al.*, 2006). The intuition being pursued in each case is that the interface between language and the sensorimotor system is more direct than is commonly assumed, and therefore that studying the sensorimotor system can yield insights into language. In this book I will argue in some detail for a particular version of this claim.

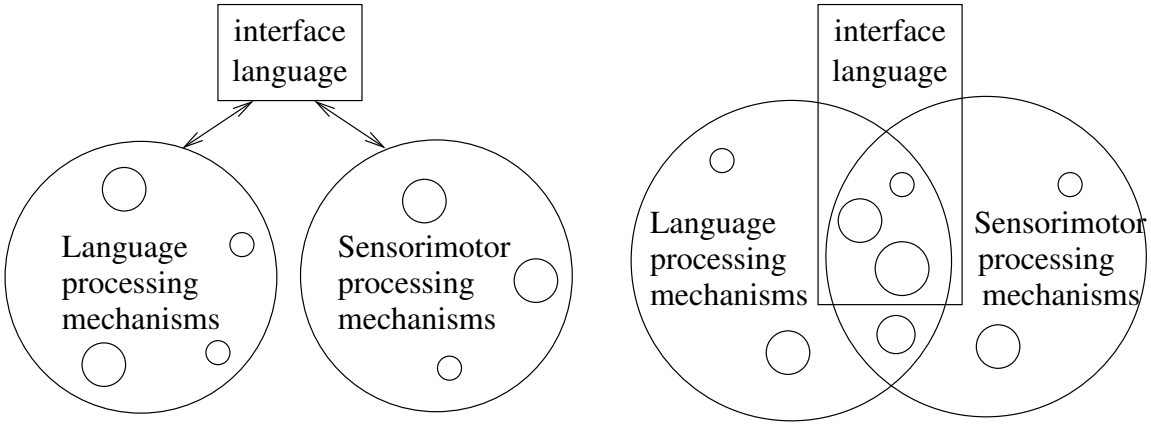
The claim I want to argue for, stated baldly, is that the syntactic structure of a sentence reporting a concrete event in the world can be understood, at least in part, as a *description of a sensorimotor process*—namely the process involved in experiencing the event. There are many caveats which I must add to this simple statement, but I basically want to make a very strong claim: that certain parts of a model of natural language syntax are also straightforwardly part of a model of sensorimotor cognition. And therefore that when we study syntax, we are unavoidably studying some aspects of the sensorimotor system as well.

In this book I want to make a particularly detailed, and hopefully falsifiable, version of the claim that there is a close relationship between language and sensorimotor system. But I will begin by introducing a very general form of the claim, and providing some arguments for it.

1.1 The shared mechanisms hypothesis

It is quite clear that the mechanism supporting natural language has to *interface* with sensorimotor processes, because otherwise it could get no input from the world, and could have no consequences on our behaviour. There are several ways this interface could work. One approach, explored by Fodor (1983), is that language and sensorimotor cognition are both **modular** mechanisms, which operate relatively autonomously, and communicate via

an interface language which abstracts away from the details of the mechanisms (see Figure 1.1(a)). An alternative approach is that language and sensorimotor cognition to some



(a) Modular mechanisms (b) Shared mechanisms

Figure 1.1: Two models of language and its relation to sensorimotor processing

extent actually involve the same mechanisms: see Figure 1.1(b). What this means is that when we generate or interpret a concrete sentence, at least some of the mechanisms which are involved are also mechanisms which operate when we perform actions or observe events in the world—and conversely, when we perform actions or make observations, some of the mechanisms we use are also involved in generating or interpreting sentences describing these actions or observations. In turn, this means that the 'interface language'—i.e. the language which allows transmission of information from linguistic to sensorimotor modalities—can make reference to these shared mechanisms. The shared mechanisms hypothesis which I will propose is in fact primarily a hypothesis about sentence representation, rather than sentence processing. Assume we have a simple sentence conveying some concrete proposition P: a proposition we can come to learn through sensorimotor experience. I will claim that the syntactic representation of this sentence—specifically, the representation which encodes its meaning, and makes this meaning available to other cognitive modalities—can be understood as a *description* of the sensorimotor processes which are involved in establishing P.

Many theoretical linguists adopt the modular mechanisms hypothesis, at least as a working method. Linguists argue about alternative formalisms for representing the semantics of sentences, and about alternative syntactic analyses from which to derive semantic

representations. But these arguments tend to be grounded in data about language, rather than about the suitability of semantic formalisms to describe sensorimotor processing. For linguists, a good syntactic theory is one which distinguishes successfully between well-formed and ill-formed sentences, and assigns plausible semantic representations to the well-formed ones. The working assumption is that data about sensorimotor mechanisms is not relevant to the task of choosing a good syntactic theory. In contrast, if the shared mechanisms hypothesis is true, then arguments about about the nature of sensorimotor cognition are likely to have a bearing on the choice between different syntactic and semantic formalisms, because syntactic representations must serve to describe these mechanisms as well as performing the functions they normally serve in linguistic theory.

There are some interesting differences between the modularity and shared mechanisms hypotheses as regards their conceptions of sentence meaning. One difference concerns the issue of the 'dynamism' of semantic representations. The shared mechanisms hypothesis requires semantic representations to be descriptions of processes, which take an agent from an initial state into a new perceptual (or sensorimotor, or cognitive) state. There is no such requirement in the modularity hypothesis. The interface representation could simply contain a description of a state or event in the world, without making any reference to the agent perceiving it or involved in it. As a matter of fact, many of the semantic formalisms currently advocated by linguists have a strong dynamic flavour: sentences are often understood as context-update operations, taking an agent from one state into another. However, there is still a very large gap between formal theories of dynamic semantics and models of cognitive processing.

A second difference between the modularity and shared mechanisms hypotheses relates to the degree of distributedness of semantic representations. The Fodorian model is centralised: there is a single location where semantic representations of sentences are provided, and both linguistic and sensorimotor modules have access to this location. The assumption in the shared mechanisms hypothesis, on the other hand, is that semantic representations are distributed across a wide range of neural media, rather than localised at a single point. On the issue of locality of representations, almost all natural language semanticists assume a Fodorian model rather than a distributed one.

1.1.1 General motivations for the shared mechanisms hypothesis

Why would we think that the linguistic analysis of a concrete sentence should make any reference to a cognitive model of perception and sensorimotor control? Both empirical and conceptual arguments can be given. From the empirical point of view, much evidence has recently emerged that processing linguistic stimuli activates sensorimotor representations. For instance, brain imaging studies have shown that when a subject listens to a sentence

describing an action of the hand, foot or mouth, this activates an area of the brain (the left fronto-parieto-temporal region) which is also activated during the observation or execution of this same action (Tettamanti *et al.*, 2005; see also Hauk *et al.*, 2004 for results from isolated action verbs). There is also evidence from a technique called transcranial magnetic stimulation, in which a strong magnetic field is locally applied to a region of the brain to amplify signals in that area. It was found that transcranial magnetic stimulation of motor areas associated with hands or feet selectively speeds up processing of verbs denoting actions using these motor systems (Pulvermüller *et al.*, 2005). Similarly, there is evidence that naming concrete objects involves areas of the brain involved in recognising these objects (see e.g. Damasio *et al.*, 1996). However, while this evidence strongly suggests a relationship between *individual words* and sensorimotor structures, it does not demonstrate that there are any sensorimotor correlates of the syntactic structures which link individual words together. And I am primarily interested in syntax, as mentioned at the start.

There are several general reasons why we might expect the syntactic analysis of a sentence to make reference to sensorimotor representations. Firstly, there is an argument from theoretical parsimony. From this point of view, it is preferable to choose a model in which the same theoretical machinery does service both in a model of language and in a model of sensorimotor cognition. It is interesting that many syntacticians, while deeply concerned about parsimonious theories within the domain of linguistics, are less concerned about the need for parsimony in an account of how language relates to other cognitive faculties. However, parsimony is just as important in this case.

Secondly, any successful theory of human linguistic competence must eventually provide an account of how the cognitive mechanisms which implement our syntactic abilities originated during human evolution. The question of how the human language faculty evolved was until recently considered unanswerable, but has become scientifically respectable again over the last ten or so years. A number of very different accounts of language evolution are currently being explored. One of these is that language emerged using sensorimotor mechanisms as a preadaptive platform. This has been fleshed out in several different ways; see e.g. accounts by Rizzolatti and Arbib (1998), Calvin and Bickerton (2000), Corballis (2002), Givón (2002), Hurford (2003) and Arbib (2005). What all of these accounts have in common is the prediction that there are *some* elements of the human language capacity which supervene on underlying sensorimotor mechanisms.

A final argument for the shared mechanisms hypothesis relates to some interesting high-level similarities between the structure of syntactic representations and the structure of sensorimotor mechanisms. Linguistic theories assume that sentences are **compositional**: that is, the meaning of a whole sentence is taken to be a function of the meaning of its individual components, plus their manner of combination. Theories of sensorimotor neuroscience also tend to assume that processes such as event perception and action execution

are organised compositionally. Neural processing is held to be organised into relatively distinct 'pathways' (see e.g. Ungerleider and Mishkin, 1982; Milner and Goodale, 1995; Jeannerod, 1996), which extract or act on different types of information. For instance, it is claimed that there are specialised pathways for categorising objects, categorising actions, and attending to locations, and specialised pathways for reaching and grasping objects. The behaviour of the sensorimotor system as a whole results from interactions between representations in these partly-autonomous pathways. If sentences and sensorimotor operations are both known to break into components, maybe they break into components in the same way. This is a proposal I will try to articulate in the book.

1.1.2 A specific model of shared mechanisms: reference to existing syntactic and sensorimotor models

As just noted, there are some good general arguments in favour of the shared mechanisms hypothesis. However, by themselves these do not provide much help in formulating a specific model of the relationship between language and sensorimotor cognition. A specific model must make reference to a detailed model of language in its own right, and to a detailed model of sensorimotor cognition. It is only when these have been given that we can express a specific hypothesis about the relationship between the two domains. In this section I will briefly describe what kind of use I will make of existing models, both in the linguistic and sensorimotor domains.

On the linguistic side, I should emphasise that the account I develop in this book will make heavy use of existing models of syntax, devised within 'traditional' (i.e. 'Chomskyan') linguistics. Cognitive linguistics and embodied cognitive science typically have a revisionist flavour, rejecting previous approaches and devising new ones. But my approach is more conciliatory. My intention is to hold on to some of the insights expressed within traditional Chomskyan syntax, though I will eventually abandon many of the formal mechanisms associated with this tradition. At the same time I will also attempt to integrate these insights with insights from newer linguistic traditions.

On the sensorimotor side, the main point to note is that the sensorimotor model I introduce in this book is quite a detailed one. In fact, the next two chapters are given over entirely to presenting this model, and make almost no reference to language at all. I make no apologies for this: I want to relate the detail of a syntactic model to the detail of a sensorimotor model, because I think that the shared mechanisms hypothesis is most interesting if it engages with the details of the models in each domain. But it does mean that the model of sensorimotor cognition developed at the start of the book is quite an edifice in its own right. In fact, any researcher who makes a proposal about how language

is implemented in the brain is probably making assumptions about several different sensorimotor and cognitive mechanisms. These might include visual object classification, action perception, spatial cognition, attention, motor control and working memory, to name just a few. Neural models of language can often be understood as proposals about how these mechanisms interact. In this book, I will try to be as explicit as I can about the cognitive mechanisms which I assume.

1.2 An overview of the argument of the book

My approach in this book is to explore the relationship between language and sensorimotor cognition by beginning with one very simple concrete event: the event of a man picking up a cup.

From the perspective of a sensorimotor psychologist or neuroscientist, the interesting questions about this event relate to the visual and motor mechanisms which allow an agent to execute a simple reach-to-grasp action like grabbing a cup, or which allow an agent to recognise such an action being performed by another agent. These mechanisms have been extensively studied, and some of them are becoming quite well understood. I will develop one particular account of these mechanisms, which is a synthesis of current research in this area.

From the perspective of a linguist, the interesting questions about the cup-grabbing event relate to sentences which describe it: *The man picked up a cup*, *I picked it up* etc, along with their equivalents in other languages of the world. A linguist must formulate an account of the syntactic structure of such sentences, as part of a larger theory of grammar. The linguist's account is motivated from empirical data describing the languages of the world. A linguist has to explain, for instance, why *The man picked up a cup* belongs to the English language (while other sequences of English words do not), and how its component words combine together to yield its meaning. I will outline one particular syntactic model of simple transitive sentences of this kind.

In the sensorimotor model which I develop, the basic suggestion is that picking up a cup, or perceiving someone else doing so, is achieved by means of a number of well-defined sensorimotor operations, occurring in strict sequence. The first is an action of attention to the agent of the action; the next is an action of attention to the cup; the next is the process of monitoring the action; and the last is a representation of the state that obtains when the action is completed. This sequence of operations is modelled as a **deictic routine** (Ballard *et al.*, 1997): each operation in the routine is defined with reference to the agent's transitory attentional state, and has the effect of changing this state to enable the next operation. The operations occur at roughly the 1/3-of-a-second time scale, mapping fairly

closely to an agent's eye movements. This model of action perception and execution can be motivated from evidence in pychology and neuroscience.

The syntactic model which I adopt is a version of the Minimalist theory of Chomsky (1995), itself the successor to Chomsky's theory of Government-and-Binding (Chomsky, 1981). In this model, a sentence is represented at two syntactic levels: a level of **phonetic form** (**PF**), which describes its surface characteristics as a sequence of words, and an underlying level of **logical form** (**LF**), which gives an account of how the sentence is constructed from its component words, and why it means what it does. The LF of a transitive sentence like *The man grabbed a cup* is a large right-branching tree, something like that shown in Figure 1.2 (simplified, for effect). This model of clause structure, while it

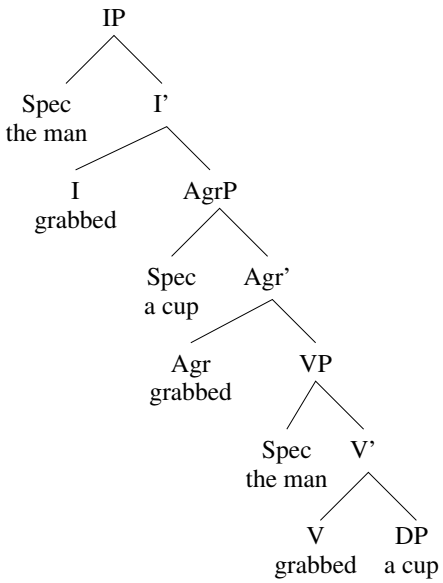


Figure 1.2: Schematic Minimalist model of the LF of *The man grabbed a cup*

might appear very odd to a non-linguist, can be motivated from linguistic argumentation.

The main proposal in the book is that the syntactic structure shown in Figure 1.2 can be understood as a *description* of the sequence of sensorimotor operations involved in experiencing the event of a man grabbing a cup. Each node in the syntactic structure is characterised as describing a sensorimotor operation, and the hierarchical syntactic relationships between nodes indicate the sequencing relationships between these operations. To outline this correspondence very roughly: the nodes between IP and AgrP collectively denote an action of attention to the man (the agent of the action), the nodes between AgrP

and VP collectively denote an action of attention to the cup (the patient of the action), and the nodes between VP and DP collectively denote the process of monitoring the action. And the hierarchical syntactic relations between these groups of nodes denote the sequencing of these operations, so that nodes higher up in the tree denote operations which precede those denoted by nodes lower down. The interpretation is illustrated in Figure 1.3. There are also interesting sensorimotor interpretations of various syntactic relationships in

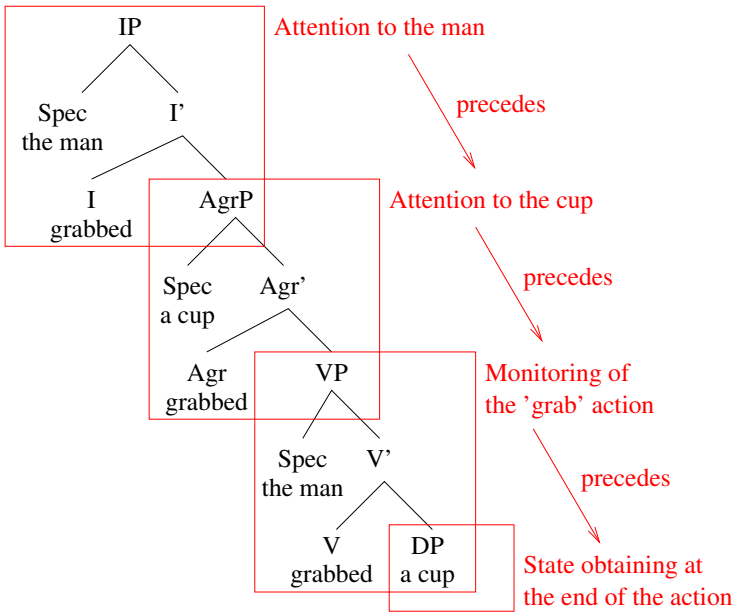


Figure 1.3: Schematic sensorimotor interpretation of the LF of *The man grabbed a cup*

the LF structure. The subject (*the man*) appears at two positions in LF, and so does the object (*a cup*), while the inflected verb (*grabbed*) appears in three positions. In Minimalism, these duplicated elements are assumed to 'move' from one position to another during the construction of LF; movement operations provide the framework for an account of cross-linguistic differences in the positions of subjects, verbs and objects. A sensorimotor interpretation of these positions allows movement operations to be understood as a reflection of general facts about sensorimotor processing, as well as as part of a mechanism for accounting for cross-linguistic distributional data. Given that the sensorimotor and syntactic models are developed using entirely separate methodologies (psychology/neuroscience vs linguistic argumentation), the detailed correspondences between them provide some good preliminary support for the shared mechanisms hypothesis.

My basic proposal is that the right-branching hierarchical structure of LF describes a temporal sequence of sensorimotor processes, and that movement operations have a natural interpretation in relation to this sequence. Obviously this is a general claim, which must be tested for many different syntactic structures before it has any credibility. In the first part of the book, I introduce and motivate the claim for the single example sentence just discussed. In the second part of the book I consider whether the claim extends to a number of other syntactic structures, including noun phrases, predication, quantification and prepositional phrases. In each case, there are some grounds to think it does.

1.2.1 Some objections

There are a few obvious objections which can be made to the proposal that 'syntactic representations describe sensorimotor processes', as set out above. It is worth considering these immediately, as it will help to explain the approach I will take to justifying the proposal.

1.2.1.1 Abstract sentences

Firstly, what does the proposal say about abstract sentences? Clearly these do not describe 'sensorimotor processes' in any direct sense at all. And yet there is nothing in *syntax* which distinguishes between concrete and abstract sentences. Syntactically, the concrete sentence *the man grabbed a cup* is identical to the abstract sentence *the company acquired a subsidiary*. Even if the former sentence does describe a sensorimotor operation, the latter sentence clearly does not. Given that the number of 'concrete' sentences is tiny in comparison to the number of 'abstract' sentences, what is the rationale in looking for sensorimotor mechanisms underlying syntactic structures?

In response: there are good reasons to suppose that the meanings of abstract sentences must be grounded in concrete experience. The concept of 'a company' is certainly abstract and complex, but its primitive components must be presumably be concrete. There has been a lot of work examining how abstract concepts can be grounded in concrete domains—see for instance the work of Lakoff and Johnson (1980). However, to me it seems that once it is accepted that abstract concepts *are* thus grounded, the place to begin our understanding of abstract concepts is with a very detailed understanding of the concrete domain, i.e. of sensorimotor cognition. Once this is in place, we will be in a good position to explore how sensorimotor concepts structure more abstract conceptual spaces. Note that if the syntax of concrete sentences *can* be understood as a description of underlying sensorimotor mechanisms, then the fact that there is no syntactic distinction between concrete and abstract sentences is far from being problematic—it is positively encouraging, because it

allows us to use the syntax of abstract sentences to frame hypotheses about the cognitive processes which they may describe. In this book, therefore, I will be concerned almost exclusively with very concrete sentences, in the hope that an account of these will later be able to shed some light on the semantics of abstract sentences.

1.2.1.2 Levels of representation

Another objection is that language operates at a level of representation which is much higher than that of sensorimotor processing. In one sense this is certainly true: detailed muscle movements, and detailed representations of the visual properties of objects are clearly below the level at which language describes the world. However, sensorimotor mechanisms are hierarchical: there are higher levels of sensorimotor cognition which cannot be rejected out of hand as being involved in an interface with language. My account of sensorimotor cognition will basically be an account of how these higher levels interact with the lower ones, and therefore make an interface with language possible. In fact, in the model I propose, all of the cognitive mechanisms involved in the interface with language are mechanisms for storing sensorimotor experience in *working memory*, rather than mechanisms directly involved in sensorimotor experience. However, the working memory mechanisms strongly reflect sensorimotor mechanisms, and to some extent overlap with these mechanisms. So while it still makes sense to think of my model as 'relating language to sensorimotor cognition', it is more accurate to think of it as relating language to working memory mechanisms, which are themselves strongly related to sensorimotor mechanisms.

1.2.1.3 Differences between languages

Another objection comes from the diversity of human languages. *The man grabbed a cup* can be expressed in a large variety of different ways in different languages: the surface order of words varies, as does the degree of inflectedness of verbs and nouns. In addition, languages differ as to what information is signalled linguistically: some languages systematically leave out words like *the* and *a*, while others leave out information about tense. On the other hand, speakers of all these languages have the same sensorimotor mechanisms. How can we possibly argue that linguistic representations can function as descriptions of sensorimotor processes?

The answer to this question can serve as a brief introduction to Minimalism, the syntactic framework which I use in this book. As already mentioned, Minimalism proposes that a sentence must be described at two levels: the level of phonetic form (PF), which describes the language-specific surface structure of the sentence, and the level of logical form (LF), which is relatively invariant across translations (at least for simple concrete sentences of

the kind I will be considering). LF is used in Minimalism to express the semantics of a sentence—or strictly speaking, to express the syntactic representation of a sentence which communicates with the semantic system. It is also used to express syntactic properties of language which are assumed to be universal, i.e. shared across all languages. My intention is to propose a sensorimotor interpretation of *LF*, not of PF.

Cognitive scientists are typically suspicious of the Minimalist notion of LF. It is often seen as unparsimonious, or methodologically dubious, to appeal to a 'hidden' level of syntactic analysis. But in fact, if the shared mechanisms hypothesis is correct, I suggest we must *expect* a model of language to contain something like LF. If we start from the assumption that speakers of different languages have the same sensorimotor apparatus, which is presumably undeniable, and we also assume that language strongly supervenes on this apparatus, then it follows that we expect to see similarities between languages. If these similarities are not manifest in the surface structure of languages, as they patently are not, the only way of maintaining the shared mechanisms hypothesis is to assume that they are present at an underlying level of analysis. I should note that this is a somewhat unconventional argument for LF. Minimalist linguists typically think of the universal properties of language manifested at LF in Fodorian terms, as reflections of a modular mechanism specific to language. But it is not necessary to think of them in this way. If language shares mechanisms with sensorimotor processing, then there are likely to be universal properties of language which reflect the fact that we all have the same sensorimotor apparatus.

1.2.1.4 Syntactic structure does not determine semantics

A final objection, which many linguists may be inclined to voice, is that giving a sensorimotor interpretation to LF (or indeed to any syntactic structure) overstates the strength of the relationship between syntax and semantics. The question of how, in general, semantic structures can (or must) be realised syntactically is a fearfully complex one. There are certainly some basic rules which appear fairly robust, but there is no doubt that a complete answer to the question involves enumeration of countless idiosyncracies and conventions as well as reference to general principles. For instance, transitive sentences can realise a wide range of semantic messages, as shown below:[1]

(1.1) Liana left the room.

(1.2) Capitalism disgusts Sam.

(1.3) Emily likes cats.

[1]These examples are taken from a more extensive discussion in Jackendoff (2002).

(1.4) John mentioned his pet elephant.

(1.5) Harry owns a BMW.

These sentences would normally be given the same syntactic analysis as *The man grabbed the cup*. But they report a range of different semantic structures. (In particular, the semantic roles played by the subject and object differ widely from one sentence to the next.) Suggesting that the LF of a transitive clause can be read as a sensorimotor trace may seem hard to square with the obvious semantic variety of transitive sentences.

I should say straight away that the model I will present does allow for idiosyncracies in the mapping between semantic messages and sentence structures. While I will adopt a Minimalist model of LF, my account of the mapping between LF and PF will be very different from the standard Minimalist account. It will include a rich notion of surface syntactic constructions, and will allow for the development of complex conventions about how these realise semantic structures (see Chapter 6). But obviously I still intend for my sensorimotor interpretation of LF to generalise usefully beyond my initial example sentence.

To start with, I want to emphasise that we do not yet have a good understanding of the processes involved in apprehending the events or states described by Sentences 1.1–1.5 (and other similar sentences). In some cases (e.g. Example 1.1), these processes are probably mainly perceptual, and we have some idea of the neural representations which are created. But they mostly involve a mixture of perceptual and inferential processes, whose neural basis we are very far from understanding. So it would certainly be premature to rule out the possibility that the syntax of these sentences encodes something about the epistemic processes involved in learning the facts they report. *The man grabbed the cup* is in one sense an unusual sentence, because the processes involved in experiencing the episode it reports have been intensively studied, and we now know quite a lot about them. At the very least, an interesting sensorimotor interpretation of the LF of this sentence may help generate hypotheses about the more complex epistemic processes associated with other transitive sentence types.

I should also note that a 'sensorimotor' interpretation of syntax is not the same as a 'semantic' interpretation of syntax. I am not suggesting that the LF of a sentence directly encodes the semantic representation associated with a sentence, but rather that it encodes something about the process involved in *obtaining* this semantic representation. For the cup-grabbing episode, I argue that this process is extended in time, and involves a number of distinct steps. The earliest steps involve direction of attention to the objects which participate in the episode; the final steps involve apprehension of the event in which they participate, and a consequent reattention to the objects. Linguists often think of certain elements of syntactic structure as *purely* syntactic—i.e. as unrelated to semantics. I want

to interpret at least some of these 'purely syntactic' aspects of syntax as reflections of contingencies in the order in which sensorimotor operations must occur, and of contingencies in the way they are stored in working memory. This goes some way beyond a simple attempt to reduce syntax to semantics. But I think it offers a helpful new way of articulating the subtle relationship between the two.

1.3 Structure of the book

The dual emphasis on sensorimotor and syntactic theory in this book means that most readers will be unfamiliar with at least one half of the material being presented. I have therefore undertaken to introduce all the necessary background material 'from scratch'.

In Chapter 2 I introduce a model of the sensorimotor processes which underlie the action of picking up a cup. I begin by surveying what is known about the sensorimotor pathways involved in the execution of simple reach-to-grasp actions, and in the perception of such actions performed by others. From this survey, I construct an account of the characteristic sequential structure of the sensorimotor events involved in these two processes. This chapter provides an introduction to the sensorimotor representations in terms of which the sentence *The man grabbed a cup* must ultimately be grounded.

In Chapter 3, I consider how this sensorimotor model interfaces with a model of procedural learning and episodic memory. The sensorimotor system is not linked *directly* to the linguistic system, because people do not speak as a side-effect of perceiving or acting. I suggest that when an episode is experienced, evoking a sequence of sensorimotor operations, this sequence of operations can be stored in working memory, in a form which allows the sequence to be *internally replayed*. I will argue that this replay operation plays a crucial role in the linguistic interface, and that during replay, sensorimotor signals can have direct linguistic side-effects. In Chapter 3, I review existing models of working memory, focussing on memory for sequences of items. I also introduce a model of longer-term episodic memory, and review recent evidence that our episodic memory system is well configured for the storage and recall of sequences.

In Chapter 4, I turn to linguistics. I introduce the Minimalist model of natural language syntax (Chomsky, 1995). In this model, as already noted, a key idea is that sentences must be analysed at two levels: a level of surface structure (PF), and an underlying level of semantic representation (LF). I will outline some of the syntactic arguments for this idea, and present a model of the LF of the example sentence, *The man grabbed a cup*.

In Chapter 5 I will finally be in a position to state some correspondences between the sensorimotor/memory model and the syntactic model. My claim is that the LF of a sentence (i.e. the representation of its meaning in the Minimalist syntactic model) can be

characterised as an encoding or trace of the operations taking place in the sensorimotor system when the event or state depicted by the sentence is simulated in working memory. I will argue that there are strong formal similarities between the Minimalist model of LF and the sensorimotor model of event/state perception, and that these similarities are deep enough and detailed enough to require explanation. My explanation is that syntactic representations (of concrete sentences) are representations of sensorimotor processes, as replayed from working memory—in other words, that linguists and sensorimotor neuroscientists are in some sense studying the same topic, albeit using different methodologies.

In Chapter 6, I consider the implications of the sensorimotor interpretation of LF for models of how language is implemented in the brain, and for models of how children learn language. In my interpretation, an LF structure describes a cognitive process (the process of internally rehearsing a sensorimotor sequence), rather than a static mental representation. While LF is still primarily to be understood as part of a declarative theory of grammar, the fact that LF representations are understood as processes also allows them to feature in models of 'language processing'—i.e. sentence interpretation and sentence generation—and also in models of the neural substrates of language. The way LF structures are defined in Minimalism makes the theory hard to deploy within such models, but I will argue that my interpretation of LF removes some of these difficulties. The first half of the chapter is another large literature review: a synopsis of work in neurolinguistics, developmental linguistics and connectionist linguistics. In this synopsis I also introduce the **empiricist** and **constructivist** models of language and language acquisition, which are frequently adopted by neurolinguists and developmental linguists, and tend to be understood as alternatives to the 'nativist' model of acquisition embodied in Minimalism. In the second half of the chapter I introduce a new model of language learning and language processing, expressed in a collection of neural networks. The model relates closely to several existing connectionist/constructivist accounts of language learning and language processing. But it can also be understood from a Minimalist perspective as a device for learning the mapping from LF to PF representations in a particular language.

Up to this point, the book focusses on the clause-level structure of the chosen example sentence, *The man grabbed a cup*. I do not consider the internal syntax of the noun phrases in the sentence (*the man* and *a cup*) in any detail in Chapters 5 or 6. In the next two chapters, I turn to the topic of noun phrase syntax, and in the remaining chapters, I consider a number of other syntactic constructions.

In Chapter 7, I suggest a sensorimotor interpretation of the internal syntax of noun phrases. I begin by providing a more detailed model of the sensorimotor processes which underlie attention to and categorisation of objects. I argue that these processes also have a strong characteristic sequential structure. I then introduce a Minimalist model of the internal syntactic structure of noun phrases. The LF of a noun phrase has the same basic

right-branching structure as the LF of a clause. I argue that the LF of a noun phrase can be understood as a description of a sequence of attentional operations. In simple cases, where noun phrases can be interpreted 'referentially', the associated attentional sequences are those which result in perceptual establishment of the object or group which is referred to.

In Chapter 8 I consider the relationship between noun phrases and the clauses in which they appear. This is a thorny issue in linguistics. For one thing, clauses can only refer indirectly to the semantic objects contributed by noun phrases. Some account of variables and variable binding is required to implement this indirection. For another thing, noun phrases do not always 'refer' to objects or groups. In a 'quantified sentence' like *The man grabbed most cups*, the semantic contribution of the noun phrase *most cups* cannot be described in isolation from the contribution of its host clause. In this chapter, I will discuss the relationship between attentional actions and episode representations, and give an initial account of the interface between noun phrases and clauses which makes reference to this relationship.

In Chapter 9, I turn from a discussion of clauses describing 'events' to a discussion of clauses describing 'states'. I begin by considering the sensory and attentional processes involved in perceiving simple concrete properties of concrete objects—for instance, the processes involves in perceiving that a cup has a certain colour. I then consider the way simple properties of objects are stored in long-term memory. Long-term memory for the properties of objects is termed 'semantic memory', and is commonly held to be distinct from long-term memory for episodes. I will discuss the structure of semantic memory, and make a suggestion about how it relates to epsisodic memory. This will lead to a proposal about how quantified propositions are stored in semantic memory, and to a 'sensorimotor' interpretation of the structure of quantified sentences, which gives some more detail about the syntactic interface between noun phrases and clauses. Based on this interpretation, I also suggest a sensorimotor interpretation of relative clauses—constructions where a clause is embedded inside a noun phrase.

I consider one other syntactic topic in detail: prepositional phrases (PPs). In Chapter 10, I begin by outlining a model of spatial cognition synthesised from the current literature, which covers how 'environments' are represented, the relationships between environments and objects, and the processes involved in monitoring (or causing) the movement of objects through environments. I then present a model of the internal syntax of PPs, and again argue that the syntactic structure of PPs can be given a sensorimotor interpretation.

The next two chapters take the discussion in a few different directions. In Chapter 11 I extend the discussion beyond transitive and predicating clauses to consider a variety of other clause types. These include intransitive clauses and passive clauses, and also clauses

whose verbs take PP arguments. And in Chapter 12 I consider the linguistic notion of 'discourse context', and how this might be interpreted in relation to the sensorimotor model.

In Chapter 13, I consider how the theory presented here relates to existing linguistic and psycholinguistic theories. There are many interesting points of contact, which it is useful to highlight. Finally, in Chapter 14, I summarise the present work, and formulate some some conclusions.

1.4 How to read the book

I assume that readers of this book will either be linguists or cognitive scientists of one variety or another. Linguists can treat Chapter 2 as a survey (by necessity somewhat biased) of everything they might ever need to know about the sensorimotor system. Chapter 3 is a similarly biased whistle-stop tour of working memory and its interfaces with the sensorimotor system and with episodic long-term memory. Neither chapter is a straight literature review, but I have tried to be explicit about which proposals come from me, and which proposals reflect a minority or controversial position in the literature. If the reader wants to skip the gory detail of Chapter 2 and take me on trust, there is a summary starting in Section 2.9. If an even more trusting reader wants to skip both Chapter 2 and Chapter 3, both these chapters are summarised at the start of Chapter 5.

Chapter 4 is an introduction to Minimalism for non-linguists. It is quite cursory, and I do not expect a reader coming to Minimalism for the first time to be particularly convinced by the presentation I give. There are many textbooks which introduce the theory more thoroughly, to which I refer the reader at the start of the chapter. (One reason I go into more detail in the sensorimotor and memory chapters is that there are not many attempts to formulate a detailed model of episode representations in the neuroscience literature.) But even after a thorough introduction, it is easy to remain skeptical about the Minimalist idea that sentences have an 'underlying' level of syntactic structure, whose form can only be inferred through syntactic argumentation. I can only urge the reader to bear in mind that my project is to propose an alternative, sensorimotor characterisation of these underlying structures which corroborates the syntactic arguments.

Chapter 5 is where the central idea of the book is presented. One way of reading the book is to start here to get the gist of the proposal. There are plenty of signposts to earlier chapters to facilitate this option.

Chapter 6 gives another extensive literature review, covering neurolinguistics, developmental linguistics and aspects of connectionist and statistical language modelling. These topics are the home territory for most researchers interested in 'psychological' aspects of

language. Readers familiar with these areas should be able to skip or skim Sections 6.1 and 6.2. The new material in this chapter is in Sections 6.3 and 6.4.

In Chapters 7–12, it is harder to point particular readers to particular places. Each chapter contains sections dealing with extensions to the sensorimotor model, sections introducing new syntactic constructions, and sections relating sensorimotor and syntactic models. These chapters can (I hope) be thought of as models of the type of argumentation which is legitimised if the shared mechanisms hypothesis is correct, in which syntactic and sensorimotor models mutually inform and constrain one another.

Chapter 2

Sensorimotor processing during the execution and perception of reach-to-grasp actions: a review

In this chapter, I provide a review of the sensorimotor processes which occur when an agent observes or participates in a simple transitive action: the action of reaching for and grasping a target object. A well-established hypothesis about the sensorimotor system is that it can be decomposed into a number of reasonably separate **neural pathways**. Some of these pathways are involved in the immediate control of motor actions, while others are involved in the creation of more conceptual representations of the agent's environment, and are hypothesised to influence the formation of higher level goals and plans. In this chapter, I will provide an overview of the main pathways which have been discovered, drawing on neurophsyiological work (mainly done on the macaque monkey) as well as on human experiments. The main question I want to address is: how is a 'high-level representation' of a reach-to-grasp episode created from the complex pattern of sensory and motor representations evoked in these pathways?

In Sections 2.1 to 2.9, I will consider several different sensorimotor pathways in turn, which deal with early vision, object classification, spatial attention, the execution of reach and grasp actions, the planning of higher-level actions, and the recognition of actions in other agents. In each case, my focus will be on the areas in the brain involved in each pathway, and on the representations which are computed in these areas. But I will also be thinking about the temporal organisation of a reach-to-grasp action. Operations in the different pathways are sometimes strongly constrained to happen in a particular order. Where ordering constraints apply, I will point them out explicitly. Then in Section 2.10, I will draw together these ordering constraints, and propose that the execution and recognition

of a reach-to-grasp action both involve a well-defined *sequence* of sensorimotor operations, in which representations of agent, patient and motor action occupy characteristic serial positions. In Chapter 3 I will argue that our mechanism for representing episodes in working memory capitalises on this sequential structure—in other words, that constraints on the ordering of sensorimotor operations are reflected in the structure of the high-level episode representations which interface with natural language.

Before I begin, a note on terminology: the terms I use to refer to brain areas mostly come from models of the macaque. (Much of the neurophysiological work on primate sensorimotor pathways has been done on macaque.) But many terms are used for both human and macaque brain areas, and there are also terms which uniquely describe human brain areas. By default, my terms refer to macaque brain areas; it should be clear from context when I am talking about the human brain.

2.1 The early visual system: LGN, V1, V2, V3 and V4

Both action execution and action observation normally involve vision. Action execution also involves a large measure of **somatosensory** input (a mixture of tactile information and **proprioceptive** information about the position of the body), and action observation can also involve audition; in fact, both processes can occur with no vision at all. However, we will begin by considering the visual mechanisms which are normally involved in each process.

The early visual system is relatively well understood. A very simple diagram is shown in Figure 2.1. Light from the visual array is received by photoreceptors on the retina, which

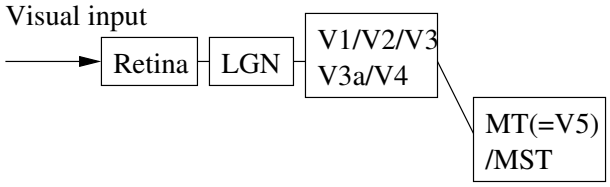


Figure 2.1: The early visual system

relay signals down the optic nerve to mid-brain structures called the lateral geniculate nuclei (LGN), of which there is one in each hemisphere. Signals associated with the left visual fields of both eyes are sent to the right LGN and vice versa. From here onwards,

processing associated with the two visual fields is carried out relatively independently in the two hemispheres, so there are two copies of each region mentioned in the pathway, but it is more convenient to refer to regions in the singular.

The LGN sends output (or **projects**) to the primary cortical visual area, called V1, which in turn projects to V2. Cells in both these areas are organised retinotopically, such that cells in neighbouring locations are sensitive to stimuli in the same or neighbouring regions of the retina (we can say that they have similar **receptive fields**). V1 and V2 can thus be thought of as maps of the retina. In each case, cells associated with the centre of the retina (the **fovea**) have very small receptive fields, and those associated with the periphery have increasingly large receptive fields.

Within each point in V1 and V2, cells are sensitive to different kinds of stimulus. There are several dimensions of variation; cells can prefer stimuli of different orientations, stimuli of different spatial frequencies, stimuli of different opposing colours, stimuli of different binocular disparities (reflecting different distances from the viewer), and moving stimuli travelling along different paths. The anatomical basis of this segregation is less clear than was once assumed (see e.g. Sincich and Horton, 2005); however, it is still possible to view V2 as providing a set of parallel **feature maps** charting the distribution of various simple visual features across the retina.

V2 projects to several other areas with retinotopic organisation, most prominently V4, V3a and MT (the medial-temporal lobe), LIP (the lateral intraparietal area), c-IPS (the caudal intraparietal sulcus) and PO (the parietal-occipital area). These areas typically show an increased specialisation for one type of feature, as well as a general increase in the complexity of their preferred stimuli. For instance, sub-areas of cells in V4 are particularly responsive to colour (Zeki, 1983), while MT cells are particularly responsive to direction of motion (Maunsell and van Essen, 1983), and c-IPS cells are particularly sensitive to binocular disparity (Tsao *et al.*, 2003).

2.2 The object classification pathway: IT

Action observation and action execution typically involve the visual categorisation of objects in the world. In order to represent a reach action, for instance, it is necessary to use vision to identify the object being reached, and probably also the agent doing the reaching (if it is not onesself). Again, vision is not necessary for object identification; somatosensory information can be used for this purpose too (see Section 2.5.4 for some more about this process). However, the normal mechanism for object identification is through vision.

A well-defined visual pathway in the ventral cortex (also referred to as the inferotemporal cortex, or IT) specialises for categorising objects. This pathway receives input primarily

from V4, and progresses through inferotemporal areas TEO and TE, as shown in Figure 2.2.

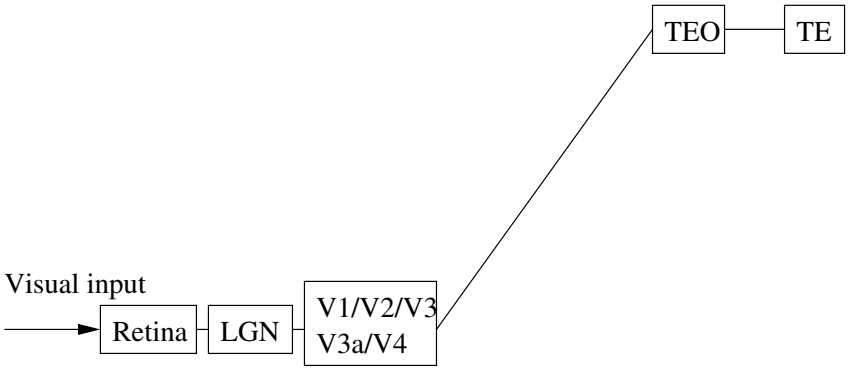


Figure 2.2: The object categorisation pathway

There are many studies which show that cells in the ventral visual pathway respond to the form of complex visual stimuli (see e.g. Logothetis *et al.*, 1995; Logothetis and Sheinberg, 1996; Tanaka, 1996; Tanaka, 1997; Riesenhuber and Poggio, 1999). As we move up the pathway from V4 to TEO to TE, cells encode progressively more complex forms, beginning with simple oriented Gaussian shapes in V4, and culminating in TE with cells sensitive to complex combinations of shapes (with an emphasis on 'natural' shapes such as faces and hands). In Logothetis *et al.*'s (1995) classic study, monkeys were trained to recognise a series of 'paperclip'-like stimuli consisting of crooked three-dimensional lines. After intensive training, TE cells were found which responded selectively to individual stimuli. Typically, these cells responded to particular *views* of a given stimulus, although there were a small number of cells which responded to a range of different views. In the traditional model of the ventral pathway, cells from each stage in the pathway respond to combinations of inputs from cells at the previous stage. Thus cells towards the end of the pathway respond to combinations of combinations of combinations of simple features.

While the complexity of the forms to which cells are tuned increases as we progress down the ventral pathway, their receptive field also increases. While cells at the start of the pathway are selective to simple forms at specific points on the retina, a typical cell at the end of the pathway is selective to a particular complex form appearing at a range of retinal locations. The degree of spatial abstraction is different for different neurons; op de Beeck and Vogels (2000) found receptive fields ranging from 2° to 25°, so abstraction is far from ubiquitous in TE (see also Kravitz *et al.*, 2007 for a note of caution on how to interpret these findings). However, it certainly seems far more prevalent than in earlier visual areas.

The need for spatial abstraction can be understood from considerations of computational complexity. The number of possible features rises exponentially in each stage in the pathway. Given that there are a finite number of cells at each stage of the pathway, it makes sense to maximise the number of possible combinations which can be encoded by sacrificing spatial resolution (see Mozer and Sitton, 1996 for a good presentation of this argument). On the other hand, this spatial abstraction creates potential problems; if two objects are presented at different points on the retina, one might imagine that their visual features would be blended together by the time a representation reached the latter stages of the ventral pathway. The problem is traditionally seen as being resolved by the mechanism of **spatial attention**; see e.g. Desimone and Duncan (1995), Reynolds and Desimone (1999). According to this model, attention can be allocated selectively to different points on the retina. Information about these points is preferentially represented in the input to the object classification pathway, so that only the attended-to stimulus is classified. These attentional mechanisms will be considered in Section 2.4.

2.2.1 Object categorisation in humans

In humans, fMRI studies have identified brain regions which are particularly selective to the shape of objects. A well known study by Kourtzi and Kanwisher (2001) found that a fairly early visual area, the **lateral occipital cortex** (or **LO**), appears to encode a fairly high-level representation of object shape. If a shape is presented partially occluded by bold lines, the representations in LO remain relatively unchanged, suggesting that the region is not representing low-level features such as image contours. Conversely, if a stimulus containing a figure-ground ambiguity is presented, manipulating which part of the stimulus is figure has an effect on LO, even though there is no change in the contour dividing figure from ground. LO is not in the temporal cortex, but it appears to create the representations which are passed into the temporal cortex.

Categorisation of an object involves more than just representation of its shape: the shape of an observed object must be matched to a shape associated with the appropriate category. In Kourtzi and Kanwisher's study, LO responds strongly to 'nonsense objects' as well as to objects of identifiable categories. An fMRI experiment by Bar *et al.* (2001) provides more detail about regions active when an object's category is identified. In their experiment, subjects were shown a stimulus for a brief period followed by a visual mask, and asked to give a graded judgement about how the degree to which they had recognised the stimulus. The time period was varied, to generate a range of responses. Activity in two areas correlated with subjects' judgements: one was the occipitotemporal sulcus (OTS), which is part of LO, and the other was the fusiform gyrus, which is part of inferotemporal cortex proper, downstream from LO. This experiment suggests that object categorisation

is a gradual process, which begins by the refinement of shape representations in LO, and then involves a temporally extended process of matching these representations to stored templates in inferior temporal cortex. The suggestion that human inferotemporal cortex is involved in object categorisation is also supported by single-cell recordings in humans (Kreiman *et al.*, 2000), and clustering analyses of fMRI data in humans (Kriegestorte *et al.*, 2008). In what follows, I will assume visual object categorisation happens in the inferior temporal cortex or IT, both in humans and in monkeys. I will discuss the pathways involved in object categorisation in more detail in Section **??**.

2.2.2 Top-down influences on object categorisation

The object categorisation mechanism in IT is driven not only by perceptual information from the retina, but also by top-down influences relating to the agent's current task. Two sorts of influence can be distinguished. Firstly, there are biases created by expectations about the object currently being classified. To give a quick example: subjects are faster at naming a visually presented object if they have recently seen that object, or a related object. This phenomenon is called **priming**: for instance, showing subjects a picture of a knife primes (or speeds) their subsequent identification of a fork (see e.g. Sperber *et al*, 1979 for an early reference). There are many models of priming; classically, it is explained as an effect of increasing the activation of the units in IT which encode elements of the primed object type (see e.g. Rumelhart and McClelland, 1995). Thus when the IT units which encode 'knife' activate, they increase the baseline activation of the IT units which encode 'fork', so that less additional activation is required in order for them to fire, and reaction time is improved.

Priming effects are typically analysed as being internal to the IT classification system. Another type of top-down influence on object categorisation comes from outside IT, and relates to an agent's goals. In certain circumstances, an agent actively looks for an object of a certain type. Clearly, the categorisation system should not be directly biased towards the sought-for type, because it must reflect the actual object being classified, rather than the agent's desires about it. However, a representation of the sought-for type must be compared with the type computed by IT, to determine whether the search goal has been achieved. There is evidence that representations of sought-for objects in a visual search task are maintained in a region called the **prefrontal cortex** (or **PFC**)—see e.g. Hasegawa *et al.* (2000); Hamker (2004); Tomita *et al.* (1999. Prefrontal cortex is involved in maintaing working memory representations relevant to the agent's current task; we will consider its function in more detail in Section 2.6, and in much of Chapter 3. There are strong links between IT and PFC; it is likely that some of these create circuits which allow a representation of the currently classified object to be compared with a goal representation

in PFC. In monkeys, the IT representation of a perceived object is enhanced if the animal is looking for that particular object, while it is attenuated if the animal is looking for a different object (Miller and Desimone, 1993).

It also appears that IT object representations are active in working memory tasks where a subject must maintain object representations over a short delay and then simply recall them; see e.g. Ranganath and D'Esposito (2005) for a review. This working memory IT activity is also likely due to top-down activation of IT by PFC, and by the hippocampal region, which also participates in some types of working memory. Thus while PFC can bias the process of visual classification in IT, it also appears able to activate object representations in IT purely internally, in the absence of any visual input.

2.3 The posterior parietal cortex: vision for attention and action

The visual pathways which provide the most direct links to motor action are found in posterior parietal cortex. This area can be broadly characterised as implementing *sensorimotor transformations*, which convert sensory information into appropriate motor actions. The inputs to the parietal system are mainly sensory signals (both visual, auditory and somatosensory), and the outputs are mainly motor signals. The motor signals can be to the motor system proper, via the primary motor cortex (F1, also called M1), or to the oculomotor system via the frontal eye fields (FEF) and the superior colliculus (SC).

Both the motor and oculomotor pathways are heavily concerned with the representation of space around the agent. A key component in the specification of any motor action is the location of the *target* of the action, whether this is a location to be fixated by the eye or a location to be reached by a limb movement. One central task of parietal cortex is to convert information about target location from sensory coordinate systems to motor ones; for instance, the position of an object arriving in a retinotopic frame of reference might need to be converted into a set of goal arm joint angles. In addition, there are likely to be several objects in the current visual field; one of these must be *selected* as the target for the next action, whether it be motor (e.g. a reach) or purely attentional (e.g. a saccade) or (as frequently happens) both. There is evidence that the parietal cortex is where this selection process takes place as well. In sum, parietal cortex is involved in deciding between alternative spatial actions, and in planning and executing these actions.

I will discuss the parietal pathways involved in attention and action in separate sections. Section 2.4 deals with pathways involved with attention, and Section 2.5 deals with those involved in reach actions.

2.4 Vision for attentional selection: LIP and the frontal eye fields

Alongside the ventral/inferotemporal visual pathway, in which cells encode the shapes or categories of objects while abstracting away from their spatial location, there is a separate pathway which appears to perform a complementary operation, encoding the location of objects in the agent's environment but not their form. In this pathway, a map of locations in the visual field is computed, which provides the framework for a decision about what to attend to in the current scene; it emphasises locations which are likely to be interesting or behaviourally relevant. This pathway has its origins in the same visual feature maps which provide input to the object classification system, but computes information about the location of visual stimuli rather than about their category or identity.

Computational models of visual attention, inspired by the spatial focus of parietal cortex, have often posited a representation called the **saliency map** (Koch and Ullman, 1985; Wolfe, 1994; Itti and Koch, 2001). The saliency map identifies regions where there is movement, or 'local contrast'—i.e. a discontinuity in the distribution of simple visual features such as colour, orientation or intensity. Computational models of the saliency map typically have two levels: one at which all salient regions are represented, and one in which regions compete, so only a single region is selected. An illustration is given in Figure 2.3.

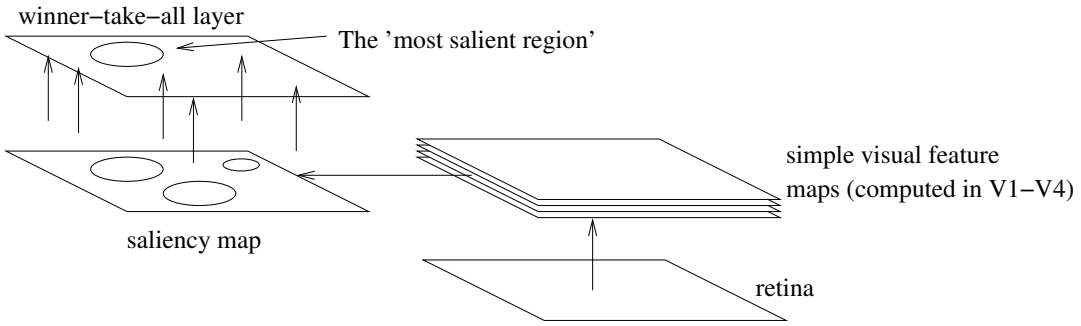


Figure 2.3: The saliency map, and the winner-take-all map representing the 'most salient region'

Neuroscientists have looked for evidence for something analogous to a saliency map in the brain. There are regions of parietal (and frontal) cortex which have been argued to deliver such a representation. These regions, which are part of the circuits which control eye movements and other covert forms of attention, are shown in Figure 2.4. As the figure shows, there are two distinct circuits for control of eye movements. An evolutionarily old

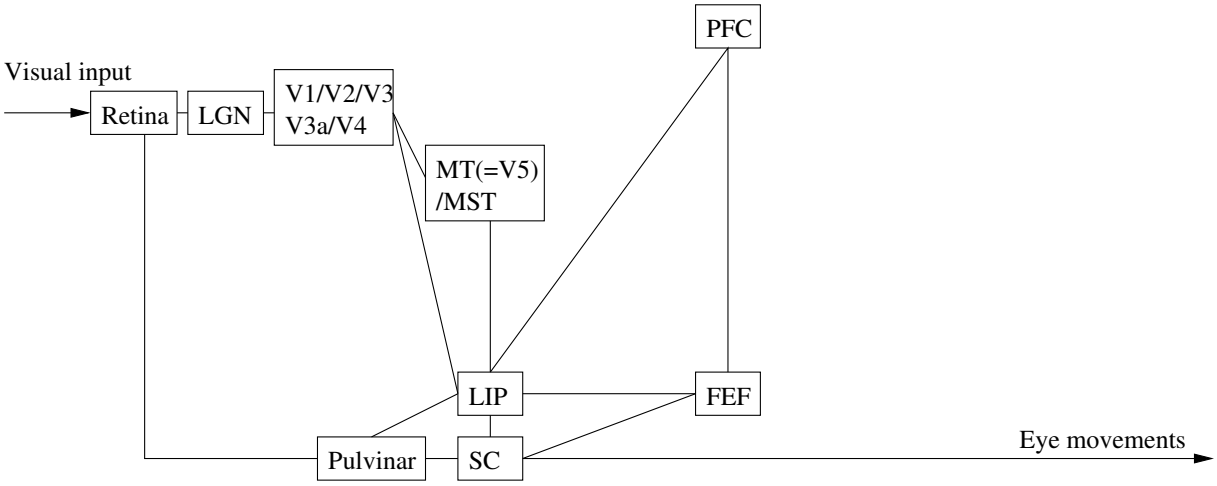


Figure 2.4: The 'attention' visual pathway

subcortical circuit leads from the retina to the superior colliculus (SC), from which direct saccade signals are generated. The SC also links to the pulvinar, another evolutionarily old structure, which is involved in the process of shifting eye position from one point to another. An evolutionary extension of this circuit involves the lateral intraparietal cortex (LIP) and an area of premotor cortex called the frontal eye field (FEF). LIP and FEF both receive input from visual areas processing form (V4) and motion (MT/MST). But cells in these areas do not typically distinguish between particular forms or motions; rather, they appear to encode the presence or absence of stimuli 'worthy of attention' at a given point in the visual field, regardless of what shape or direction it has (Colby and Goldberg, 1999; Thompson and Bichot, 2005). Interestingly, the map of the visual field generated in LIP and FEF appears to be three-dimensional; cells in both LIP and FEF are sensitive to stimulus distance, as diagnosed from both retinal disparity and vergence cues (Gnadt and Mays, 1995; Ferraina *et al.*, 2000).

2.4.1 LIP and FEF cells encode salient visual stimuli and associated eye movements

The typical LIP or FEF cell fires when a stimulus appears in its receptive field which the animal subsequently looks at. This firing is ambiguous—it could simply be encoding the presence of a stimulus at a certain location (as cells in early visual areas do), or it could be encoding the presence of an 'important' stimulus at this location, or it could be encoding

information about the preparation or the execution of the saccade that subsequently takes place. Various tasks can be devised which distinguish between these possibilities. Firstly, we can manipulate the salience of stimuli. For instance, we can compare a situation where a stimulus arrives *de novo* in a cell's receptive field (a salient event) with one where the stimulus already exists in the display and is brought into the receptive field by a saccade. LIP cells are much more responsive in the first case than the second, suggesting that they encode the locations of important stimuli, rather than the locations of all stimuli (Gottlieb *et al.*, 1998). Another way of manipulating salience is to present an object which 'pops out' of a visual display—for instance, a single red object in a field of green objects. FEF cells respond more strongly to such an 'oddball' item than to a stimulus in a field of identical stimuli (Bichot *et al.*, 2001b). Such findings suggest that LIP and FEF cells encode salient objects, rather than all objects in the scene.

To distinguish between cells encoding location and cells encoding eye movements, tasks can be devised in which animals must locate a target object and then make a saccade *away* from this object, or in which animals must make a saccade to a location containing no visual stimulus. Studies using such tasks have found LIP neurons which primarily encode location rather than eye movements (Kusunoki *et al*, 2000), and LIP neurons which primarily encode movements rather than location (Snyder *et al.*, 2000). Cells in FEF have a similar distribution of preferences. In one study (Thompson *et al.*, 1997), animals were trained to locate an oddball stimulus, in one condition making a saccade to the stimulus and in another condition witholding a saccade. The response of FEF cells was the same in both conditions, suggesting they encode object location rather than saccade preparation. In another study (Sato and Schall, 2003), some FEF cells encoded stimulus location and some encoded saccade endpoint.

In fact, if we assume that LIP and FEF are both involved in transforming visual stimuli into attentional actions, it is unsurprising that there are some cells selective for location and some selective for eye movement in each area. However, there are two points worth making. Firstly, it does seem that both LIP and FEF can encode a salient stimulus to be attended to even if overt eye movements are suppressed; in other words, they are involved in **covert attention** as well as overt eye movements. There is considerable evidence that the pathways for overt and covert attention overlap extensively (see e.g. Kowler *et al.*, 1995; Nobre *et al.*, 2000). Secondly, in the normal situation in which a saccade is made to an attended-to location, we can think of LIP/FEF neurons as implementing the process of *deciding* where the animal looks next (Schall, 2001). The animal's overt action is determined by a competition between neurons representing different locations and/or their associated eye movements.

2.4.2 LIP/FEF cells also encode top-down attentional influences

The activity of LIP and FEF cells is modulated by top-down influences as well as bottom-up ones. For instance, LIP cells respond more to an object in their receptive field if it is relevant to the task the monkey is currently performing (Kusunoki *et al.*, 2000). Similarly, if a monkey is searching for a target object with a particular feature, FEF cells respond more to this object than to other objects, even if they are more salient in other ways (Bichot *et al.*, 2001a), and respond more to distractors which share this feature than to those which do not (Bichot and Schall, 1999).

It seems likely that top-down influences on attention originate in frontal cortex, while bottom-up influences originate in parietal cortex. Buschman and Miller (2007) found that when monkeys attended to an object top down, the location attended to become active in FEF before becoming active in LIP, while for actions of attention driven by bottom-up salience, the opposite pattern was found. There are direct connections between PFC and both LIP and FEF, as indicated in Figure 2.4. These connections are likely to implement a bias on the saliency map towards objects of certain types. Given that this bias is applied in parallel across the whole retina, it is likely to be rather crude—for instance, a PFC preference for attention to a can of coke might manifest itself as a bias in the saliency map towards stimuli which are red, vertically oriented and of a certain size (see e.g. Walther *et al.*, 2002). Having created a shortlist of likely locations in the saliency map, it will then be necessary to search serially through these locations, attending to each in turn and processing it in the object categorisation system, until the search target is found. (A likely mechanism for performing this cycling operation is inhibition-of-return in the saliency map, which will be discussed in Section 2.4.5.) When each candidate object is attended to, it will be compared to a target stored in PFC (see Section 2.2.2): a match indicates a target has been found at the currently attended location, while a mismatch trigger inhibition-of-return, to move attention to another candidate object.

2.4.3 Spatial attention and object classification

There is a considerable amount of evidence emerging that visual attention modulates the activity of cells in all areas of the visual system (see e.g. Treue, 2001; Maunsell and Cook, 2002). While there remains debate about the source of this modulation, it is likely that LIP and FEF are major contributors. Cells in LIP and FEF project back by indirect cortical relays to earlier visual areas such as V4, and they also link back to these areas by a direct path via the pulvinar (Shipp, 2004). Stimulation of FEF has been shown to modulate activity in retinotopically corresponding areas of V4 (Moore and Armstrong, 2003). In a similar vein, it has been found that transcranial magnetic stimulation of the human FEF

facilitates detection of near-threshold visual stimuli (Grosbras and Paus, 2003). Recall from Section 2.2 that V4 is the main source of input to the object classification pathway. As noted in Section 2.2, cells in that pathway progressively lose information about spatial location while deriving progressively more complex representations of visual shape. If the retinal input to IT were unmodulated, we might expect all kinds of illusory conjunctions of visual features. Spatial attention has commonly been invoked as a mechanism for solving this feature binding problem; see e.g. Mozer and Sitton (1996), Reynolds and Desimone (1999). The contribution of LIP and FEF in modulating the responses of earlier visual areas can be understood as a central component of the mechanism which prevents such spurious conjunctions from arising. Indeed, given that LIP/FEF activity need not result in an overt saccade, this internal modulatory action of LIP/FEF cells is arguably their most central role. It is no less an 'action of attention' for having no overt motor component.

The modulatory effect of LIP/FEF on object categorisation is illustrated in Figure 2.5. As the figure shows, activity in the winner-take-all region of the saliency map gates the

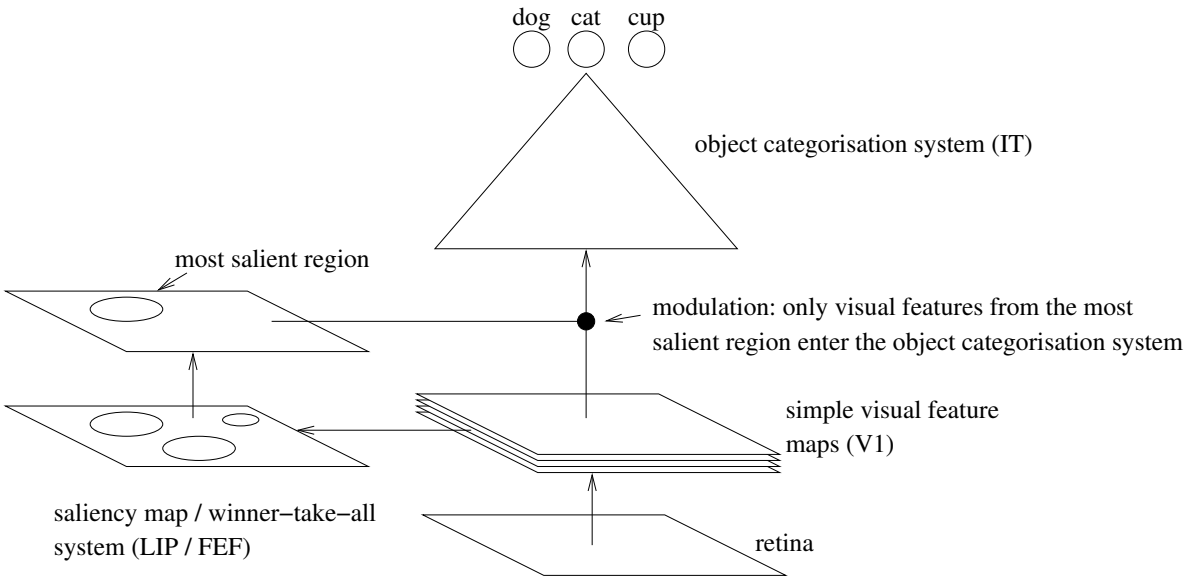


Figure 2.5: Attentional modulation of object categorisation

input to the object categorisation system, so that only visual features from this selected region are categorised.

The suggestion that attention modulates activity in the object categorisation pathway allows us to propose an initial ordering constraint:

Ordering constraint 1 *An object must be attended to (i.e its location must be activated in LIP/FEF) before it can be categorised (in IT).*

This constraint as stated is actually too simplistic. There is a famous proposal that focal attention is needed to properly classify an object (Treisman and Gelade, 1980), but much subsequent work has found that some degree of parallel classification of objects is possible. For instance, there are experiments indicating that subjects can categorise unattended objects in parallel with an attended object even if they are not aware of doing so (see e.g. Tipper, 1985). (There is also evidence that IT can encode multiple object types simultaneously; see e.g. Zoccolan *et al.*, 2005, and MacEvoy and Epstein, 2009 for similar evidence in humans.) However, the number of objects which can be classified in parallel appears to be quite small. The default procedure for categorising an object, especially in complex natural visual environments, is to attend to it first, probably overtly (see e.g. Rolls *et al.*, 2003).

2.4.4 The coordinate systems of LIP and FEF cells

Both LIP and FEF cells represent location using a retinotopic coordinate system. However, many LIP cells exhibit **gain field modulation**, in which the strength of firing of a cell is dependent on the position of the eye in relation to the head, or on the head in relation to the trunk (Andersen, 1997). This mechanism means that *populations* of LIP cells can provide unambiguous information about object location in coordinate systems centred both on the head and on the trunk. LIP thus delivers representations of location which are stable over movements of the eyes and head. FEF cells are also modulated by eye position (Balan *et al.*, 2003), so it is possible that a gain field mechanism operates here too.

2.4.5 Visual search by inhibition-of-return

One interesting phenomenon in visual attention is known as **inhibition-of-return** (**IOR**). IOR is manifested in a slowed response to recently-attended stimuli (Posner *et al.*, 1984); its purpose appears to be to prevent attention from becoming 'stuck' on a single salient stimulus, but instead to survey a range of different locations. There are several computational models of visual search which make use of an IOR mechanism; see in particular that of Itti and Koch (2001). Experiments demonstrating IOR typically involve presentation of a salient cue stimulus, followed after an interval by presentation of a target stimulus which the subject has to respond to. If the target appears in the same location as the cue, then

for certain interstimulus intervals, the subject's response to the target is slowed. There appears to be a mechanism for delivering a pulse of inhibition to recently-attended-to locations, at some fixed time after they were attended to. In models of the saliency map, this is often implemented as a set of inhibitory connections from points in the winner-take-all map back to corresponding points in the saliency map, as shown in Figure 2.6.

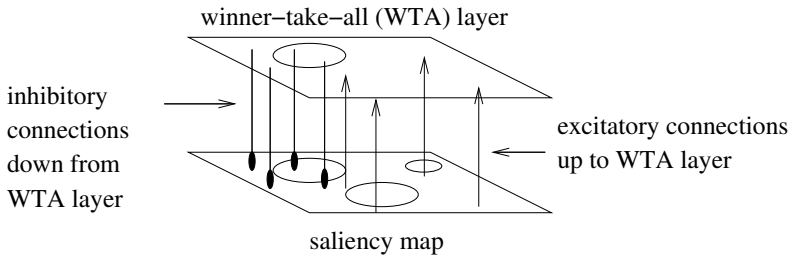


Figure 2.6: The saliency map, augmented with inhibitory connections implementing inhibition-of-return

Several brain regions have been shown to be involved in IOR. There is good evidence that it involves the superior colliculus (see e.g. Sapir *et al*, 1999). There is also evidence that it involves another subcortical region called the **basal ganglia** (see Poliakoff *et al.*, 2003): these authors found evidence of disrupted IOR in Parkison's disease, which is a disorder of the basal ganglia. It has also been found that FEF plays a role in IOR. Ro *et al.* (2003) found that transcranial magnetic stimulation applied to FEF midway through the interstimulus interval eliminated the IOR effect, while stimulation early during the interval had no effect; this suggests that FEF is involved in delivering the inhibitory pulse responsible for IOR.

It also appears that intraparietal cortex is involved in the IOR mechanism. Normally, IOR operates in a coordinate system which is stable over eye movements (Maylor and Hockey, 1985). The superior colliculus operates in retinal coordinates, so cannot provide this stability itself. However, LIP representations are stable over eye movements, as just discussed. It has been suggested that the remapping involves links between the superior colliculus and parietal cortex (Tipper *et al.*, 1997). In support of this, Sapir *et al.* (2004) found that patients with lesions to the intraparietal sulcus only generated IOR in retinal coordinates. This suggests that the circuit involving the superior colliculus and the LIP implements the kind of IOR function associated with visual search.

2.5 Vision for action: the reach and grasp motor circuits

I will now consider the remainder of the posteror parietal cortex, whose function is to generate motor actions as opposed to purely attentional ones. I will focus on a single kind of motor action, namely reaching to grasp a visually presented target object.

Reach-to-grasp actions have two separate components. The **reach** component comprises movements of the shoulder, elbow and wrist joints, to transport the hand to an appropriate position in relation to the target object. The **grasp** component comprises movements of the fingers and thumb, which preshape the hand and control the contacts it makes with the object. While these two components must clearly interact with one another, there are quite distinct parallel neural pathways for controlling reaching and grasping (Jeannerod, 1996). These two pathways are relatively separate until they converge in the primary motor cortex (F1), which is responsible for the low-level execution of all motor actions. I will begin by summarising the characteristics of F1. Then I will consider the reach and grasp pathways in turn.

2.5.1 The primary motor cortex (F1)

The primary motor cortex is organised **somatotopically**; that is, cells which are close to each other in this area elicit movements in functionally related muscle groups (See Donoghue *et al.*, 1992 for arm-related areas of motor cortex, and Schieber and Hibbard, 1993 for hand-related areas.) F1 cells encode a mixture of kinematic and dynamic parameters of movement—variables such as force, speed and direction of movement. These parameters are specified by populations of cells rather than individual cells, a scheme which is known as **coarse coding** (Georgopoulos *et al.*, 1982). The important point for the present review is that F1 provides relatively low-level instructions to muscle groups.

2.5.2 The reach pathway

The reach pathway which leads to F1 is internally complex, involving a number of interacting pathways in a region called the superior parietal lobule. Inputs to the system come from vision, but also from the somatosensory system, which contributes information about limb positions (derived from proprioception or from **efferent copies** of instructions sent to the motor system) and information about objects in the world (derived from touch). The output of the system consists of a relatively high-level specification of an arm movement, which is converted to dynamic and kinematic signals in F1 (see e.g. Kakei *et al.*, 2001).

This arm movement specification is a pivotal construct in understanding processes in the reach pathway; I will begin by introducing the construct more fully.

2.5.2.1 Movement vectors: high-level specification of reach actions

The high-level representation of a movement sent to F1 appears to take the form of a **movement vector**, which specifies the direction and distance of the target from the current position of the hand (Flanders *et al.*, 1992; Bullock *et al.*, 1998). To obtain a movement vector, we need to map the position of the hand and of the target into a common frame of reference; at this point, we can obtain the vector simply by subtracting hand position from target position. Much of the reach pathway is concerned with converting sensory data about the hand and the target into an appropriate reference frame, though it is still unclear exactly how this process occurs. It seems likely that multiple different frames of reference are used, and that different frames dominate in different tasks or contexts (see Battaglia-Mayer *et al.*, 2003 for a good summary).

Movement vectors appear to play a role both in the *selection* of a reach action (i.e. the decision about which target to reach for) and in the *execution* of the selected action. In the remainder of this section, I will sketch the basic algorithm, and then link different stages in the process to different points in the reach pathway.

2.5.2.2 Movement vectors for action selection and action execution

To begin with, in parallel, the locations of several different objects in the visual scene are transformed into a coordinate system permitting direct comparison with the current hand location, and a number of *alternative* movement vectors are computed. These movement vectors compete with one another: competition is biased 'bottom-up' by the properties of the movements themselves, with short movements being given an advantage (see e.g. Tipper *et al.*, 1998) and 'top-down' by task-based considerations about the relative benefits of different reach actions. Eventually, one movement vector is selected, and prepared, pending an internal 'GO' signal.

When the 'GO' signal is given, the selected movement vector takes on a dynamic role in determining the **trajectory** of the hand as the action unfolds. The trajectory of an action is a specification of the **motor state** of the effector (its position and velocity) at each moment during its execution. Robotic control systems typically plan a trajectory in advance of initiating movement, but humans and primates are more likely to generate action trajectories on the fly, by deriving an instantaneous motor signal directly from the movement vector in real time, while dynamically updating the movement vector as the position of the hand changes (Bullock *et al.*, 1998; Desmurget *et al.*, 1999; Sabes, 2000).

Following the 'GO' signal, the selected movement vector in parietal cortex is communicated to primary motor cortex, resulting in the initiation of a motor movement. We can think of the circuit which generates the movement as implementing a **motor controller** function. A motor controller takes as input the current motor state and the desired motor state (which in our case are jointly specified by the movement vector) and it delivers an appropriate motor signal as output. The motor signal alters the agent's hand position, taking it towards the target. Throughout the movement, parietal cortex maintains a representation of hand state (i.e. hand position and velocity) in real time derived from visual and proprioceptive mechanisms. The dynamically changing representation of hand state results in turn in a dynamically changing movement vector, with on-line consequences for the motor signal delivered at each moment. As the hand approaches the target, the movement vector reduces to zero, and so therefore does the motor signal, and the movement ceases.

Sensory feedback about the consequences of a current motor signal on hand state (also called **reafferent feedback**) takes time to arrive; if motor control depended solely on such feedback, control of the arm would be unstable (see Jordan and Wolpert, 2000 for a good review). To circumvent this problem, the sensorimotor system makes use of a **forward model** of the arm's dynamics, which allows it to generate a fast prediction about the hand state which will result from the current motor signal, to use as the basis for the next motor signal. Forward models also feature in action perception, and we will discuss them in some detail in Section 2.7.

We do not need to discuss the motor controller function in detail, but it is worth mentioning two points. Firstly, the function is likely to involve a combination of **feedback control** and **feedforward control**. In feedback control, the motor signal is a simple function of the difference between the current hand state and the desired state (at the target position with zero velocity). In feedforward control, the controller function incorporates a learned **inverse model** of the dynamics of the arm, which specifies for each pair of actual and desired hand states what motor signal to deliver. The feedforward controller can be learned by using the output of the feedback controller as an error term (Kawato *et al.*, 1987; Kawato and Gomi, 1992). The controller function generates movements whose trajectories have a smooth, bell-shaped velocity profile and are gently curved at normal speeds (Kelso *et al.*, 1979; Uno *et al.*, 1989). These characteristic trajectory shapes are likely to be due to the controller's being optimised to generate movements with minimum variance at the endpoints (Harris and Wolpert, 1998).

2.5.2.3 The components of the reach neural pathway

Having given an overview of the processes involved in selecting and executing reach actions, we can consider their neural implementation. A diagram of the reach pathway is given in Figure 2.7. The basic pathway involves a flow of information from visual and somatosensory

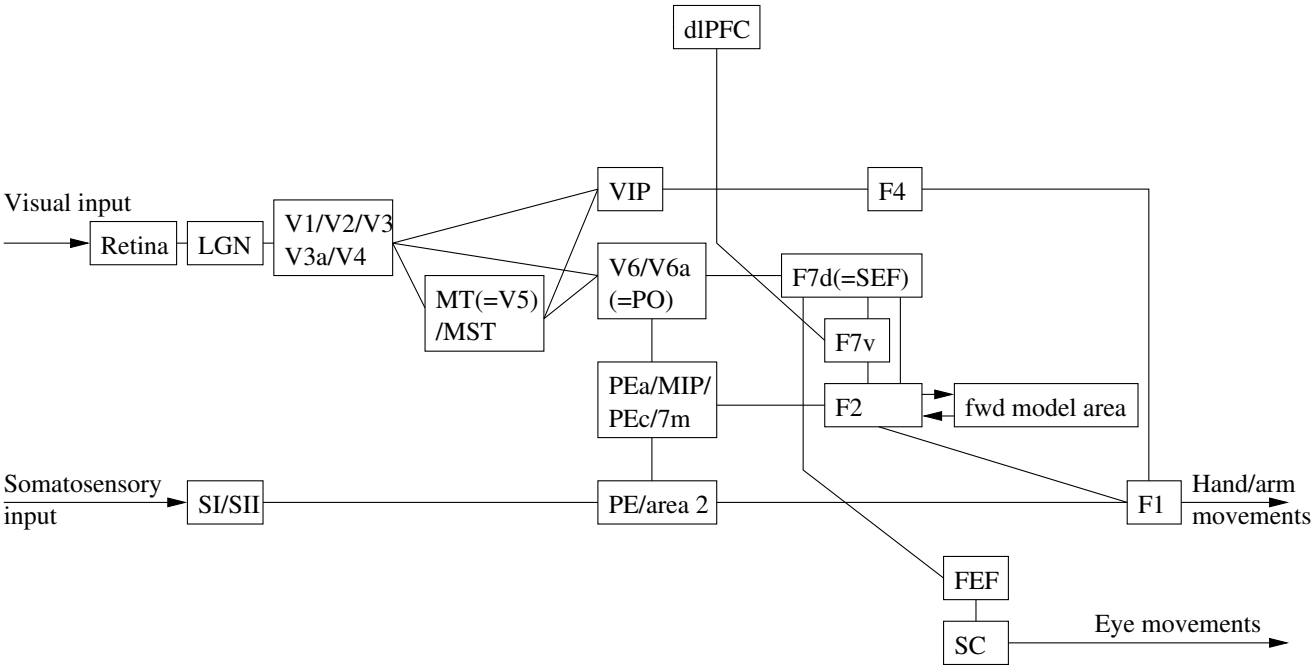


Figure 2.7: The 'reach' visual pathway

cortices to parietal cortex (VIP, V6, PE), and then to **premotor cortex** (F4, F7, F2, FEF), and then finally to primary motor cortex (F1).

The two main premotor areas involved in the reach pathway are F2 and F4. Both these areas are directly connected to F1. Cells in F2 and F4 provide a representation of the animal's **peripersonal space**; i.e. the space which is close enough to reach with with a movement of the limb or head (Colby and Goldberg, 1999). F4 cells provide a representation of the space around the head and body (Fogassi *et al.*, 1996); F2 cells focus on body rather than face regions (Fogassi *et al.*, 1999). Some cells in both regions are purely visual, responding to the sight of objects in peripersonal space. Others are purely somatosensory, with tactile receptive fields on particular parts of the hand, arm or face. A final group of cells are bimodal, responding either to the sight of an object very near

a particular arm/face region or to touch in this same region. F2 and F4 cells code visual and tactile information in a predominantly **effector-centred** coordinate system, in which (for instance) the location of an object in the monkey's perispace is given in relation to its current hand position. So these cells are encoding something like the movement vector introduced before.

The projections to F2 and F4 are from clearly segregated parietal regions (see Geyer *et al.*, 2000). F4 receives input from the ventral intraparietal cortex (VIP), whose neurons are similar to those in F4, though with a larger number of bimodal (visual/tactile) than unimodal neurons. The VIP-F4 pathway appears to be fairly specialised for recognising the position and trajectory of objects close to or looming towards the agent's head, integrating stimuli from several sensory modalities (Schlack *et al.*, 2005) and controlling reflexive defence behaviour (Cooke *et al.*, 2003) or reaching movements made with the head (Colby *et al.*, 1993; Colby and Goldberg, 1999). F2, on the other hand, appears more concerned with planned reach actions to manipulable objects in the animal's perispace. I will therefore focus on the pathways leading to F2 in this review. I will draw largely on the model of the reach pathway given in Burnod *et al.* (1999), which accounts for data from many studies of the relevant regions.

The pathway leading to F2 receives input from a mixture of somatosensory and visual areas. Purely visual input arrives via V1–V4 and from MT/MST; the former input provides information about the location of retinal stimuli, while the latter provides information about their motion. This information is combined in parietal area PO. Purely somatic information originates in the somatosensory centres SI and SII, and arrives in parietal areas PE and area 2. Visual and somatic information is combined in a set of adjacent parietal areas PEa, MIP, PEc and 7m, which I will refer to as the PEa complex.

These parietal areas communicate with a range of frontal areas with similar properties. PO is strongly linked to F7d, also known as the supplementary eye field (SEF), which and is another source of eye movement commands (Pierrot-Desseilligny *et al.*, 1995); note that it projects strongly to FEF, the frontal eye field. PE and area 2 are strongly linked to F1, providing direct proprioceptive feedback to modulate the motor signals developed in F1. Finally, the PEa complex is strongly linked to F2. While the PEa complex receives a mixture of inputs from PO and PE, F2 receives a mixture of inputs from F7d and F1. Thus while PO and F7d primarily encode visual and gaze information, and PE and F1 primarily encode somatic information, the PEa complex and F2 both integrate these two sources of information.

2.5.2.4 Mechanisms for action selection in the reach pathway

A key finding about the parietal and frontal regions just discussed is that the response properties of cells vary gradually between regions, so that region boundaries are in fact fairly blurred. Burnod *et al.* propose two dimensions of variation for cells in this complex of regions. Firstly, there is a **visual-to-somatic** dimension, which has already been described. Secondly, there is a **sensory-motor** dimension. Cells in the visual region PO and the somatic region PE are time-locked to sensory signals. Cells in F1 are time-locked to motor movements. Cells in intervening regions can either have pure sensory or pure movement-related responses, but there are also two interesting types of intermediate response.

Firstly, **matching cells** encode *associations* between sensory inputs and motor commands (represented as movement vectors); in other words, they allow a sensory input to be converted into a movement vector. In the account of Burnod *et al.*, matching cells acquire their sensitivity through simple **Hebbian learning**. In this type of learning, the strength of the connection between two cells is increased in proportion to the frequency with which they fire simultaneously during a training period. After training, the strengthened connection means that activity in either cell will tend to cause activity in the other. In the case of sensory cells and motor cells, the matching cells are trained when the agent issues arbitrary motor commands resulting in arbitrary movements (during a developmental phase called 'motor babbling'). While a given movement is under way, the agent experiences its visual consequences, and the matching cells therefore learn to associate the current motor command with a visual representation. This association is bidirectional, so that after training, when a visual stimulus is presented, the same matching cells evoke an appropriate motor command.

A second type of intermediate cells in the reach pathway are **condition cells**. These cells encode learned mappings from arbitrary sensory stimuli to motor commands; in other words, they provide an indication of which motor commands are useful to perform in which circumstances. The sensititity of condition cells is established through **reinforcement learning**. During training, arbitrary stimuli are paired with arbitrary motor commands; whenever the motor command results in a 'reward state' (whatever that might be), the association between the current stimulus and the current command is strengthened.

Matching cells and condition cells are found in several intermediate regions, but matching cells are particularly common in the PEa complex, while condition cells are particularly common in a frontal region called F7v, which neighbours and connects to F2, and also has strong projections to prefrontal cortex PFC (Geyer *et al.*, 2000). As already mentioned, PFC holds information about actions which have been learned to have beneficial consequences for the agent (see e.g. Freedman *et al.*, 2003). An important feature of both types

of cell is that they encode *prepared* motor commands, in advance of any overt action. A matching cell is typically responsive during movement in some preferred direction, even if visual signals are suppressed, so it certainly encodes a motor action; however, it also becomes active when an object affording this action is presented. So match cells appear to encode *posssible* motor commands. A condition cell is also active during movement in some preferred direction, but it is also activated by presentation of an appropriate contextual stimulus indicating that the movement is beneficial. So condition cells appear to encode *desirable*, or *prepared*, motor commands. In summary, it seems likely that the reach-related pathway up to and including F2 involves the computation of multiple alternative motor commands (each represented as a movement vector) followed by the selection of one of these commands, which is transmitted to F1. F1 and associated regions implement the motor controller function, which turns this command into a low-level motor signal, and an action is initiated. The monitoring of this action in turn involves visual and somatic representations which are distributed throughout the same network responsible for its selection. This network is now modulated so as to encode a single dynamically changing action rather than a set of potential actions.

There is considerable evidence that human agents compute a number of alternative movement vectors, and that the decision of which action to execute is partly determined by the properties of these different actions. For instance, Tipper *et al.* (1992) found that the effect of a distractor object in a reach task depends on its relationship to the actor's hand; distractors close to the hand, and hence more 'reachable', slow reach time more than distractors further away. Consistent with this study, Cisek and Kalaska (2005) have found neurons in the primate F2 and F7 which simultaneously encode the direction of reaching actions to two possible targets before the animal has selected between these two actions. (These neurons are particularly numerous in F7.)

In fact, activity in F7 is reminiscent of activity in LIP/FEF associated with choosing the target of an upcoming saccade (Section 2.4). Different neurons encode different objects and their associated reach actions. Neurons are activated bottom-up in measure of the easiness of their associated action, and top-down in measure of its projected usefulness. Competition between neurons results in a single action being selected; if the neurons have similar levels of activation, competition will take longer, with a corresponding delay before movement execution begins.

2.5.2.5 The location of the forward model in the reach pathway

As already mentioned, during the execution of a reach action, the state of the hand is likely to be updated by proprioceptive information, but also by a forward model of hand location driven by efferent copy of previous motor signals. Forward models are likely

to be contributed by several different brain regions. Cells in the reach pathway itself (particularly F2 and the PEa complex) appear to hold predictive information about the visual consequences of forthcoming actions (see e.g. Battaglia-Mayer *et al.*, 2003), as in fact do cells in LIP (Eskandar and Asaad, 1999; 2002). Also, there is considerable evidence that the cerebellum functions as a forward model during manipulation actions (see e.g. Kawato *et al.*, 2003); consistent with this role, there are strong connections from premotor cortex to the cerebellum and back (see Miall, 2003). In summary, like most other components of visually guided reaching, predictions about the current motor state appear to be distributed across several neural regions.

2.5.2.6 The location of the motor controller in the reach pathway

The motor controller function, which combines feedforward with feedback control, is also likely to be distributed over several neural regions, including the primary motor cortex, the basal ganglia and the spinal cord. However, the cerebellum has again been singled out as particularly important in the operation of this function, separately from its role in computing a forward model. In the model of the controller already mentioned (Kawato *et al.*, 1987), a feedback signal is combined with a feedforward signal generated by application of an inverse model of the dynamics of the effector; the inverse model is trained using the feedback command as an error term. There is good evidence that the cerebellum is involved in all of these processes (at least for eye movement control in monkeys); see especially Gomi *et al.* (1998); Kobayashi *et al.* (1998). There is also evidence that the cerebellum is involved in inverse models in humans (Imamizu *et al*, 2000). If the cerebellum is damaged, movement is typically still possible, but agility and efficiency are lost; for instance, arm movements are slower and more jerky, with oscillations during motion, especially as the target is approached (Holmes, 1939).

2.5.2.7 Integration of motor and attentional actions in the reach pathway

In the final two parts of this section, I will consider the role of the F7d (supplementary eye field) region. As already mentioned, F7d receives projections from PO, whose cells encode the location of visual stimuli, and F2/F7v, whose cells encode possible movement commands. It appears that the role of F7d is to generate eye movements subserving the planning or execution of motor commands (Pierrot-Desseilligny *et al.*, 1995).

There is indeed considerable evidence that the pattern of saccades which we execute when performing an action is very tightly integrated with the motor components of the action (Johansson *et al.*, 2001; Land *et al.*, 1999). For instance, in actions involving grasping objects or moving objects from one point to another, anticipatory saccades are

made to well-defined **control points**, such as the point where the hand will contact a reached-for object, or the location where an object is being moved to, or the edge of an obstacle being avoided. The great majority of saccades during transitive or ditransitive action execution are to these fixed control points; strikingly, the moving hand is never tracked by the eye (Johansson *et al.*, 2001). In a reach-to-grasp action, the agent typically fixates the target in the very early stages of the movement, before it is properly underway. We can state a constraint to this effect:

Ordering constraint 2 *During execution of a reach-to-grasp action, the agent typically attends to the target object in the early stages of the action.*

There are several reasons why it is useful to attend to the target of a hand action in advance. Firstly, while the eye is tracking the target, the orientation of the eye (in relation to the body) provides good online information about the location of the target. Secondly, attending to the target allows the agent to compute the movement vector directly from vision in the latter stages of the reach, when his hand appears in the periphery of the retina. If the target is guaranteed to be at the fovea, then the movement vector can be computed as a direct function of the retinal position (and speed) of his hand, permitting efficient visual guiding of the hand onto the target in the final stages of the action. Thirdly, the high resolution of the fovea allows detailed information about the form of the target to be computed, which is useful for computing a grasp, as will be discussed below. (Ballard *et al.*, 1997 provide detail on all of these reasons.) Finally, attention to the target creates a stable *positive feedback loop*, reinforcing its selection as the target of the reach action. There is thus an element of **recurrence** in the reach pathway; an object which is slightly preferred as the target of a reach action will receive increased attention, which in turn strengthens its representation as the preferred reach target.

A tight linkage with saccades is likely to be characteristic of all motor actions, or at least all highly practiced actions. Actions become more stereotyped and efficient over the course of practice; rehearsal of an action is frequently modelled as the development of an **action schema**, which coordinates the various different motor movements necessary to achieve it. Land and Furneaux (1997) demonstrate that well-practiced actions are also associated with characteristic patterns of eye movements; for instance, there are stereotypical patterns of eye movements associated with playing a stroke in table tennis or steering a car round a smooth corner. Following an idea originated by Norman and Shallice (1986), they propose a model in which an action schema integrates not only motor movements, but eye movements designed to provide relevant parameters for these movements at the right moments. In

other words, an action schema is a stereotypical *sensorimotor* construct, rather than just a motor one. Reaching for a target object is a sensorimotor process, not just a motor one. It seems likely that the F7d and SEF areas contribute to the attentional actions which are involved.

2.5.2.8 Attentional selection for perception and action

As already mentioned, the supplementary eye field F7d is heavily interconnected with FEF. Thus selection-for-action contributes to regular attentional processes, including modulation of activity in the early visual pathway to give a preference to the attended-to stimulus in the object categorisation pathway. The model being developed thus predicts that the kind of attentional selection involved in choice of a reach target biases selection for object categorisation—which indeed seems to be the case (see Schneider and Deubel, 2002 for very convincing evidence). Conversely, we predict that attentional selection of an object in FEF biases the motor system towards the execution of an action on this object.

2.5.3 The grasp pathway

While the reach pathway is mainly concerned with converting a visual representation of object *location* into *arm* movements, the grasp pathway is mainly concerned with converting a visual representation of object *shape* into *hand* movements. To clearly distinguish these two concepts, it is useful to note an ambiguity in the term 'movement', which can either mean 'articulation' or 'travel in a certain direction'. Thus the reach pathway generates an articulation of the arm which causes the hand to travel to an appropriate position, while the grasp pathway generates an articulation of the hand and fingers which causes the fingertips to travel to appropriate positions. We have already seen that the target position of the hand is given in a coordinate system representing the agent's peripersonal space. The target positions of the fingertips, on the other hand, are given in relation to the shape of the object. A useful model is given by Iberall and Arbib (1990), in which each kind of grasp in the agent's repertoire is analysed as involving two **virtual fingers**, whose fingertips apply opposing forces along an **opposition axis**. The position and orientation of the opposition axis depends on the location and orientation of the hand. The object, in turn, affords various different kinds of grasp, each of which can be defined as an opposition axis associated with the object itself, linking the points on its surface where the virtual fingers would have to make contact. The process of grasping the object can then be defined as the process of bringing one of the opposition axes of the hand into line with one of the opposition axes of the object (Iberall and Arbib, 1990).

The grasp pathway receives input from neural regions concerned with classifying the

shape of observed objects, and from somatosensory areas encoding the current shape of the hand as well as tactile sensations on the hand. The pathway is illustrated in Figure 2.8. It begins in the caudal intraparietal sulcus (c-IPS). Cells in this area produce a

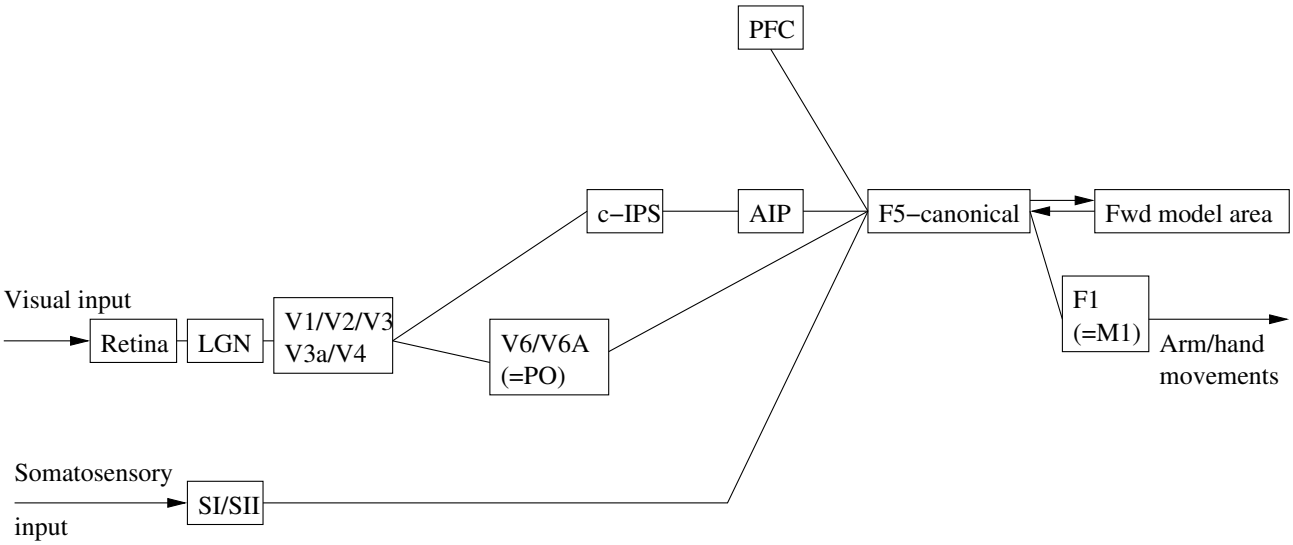


Figure 2.8: The 'grasp' visual pathway

representation of the three-dimensional shape and orientation of an observed object (see Shikata *et al.*, 2003 for evidence in humans). This representation is derived primarily from stereopsis: c-IPS receives input from cells in early visual areas (especially in V3) which encode **binocular disparity** (Sakata *et al.*, 1998; Tsao *et al.*, 2003). The caudal IPS projects to the anterior intraparietal area (AIP). Cells in the AIP are described by Taira *et al.* (1990) and in more detail by Murata *et al.* (2000). A typical cell in this area shows sensitivity a specific type of grasp action, which we can term its 'associated grasp'. Some cells are sensitive to a particular grasp when this grasp is executed; others are sensitive to a particular grasp when the agent attends to an object for which this grasp would be appropriate. **Motor-dominant** cells respond to their associated grasp action whenever it is executed (including if it was executed in the dark). **Visual-motor** cells have an attenuated response in the dark. **Visual** cells respond only in the light. A majority of visual cells respond not only when their associated grasp action is executed, but also when the monkey fixates a target whose geometric properties are such as to evoke their associated grasp. Several theorists have suggested that these latter cells code for the **motor affordances** of graspable objects (see especially Fagg and Arbib, 1998; Murata *et al.*, 2000). In

summary, some cells in AIP code for the actions afforded by observed objects, while others are involved in the planning or execution of these afforded actions.

There are strong reciprocal projections between the AIP and area F5. F5 neurons have been studied extensively (di Pellegrino *et al.*, 1992; Gallese *et al.*, 1996); see Rizzolatti *et al.* (2000) for a review. They appear to represent different actions of the hand and mouth, with the most common actions being different types of grasp (e.g. a precision two-finger pinch or a full-hand power grasp) or manipulation (e.g. hold or tear). F5 neurons are different from neurons in the primary motor cortex F1. Firstly, they appear to encode at a higher level of abstraction than motor cortex neurons. For instance, there are F5 neurons which respond equally to 'grasping with the hand' and to 'grasping with the mouth'. Secondly, many F5 neurons respond not only when the animal executes its preferred action, but also when a visual stimulus which *evokes* its preferred action is presented (similar to the 'match' cells in the reach motor pathway, which associate visual stimuli with motor responses). There are two main types of 'visual' F5 neuron. **Canonical neurons** are similar to AIP visual neurons; they respond when the animal observes an object for which their preferred action would be appropriate, even if the animal does not execute this action (Murata *et al.*, 1997). For instance, passive viewing of a pencil might evoke a 'precision pinch' canonical neuron. **Mirror neurons** respond when the animal observes *another agent* performing their preferred action. For instance, a particular mirror neuron might respond to a precision pinch rather than a power grasp whether the animal is performing the action itself, or watching some other agent performing the action. Mirror neurons will be discussed further in Section 2.7. For the present, the important point is that canonical neurons send output to primary motor cortex F1, resulting in execution of an appropriate motor action.

2.5.3.1 Canonical F5 neurons require attention to the target object

Recall from Section 2.5.2 that several alternative reach actions are computed in parallel in the reach pathway. When one is selected, a saccade will be generated to the chosen target, and a reach movement will be initiated. On the other hand, it seems that canonical F5 neurons only fire after the animal has attended to a particular target object. In other words, it appears that the animal first decides which object to reach towards—or at least, entertains a reach to a particular object—and only after this computes a suitable grasp. Note that this is consistent with what has already been said about the need for spatial attention to gate input to the object classification system. Attention needs to be allocated to a single object in order to determine its semantic type (see Section 2.2), and attention is presumably also necessary to determine its three-dimensional shape for the purposes of computing grasp affordances. We can therefore state an ordering constraint for execution

of a reach/grasp action:

> **Ordering constraint 3** *During execution of a reach-to-grasp action, the agent must attend to the target object before activating a detailed motor programme.*

This constraint is consistent with the observation that the agent attends to the target in the early stages of action execution (Constraint 2, Section 2.5.2.7.)

2.5.3.2 Influence of the reach pathway on F5 neurons

As already mentioned, while the reach and grasp pathways are relatively separate, there must be some communication between the two. This is especially true during the final stages of the reach-to-grasp action, when aligning the selected opposition axes of the hand and target object involves adjustments to the position of the hand in space. We therefore expect that computations in the 'grasp' pathway will exert an influence on the movement vector responsible for controlling the position of the hand. Indeed, there is evidence for links between the reach and grasp pathways; see Galletti *et al.* (2003) for a review. The two pathways make contact at various points; in particular, there are quite strong projections between F2 and F5 (see e.g. Marconi *et al.*, 2001), and both pathways receive input from V6A, part of the PO region (see Caminiti *et al.*, 1999). The upshot of this is that F5 canonical neurons encode not only grasp preshapes, but also aspects of the hand trajectory (Caminiti *et al.*, 1991). This point will be corroborated in our discussion of mirror neurons in Section 2.7.3.

2.5.3.3 Grasp selection and execution in the FARS model

A valuable model of the relationship between AIP and F5 canonical neurons is the FARS model of Fagg and Arbib (1998).[1] The model accounts for the finding that AIP neurons tend to encode all phases of a grasp action, while F5 neurons tend to be selective for a single phase (for instance finger preshaping, finger closing, holding or releasing). According to the FARS model, AIP computes affordances for a variety of different grasp actions on the object selected as the reach target, and communicates them to F5, which selects one of them and communicates the result back to AIP. As in the reach pathway, this **recurrent** connection from F5 to AIP creates a stable loop reinforcing the action which has been selected. AIP is then in charge of executing the phases of the action in sequence, by

[1] FARS stands for 'Fagg, Arbib, Rizzolatti and Sakata.'

selectively priming a succession of F5 actions. In short: following the selection of a target object, a variety of alternative grasps compete amongst each other until one is selected and then executed.

Each F5 action is monitored using sensory feedback about the hand, reporting factors such as 'maximum aperture reached' and 'contact with the object' (Fagg and Arbib, 1998). In the Fagg/Arbib model, this feedback is largely somatosensory, while in a subsequent model by Oztop and Arbib (2002), it is largely visual, deriving from a temporal region called the superior temporal sulcus (STS). The STS is also centrally involved in action perception, and its role in action execution is debated; see e.g. Iacoboni *et al.*, 2001. STS is certainly involved in the control of oculomotor actions (Maioli *et al.*, 1998), but its role in the control of hand/arm actions is not established. The action representations evoked in STS will be discussed further in Section 2.7.2. In the meantime, it is worth noting that the motor system may also receive visual information about hand shape from parietal areas more traditionally implicated in motor control. A likely candidate is V6A, which is sensitive to the form as well as the location of visual stimuli, and which causes impairments in grasping as well a in reaching when damaged (Galletti *et al.*, 2003).

An interesting feature of the FARS model is the influence of the inferotemporal cortex (TE) and prefrontal cortex (PFC) on F5 neurons. TE computes the category of the currently-attended-to object (Section 2.2), and PFC holds information about stimulus-response pairings learned through reinforcement. FARS suggests that TE and PFC exert an influence on grasp selection which is independent of the affordances computed in AIP. The main evidence for this is a study of a patient with lesions to the parietal visual pathway but with relatively intact inferotemporal and prefrontal cortex (Castiello and Jeannerod, 1991). This patient could preshape quite accurately when reaching for objects whose type allows an accurate estimation of size and shape (e.g. a pencil or a cup) but not when reaching for objects of other types (e.g. a cylinder). Fagg and Arbib suggest that inferotemporal cortex communicates an object category to PFC, which then conveys to F5 roughly what kinds of grasp are suitable for objects of the observed category. But detailed information about the *individual* object being reached for (including information about its orientation and other peculiarities) has to come from parietal cortex.

2.5.3.4 The role of forward models in grasp execution

A final note about reaching is that the execution of grasp actions is likely to involve a predictive forward model of the effects of motor commands to the hand and fingers; see Kawato *et al.* (2003) for some very solid evidence to this effect. As with reach actions, the cerebellum is likely to be involved in computing this forward model; note that F5 is reciprocally connected to the cerebellum.

2.5.4 Endpoint of the reach-to-grasp action: the haptic interface

At the end of a successful reach-to-grasp action, the agent is holding the target object in a stable grasp. There are two ways of characterising the state at this point. Firstly, as with any successful motor action, the agent has brought about a change in the state of the world; certain facts about the world become true which were not true before. Secondly, and less obviously, there is a change in the agent's *epistemic* state. A reach-to-grasp action can be characterised as an attentional operation as well as a motor one. When an object is held in a stable grasp, new sources of information about its location and identity become available: its location can be derived from proprioceptive information about the location of the agent's hand, and its shape (and frequently type) can be derived from a combination of tactile and proprioceptive information from the fingers and hand (see e.g. Goodwin and Wheat, 2004; Henriques *et al.*, 2004). A good example of this phenomenon is the 'Pinocchio illusion' (Lackner, 1988). Lackner found that vibrating a subject's bicep when it was flexed caused an illusion that it was extending. If subjects were holding their nose while the vibration was applied, they experienced the illusion that their nose was stretching.

There is a strong analogy between the state of 'haptic attention' to an object and the state in which the agent visually attends to it. Just as spatial attention gates the input to the visual object classification system, so the tactile sensation of a stable grasp allows information about the shape of an object at the associated location to be read from the current configuration of the hand. Indeed, there is considerable evidence that haptic and visual attention access a common reference frame (see in particular Driver and Spence, 1998; Spence, 2002; Macaluso *et al.*, 2002). We have already seen that an agent initiating a reach action typically attends to the target object. During the action, it is likely that his visual attention is spread to some extent between the target object and his own body, especially in the closing phases of the action. However, the endpoint of the action is characterised by strong *reattention* to the target object, in a motor modality. We can state another constraint to this effect:

Ordering constraint 4 *After completion of a reach-to-grasp action, the agent reattends to the target object (in a haptic modality).*

Note that haptic attention the target at the end of a reach-to-grasp action provides an important opportunity for learning cross-modal representations of manipulable objects. As discussed in Section 2.5.3, when an agent initially looks at a target object, he evokes its grasp affordances—i.e. the motor state he must achieve in order to obtain a grasp it. The

state in which an agent is holding an object and also visually attending to it constitutes an ideal opportunity—perhaps the only opportunity—to train the system which maps from the visual form of an object to an associated motor state. So it is likely that haptic reattention to the target object plays a crucial role in the formation of our concepts of manipulable objects.

One final question: how can we understand the notion of 'selective attention' within the motor modality? A useful idea is that the body's motor system has several relatively distinct components—i.e. motor subsystems which are relatively independent. Thus there could be a motor system for each limb, one for the torso, and also motor systems for combinations of limbs (e.g. a system which controls both arms, and which is involved in picking up large or heavy objects, or a system which controls both legs, and is involved in walking). Haptic attention could be the activation of one particular motor system. If the motor system is involved in controlling a tool, then activation of the motor system is a form of motor attention to the tool. If no tool is being controlled, then activation of a particular motor system could be a special way of attending *to parts of one's own body*. There is indeed evidence that agents have a special modality for attending to parts of their own body, which exploits representations used in the control of specific body parts or ensembles. This modality is often referred to as the **body schema** (see e.g. Schwoebel and Coslett, 2005). In our example action, then, the haptic establishment of the cup could be thought of as a form of motor re-attention to the cup, which was originally established in the visual modality.

2.6 Planning higher-level actions: prefrontal cortex and 'higher' motor areas

In the above review, we have only been considering simple reach-to-grasp actions: once an object has been selected as a target, the only choice to be made is how to grasp it. In a natural situation, of course, the question of 'what action to do' is far less constrained. When the agent perceives an object in a scene, we might expect a range of alternative actions on this object to be evoked, some of which are more complex than reach-to-grasp, such as lifting, pushing, putting and so on. Two questions then arise. Firstly, how are distinct actions such as 'lift', 'push', 'put' etc represented in the motor system? Secondly, what mechanism decides which of these actions to perform in a given situation? I will address both questions in this section.

2.6.1 The representation of 'action categories' in the motor system

There are a range of distinct hand actions an agent can perform on a target object. In language, these distinct actions are categorised using verbs like 'take', 'hit', 'snatch', 'push' and so on. Action categories are defined functionally—that is, in relation to the effect they bring about on the object (see e.g. Hommel *et al.*, 2001; Hauf *et al.*, 2004). However, the motor system must also have a lower level representation of action categories, encoding the mechanisms by which these effects are achieved.

At this lower level, action categories vary along two dimensions. Firstly, they require that the hand takes different trajectories in relation to the target object. (For instance, a 'squash' action will involve a trajectory which brings the hand to a point above the object and then down onto it, a 'hit' aims to have the hand contact the object with a certain velocity, and so on.) Secondly, they require different hand shape sequences. (For instance, a 'slap' requires a rigid open hand, while a 'punch' requires a closed fist.) Note that hand trajectory and hand shape must often be defined for a period after contact is made with the target as well as before it; for instance 'pick up' requires a grasping hand posture to be maintained while the hand is raised; 'push' requires the hand to maintain contact with the target while the hand is moved. When producing a hand action, the agent has to select an appropriate action category as well as an appropriate target.

2.6.1.1 Parameterised action categories

The characteristic hand trajectory associated with a transitive action category must be defined relative to the location of the target object. Until now, we have assumed that the motor system's representation of the target location (which we termed the 'movement vector') is sufficient to determine the hand trajectory. This may be true in simple 'point-to-point' movements, but more generally, hand trajectory must probably be defined by a combination of a movement vector (specifying the location of the target) and an action category (specifying a sequence of sub-movements in relation to this target position). There are several ways the required 'relative trajectory' could be implemented. For concreteness, I will assume that the action category delivers a sequence of biases to the movement vector, so that—for instance—in a 'hit' action, the actual position reached for is a point beyond the location of the target, while in a 'squash' action, the position reached for is initially some distance above the actual target, and then some distance below it.

Similar points can be made for hand preshapes. While some action categories specify a preshape without regard for the shape of the target object (e.g. a slap always involves an open palm), others specify parameters on the grasp type which is afforded by the object.

For instance, 'pick up' involves a hand preshape which is defined by one of the opposition axes associated with the target; 'squeeze' presumably involves a similar hand preshape but with relatively more force applied.

2.6.1.2 Where are action categories represented?

Hand/arm action categories must be represented in an area with four properties. Firstly it must project to both the reach and the grasp pathways, so that the appropriate movements are planned in each pathway. Secondly, it must support the planning of *sequences* of simple actions. Thirdly, it must represent high-level *categories* of action, rather than specific actions. Fourthly, it must represent actions functionally, in terms of their effects, rather than just as movements.

There are several plausible candidate locations. In Fagg and Arbib's (1998) model of grasping, units which represent sequences of hand gestures (opening, closing etc) are present in AIP, in the 'grasp' pathway (see e.g. Taira, 1990). There is also evidence of generalisation; for instance, some F5 cells represent grasp actions whether they are performed by the hand or by the mouth (Gallese *et al.*, 1996). However, action categories involve both hand gestures and arm movements, and often specific attentional movements as well; it is thus likely that they are planned at a level which interfaces with both grasp and reach pathways, as well as attentional visual pathways. And the amount of generalisation in the AIP/F5 system is quite limited.

The most plausible location for action categories is prefrontal cortex—especially dorsolateral prefrontal cortex. PFC has strong projections to reach and grasp pathways, and also to the pathways which control visual attention. In addition, it is well established that PFC is involved in high-level executive control and supervision of behaviour (see e.g. Miller and Cohen, 2001). Moreover, there is good evidence that the dorsolateral PFC is involved in the preparation of action sequences, which I will review in detail in Section 3.2. There is also evidence that actions are represented in lateral PFC by their effects, rather than as movements; see e.g. Saito *et al.* (2005) for evidence from single-cell recordings in monkeys. Finally there is evidence that lateral PFC encodes high-level generalisations which group sequentially structured action actions together (Shima *et al.*, 2007). Tanji *et al.* (2007) conclude that lateral PFC is where 'action categories' (understood as high-level, abstract, sequence-encoding, functionally defined movement plans) are stored. Of course it is still a jump to say that this area is where 'lexicalised' hand/arm action categories like 'grab', *hit*, *squash* etc are stored. But it is certainly the most likely candidate, based on sensorimotor neuroscience.

Two other areas involved in the planning of higher-level actions are the **supplementary motor area** (SMA) and the **pre-supplementary motor area** (pre-SMA). Cells in these

areas have many of the same sequence-encoding properties as PFC cells (see e.g. Shima and Tanji, 2000; Tanji *et al.*, 2007). It is quite likely that high-level action representations are held here too. However, for the moment I will focus on the role of PFC; I will consider the relationship between PFC and SMA/pre-SMA in Section 3.3.2.

2.6.2 Top-down action biasing in PFC: Miller and Cohen's model

Deciding on a suitable action category is partly a matter of bottom-up affordances (creating a preference for the actions which can be performed with least effort), but also crucially a matter of learned outcomes. The agent will attempt to perform an action, or a sequence of actions, which has led to beneficial consequences in the past. In our cup-grasping scenario, it is likely that the agent performs the action of grasping the cup rather than some other action because it achieves, or leads to, some positive outcome for him. This predisposition towards a particular action is often referred to as **task set**.

How and where is task set represented? Again, there is evidence that PFC plays a central role. Damage to PFC does not tend to prevent an agent from responding quite naturally to stimuli, but it does make it hard for an agent to suppress automatic or prepotent responses to stimuli in situations where they are not appropriate, and to switch from one task to another (see Miller, 2000 for a review). In this section I will outline an influential model of how the PFC imposes a particular task set, due to Miller and Cohen (2001).

The key idea in Miller and Cohen's model is that the PFC holds a set of biases about how sensory stimuli should trigger motor responses in the agent's current circumstances. The model is illustrated in Figure 2.9. This figure shows a schematic stimulus-response

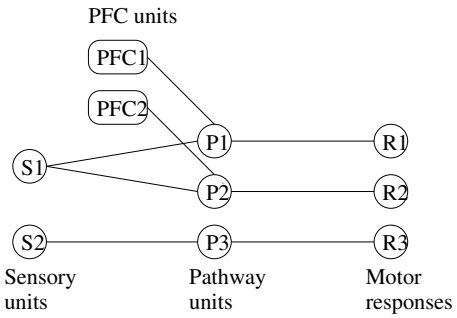


Figure 2.9: Miller and Cohen's (2001) model of PFC neurons

pathway, comprising a sensory layer, a response layer and one intermediate layer. The

sensory unit S1 permits two responses: R1 (via pathway unit P1) and R2 (via P2). Pathway units compete with one another; which response is chosen depends on which intermediate unit is most strongly activated. In the absence of top-down control, the winner will be the pathway unit which is most strongly activated by S1; i.e. the unit with the strongest connection to S1. In Miller and Cohen's model of the PFC, top-down control is provided by PFC units activating specific intermediate units, and thereby influencing the competition between them. Thus if PFC1 is active, there is a bias towards P1 (and therefore R1), and if PFC2 is active, there is a bias towards P2 (and therefore R2). (Clearly it is unrealistic to assume a single pathway unit for each possible mapping from stimulus to response which the animal might want to learn, but the localist model of pathway units is at least useful for presenting the basic idea.)

In Miller and Cohen's model, active PFC units do not generate responses by themselves—rather, they determine under what circumstances different responses will occur. A situation where PFC1 is active can be interpreted as 'a state of readiness to perform the action R1 upon presentation of S1'. This state of readiness can be initiated some time before the R1 action is actually produced. Note that a PFC state can also be understood as one in which established responses to certain stimuli are *suppressed*. For instance in Figure 2.9 neither PFC unit introduces a bias towards P3. If either PFC unit is active, the result will be a blocking of the established pathway from stimulus S2 to response R3, since P1 and P2 both compete with P3. In either case, the stimulus S2 will effectively be ignored.

The schematic pathway in Figure 2.9 can be quite effectively mapped onto Burnod *et al.*'s (1999) model of the parieto-frontal pathway for reaching as described in Section 2.5.2. In Burnod *et al.*'s model, there are four types of cell: sensory cells, motor cells, matching cells (which encode the responses afforded by different sensory stimuli) and condition cells (which encode learned mappings between arbitrary sensory stimuli and motor responses). Matching cells correspond well to pathway cells in Figure 2.9, while condition cells correspond well to PFC cells. Note that condition cells are particularly common in F7v, which projects strongly to PFC. It is therefore possible to think of Miller and Cohen's PFC cells as carrying the same sort of prepared responses as Burnod *et al.*'s condition cells.

There is also good evidence for cells in PFC itself which encode mappings between stimuli and responses learned through reinforcement. For instance, Asaad *et al.* (2000) trained monkeys on two separate tasks, which used the same sets of cue stimuli and response actions, but required different mappings between stimuli and responses. They found a large number of PFC neurons whose activity was higher for one task than for the other, both at the time of cue onset and during the period prior to cue onset when the monkey was preparing suitable conditional responses. These PFC neurons can plausibly be interpreted as having a role in selecting a particular stimulus-response pathway.

Note that the PFC is assumed to be a *high-level* executive controller. In Figure 2.9,

the stimuli S1 and S2 are not low-level sensory stimuli; rather they are high-level representations, of the kind which are evoked in inferotemporal cortex. Similarly, R1–R3 are high-level motor programmes; perhaps assemblies in premotor cortex, or perhaps the PFC-based action categories introduced in Section 2.6.1. (Note that since action categories are defined functionally, by their effects, the PFC's role in representing action categories is intimately related to its role in mapping stimuli to useful responses.) A PFC stimulus-response rule might specify: 'if you are thirsty and you are attending to a cup, execute a 'grasp' action'. This means that a grasp action is only likely to be triggered *after* a fairly high-level representation of a candidate target has been produced—i.e. after the candidate target has been categorised in inferotemporal cortex.

2.6.3 Summary

This section has proposed two related roles for PFC in the preparation of actions. Firstly, PFC is involved in defining individual transitive actions such as 'grab', 'hit' and 'push'. These actions are defined as sequences of basic motor commands, parameterised by the affordances of the target object, and their implementation is partly devolved to effector-specific regions such as AIP. Secondly, PFC is involved in specifying top-down preferences for particular actions in particular circumstances, learned through reinforcement. In Miller and Cohen's model, these preferences are defined as activity-based biases on particular stimulus-response pathways.

2.7 The action recognition pathway

There has been an enormous amount of interest in the cognitive processes underlying action recognition in recent years. This interest was sparked by the discovery of mirror neurons in the premotor area F5 of monkeys (di Pellegrino *et al.*, 1992; Gallese *et al.*, 1996). As already mentioned in Section 2.5.3, F5 mirror neurons respond selectively to particular types of grasp or manipulation action, whether the action is performed by the animal itself, or by an agent the animal is observing. The existence of such neurons suggests that the representations which underlie the perception of actions in others are at some level the same as those involved in the execution of actions. From this it can be argued that we recognise the actions of another person by simulating their execution ourselves at some sub-threshold level. The key attraction of this model is that it creates a link between a relatively superficial *visual* representation of the movements of an observed agent and a much more profound representation of the intentions and goals which underlie the execution of our own actions. Humans can clearly make this link; we are able to develop very complex

representations of the mental and motivational states of agents whom we observe. Recall from Section 2.6 that our own actions are strongly influenced by representations in the prefrontal cortex, a network of learned mappings between highly derived sensory stimuli and complex patterns of conditional motor responses. Mirror neurons suggest a route by which the perceptual system can access these derived representations, and hence permit the rich analyses of the mental state of observed agents which humans so obviously perform.

The discovery of mirror neurons led to a veritable industry of research into possible commonalities between the representations underlying action execution and observation in humans. In humans, evidence from imaging studies shows that there are overlaps in the areas of the brain active during action perception and during action execution (see e.g. Rizzolatti *et al.*, 1996; Jeannerod, 1999; Iacoboni *et al.*, 1999; Iacoboni *et al.*, 2001; Buccino *et al.*, 2001). There are also studies demonstrating that perception of an action results in sub-threshold activation of the motor systems which would be activated if this action was executed by the observer. The most sophisticated of these use transcranial magnetic stimulation to amplify the signals in the motor cortex of human subjects while they observe actions (see Fadiga *et al.*, 2005) for a review. In the first such study (Fadiga *et al.*, 1995), subjects watched reach/grasp actions and simple arm actions. It was found that the muscles activated by the amplified motor signals were indeed the same as those activated during performance of similar actions. More recently, it has been found that the temporal dynamics of evoked muscle activation during observation of a reach action mirrors the dynamics found during performance of the same reach action (Gangitano *et al.*, 2001). This is strong evidence for a mirror system for action recognition in humans. It is corroborated in a wide range of studies using different techniques; for instance, it has been found that observers breathe more deeply when watching effortful actions (Paccalin and Jeannerod, 2000); and that when people watch themselves perform an action, they are better at predicting its effects than if they watch someone else performing the same action (Knoblich and Flach, 2001).

All these findings have led to numerous theories about the common neural representations involved in action execution and action observation; see e.g. Gallese and Goldman (1998), Jeannerod (2001), Hommel *et al.* (2001), Iacoboni *et al.* (2001), Keysers and Perrett (2004). Recently, many of these theories have been presented as computational simulations of various aspects of the mirror system for action representation; see in particular Oztop and Arbib (2002); Demiris and Hayes (2002); Schaal *et al.* (2003); Oztop *et al.* (2005). I will draw on several of these theories in the model I present.

To set the stage for an account of the neural pathways involved in action observation, I will begin in Section 2.7.1 by describing some of the observable temporal structure of the action observation process, focussing on studies of eye movements during observation of reach-to-grasp actions. In Section 2.7.2 I will discuss the representations of actions

which are developed in the superior temporal sulcus (STS) during action recognition. In Sections 2.7.3 and 2.7.4, I will review the properties of mirror neurons in F5, and in a second area in inferior parietal cortex called PF/PFG, which links F5 to STS. In Sections 2.7.5–2.7.7, I will present a model of the mirror neuron circuit and compare it to other existing models.

2.7.1 The attentional structure of observation of reach-to-grasp action observation

It has recently been found that when an observer watches another agent perform a reach-to-grasp action, the observer's eye movements closely resemble those of the agent. Recall that when an agent executes a reach-to-grasp action, he attends to the target object early in the process (see Section 2.5.2.7), before a detailed motor programme has been activated (see Section 2.5.3.1). Experiments by Flanagan and Johansson (2003) and Rotman *et al.* (2006) have shown that the observer of a reach-to-grasp action executes a similar anticipatory reach to the target object. These authors suggest that the mirror system for action recognition involves attentional operations as well as motor schemas. On this view, the observer of an action establishes the same attentional state as the agent, so that the same couplings between attentional and motor processes are in force during observation of the action as during its execution. I will refer back to this anticipatory saccade to the target several times in the forthcoming discussion.

It is also interesting to consider when the observer of a reach-to-grasp action attends to the agent of the action. We know this must happen at some point. For one thing, the agent needs to be identified or categorised, and we have already seen that object categorisation typically requires attention (see Section 2.4.3). For another thing, the agent is the one whose action must be identified, and it is likely that this process requires attention of some sort on the agent. But we do not know at what point the agent is attended to.

There are two reasons for expecting the agent to be attended to before the target. Firstly, agent is often a more salient cue for bottom-up attention than the target. The agent must be animate, and is likely to be moving more, or at least earlier, than the target. It is known that movement is a strong bottom-up attentional cue (see e.g. Itti and Koch, 2001), and that initiation of movement captures atttention (see e.g. Abrahams *et al.*, 2003). Secondly, we have just seen that the observer can anticipate the agent's intended target: inferring the agent's intentions is likely to require some kind of attention to the agent.[2] A recent study by Webb *et al.* (in press) found that observers of a reach-to-grasp

[2]Sometimes the observer may be able to guess the target from context or world knowledge. But it has been shown that the observer can anticipate the target even if there is no top-down information to help

action do indeed tend to saccade to the agent before making an anticipatory saccade to the target. In fact, observers do not always need to see a complete agent in order to make an anticipatory saccade; extrapolation from hand trajectory appears to provide enough information. But in a situation where observers must identify the agent of an action as well as monitor the action itself, we can say that they typically attend to the agent before the patient.

Ordering constraint 5 *During observation of a reach-to-grasp action, the observer typically attends to the agent before attending to the patient.*

2.7.2 STS: biological motion recognition, joint attention and target anticipation

The superior temporal sulcus is well suited for the task of action recognition. Firstly, action recognition is thought to involve the recognition of characteristic body positions, but also of characteristic patterns of movement. An area in the STS called STP receives projections from the latter stages of the ventral object recognition pathway (TE) and also from regions in the dorsal pathway concerned with motion processing (MT and MST)—see Oram and Perrett (1996). Secondly, the recognition of grasp actions is thought to involve a mechanism called **joint attention**, in which the observer monitors where the observed agent is looking, and directs his attention to the same point. STS is also involved in controlling joint attention. In this section, we will consider both of these functions of the STS in more detail.

2.7.2.1 STP and biological motion recognition

The starting point for much modern work in action recognition was the finding by Johansson (1973) that subjects can recognise actions purely on the basis of patterns of motion, without any information about form. Johansson created films of an agent in which the only visible points were lights attached to the agent's limb joints. Under these conditions, if the agent remained motionless, subjects perceived only a cluster of lights, but if the agent executed an action, subjects instantly recognised the action. The strong percept of a walking person created by these point-light stimuli appears to be created by an analysis of the relative motions of the joints; certainly, no such percept emerges from any single

(see Rotman *et al.*, 2006; Webb *et al.*, in press). In this case, the information about the intended target can only come from the agent.

frame of the stimuli. It therefore appears that there is circuitry in the brain specialised for recognising 'biological motion', or perhaps more generally, the motions of articulated objects.

Most work with point-light stimuli has focussed on whole-body actions such as walking and running. However, we have also found that transitive action categories can be reliably identified from point-light stimuli (Knott, Mackie and McCane, unpublished data). We showed subjects point-light displays of an actor performing three different transitive actions ('grab', 'squash', 'push') on a target object (which was also identified by a single point in the display). Subjects were able to identify the action with over 90% accuracy. One possible explanation of this result is that subjects were ignoring the biological motion signal in the stimuli and simply monitoring the trajectory of the actor's hand in relation to the target. However, subjects' categorisation performance did not fall off much even if they were shown point-light stimuli in which the point on the actor's hand was removed. It thus seems that biological motion processing accounts for at least one component of transitive action recognition.

There is a strong consensus that STP is involved in biological motion recognition. Most directly, Oram and Perrett (1994) investigated the responses of STP cells to action stimuli in macaque. They found cells which were active during specific actions, whether these were presented in full illumination or as point-light stimuli. Other studies found cells in STP which responded to specific mouth and finger movements (Allison *et al.*, 2000) and to specific movements of the arm (Jellema *et al.*, 2000). In humans, there is also good evidence from imaging studies that STS is involved in the processing of biological motion; see e.g. Bonda *et al.* (1996), Servos *et al* (2002), Grossman (2000). Given our emphasis on transitive hand actions, it is also important to note that STS is sensitive to isolated hand actions, not just whole body actions (Pelphrey *et al.*, 2005).

As already mentioned, STP receives strong projections from TE (which encodes the form of observed objects) as well as from motion areas MT and MST, so it is sensitive to form as well as to motion. The relevant form cues are presumably transitory view-specific body poses, which when viewed in an appropriate order provide information about actions. There is evidence that STP cells encode such poses (see e.g. Oram and Perrett, 1996), and that cells in STP integrate form and motion stimuli (Oram and Perrett, 1996). Consistent with this, Grossman and Blake (2002) found that while only STS was activated by point-light stimuli, during normal observation of actions, both STS and TE were activated. In summary, it appears that 'normal' action recognition uses both form and motion cues. The resulting picture of STS is as a region performing a classificatory function similar to that performed by TE, but integrating information about form and motion across a succession of time points (see Giese and Poggio, 2000 for further discussion and a computational model).

2.7.2.2 STS, joint attention and early attention to the agent

Both humans and monkeys have a well-developed facility for **joint attention**; i.e. for monitoring the gaze of an observed agent to identify what the agent is attending to, and attending to this same object themselves. In fact, humans find it impossible to ignore the eye gaze of an observed agent when deciding how to allocate attention when performing a task (Driver *et al.*, 1999), and early visual processing is improved at the point the observed agent is attending to (Schuller and Rossion, 2004). Joint attention is one likely mechanism by which the observer of a reach-to-grasp action can infer the agent's intention, and saccade to the target in advance (as discussed in Section 2.7.1).

Joint attention is not the only means by which observers can identify the intended target. In fact, in the experiments by Flanagan and Johansson and Rotman *et al.*, observers appeared to maintain attention on the actor's 'workspace', and did not attend to the actor's eyes at all. Even in infants, the intended target of an action can be identified without any information about eye gaze (Woodward, 1998). In all these experiments, it is likely that the intended target was established by extrapolating the hand's future position from its current trajectory; several models of action recognition in fact assume this mechanism (Oztop and Arbib, 2002; Oztop *et al.*, 2005). However, normal action recognition probably involves monitoring of the agent's eye gaze. Webb *et al.* (in press) found that observers execute a saccade to the agent, often to the agent's face. And frequently they saccade to the correct target before the agent's hand has moved far enough to provide accurate information about it. Castiello (2003) also found that adult observers of human reach actions monitor the agent's gaze, and derive information about its intended target, as well as about distractor objects which the agent is ignoring. If subsequently the observers have to imitate the action themselves, their reach trajectory is influenced by the observed action. Interestingly, this influence is present even if the observer just sees the agent *look* at the target, without reaching for it. The important cue for determining the intended target seems to be the agent's eye gaze rather than the orientation of the agent's torso. In summary, there is good evidence that joint attention plays a role in the recognition of reach actions.

There is considerable evidence for specialised mechanisms for gaze monitoring in the STS. For instance, many STS cells are responsive to particular views of the head, and to particular gaze directions (see especially Perrett *et al.*, 1985; Perrett *et al.*, 1992). Similar observations have been obtained from fMRI studies of humans (Hooker *et al.*, 2003; Pelphrey *et al.*, 2005). There are projections from STS to parietal regions within the attention pathway, which allow the generation of appropriate reflexive biases on spatial attention (Maioli *et al.*, 1998; Kingstone *et al.*, 2000). In summary, STS cells are involved in establishing joint attention as well as in classifying actions.

2.7.2.3 Joint attention interacts with biological motion recognition

Clearly, STS is in a good position to integrate information about the movements of an observed agent with information about the agent's direction of attention. Indeed, many of the STP cells which encode action representations are also modulated by joint attention. Jellema *et al.* (2000) found neurons in STP which respond to an observed hand motion towards a target object, but only if the observed agent is *looking* at the target while executing the motion. Perrett *et al.* (1990) report many STP cells which are sensitive to hand actions involved in manipulating objects, but only in the presence of these objects; they are not activated by the hand action alone (or the object alone). These findings provide clear evidence that the biological motion system in STP does not simply classify 'reach movements'. Instead, they rely on an intended target having been identified, and monitor the progress of the observed agent's hand towards this target. We can thus note another another ordering constraint:

Ordering constraint 6 *During observation of a reach-to-grasp action, the observer must attend to the target before the movement can be monitored in STS.*

Note an analogy with visual processing during the execution of reach actions; here too, the agent's visual attention is on the target object while the action is underway. As we have already seen, the visual computations that occur during the monitoring of one's own reach actions are to do with the position and direction of the hand *in relation to the target.* There is evidence that STP computes similar information when monitoring the reach actions of another person.

Given that STS computes information similar to that used to control one's own reach movements, it is worth asking whether STS is in fact involved in motor control. Some theorists have seen STS as the region providing information about the motion of both arm and fingers (Oztop and Arbib, 2002). However, as already discussed in Section 2.5, the prevailing view is that these functions are mainly provided by visual and somatosensory parietal areas. Certainly, the great majority of action-sensitive cells in STP only respond to the actions of observed agents, and are not sensitive to similar actions in the agent (see e.g. Carey *et al.*, 1997, Keysers and Perrett, 2004). Recently, a region of the STS has been discovered containing cells which respond during the agent's own hand actions, even in conditions where the agent has no sight of his own hand (Iacoboni *et al.*, 2001). However, Iacoboni *et al.* argue that these cells are not involved in motor control, but rather in training the mirror system; they are suggested to provide the STS with a copy of the visual stimulus expected as the result of one's own motor commands, from which it can

learn to map the observed movements of other agents onto its own motor representations. The suggestion that STS plays a relatively minor role in motor control is corroborated by evidence from Rizzolatti *et al.* (1996), which shows that STS is activated in humans during observation of grasp actions, but not during execution of similar actions.[3]

2.7.2.4 Action monitoring and attention

It is interesting to consider the time course of attention while 'action monitoring' is under way in STS. As already mentioned, the observer typically makes a saccade to the target of the agent's reach in anticipation of the hand reaching it (Flanagan and Johansson, 2003), and maintains this fixation until the target is reached. Given that the agent's arm and hand are spatially separated from the target object, this means that for most of the observed reach, biological motion processing must rely on somewhat peripheral visual stimuli, which are not the primary focus of spatial attention. It is thus interesting to consider how biological motion processing operates in relation to attention.

A number of different strands of evidence suggest that attentional constraints are somewhat more relaxed in the biological motion pathway (STS) than in the object categorisation pathway (TEO-TE). Firstly, biological motion processing is not interrupted by saccades to different points within the stimulus (Verfaillie *et al.*, 1994). Secondly, it appears that STS cells are sensitive to stimulus location as well as to form and motion (Jellema *et al.*, 2004; Baker *et al.*, 2001); this distinguishes them from cells in TE, and may facilitate parallel processing at more than one location. Most directly, smooth point-light stimuli can be processed quite well even when combined with a secondary attention-demanding task (Thornton *et al.*, 1999).[4] On the other hand, given that normal action recognition relies on form information provided by TE as well as motion information, and that object categorisation is held to require attention, normal action recognition probably requires some

[3]I agree with Iacoboni *et al.* that the STS neurons they observed are not likely to be involved in the motor control circuit. However, I would like to leave open the possibility of a role for STS neurons in certain specific types of motor control, namely in actions whose object is to bring about a certain type of movement in a manipulated object. For instance (anticipating a much later discussion in Section 11.2.1.2), the transitive verb *to open X* really means 'to cause X to open', and the transitive verb *to crumple X* really means 'to cause X to crumple'. In such causative actions, it seems plausible that the agent adopts a different source of sensory feedback to control his motor system than is used during regular transitive actions. The biological motion system in STS seems a good candidate for providing the relevant information for actions whose goal is to bring about characteristic changes in the articulation of a manipulated object, such as *open*, *crumple*, *fold*, *roll* and so on. See Section 11.2.1.2 for more discussion of this point.

[4]Point-light stimuli can be manipulated so as to require more active 'tracking', for instance, by interleaving the frames of the display with blank frames; under these circumstances, a secondary attention-demanding task interferes seriously with biological motion processing (Thornton *et al.*, 2000). However, these displays are not characteristic of natural action stimuli.

measure of attention. It thus seems likely that while *overt* attention is on the target object during a reach action, some measure of *covert* attention is directed back to the agent while the action is underway. We can state this as another ordering constraint:

Ordering constraint 7 *The observer of a reach-to-grasp action reattends to the agent to some degree when monitoring/classifying the action.*

To summarise, the normal order of events for perception of a reach-to-grasp action is as follows: the observer first attends to the agent (to work out who the agent is, and to gather information about his intentions), then attends to the target (at a fairly early stage in the agent's reach), and only then begins to categorise the agent's action. At this point, his attention is in some measure *redirected* to the agent, probably covertly, as a result of biological motion processing.

2.7.3 Mirror neurons in F5

The properties of F5 mirror neurons have already been discussed; I will review them in somewhat more detail here. The defining properties of a mirror neuron are that it is selective for a specific kind of action (for instance that it is activated during a precision pinch, but not during other kinds of grasp actions) and that it fires both when this action is performed by the agent and also when the agent observes this action being performed by another agent. There are two other important points worth mentioning about F5 mirror neurons.

Firstly, many F5 mirror neurons are sensitive to combinations of arm and hand/finger movements. Recall from Section 2.5.3 that the AIP-F5 regions are traditionally seen as specialising in the grasp component of reach-to-grasp actions; however, as mentioned in that section, there are plenty of interactions between the reach and grasp pathways, and it is likely that F5 canonical neurons encode information about the hand trajectory as well as the grasp preshape. F5 mirror neurons likewise encode a mixture of hand trajectory and grasp preshape information. For instance, a typical F5 mirror neuron might fire when a particular grasp action is observed, but only if this grasp is accompanied by an appropriate hand trajectory.

Secondly, many F5 mirror neurons are sensitive to interactions between arm/hand actions and the objects of these actions. Thus an F5 mirror neuron which fires when an observed agent manipulates an object with the hand is not typically sensitive to the sight of the manipulating hand action in the absence of the object (Gallese *et al.*, 1996). We can note a constraint similar to Constraint 6:

Ordering constraint 8 *During observation of a reach-to-grasp action, the observer must attend to the target before the movement can be monitored in F5.*

In fact, there is an interesting qualification to make to this constraint. A study by Umiltà *et al.* (2001) found that many F5 mirror neurons respond to their associated grasp if the monkey observes an agent reach for an object hidden behind a screen. In this case the monkey cannot 'attend' to the target object. However, Umiltà *et al.* found that these mirror neurons only fired if the monkey knew there was an appropriate target object behind the screen. So the firing of F5 mirror neurons seems to be contingent on the presence of a target, whether this is established visually or through some form of inference. I will stick to the simple visual scenario for now, but I will consider working memory object representations in some detail in Chapter 7.

Note that the patterns of response shown by F5 mirror neurons during action perception are very similar to those shown by STP neurons. In each case, cells encode patterns of movement, frequently in relation to identified taarget objects. The key difference is that STP neurons seem to be principally involved in action observation, while F5 neurons are involved in both action observation and action execution.

Finally, many F5 mirror neurons are only triggered by observation of an action if the the observer can see the whole agent. Nelissen *et al.* (2005) presented monkeys with two types of hand action stimuli. In one, they saw an agent performing a hand action. In another, they saw an isolated hand performing the same action. They used fMRI imaging to measure brain activity, rather than single-cell recordings. They found that one area of F5 (called F5c) responded to hand actions in the 'full agent' condition, but not in the the 'isolated hand' condition. Another area of F5 (F5a) responded to hand actions in both conditions, and also responded to 'mimed' actions, reaching for a nonexistent object. However, most mirror neurons are found in F5c, so we can say that mirror neurons typically require a visible agent as well as a visible target in order to be triggered visually.

There have been some claims that neurons in F1 have mirror properties as well (see e.g. Hari *et al.*, 1998; Stefan *et al.*, 2005). These claims come not from single cell recordings, but from magnetic stimulation techniques which record activity of larger brain areas. F1 activity is considerably lower during action observation than during action recognition; either this activity is below the threshold required to generate overt actions, or further gating of the F1 signal is required before it influences motor systems. I will assume the former hypothesis in the model I present.

2.7.4 Mirror neurons in inferior parietal cortex

Given the similarities between cells in STS and F5, we expect there to be links between them. In fact, there seem to be no direct links; however, they are strongly linked by intermediate regions in the inferior parietal cortex. There are strong projections from STS and TE to inferior parietal regions PF and PFG (Perrett *et al.*, 1989; Tanaka, 1997), and further projections from PF/PFG to F5. PF/PFG cells can thus be expected to have response properties similar to those of STS and/or F5. Indeed, mirror neurons have recently been found in PF/PFG (Gallese *et al.*, 2002; Fogassi *et al.*, 2005). Thus while posterior parietal cortex appears to be mainly involved in action execution (as discussed in Section 2.3), inferior parietal cortex appears to have a role in action observation.

The experiments which examined the mirror properties of PF/PFG neurons were somewhat more elaborate than the original studies on F5, in that they examined responses to sequences of arm movements, rather than to individual movements. The monkey picked up a piece of food and then carried it either to its mouth (an untrained response) or to a container (a trained response). Both sequences ('eating' and 'putting') are part of the monkey's natural repertoire of actions. PF/PFG neurons responded to specific types of initial grasp made by the monkey. Many neurons had mirror properties, responding to a specific grasp whether performed by the monkey itself or by an agent it was observing. Interestingly, the great majority of PF/PFG neurons also responded differentially to the initial grasp action depending on whether the second phase of the movement involved moving the food to the mouth or the container. These neurons thus seem to carry information about the forthcoming phase of the movement as well as about the current movement. The existence of neurons in the motor control pathway with sensitivity to the preparation of subsequent actions has already been noted; see the discussion of 'set-related' neurons in the reach pathway (Section 2.5.2) and of sequence encoding in the prefrontal cortex (Section 2.6). However, the discovery of neurons sensitive to forthcoming actions in an observed agent is new, and highlights another similarity between the representations subserving action execution and action perception. Fogassi *et al.* state that while such neurons have not been reported in F5, they would probably be found if looked for.

Note that PF/PFG neurons appear to be sensitive to a fairly high-level representation of the subsequent action, rather than just to its kinematics. In a followup experiment, the container could appear in two positions—either in front of the monkey, or adjacent to the monkey's face, close to the mouth. In the latter condition, a reach-and-put action was kinematically very similar to a reach-and-eat action. The firing of PF/PFG cells differed between these two conditions, which suggests that they are encoding the goal of the action at a relatively high level ('put' versus 'eat') rather than simply its direction. We have already noted the sensitivity of F5 and STS neuron firing to the intended target

of observed reach actions; this finding about PF/PFG neurons emphasises that the action recognition system is sensitive not only to the location of the target, but also to its type.

Since PF/PFG neurons are sensitive to the second phase of the action before it is initiated, the monkey is clearly able to predict this second action when observing the first phase. It is interesting to consider what cues are available to generate this prediction. The most obvious cue is built into the design of the experiment: the container is only present in contexts where the food will be placed in it, so the presence of the container should be sufficient to prime the 'put' action over the 'eat' action. It seems likely that the presence of the container activates a plan, or cognitive set, in the prefrontal cortex of the observing monkey, which causes it to expect a 'put' action rather than an 'eat' action; in other words, the predictive coding of the second action in PF/PFG probably has its origin in prefrontal cortex.[5]

There is also good evidence that inferior parietal cortex has mirror properties in humans. The best evidence comes from an fMRI habituation paradigm. In a study by Chong *et al.* (2008), subjects executed a series of hand actions, and then watched videos of another agent performing a series of hand actions, which in one condition were the same as the executed ones, and in another condition were different. During action observation, fMRI activity in the subjects' right inferior parietal lobe was attenuated when the observed actions were the same as those the subject had just executed. This suggests firstly that different types of hand action are represented in this area. Furthermore, if we assume that repeated activation of a neural representation causes habituation, it also suggests that the representations of hand actions in this area are active both when they are executed and when they are observed.

Another habituation study by Hamilton and Grafton (2006) has found evidence that human intraparietal cortex encodes fairly high-level representations of the goals of an observed grasp action. Subjects watched a series of video stimuli in which a hand approaches and grasps one of two target objects. The fMRI signal in the anterior intraparietal sulcus decreased if the grasp action was to the same object in each video, even if the object appeared in different positions in different videos. This suggests that this region encodes the target of the grasp action rather than just its absolute trajectory. After habituation, a reach to the other object caused a renewed fMRI signal in this region, even if the trajectory was unchanged, which reinforces this conclusion. In summary, inferior parietal (and per-

[5]Given that joint attention is important in determining the target of the current action, it might be thought that the observed agent's gaze also plays a role in predicting the second action. However, this seems unlikely; if the target of the second action is a region close to the face, it cannot be fixated anyway, and even in the conditions where it can be fixated, gaze is unlikely to be a helpful cue; it has been found in other experiments on reach-and-place actions that gaze only moves to the target of the place action after the grasp action is almost completed (Johansson *et al.*, 2001).

haps intraparietal) cortex appears to be involved in the mirror system circuit in humans, and to be sensitive to the intentions of observed actions, just as it is in macaque.

2.7.5 A model of the mirror neuron circuit

I will now draw the above observations together in a model of the mirror neuron circuit responsible for action recognition. From a physiological perspective, the model draws mainly on a proposal by Rizzolatti and Matelli (2003) that there is a dorsal visual pathway which is specialised for action recognition, which proceeds through inferior parietal cortex, and is distinct from the dorsal pathway through posterior parietal cortex which is involved in action execution. More specifically, the model draws on a proposal by Iacoboni *et al.* (2001), which is echoed in more detail by Keysers and Perrett (2004), but it also incorporates elements of the detailed models of Oztop and Arbib (2002) and Oztop *et al.* (2005). A diagram showing the regions involved is given in Figure 2.10.

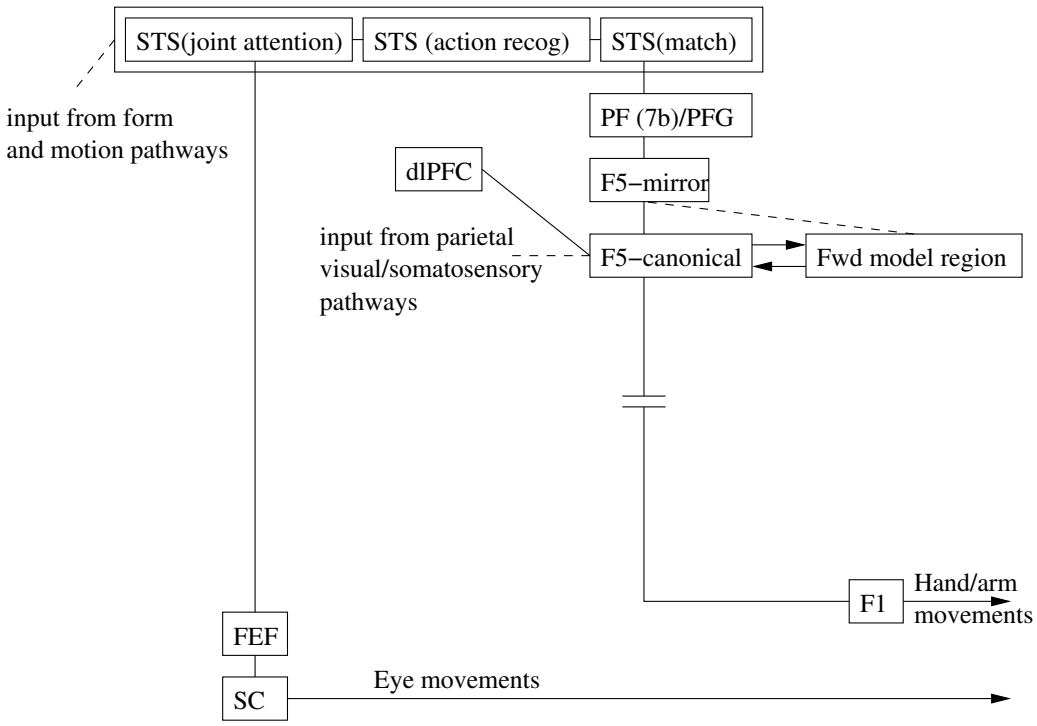


Figure 2.10: The action recognition visual pathway

The action recognition network operates in two different **modes**. In **training mode**,

the agent executes a reach action, selected by high-level goal representations in the PFC, and controlled by representations in the posterior-parietal/premotor reach/grasp pathways described in Section 2.5 (not shown in Figure 2.10) culminating in activity in the F5 canonical neurons. During such actions, the agent is typically attending to the target object, and visually monitoring the position of his hand in relation to this object; consequently, activity will be generated in the 'action recognition' region of STS, which is configured to recognise visual stimuli combining characteristic patterns of form and motion. This region is not involved in *controlling* the ongoing movement, but its activity is nonetheless *correlated* with the motor representations in F5; these motor representations deterministically generate certain patterns of overt behaviour, which are deterministically analysed by STS.

The network of inferior-parietal areas in between this region of STS and F5 learns to create *associations* between the visual representations in STS and the motor representations in F5. The regions involved here are the STS 'match' region (the region found by Iacoboni *et al.* to respond during both action execution and action recognition), the PF/PFG mirror neuron region discovered by Gallese *et al.*, and the F5 mirror neurons. The mode of learning is simple Hebbian association, just as in Burnod *et al*'s model of matching cells in the reach pathway. Essentially, the areas in between STS 'action recognition' cells and F5 canonical cells all function as matching cells, whose response properties are trained during the agent's own movements, just as the reach pathway matching cells are trained during motor babbling. All of these intermediary areas will therefore contain neurons with responses to both executed and observed actions—i.e. with mirror properties. Motor activity in F5 canonical neurons will thus generate activity in F5 mirror neurons, in PF/PFG mirror neurons, and in the 'match' neurons of the STS.

As for the reach pathway match neurons, the associations in these intermediate regions are *bidirectional*—thus activity in the STS action recognition area will trigger activity in the intermediate regions, culminating in activity in the agent's motor system proper in the F5 canonical neurons. During the agent's own motion, this reverse pathway presumably results in a positive feedback loop: motor activity in F5 generates an action, which generates a representation in STS which in turn strengthens the motor signal in F5. But more importantly, the reverse pathway also has a role in the recognition of the actions of other agents. The basic idea is that the STS signal evoked by the observation of someone else's action triggers the same kind of activity in F5 which would be generated if the observer were executing the observed action himself. Naturally, during action observation mode, this premotor activity must not necessarily lead to overt movement execution—we do not imitate observed actions as a reflex. (Note that in the diagram in Figure 2.10, the direct connection between F5 and F1 is turned off or inhibited.) But the STS signal nonetheless evokes structures involved in motor control.

2.7.5.1 Invariances between STS encodings of self- and other-generated actions: the role of joint attention

If the associations created in the STS-to-F5 pathway during the observation of one's own actions are to be helpful for action recognition, the representations generated in STS during one's own actions must obviously be *similar* to those generated in STS during the actions of an observed third party. Clearly, if STS is a region which encodes characteristic patterns of form and motion, both one's own actions and the actions of others will generate *some* activity in STS. But why should we expect a given action (say a reach-to-grasp) to elicit the *same* activity in STS whether it is executed by ourselves or by someone else?

For a transitive action like reaching to grasp, one consideration which increases the likelihood of invariant representations is that during action execution, the agent's attention is on the target object (as discussed at length in Section 2.5.2). If there is some means for ensuring that attention is directed to the target object during the *observation* of a reach action, then there is a strong likelihood that the visual representations generated in STS during an observed reach are similar to those generated during one's own reach actions. Of course, as we have already seen, attention *is* directed to the target of an observed reaching action; in fact the pattern of saccades executed by the observer of an action is essentially identical to that executed by the agent (Flanagan and Johansson, 2003). And as we have already proposed, the ability of the observer to identify the intended target at an early point during action observation is likely to be due at least in part to a mechanism for establishing joint attention implemented in a region of the STS.

It is interesting to consider the means by which the STS joint attention circuit is learned. It may be that the learning mechanism has nothing to do with action recognition—for instance, it has been proposed that the joint attention ability emerges through reinforcement learning, because the gaze of an observed agent is a good indicator of interesting objects in the observer's environment (see e.g. Carlson and Triesch, 2003). However, one might also speculate that the joint-attention capacity in STS emerges as the result of a pressure to create visual representations of actions which are the same for observed and executed actions. I will not attempt to decide between these alternative explanations.

2.7.5.2 A role for forward models during action recognition

It is important to consider the nature of the match made between STS visual representations and F5 motor representations. The processes of action execution and action recognition are both extended in time, and their associated patterns of motor and visual activity likewise each have a characteristic temporal structure. So the mapping to be learned is between two signals which are extended in time. This point is particularly clear in Oztop

and Arbib's (2002) model of the mirror system.

It is not clear whether the integration of signals over time is carried out separately within STS and within F5, to create relatively static representations which can be matched in a pointwise fashion, or whether the matching process deals directly with the time-varying signals, and computes an explicit correlation between them. But in either case, the matching process must result in the generation of an extended temporal signal in the F5 region during action recognition. How might this come about? One mechanism which might be involved is the circuit which computes an anticipatory forward model of the sensory consequences of the current motor signal during action execution. As discussed in Section 2.5.2, there is plenty of evidence that the control signals generated in premotor areas like F5 inform (and in turn are influenced by) a forward model of hand/arm state, whose purpose is to minimise the instabilities associated with delayed sensory feedback. The forward model circuit allows a motor signal in F5 at any given point in time to be projected forward as a simulation, without any overt action taking place. During action recognition, when overt actions are suppressed, the forward model allows the evocation of a temporally extended action.

Another benefit of the forward model could be to simplify the matching to be learned between F5 and STS. Recall that STS holds visual representations, while F5 holds motor signals. However, the forward model circuit associated with F5 converts F5's motor signals into anticipated sensory consequences; these sensory representations may be easier to match with STS signals than the motor representations from which they derive. In the diagram in Figure 2.10, I have allowed for both of these options, by giving a connection between the 'forward model region' and F5 mirror neurons as well as the connection between canonical and mirror neurons, which may be somewhat harder to acquire.

2.7.5.3 Visual recognition of whole-body actions

Not all actions will have visual representations which remain invariant regardless of whether the agent is onesself or a third party. Actions of the whole body, in particular, will have very different visual profiles when performed by onesself or by someone else. Consider a walking action: when one walks onesself, the movements of one's own limbs are hardly seen at all, while the limb movements of an observed walker constitute a very clear visual stimulus. STS is particularly good at encoding whole-body movements such as walking, and while mirror neurons for walking have not to my knowledge been found, it certainly appears that such movements evoke premotor representations in an observer just as strongly as movements with more invariant visual representations—witness for instance the involuntary movements made by people watching a football match. How, then, can an agent learn the association between a visual representation of a whole-body action and its motor

representation?

One suggestion is that there are similarities in the temporal profiles of activity developed independently within the motor system and the STS's form/motion categorisation system. Thus, for instance, the activity of walking is characterised at the motor level by a set of motor signals of a given periodicity and a given phase relationship to one another, and at the visual level by a set of signals with similar periodicity and phase relationship. Given that the match between visual and motor modalities involves signals extended in time, it may be that associations are possible purely due to these similarities in temporal profiles. It may be, then, that there are two relatively independent mechanisms for linking STS activity to premotor representations: one for actions with invariant visual profiles, which relies on associations learned during visual monitoring of one's own actions, and one which relies on similarities in the temporal profiles of independently-generated visual and motor signals.

2.7.6 The activation of goal representations during action recognition

The profoundest level of action recognition requires activation of the intentional representations underlying the observed action. When observers watch an action being performed, they can typically work out not just what the action is, but why the agent is performing it. We have already seen some evidence that the action recognition system can use inferred intentions to anticipate an observed agent's actions (see Section 2.7.4). But we have not yet considered the mechanism which allows inference of an observed agent's intentions.

During action execution, the agent's 'task set' is maintained in prefrontal cortex (see Section 2.6). The mirror system hypothesis would lead us to expect that the PFC is also involved in representing the intentions of an observed agent. In fact, imaging experiments on humans have only partially borne out this prediction. Grafton and Hamilton (2007) showed subjects sequences of video hand action stimuli, in which either the goal of the action or the physical movement of the action was repeated, and examined the brain regions which showed habituation effects in the two conditions. They found habituation effects due to repeated action goals in the posterior part of inferior frontal cortex and in inferior parietal cortex, both predominantly in the right hemisphere, but no properly 'prefrontal' effects. (Posterior inferior frontal cortex is premotor cortex, just posterior to prefrontal cortex.) Another experiment by Iacoboni *et al.* (2005) identified the same area of premotor cortex as involved in coding the intentions of observed actions. In this study, subjects saw a video stimulus of a hand reaching for a cup on a breakfast table in two different contexts, one favouring inference of a 'reach-to-drink' action, and the other favouring inference

of 'reach-to-clean'. Inferior frontal/premotor cortex responded differentially to these two cases, suggesting that it encodes not only the physical characteristics of the action but also its underlying intentions. But again, PFC proper was not differentially activated. There are certainly many studies which show an involvement of PFC in tasks which involve inference of an observed agent's intentions; for a good summary see Frith and Frith (2006). But these typically involve situations where the observer must work quite hard to construct an intentional explanation. For instance, Brunet *et al.* (2000) showed subjects cartoon strips involving an agent in different situations; some strips required an intentional analysis of the agent's behaviour, and others required an understanding of physical causality. More PFC activity was found in the former case, suggesting that PFC is involved in representing the intentions of an observed agent. We might conclude that the simple intentional explanations required to understand simple reach-to-grasp actions in different contexts do not heavily involve PFC, but can be somewhat automatically computed in premotor cortex. (Iacoboni *et al.*, 2005 reach a similar conclusion.) However, more elaborate intentional explanations certainly appear to recruit PFC.

In the remainder of this section, I will first discuss the differences between the intentional representations involved in action observation and action execution, and then consider the mechanisms which must be involved in learning how to activate appropriate intentional representations during action execution.

2.7.6.1 Differences between PFC representations in action observation and execution

Note that the connection between PFC and premotor cortex is quite different in action observation than in action execution. During execution, PFC influences premotor cortex, helping the agent decide which action to perform (see Sections 2.5.2 and 2.5.3). During observation, the influence runs partly, perhaps mainly, in the other direction. The observer may begin with a set of hypotheses about what the observed agent might do in the current situation. But it is only when the agent starts to act that these hypotheses can be confirmed or rejected. The agent's actions create representations in the observer's STS, which propagate first to inferior parietal and then to premotor areas. These premotor action representations then exert an influence on intentional representations in PFC. Thus while intentional representations in PFC are some of the first representations active during action execution, they are among the last to be active during action observation. Note also that PFC representations generated during action observation do not encode the goals of the observer, but those of the observed agent. These hypothesised goals are likely to be helpful primarily in creating appropriate expectations about the observed agent's future behaviour, which can exert a top-down influence on subsequent perceptual process-

ing. PFC representations thus encode very different types of information during action execution and action observation. Borrowing some terminology from logic, we can say that they encode different **modalities**: during action execution, they encode the agent's desires, while during action observation, they encode the agent's expectations.

2.7.6.2 Building representations of other agents: integrating object form, movement and intention

I have just suggested that PFC has a role during action observation: the intentions of an observed agent can be inferred from the premotor representations activated during action observation, together with a representation of the situation the agent is in. But note that the observer and the observed agent might have different strategies in a given situation, so the observer must have some way of storing different sets of condition-response links for different agents, and enabling the appropriate set when observing an action or generating an action himself. An observer's representations of 'individual agents' must include representations of their individual predispositions to act, as well as representations of their physical shape. When he watches another agent acting, he must presumably activate the appropriate set of predispositions into PFC, so that PFC represents the observed agent rather than himself, and generates the right expectations. Some convincing evidence that individuals with different predispositions to act are represented differently in PFC is given in a study by Mitchell *et al.* (2006), who found activity in different areas of PFC when subjects had to anticipate the behaviour and attitudes of people with similar or different predispositions to their own.

How does the agent know which predispositions to activate? He has already attended to the agent, and so has classified the agent 'as an object'. Presumably, the concept of an 'agent' somehow combines a representation of an object (probably held in IT) with a representation of a dynamic entity, with characteristic patterns of movement, and with a characteristic set of higher-level cognitive strategies. The patterns of movement are probably held in STS, as discussed in Section 2.7.2.1, while the cognitive strategies are held in PFC. A good account of the need to combine these different types of information about people is given in Frith and Frith (2006). There is considerable evidence that monitoring an observed agent's intentions activates STS as well as temporal areas; see e.g. Amodio and Frith (2006) for a review.

It is important to explain how these cross-modal concepts of agents are learned. The dynamic and intentional components of an agent representation must presumably be acquired when monitoring the agent's actions. Recall that observers of a reach-to-grasp action typically attend first to the agent and then to the patient (Constraint 5, Section 2.7.1), and only then begin monitoring the action (Constraint 6, Section 2.7.2.3). So in order to

learn a cross-modal representation of an agent when monitoring the agent's action, the observer must to some degree *reattend* to the agent as an object, evoking a representation of the agent rather than the patient in IT. In fact, we have already seen evidence for this action of reattention. As described in Section 2.7.2.4, biological motion monitoring probably requires some degree of reattention to the agent (see Constraint 7). There are now two reasons for thinking that the observer must reattend to the agent during action monitoring. One is that action classification relies partly on static shape recognition, which is known to require attention. The other is that reattention seems essential for the formation of cross-modal conceptions of agents, combining combining static shape/form representations with dynamic representations of articulated objects in STS and of intentional representations in PFC.

Ordering constraint 9 *During perception of a reach-to-grasp action, the observer must reattend to the agent to allow the formation of a crossmodal agent representation, combining shape, motion and intentional attributes.*

How can an observer learn the PFC-based stimulus-response pathways which characterise a particular external agent? The reinforcement learning scheme mentioned in Section 2.6.2 works fine for the agent's own actions, but it will not work for action observation. A reward state experienced by an observed agent will not necessarily trigger a reward state in the observer (in fact, if the two agents are competing, the opposite might be true). A separate mechanism must be proposed. One suggestion is that during action observation, a reward is associated with a successful *prediction* about the observed agent's next action.[6] Thus any condition-response links which generate correct predictions about the observed agent's next action will be strengthened. Again, note that this learning scheme encourages the development of expectations in PFC during action observation, while the learning scheme used in action execution develops 'desires', or 'intentions'.

2.7.7 Comparison with other models of mirror neurons

By way of summary: the basic flavour of the mirror neuron model given above is that action execution involves propagation of activity *forward* from intentional representations in PFC to the premotor area F5, out into the world, and thence to STS, while action recognition involves propagation of activity in STS in the reverse direction, to the F5 premotor area and thence to PFC. To borrow some more terminology from logic, we can think of

[6]This idea in fact links to an idea about how infants learn words; see Section 6.3.2 for details.

the forward propagation as implementing a process of **deduction**, reasoning from causes to effects, and of the backward propagation as implementing a process of **abduction**, reasoning (nonmonotonically) from effects to their likely causes.[7] This basic model is the one proposed by Iacoboni *et al.* (2001) and Keysers and Perrett (2004). However, there are a number of other related models of the mirror system with which it is useful to make comparisons.

2.7.7.1 Oztop and Arbib (2002)

Oztop and Arbib (2002) provided one of the first neurally informed computational models of the mirror neuron system. In their account, the action recognition system is an **exaptation** of a circuit whose primary function is to provide visual feedback about the hand during the control of the animal's own movements. A key construct in their model is the **hand state**: a vector describing the sequence of locations and grasp shapes of the hand in relation to the location of the target object and one of its opposition axes. Hand state is computed in STS and another posterior parietal area called 7a. Hand state is involved in the control of the animal's own actions, and because it is defined in relation to the target object, it can also be computed for the hand of an observed agent performing a reach action. When the animal performs a reach action itself, area PF (which they term 7b) and the F5 mirror neurons learn an association between between the evolving hand state and the sequence of motor signals it is generating in F5 canonical neurons, much as in Iacoboni *et al*'s model. These associations can then be used during action recognition to evoke F5 activity during an observed action.

Oztop and Arbib's model is fully implemented, and thus much more detailed than the one I have sketched. But the basic ideas are similar. The only real difference is that in Oztop and Arbib's model, STS is directly involved during the execution of reach/grasp actions, while in the one I sketched it is not.

2.7.7.2 Oztop *et al.* (2005)

Oztop *et al.* (2005) present a somewhat different model of the mirror system. In their account, forward models play a crucial role. The main difference about this account is

[7] We can also use terminology from motor control, and say that action recognition involves implementation of an *inverse model*, which takes the position of the observed agent's body at two points in time and computes the motor signal which would have been responsible for the change in position. Note that this inverse model takes *two observed* body states, whereas an inverse model used in motor control takes one observed and one desired body state. We can also use terminology from statistics, and say that the action recognition system implements a **Bayesian** inference procedure, again reasoning from observed effects to their most likely causes. An explicitly Bayesian model of the mirror system is given by Kilner *et al.*, 2007.

that there is no method whereby representations in STS can actively *trigger* associated representations in F5 and in turn in PFC. Rather, activation leading to F5 comes from a hypothesis about the observed agent's reach target, which is generated in PFC from contextual cues, and from an initial assessment of the position of the agent's hand in relation to this target. The activation created in F5 by these signals is thus a hypothesis about the agent's likely motor signal. This hypothesised motor signal can be played forward in simulation, by means of the circuit involving the forward model of F5 motor signals. The evolving signal in this forward model generates activity in STS, which represents the expected visual stimulus if the hypothesis is correct. At the same time the agent is calculating the *actual* visual stimulus in the parietal visual areas which normally contribute to motor control. This signal is also sent to STS, where a match operation is carried out between the expected and actual visual signals. If there is a match, the hypothesis about the agent's action is accepted; if not, a new hypothesised goal is generated.

The key element of this model which I have incorporated is the suggestion that the forward model generated in premotor areas could be involved in generating a visual signal to be matched with the signal derived from vision (see also Miall, 2003). However, there are also two significant differences.

Firstly, in Oztop *et al.*'s model, the role of STS in action recognition is somewhat reduced: representations of the observed agent's hand are primarily computed in the parietal sections of the reach/grasp pathways, and then sent to STS for matching with hypothesised representations. In the model I presented, STS receives some input from areas of the reach/grasp pathway during action recognition (namely MT and MST), but it also receives input from the form classification pathway, and in general is more autonomous in generating visual representations of actions. My conception of STS is of a classification area, performing quite similar computations to the object classification pathway, but using motion information as well as form information, and integrating stimuli over a range of successive time points, as in the model of Giese and Poggio (2000). It is thus able to develop its own analysis of the characteristic forms and motions involved in various different kinds of action.

The second difference in Oztop *et al*'s scheme is that the learned association between F5 and STS only runs in one direction, from F5 to STS, even during action recognition. For Oztop *et al.*, the hypothesis about the observed agent's motor programme is not generated by STS, only tested by it. This decision probably reflects the fact that STS has less of an active role in generating visual representations of actions in the Oztop *et al.* model. However, if STS is assumed to generate its own autonomous visual representations of actions, there seems no reason why the associations learned during monitored self-motion should not cause the triggering of F5 activity, as Iacoboni *et al.* and Arbib and Oztop both suggest. The model of sensorimotor learning in the reach pathway developed by Burnod

et al. trades on exactly the same kind of bidirectionality. It is in fact attractive to assume that the same mechanism which creates links from the parietal cortex to the premotor cortex also links premotor cortex to the STS.

2.7.7.3 The order of target selection and action monitoring

There is one point on which all the models of mirror neurons are agreed: the visual representations which inform action recognition are representations which track the agent's hand in relation to the location (and shape) of a hypothesised target object. This is crucial in order to create the invariances between visual representations of action recognition and action execution which support development of the mirror system. This means that the intended target must be selected *prior* to the visual monitoring of the action. We can state another ordering constraint:

> **Ordering constraint 10** *Biologically-inspired computational models of action recognition assume that the agent's intended target is identified before the process of action monitoring/classification begins.*

This constraint accords with Constraints 6 and 8, concerning the dependence of action classification activity in STS and F5 on the identification of a target object.

It is interesting to compare how target selection interacts with action monitoring in action execution and in action observation. There are some similarities in the ordering constraints involved in each case. As discussed in Section 2.5.2, the selection of a target for one's own reach action is initially parallel, and later serial: the agent begins by computing movement vectors for several candidate targets, aided by a bias in the SEF generated from PFC, and then selects the best of these, resulting in a direction of attention to this target generated from the supplementary eye field. This enables the selected object to be classified; its type then is matched to a representation of the desired target in PFC; if there is a match, an action ensues, otherwise the target location is inhibited and the next-best target is selected. By the time the action begins, a single target has naturally been selected (see Constraints 2 and 3, Sections 2.5.2.7 and 2.5.3.1).

The selection of a target object during action recognition also has two phases. In the first phase, parallel mechanisms select a candidate target object, using cues from joint attention, PFC bias and so on. In the second phase, the agent's motion is tracked in relation to this target. If the observer is wrong about the chosen target, activity in the STS will not trigger activity in F5; at this point, the next-best candidate target is presumably selected.

Thus while the sequential element of target selection in action execution happens prior to action initiation, the sequential element of target selection in action recognition happens after action initiation, during action monitoring.

2.7.8 Endpoint of grasp observation: visual perception of contact

At the end of a successful reach-to-grasp action, the agent is holding the target object, and thus establishing it in the haptic modality, as discussed in Section 2.5.4. Presumably, the observer of a reach action also needs a way of recognising that the action has been successfully completed. Note that haptic information will not be available in this case; a separate purely visual mechanism must be involved, which supports inferences about the contact relationships between solid objects.

There is good evidence for a visual mechanism which makes contact with the relevant haptic circuits. For instance, when a human agent observes another person being touched, this generates the same kind of activity in their secondary somatosensory cortex as is generated when they are touched themselves (Keysers *et al.*, 2004). We can therefore state an ordering constraint about haptic reattention to the target in action recognition:

Ordering constraint 11 *When an observer watches an agent complete a reach-to-grasp action, the observer reattends to the target object (in the haptic modality).*

This constraint is the action recognition analogue of Constraint 4 (Section 2.5.4).

Keysers and Perrett (2004) suggest that the relevant connections between visual and haptic circuits could be learned by a Hebbian mechanism operative during experience of one's own actions. The sight of one's own hand touching an object will be correlated with a touch sensation; after training on one's own actions, it is plausible that the same sensation is evoked by the sight of another person's hand touching an object. The account is thus analogous to the Iacoboni (2001) / Keysers and Perrett (2004) model of the mirror system.

It is likely that learning about the haptic correlates of visual properties of objects begins in early infancy; for reviews, see Baillargeon (1994; 1998). Baillargeon's experiments show that infants are able to infer relationships of support and containment between observed objects. For instance, if the experimenter holds an object at a particular angle in relation to a supporting surface, infants as young as six months are able to judge whether the object will be stable when released. Young infants can also infer that two objects cannot pass through one another (Spelke *et al.*, 1992). While there is still some debate about how these fundamental visual abilities are acquired, it is likely that they underlie the mature ability to recognise a stable grasp.

2.8 Distinctions between executed and observed actions: representation of self versus other

The finding of common representations underlying action perception and action execution raises the question of how an agent knows whether a currently represented action is being performed by himself or being observed in someone else. This question has recently been receiving considerable attention. The consensus is that there must be areas of the brain distinct from the areas considered up to now, which are involved in determining who the **agent** of the currently monitored action is, and in distinguishing onesself from an observed agent. These areas are sometimes referred to as comprising a 'who' pathway, analogous to the 'what' inferotemporal and 'how' parietal pathways (c.f. Georgieff and Jeannerod, 1998). In the case of an observed agent, it is presumably also necessary to identify the agent, even if this simply involves assigning a basic level category (as in our example *the man grabbed the cup*). I will begin by considering evidence from brain imaging studies investigating differences between action observation and action execution. I will then consider two hypotheses about the mechanisms which result in these differences.

2.8.1 Brain regions with differential activation during observed and executed actions

While there is considerable overlap between brain activity during action execution and action recognition, brain imaging studies explicitly comparing these two processes have found some differences. In one type of study, subjects either perform an action or watch a similar action being performed. For instance, Rizzolatti *et al.* (1996) had subjects watching or performing grasp actions, and compared brain activity in each case to a baseline where the subjects observed the grasp target with no action. It was found that grasp observation activated an inferotemporal region homologous to monkey STS, which was not activated during grasp execution. Execution activated left somatomotor cortical areas and the cerebellum. However, there are so many differences between the observation and action conditions, it is hard to draw conclusions about the roles of these regions.

Another type of study attempts to manipulate a subject's 'sense of agency' while controlling for all other sensory and motor factors. For instance, in a study by Farrer and Frith (2002), subjects used a joystick to move a circle on a screen; they were told that in some trials, their movements would control the circle, while in others, it would be controlled by the experimenter. The subjects made similar movements in both conditions, and saw similar stimuli; the only difference was whether they experienced a feeling of control. A feeling of control was associated with activity in an area called the anterior insula, while a feeling

of another agent being in control activated a region in inferior parietal cortex (particularly in the right hemisphere). Other imaging studies have found similar conclusions (see e.g. Fink *et al.*, 1999; Farrer *et al.*, 2003).

These results are corroborated by studies of neurological patients. For instance, schizophrenic patients who feel that another agent is controlling their actions show abnormal hyperactivity in the right inferior parietal cortex (Spence *et al.*, 1997), while a synesthetic patient who experiences tactile sensations in her own body when observing another person being touched shows abnormally high activity in the anterior insula (Blakemore *et al.*, 2005).

2.8.2 The match model of agency

What information processing is taking place in the anterior insula and the right posterior parietal cortex? One suggestion is that the sense of being the agent of an action currently being monitored results from the integration of information generated in one's own motor system and information arriving from sensory sources. Forward models of the motor system are often given a role in this process; specifically, it is often claimed that an efferent copy of the current motor signals is given to a forward model, which calculates the expected sensory consequences of the action, and compares these to the actual sensory signals being received (see e.g. Blakemore *et al.*, 2002). This comparison process is likely to be necessary for independent reasons, as it allows a cancellation of predicted reafferent sensory signals resulting from one's own movement, and a consequent focus on signals indicating deviation of the action from its intended course. (Such cancellation operations are evoked to explain why it is hard to tickle onesself—see Blakemore *et al.*, 1998.)

However, as discussed in Section 2.7, an almost identical comparison process is also held to underlie the recognition of actions produced by another agent. According to the model put forward before, the action recognition process involves a match being made between visual information about a perceived action and the predicted sensory consequences of a simulated motor signal. So if a matching process *is* involved in creating the experience of agency, it must be separate from the matching process involved in action recognition. Specifically, it must generate a match only during action execution.

One proposal is that the match process which creates the sense of agency could involve *somatosensory* reafferent signals (which can only be due to one's own actions) rather than visual ones (which can be due to onesself or an observed agent). On this model, the agent generates predictions about the somatosensory consequences of the current motor signal, and relays these to a separate match region, where they are compared with actual somatosensory information. A match in this region indicates that the motor signals are due to an action executed by the agent, rather than to activity in the mirror system arising from an observed action. Consistent with this proposal, the anterior insula is known to be an

association area receiving somatosensory information. Increased activation in the inferior parietal cortex during action observation might in turn be associated with the mechanism which inhibits the regular parietal pathways converting sensory stimuli to motor actions and puts STS in control of activity in the premotor cortex.

There are still difficulties for this match theory, however. Firstly, the issue of whether perceptual information matches with predictions generated from motor signals feels more like the *consequence* of a decision about who the agent is, rather than its cause. Prior to action recognition, there must be some operation which puts the premotor system under the control of STS and suppresses overt motor execution of premotor signals; likewise, prior to action execution, there must be some operation which links premotor activity to overt motor behaviour. Models of the mirror system all feature two different *modes*: one for action observation, and one for action execution. This presupposes that there is a mechanism for *selecting* the mode in any given context and making appropriate switches in the circuitry. There are some situations in which it is appropriate to act, and other situations in which it is appropriate to observe rather than act; the appropriate rules are the kind of stimulus-response mapping we expect to find encoded in prefrontal cortex as a result of reinforcement learning.

A second problem with the match theory of agent attribution is that while it provides a means for identifying oneself as the agent of an action, it does not provide a means for identifying different observed agents. Knowing that onesself is *not* the agent of a monitored action is obviously insufficient; we must be able to distinguish between different external agents. Presumably, identifying the agent involves generating a representation in the object categorisation pathway. But since this pathway is separate from the pathways involved in action monitoring, it remains to be specified how this representation is associated with (i.e. bound to) the agent of an observed action.

A final problem with the match theory of attribution is that an agent can use mental simulation not only to recognise actions in others, but also to *rehearse* possible scenarios in which he executes an action, or other agents execute actions. In such situations, presumably the agent is able to attribute the simulated action to himself or to another agent as appropriate, even when there is no overt action to generate proprioceptive feedback to use in a match operation. This problem is noted by Jeannerod (2003), but no satisfactory solution is given. In fact, it seems pretty clear that when we simulate the performance of an action ourselves, we do not experience anything like the 'sense of agency' which accompanies actions we actually execute. In summary, while the match hypothesis provides a useful model of the phenomenal *experience* of agency, something else must be responsible for the *representation* of onesself or of someone else as the agent of the currently monitored action.

2.8.3 The mode-setting model of agency

Another suggestion about the origin of the self-other distinction relates not to processes *during* action monitoring, but to operations occurring *prior* to action monitoring, which define whether the agent's motor system encodes an active movement or an observed one. A model of this sort is envisaged by Farrer and Frith (2002), who suggest that the differential activations of the insula and parietal cortex result from prior *attentional* operations (p602). They do not go into any detail about what these attentional operations might be; in the remainder of this section, I will make some suggestions myself.

If the agent is observing another person executing an action, presumably the prior attentional operation is simply a direction of attention towards this person. On this hypothesis, an initial direction of attention to another person results in the observer's motor system being configured to respond to STS activity—what we might call **action perception mode**. Any evoked motor signal in this system will then refer to an action executed by the attended-to agent.

If the agent is executing an action himself, presumably the prior operation is a 'decision to act'. On this hypothesis, a 'decision to act' is a well-defined cognitive operation in its own right, which results in the agent's motor system being configured to respond to parietal activity in the reach and grasp pathways, and to generate overt actions—what we might call **action execution mode**. Any evoked motor signal will then refer to an action performed by the agent. (Clearly, to call this operation 'attentional' is to broaden the definition of attention quite considerably, but 'attention to onesself' is always going to be a very different mechanism from attention to external stimuli, so a broad definition seems unavoidable.)

A model of agency involving these two mode-setting operations addresses all of the problems with the match theory of agency just outlined. Firstly, it is plausible that such operations can be explicitly scheduled by prefrontal cortex as a result of reinforcement learning. They are both cognitive operations in their own right, it has already been established that prefrontal cortex is able to initiate attentional operations as well as motor actions. Secondly, an action of attention to an observed agent is exactly what is needed in order to categorise the agent as an object, and thus identify the agent. A prior action of attention to the agent thus serves a double purpose, both identifying the observed agent and configuring the action monitoring system to respond to this agent's actions. Finally, the hypothesised mode-setting operations can themselves presumably be simulated, allowing the possibility of distinctions between self- and other-generated actions even for simulated actions.

What evidence is there for mode-setting operations occurring prior to action observation and action execution? For action observation, we have already noted in Section 2.7.1 that

observers of a reach-to-grasp action typically saccade to the agent first, in order to recognise the agent and gather to information about his intentions. We can now suggest that this early action of attention to the agent is what puts the observer's mirror neuron circuit into action recognition mode. On this interpretation, it is not just a contingent fact that observers attend first to the agent, but something which *must* happen, overtly or covertly, in order to begin the process of action recognition. Another constraint can be stated to this effect:

Ordering constraint 12 *During observation of a reach-to-grasp action, the observer's first action is to attend to the agent, an operation which puts his mirror system into action perception mode.*

What evidence is there for a 'decision to act' occuring as the first operation in action execution, in advance of any detailed computation in the premotor system about what action to perform? A candidate for this operation is the cognitive operation which gives rise to the cortical **readiness potential**, a wave of activity in the precentral and parietal areas which precedes voluntary action, as detected by electrodes attached to the subject's scalp (see e.g. Shibasaki, 1992; Shibasaki and Hallett, 2006). Trevena and Miller (2002) summarise the readiness potential as reflecting motor preparation, but also 'cognitive or perceptual aspects of the preparation to move, such as anticipation or motivation' (p169). The readiness potential is distinguished from the **lateralised readiness potential**, which occurs later in self-initiated movements, and reflects activity in the premotor cortex relating to the specific movement which is planned. The general readiness potential reflects activity in the supplementary motor area F6 (Jenkins *et al.*, 2000), which has sometimes been given a role in the generation of an internal 'go'-signal for actions being prepared (see e.g. Fagg and Arbib, 1998).

It is interesting to note that the event-related potential evoked by observation of an action is very different from the readiness potential found during execution of the same action (Babiloni *et al.*, 2003). On the other hand, the preparatory phase of the ERP found during self-paced action execution is quite similar to that found during action simulation (Jankelowitz and Colebatch, 2002). The ERP from the sensorimotor hand areas is likewise similar during actual and simulated actions (Beisteiner *et al.*, 1995; see also Lotze *et al.*, 1999 for fMRI evidence that the SMA is equally active during actual and imagined actions). These results provide some preliminary support for the idea that the general readiness potential could function as the mode-setting operation which establishes onesself as the agent of a forthcoming action. However, it is important to note that the general

readiness potential only occurs before voluntary, planned actions. An agent sometimes executes involuntary actions, for instance reflexive avoidance or defensive actions (see Section 2.5.2.3). Agents can still presumably attribute these actions to themselves. We would have to posit some other mechanism for allowing this attribution, which might have more to do with an agent's attention being drawn bottom-up to his own body. For the moment, I will just assert that the first operation during the process of action execution must be an operation which configures the mirror system for action execution rather than action observation.

Ordering constraint 13 *The first operation in the execution of a reach-to-grasp action must be one which puts the agent's mirror system into action execution mode.*

Note that this internal operation also has the character of an 'action of attention to the agent'. But in this case, the agent is attending to himself. Naturally, this operation is not a saccade to an external point in space; it is a cognitive operation. But it is nonetheless an operation which establishes a certain sensorimotor state, in which the agent can monitor his own actions.

A diagram showing the role of the mode-setting operations of 'attention to self' and 'attention to other' is given in Figure 2.11. Note firstly that the two operations are mutually exclusive, and should be understood as competing with one another. The competition can be influenced by both top-down and bottom-up processes, as I will now describe.

Both 'attention-to-self' and 'attention-to-other' can be activated top-down by the prefrontal cortex. If attention-to-self is triggered, the link from STS to the F5 canonical neurons (via PF/PFG and the F5 mirror neurons—not shown) is blocked, but the parietal sources of input to F5 and F2 are left open, as are the links between these regions and F1. If attention-to-other is triggered, there are three consequences. Firstly, a top-down attentional operation is evoked in the frontal eye field (FEF), moving attention to an external entity. The entity will consequently be categorised in the inferotemporal pathway; moreover, if it is animate, any actions it executes will result in activity in STS. Secondly, the parietal links to F2 and F5 are blocked, while the route from STS to F5 is left open. Thus any STS activity will result in F5 activity. Finally, the routes from F2 and F5 to F1 are blocked, so that no overt activity is evoked in the observer as a result of an action being observed.

The attention-to-self or attention-to-other operations can both be activated by bottom-up signals as well as top-down ones. A strong enough external attentional stimulus might put an agent into action perception mode, even if he intended to perform an action himself.

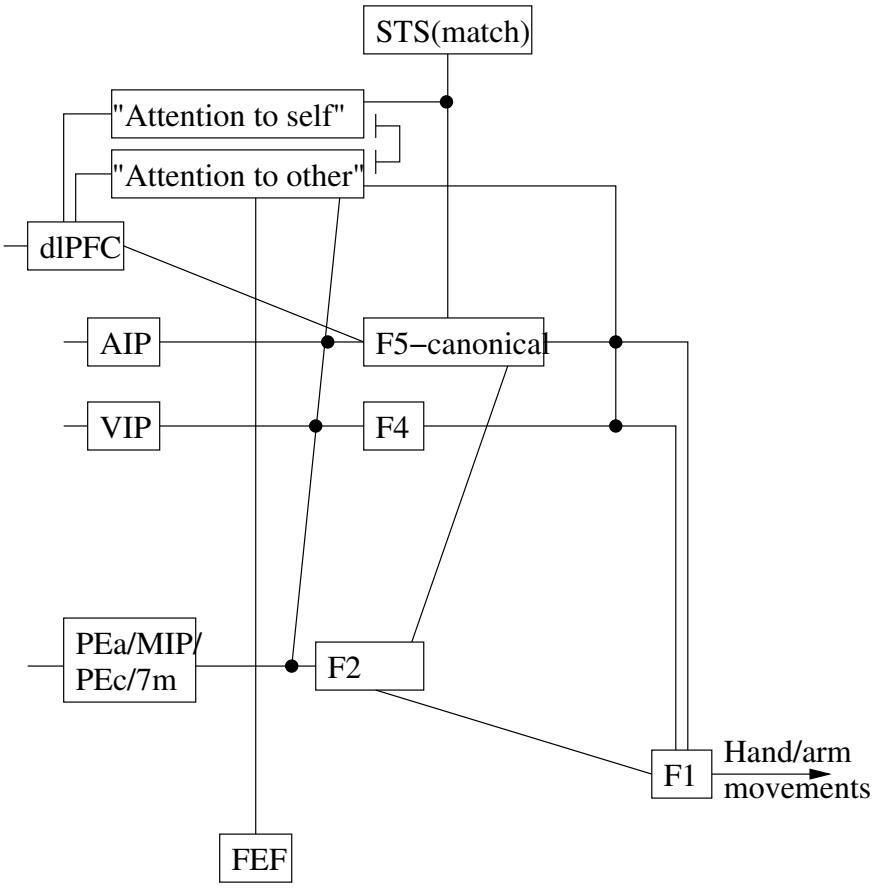


Figure 2.11: Pathways involved in distinguishing between self-generated and other-generated actions. (Note a new type of link in this diagram, which terminates with a black circle on another link. This is understood to denote a gating relationship on the second link.)

Conversely, a stimulus which evokes a reflexive action might put an agent into action execution mode, even if he was planning a perceptual action. What this means is that executing a top-down 'attend-to-self' or an 'attend-to-other' action is no guarantee of it being successful. We must envisage a role for reafferent feedback informing the agent about the success or failure of such actions. In the case of 'attend-to-other', the reafferent feedback may simply amount to evidence that one has classified an object in the world, but reafferent evidence for 'attend-to-self' is probably quite specific to this particular action. Perhaps the activity in the anterior insula discussed in Section 2.8.2 is one component of the state which results from a successful attend-to-self operation. For concreteness, I will assume this is the case.

2.9 Summary: the pathways involved in perception and execution of reach-to-grasp actions

We have now concluded our tour of the neural pathways involved in the perception and execution of reach-to-grasp actions. The process involves several interacting pathways processing sensory and motor information, which have been illustrated along the way in circuit diagrams. To complete the picture, Figure 2.12 provides a circuit diagram for the whole process, including attentional, visual and motor functions. I have left out some connections with PFC, in the interests of minimising spideriness. The diagram is still quite hard to read, but is useful in situating each of the individual pathways in its wider context.

2.10 The order of sensorimotor events during the execution and perception of reach actions

What story can we abstract from the mass of detail which has just been given? One useful approach is to consider the sequential order of processes occurring during reach execution and reach perception. During the above survey, several constraints were noted about the ordering of different sensorimotor processes and operations. Considering all these constraints together provides an interesting global picture of the temporal structure of execution and perception of a reach-to-grasp action. If there are strong constraints on the ordering of individual processes during action execution and perception, then it may be that an observed or executed action can be *represented* (e.g. in working or long-term memory) by a characteristic sequence of sensorimotor events. For instance, if the process

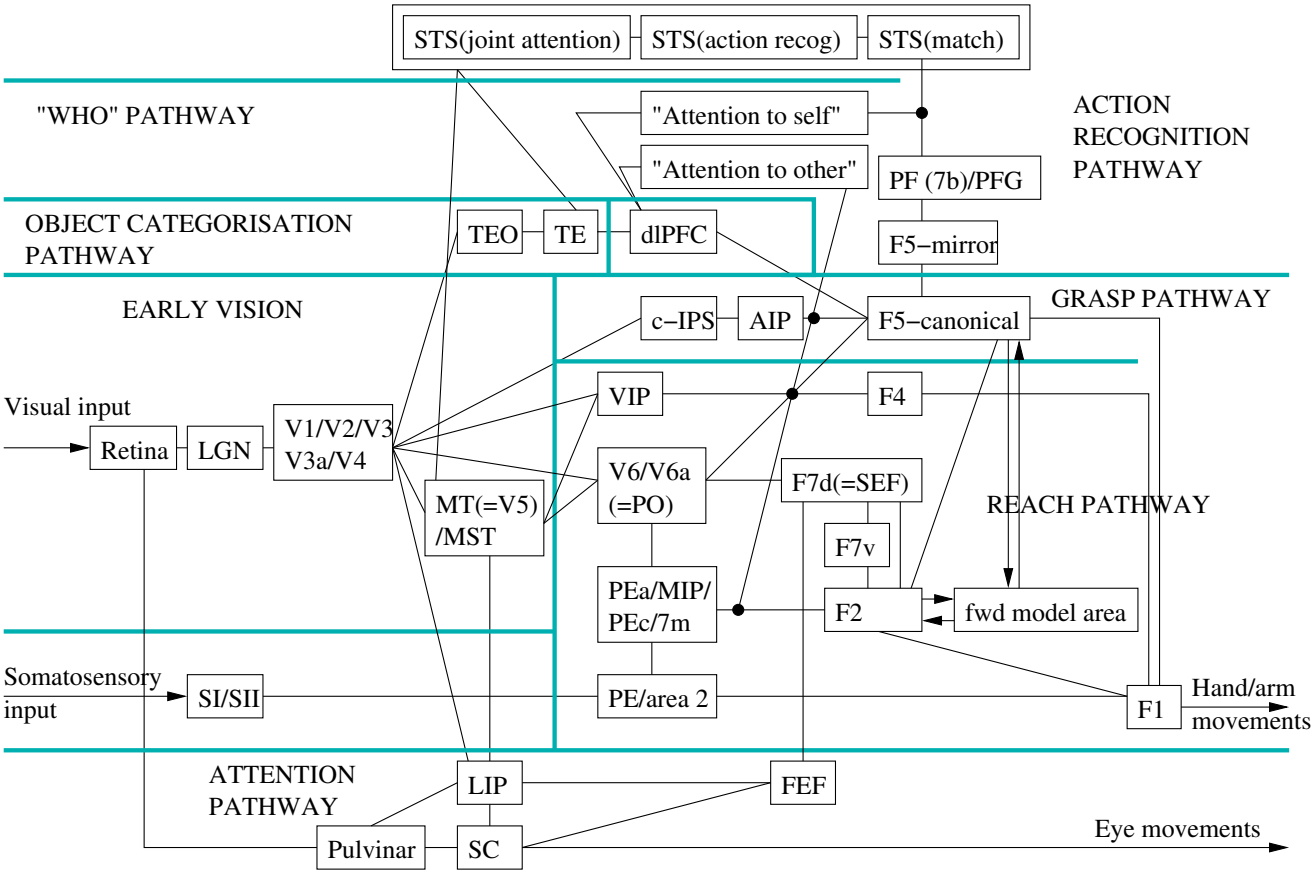


Figure 2.12: Neural pathways involved in the perception and execution of reach-to-grasp actions

of attending to the agent always precedes the process of attending to the patient, then the concepts of 'agent' and 'patient' can perhaps be defined in relation to the relative position of their associated sensorimotor events in such a sequence.

The suggestion that sequences of operations constitute a central organising principle of cognition is not new; see e.g. Ullman (1984) and Newell (1990) for some well-known formulations of this idea, and Anderson and Lebiere (1998) for a recent exposition. In Section 2.10.1, I will introduce one particular theoretical sequencing framework, which is due to Ballard *et al* (1997). In Sections 2.10.2 and 2.10.3, I will consider the sequences of processes involved in execution and perception of reach actions respectively. I will argue that both reach execution and reach perception can be characterised by a precise global order of sensorimotor processes.

2.10.1 A theoretical framework: deictic routines

A useful model of sequential cognitive processes was recently proposed by Dana Ballard and colleagues (Ballard *et al.*, 1997), drawing on earlier work by Arbib (1981) and Agre and Chapman (1987). Ballard *et al.*'s account accords a special role to *attentional* actions in sequences of cognitive processes. Sequences frequently consist of chains of attentional actions, in which each attentional action creates a transitory sensory state which enables the execution of the next action. Ballard *et al.* give several examples of such sequences. One example is the process of categorising an object. To do this an agent must first fixate the object to create a (very transitory) image of the object on the retina, and then gate open the link from the retina to the object classification pathway. Another example is reaching for a target object. To do this the agent must first attend to the target object, creating transitory representations in the parietal sensorimotor pathways. He can then activate a general 'grasp' motor programme, whose parameters are instantiated by these transitory parietal representations. In Ballard *et al.*'s account, executing a reach action involves executing a sequence of operations, the first of which is an attentional action to create the sensorimotor representations which support the execution of the motor action itself.

Ballard *et al.* call the transitory sensorimotor representations in sequences of this kind **deictic representations**, and the attentional operations which give rise to them **deictic operations**. Sequences of deictic operations are termed **deictic routines**. In general it is hard to characterise a deictic representation by itself; the significance of such a representation is given by the nature of the operation which caused it to arise. But note that this prior attentional operation may itself have been determined by the previous transitory representation. Inductively, then, any given deictic representation is given meaning by the *sequence* of attentional operations which precedes it, interleaved with the representations

these operations result in.

Our account of execution and perception of the reach-to-grasp action contains many ordered attentional operations and many transitory sensorimotor states; it is interesting to try and express the processes of action and perception of the cup-grabbing action using Ballard *et al.*'s notion of deictic routines. I will consider action execution in Section 2.10.2 and action perception in Section 2.10.3. In each case, the process will be described as an alternating sequence of deictic operations and resulting transitory sensorimotor states.

2.10.2 The sequence of processes during execution of a reach action

2.10.2.1 Operation 1: action of attention to the agent

Consider an agent about to execute the cup-grab action. According to the mode-setting model of agency outlined in Section 2.8.3, the agent's first operation is a decision to perform an action rather than to make an observation about the world, i.e. to put himself into 'action execution mode'. I called this operation 'attend-to-self': it is the attentional operation hypothesised by Farrer and Frith to be responsible for the differential activations of the insula and parietal cortex during action execution and action observation. I suggested that this operation is needed to configure the mirror system circuitry for execution rather than recognition. In the model I outlined, this operation strictly precedes any other activity in the mirror system; it has to be configured for execution before the agent can begin to execute an action (see Constraint 13).

2.10.2.2 State 1: attending to the agent

As described at the end of Section 2.8.3, top-down activation of the 'attend-to-self' operation is no guarantee that the system will enter action execution mode, because there are also bottom-up influences on which mode is entered. The agent must therefore obtain feedback about the success of this operation, which will take the form of a sensorimotor state signalling that the agent has successfully 'attended to himself', i.e. is indeed in action execution mode. I suggested that the activity in the anterior insula described in Section 2.8.2 which is associated with executed rather than perceived actions might be one component of this state.

2.10.2.3 Operation 2: action of attention to the cup

Now that the agent is in execution mode, the visual information he receives is passed into the parietal pathway and used to select (and later execute) an action, as described in Sec-

tion 2.5. The first thing that happens is that the agent attends to the target (Constraint 2). In the reach pathway, as described in Section 2.5.2, the objects in the agent's peripersonal space are initially represented in motor coordinates, as movement vectors. One of these vectors is selected based on a mixture of bottom-up and top-down factors; this *motor* selection results in a corresponding action of *visual* attention to the selected object, through projections from the reach pathway to the attentional area F7d and thence to the other attentional areas described in Section 2.4. In other words, the agent executes an action of attention to the cup. As described in Section 2.5.2, attention-for-target-selection and attention-for-action appear to involve the same basic neural mechanism (see Schneider and Deubel, 2002).

2.10.2.4 State 2: attending to the cup

Attention to a retinal location modulates activity in primary visual cortex at that location, and allows classification of the stimulus projected onto that location, as described in Section 2.4.3 (Constraint 1). Now that the agent is attending to the cup, the representation of the cup in primary visual cortex is communicated to the visual classification pathway described in Section 2.2, and a representation of the cup is produced in inferotemporal cortex. An affordance-based representation of the cup is also produced in the caudal intraparietal sulcus, as described in Section 2.5.3. Visual attention to the cup also strengthens and maintains the motor representation of the cup in the parietal reach pathway.

2.10.2.5 Operation 3: activation of the 'grasp' action

Once the cup has been attended to, both as a template in inferotemporal cortex and a shape in caudal IPS, the agent selects an action to execute on it. There are good reasons to suppose that the complete motor programme is not selected until the cup has been visually attended to. As described in Section 2.5.3.1, F5 canonical neurons representing the specific grasp to perform in an upcoming action do not fire until the agent visually fixates the target object to be grasped (Constraint 3). This is good empirical evidence that attention to the target must precede activation of the grasp component of the motor programme. Moreover, recall from Section 2.6 that low-level grasp and reach trajectories are probably selected as components of high-level action categories, which are triggered by semantically rich representations of the candidate target object. The selection of an action category in PFC is unlikely to occur before the target attended to has been classified in inferotemporal cortex.

2.10.2.6 State 3: re-attention to the agent

Once an action category has been selected, the process of executing an action has a dynamical character, as described in Sections 2.5.2 and 2.5.3. The movement vector which generates a motor signal for the shoulder and arm muscles is updated in real time (with the aid of forward and inverse models), and the hand moves through a sequence of preshapes controlled through similar mechanisms. The experience of executing a high-level action is thus a gestalt of motion signals, proprioceptive information and activity in cerebellar forward and inverse models. This gestalt constitutes a dynamic representation of the agent, which is part of the agent's conception of himself. In other words, while the action is being executed, the agent is attending to himself, in the motor modality.

2.10.2.7 State 4: re-attention to the cup

The reach-to-grasp action terminates when the agent achieves a stable grasp on the cup. As argued in Section 2.7.8, this terminal state can be described as one in which the agent is holding the cup, but from the agent's perspective it can also be described epistemically, as a state of haptic re-attention to the cup: when a stable grasp has been achieved, the agent has a new source of information about the location and shape of the cup, delivered by the position of his arm and the shape of his fingers. At this point, the agent is reattending to the cup, in the haptic modality (Constraint 4). Given the sequence of confirmatory signals he has received so far, this new attention to the cup is all the confirmation the agent needs to know that the action is completed. Note that this final state does not have an associated attentional action; it is brought about by the dynamic process initiated when the agent activated the 'grasp' action.

2.10.3 The sequence of processes during perception of a reach action

2.10.3.1 Operation 1: Action of attention to the agent

Now consider a scenario in which an observer perceives another agent grab a cup. As suggested in Section 2.8.3, the first operation the observer must perform is to put himself into action perception mode (Constraint 12). In that section I suggested that the action which has this effect is simply the action of attending to the other agent; the action of attending to the most salient visual entity was seen as competing with the alternative action of putting onesself into action execution mode. In the current scenario, the former action wins, and the observer attends to the external agent, putting the mirror system into action perception mode.

The mode-setting model of agency outlined in Section 2.8.3 predicts that the observer's first attentional action when observing a reach-to-grasp action will be to the agent (rather than to the target). Webb *et al.*'s (in press) study of eye movements during observation of reach-to-grasp actions confirms that this is the case; observers' initial saccades were overwhelmingly to the agent (Constraint 5).

2.10.3.2 State 1: Attending to the agent

The sensory consequence of attending to the agent is a state where LIP and frontal eye fields encode a certain location (the location of the agent) which allows IT to compute a representation of the agent (in our case, 'man')—see again Constraint 1. Note that the observer will not be representing the attended-to man in parietal cortex as a potential target object to be reached for, since he is in action observation mode. Instead, as argued in Section 2.7.2, representations develop in the observer's STS which encode the observed agent's direction of eye gaze, and begin to encode a pattern of biological motion. (Note that Webb *et al.* found that observers' initial saccade tended to be to the agent's face, which facilitates the extraction of gaze information.)

2.10.3.3 Operation 2: Action of attention to the cup

In Webb *et al.*'s eye movement study, after an observer's initial saccade to the agent, there was a strong tendency for the next saccade to be to the target object (Constraint 2). As in Flanagan and Johansson's (2003) experiment, this saccade to the target object occurred in advance of the agent's hand reaching it; in other words, the observer *anticipated* the agent's intended target. In ordinary action recognition, this anticipation is due in part to the establishment of joint attention, and in part to an extrapolation of the trajectory of the observed agent's hand. As argued in Section 2.7.2, both these processes are likely to involve computations in STS.

2.10.3.4 State 2: Attending to the cup

Once the observer has attended to the cup, it is categorised, allowing a representation of its category ('cup') to be computed in the observer's IT (see again Constraint 1).

2.10.3.5 Operation 3: Activation of the 'grasp' action

Once the observer has attended to and classified the agent's intended target object, action recognition proper can begin. There is good evidence that the observer needs to compute the location of the agent's intended target before he can identify the agent's action. As

we have already seen, many cells in STS which respond to a particular hand action on an object do not respond if the object is not present (Section 2.7.2; Constraint 6). Most mirror neurons in F5 similarly require the presence of the target object (see Section 2.7.3; Constraint 8). These findings suggest that the process of activating an action category in F5 and STS requires a representation of the location of the target object—a principle which is echoed in computational models of action recognition (see Constraint 10).

2.10.3.6 State 3: Re-attention to the agent

We saw in Section 2.7.2 that classifying an observed hand action probably involves the 'biological motion' system as well as a computation of the trajectory of the hand onto the target. As discussed at the end of that section, biological motion requires an analysis of the observed agent's body movements as a whole, rather than just attention to the agent's hand or arm. Biological motion appears to require some limited form of visual attention. So, while the observer's overt attention is definitely on the target during the latter stages of the agent's action, it is likely that the observer allocates some attention back to the agent while the action is being classified (Constraint 7), and while a cross-modal representation of the agent is being created (Constraint 9).

2.10.3.7 State 4: Re-attention to the cup

Once the agent's action is completed, the observer must register this fact. As outlined in Section 2.7.8, the visual perception of a stable grasp triggers matching somatosensory representations in the observer—another example of the motor basis for action recognition. Thus in the observer, as in the agent, the endpoint of the cup-grabbing action is characterised by re-attention to the cup in the haptic modality (Constraint 11).

2.11 Summary

In this chapter I have given a very detailed overview of the sensorimotor pathways involved in execution and perception of a simple reach-to-grasp action. In one sense, the story is extremely complicated. But in Section 2.10 I argued that at the level of temporal organisation, the structure of the sensorimotor processes is fairly simple. I proposed that both the execution and the perception of a cup-grasping action involve the same characteristic sequence of attentional and motor operations, interleaved with the sensory consequences of these operations. The sequence in each case is summarised in Table 2.1. The operations and states can be thought of as a deictic routine, in the sense defined by Ballard *et al.* (1997). The 'attend to agent' operation brings about the state 'attending to agent'; this

Deictic operation	Sensory consequence
Attend to the agent	Attending to the agent
Attend to the cup	Attending to the cup
Activate 'grasp' action	Re-attention the agent
	Re-attention to the cup

Table 2.1: Deictic routine underlying the execution or perception of a cup-grasping action

state in turn enables the 'attend to cup' operation, and so on. The last state is an exception; it arises as a result of the dynamic processes initiated by activation of the 'grasp' action, rather than because of a separate deictic operation.

Of course it is somewhat suspect to derive so clean a model from the mass of complex (and often controversial) data outlined in this chapter. But recall that our ultimate goal is to describe how information in the sensorimotor system is encoded in language. Obviously a huge amount of abstraction is needed in order to create the semantic episode representations which are associated with sentences. And in fact I think the basic sequence of sensorimotor operations outlined in Table 2.1 does emerge quite clearly from the data reviewed. In summary, I suggest that the canonical temporal structure of sensorimotor processing described in Table 2.1 provides a plausible basis for the formation of more abstract representations of reach-to-grasp episodes in working memory and longer-term memory. It is to this topic which I will turn in Chapter 3.

Before I continue, there is an important question to ask about the status of the sequence just proposed. Does experience of a reach-to-grasp action *always* have the sequential form proposed in Table 2.1? Or is this just the default or canonical order of sensorimotor events? Some of the constraints I have proposed (e.g. Constraints 2 and 5) refer to the order in which events *normally* happen. Others I have argued *must* hold. Some of these (e.g. Constraint 11) follow fairly directly the definition of a reach-to-grasp action. Some (e.g. Constraints 12 and 13) are corrolaries of the model of the mirror system I have adopted. Finally, there are some constraints about the formation of cross-modal object representations (Constraints 4 and 9) which I suggest must hold at some point during development, while the concepts of agents and manipulable objects are being formed, but perhaps need not hold so rigidly thereafter. The main idea I want to propose is

that the canonical sequential order is what is used to structure episode *representations*. I certainly want to allow for the possibility that a mature observer can *infer* a reach-to-grasp actions from a sensorimotor process which does not have exactly the sequential structure in Table 2.1—for instance, we can identify such actions from a glimpse, or a static drawing. (It is not obvious where the boundary between perception and inference is, but I want to stay clearly on the perceptual side of the boundary for the moment.) I also want to allow that a mature agent might have ways of optimising the performance of a reach-to-grasp action which omit or combine certain steps of the sequence. But if the structure of episode representations is established early enough during development, these actions will still be *represented* using the canonical sequence I have just described.

Chapter 3

Models of learning and memory for sensorimotor sequences

The model presented in the previous chapter describes the sequence of sensorimotor events that occur in an agent during the observation or execution of a grasp action. Recall that our ultimate aim is to provide a theory of syntactic structure which is grounded in sensorimotor processes. But note that this grounding cannot be too direct, since executing or observing a motor action does not trigger a linguistic utterance as a reflex, and many utterances encode actions or events which occurred some time in the past. As forcefully advocated by Itti and Arbib (2006), the semantic representations which underlie an agent's utterance must come from the agent's *memory* rather than from immediate sense perception.

How is the sensorimotor material associated with our cup-grabbing event stored in memory? There are two forms of memory to consider. The first is **working memory**: a short-term store of information which an agent can actively maintain and manipulate, which can be used to buffer perceived events or plan forthcoming actions. The second is **episodic memory**: a longer-term store of events the agent has experienced. Speakers can describe events retrieved from episodic memory as well as those which have just occurred, so this form of memory must be considered as well.

The distinction between working memory and episodic memory is well established in psychology. Damage to frontal areas can impair immediate recall while leaving longer term memory for events relatively intact, while damage to the hippocampus and related areas can impair the ability to form long-term memories while preserving immediate recall (Baddeley and Warrington, 1970; Shallice, 1988). I will argue that both forms of memory preserve the sensorimotor sequence which characterises the cup-grabbing action when it is first presented to the agent—in other words, that both working memory and episodic memory are able to store this sequence. In the model I propose, the sequence is initially stored

in working memory; it can subsequently be *replayed* to longer-term storage in episodic memory. When it is retrieved from episodic memory, it is replayed back into working memory, to recreate a working memory representation very similar to the one created after the original experience of the event.

In the first part of this chapter I will consider the working memory representation of a sensorimotor sequence. In Section 3.1 I summarise a well-known account of working memory, developed over several years by Alan Baddeley and colleagues. In Sections 3.2 and 3.3 I outline a model of the role of working memory representations in the planning and execution of action sequences. In Section 3.4 I extend this account to the observation of action sequences. In Section 3.5 I outline a model of how sequences held in working memory can be internally replayed.

In the second half of the chapter, I look at how events are stored in and recalled from episodic memory. In Section 3.6 I introduce a basic model of episodic memory, and of the key neural system involved, the hippocampal system. In Section 3.7 I argue that events are stored in episodic memory in a form which retains their characteristic sequential structure, and survey evidence that the hippocampus is in fact specialised for storing information in the form of sequences. In Section 3.8 I outline the role of working memory in the encoding and retrieval of episodic memory sequences. There is evidence that sequences are buffered in working memory and then replayed to the hippocampus for longer-term storage, and that hippocampal sequences are played back to this same working memory buffer when they are retrieved.

In the model of memory I present, events recalled from episodic memory create working memory representations which are very similar to those created immediately after an event is experienced. In each case, the working memory is of the sensorimotor sequence which characterised the original experience of the event, and in each case the agent is able to internally replay this sequence to re-evoke this original experience. In the account of the relationship between syntax and sensorimotor structure which I develop in Chapter 5, the process of replaying a sensorimotor sequence buffered in working memory will function as the key interface between language and the sensorimotor system.

3.1 Baddeley's model of working memory

The classic model of working memory in cognitive science is that of Baddeley and Hitch (1974). In this model, working memory is defined as a short-term store of information which subserves 'cognitive' operations: language processing, learning and reasoning. As Baddeley (2000) notes, the term 'working memory' is also used in behavioural psychology in relation to models of action preparation. A prepared action or a predisposition to respond to stimuli

in a particular way is often referred to as being maintained in working memory: in other words, it is where an animal's 'task set' is stored. We have already encountered this sense of working memory: in Section 2.6 I suggested that the PFC is responsible for imposing learned top-down biases on an agent's actions, and outlined a model by Miller and Cohen (2001) proposing how these biases are delivered. Baddeley (2000) is at pains to distinguish these two senses of working memory. He notes in particular that 'task sets' can endure for quite long periods, whereas the memories he is concerned to model last on the order of seconds before decaying. However, given our project of looking for commonalities between language and sensorimotor processing, it is useful to look closely at these two conceptions of working memory, to see if they are really as distinct as Baddeley makes out. In this section I will introduce Baddeley and Hitch's model of working memory, as recently revised and extended by Baddeley (2000). In Sections 3.2–3.5 I will give an account of working memory for sensorimotor sequences, and argue that this notion of working memory can in fact be quite closely integrated into Baddeley's model.

Baddeley and Hitch's account of working memory is modular in conception. It involves two 'slave' modules, which can maintain small amounts of information over short periods, and one 'central executive' module, which controls the operation of these slave modules and mediates exchange of information between them. Baddeley (2000) added a third slave module to the model. A diagram illustrating the extended model is given in Figure 3.1. I

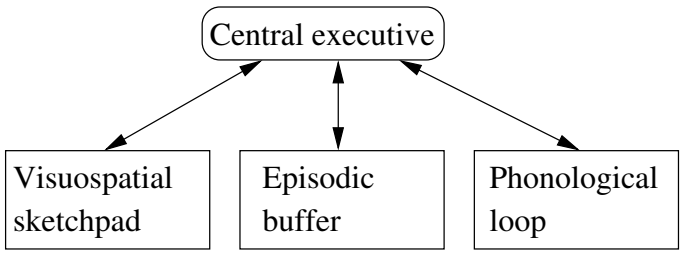


Figure 3.1: Baddeley's (2000) revised model of working memory. (Interfaces with long-term memory not shown.)

will describe each of the slave modules in turn.

3.1.1 The visuospatial sketchpad

In Baddeley's model, a special form of short-term memory underlies our ability to retain static visual shapes and patterns. It is indeed likely that memory for visual patterns has its own separate neural substrate, making use of representations in inferotemporal cortex (see

e.g. Miller *et al*, 1993). The patterns held in this form of storage can be spatially complex, but not temporally complex; the visuospatial sketchpad is assumed to be unsuited for the encoding of sequences of patterns, which are hard to remember without special mnemonic strategies (see e.g. Phillips and Christie, 1977).

3.1.2 The phonological loop

The phonological loop is another modality-specific short-term store. It allows the maintenance of a short sequence of words or phonemes for a short period of time. It is often described as the form of storage we use to retain a phone number during the time between reading it and using it. Evidence for this form of storage comes mainly from experiments on short-term memory for phonological sequences, in which subjects are given a sequence to remember, and asked to recall it a short time later. A common quantitative measure of phonological memory is a subject's **digit span**: the largest number of digits which the subject can recall perfectly on 50% of trials.

The phonological loop is inherently sequential: it is a memory for sequences of items. One of its main characteristics is support for **rehearsal**: a sequence in phonological working memory can be replayed (vocally or subvocally), which refreshes it and allows a sequence to be stored for longer periods. Items in the loop are stored as sounds (or articulatory representations), rather than as semantic structures. Evidence for this comes from the **phonological similarity effect** (a sequence of phonologically similar items is harder to remember than a sequence of phonologically distinct items) and the **word-length effect** (a sequence of long words is harder to remember than a sequence of short words, suggesting they are encoded at least in part using phonological representations). Evidence for the idea that phonological sequences are retained by rehearsal comes from studies of **articulatory suppression**: if subjects are asked to retain a phonological sequence while reciting unrelated phonological material (e.g. while repeating the word *the*), their performance deteriorates considerably.

I will discuss the phonological loop in more detail when I consider the neural substrate of natural language in Section 6.1.

3.1.3 The episodic buffer

The episodic buffer is a semantic form of working memory, specialised for encoding 'events'. Baddeley (2000) mentions several reasons why this module is needed as an extension to his original theory. One argument concerns the mechanism via which episodes are stored in long-term memory; this will be discussed in detail in Section 3.8.1.3. The other arguments

relate to findings about working memory which are hard to explain using purely phonological or visuospatial representations. I will consider two of these, which both relate to memory of material presented linguistically.

The first finding is that sentences are easier to recall than unrelated sequences of words. If a word sequence to be retained takes the form of a sentence, subjects' span is typically around 16—much higher than their span for unrelated words. The second finding is that subjects are able to report the 'gist' of a paragraph consisting of over fifteen proposition-sized elements quite reliably. It appears this ability does not require long-term memory, as it can be quite well preserved in severely amnesic patients (see for instance the case study in Wilson and Baddeley, 1988).

To explain how the episodic buffer helps account for these findings, it is useful to begin by looking at a simple case in which semantic structure can help in encoding a memory stimulus. Consider two sequences of ten phonemes: the first is a sequence of five two-syllable words; the other is a sequence of five two-syllable nonwords. Unsurprisingly, the former sequence is easier to learn (see e.g. Hulme *et al.*, 1991). It is frequently assumed that this sequence is learned by a **chunking** mechanism, in which the phonological buffer stores a sequence of pointers to semantic items encoded in more permanent storage (in this case words). When each item in the sequence of words is recalled, the associated sequence of phonemes can be recovered. (Some residual component of purely phonological storage must nonetheless be assumed, to account for the word length effect described above.) In Baddeley's account of the episodic buffer, our improved working memory for semantically structured material requires the postulation of a short-term semantic store, which interacts with the phonological store using a mechanism similar to chunking. If a sequence of words stored in the phonological buffer is a sentence which describes an event, this event can be stored in the episodic buffer (in a compressed form). When recovering the sequence, a subject can rely on this semantic representation alongside the phonological sequence. If a sequence of sentences is to be recalled, each episode can be encoded in semantic working memory as a chunk (again in compressed form), and the phonological buffer can hold pointers to these chunks.

What is the format of semantic information stored in the episodic buffer? Baddeley argues that the buffer has an integrative role, combining information from different sensorimotor short-term stores with phonological representations and with semantic representations in long-term memory. He also suggests that information in the buffer must be rehearsed to be retained, just like information in the phonological buffer. Evidence for both of these claims comes from a study by Baddeley and Andrade (2000). In this study, subjects were shown stimuli varying in two dimensions: 'meaningfulness' (high vs low) and modality (visual vs phonological), and after a short interval asked to rate the vividness of these stimuli. During the interval they had to perform a modality-specific distractor task,

of the kind known to interfere with working memory rehearsal. They found that a visuospatial distractor task impaired the perceived vividness of visual stimuli but not auditory stimuli, while a phonological distractor task had the converse effect. What is more, these modality-specific distractor effects also obtained for 'meaningful' stimuli which presumably evoke representations in semantic memory. For Baddeley, these results suggest the existence of a working memory store which is maintained via rehearsal, which integrates information across modalities, and which interfaces with phonological representations. He identifies this store with the episodic buffer.

A final suggestion by Baddeley (2000:420) is that the multimodal material making up a single episode in the episodic buffer is rehearsed by a process of 'sequential attention' to the different components of information involved. He does not address the question of what form this sequence takes. However, the sensorimotor model presented in Chapter 2 suggests a specific answer to this question, at least as regards a cup-grabbing episode. In Sections 3.2–3.4 I will return to our cup-grabbing action, and outline a model of the working memory representations which underlie the execution and perception of this action. Later in the chapter I will argue that these representations can be quite closely identified with Baddeley's episodic buffer: they support internal rehearsal and integrate information across modalities (see Section 3.5), and they can function to buffer and replay information to long-term memory (see Section 3.8.1.3).

3.2 Working memory representations of action sequences in PFC

If an action like grasping a cup is decomposed into a sequence of smaller sensorimotor actions (attend-to-self, attend-to-target, activate-grasp), the question arises as to how the action is planned as a single unit. In situations where cup-grasping is required, the agent must be able to prepare the sequence in advance, so the actions occur in the right order. We have already looked at models of action preparation or 'task set'; in Section 2.6.2 we saw how PFC is involved in biasing an agent towards a particular response, or a particular stimulus-response pathway. (This account was of the *other* sense of 'working memory', as used by behavioural psychologists.) However, we have not yet considered the interesting question of how an agent prepares a *sequence* of actions. How must the notion of task set be extended to account for the preparation of sensorimotor sequences?

There is good evidence that the dorsolateral PFC is also heavily involved in sequence planning; see e.g. Tanji (2001), Tanji *et al.* (2007) for reviews. Lesion studies both in monkeys (e.g. Petrides, 1991) and in humans (e.g. Petrides and Milner, 1982) show that

damage to the PFC results in impairments in the performance of sequential tasks. In this section, I introduce two models of the PFC's role in the planning and execution of motor sequences, and summarise evidence for both models.

3.2.1 Competitive queueing

One popular model of action sequencing is **competitive queueing** (see e.g. Grossberg, 1978; Houghton, 1995, Rhodes *et al.*, 2004). In this type of model, the individual movements in a planned sequence are activated in parallel during a preparatory phase, and the order in which they are executed is determined by their relative levels of activation, with the movement activated most strongly being executed first.

Competitive queueing models typically involve two levels of action representation—see Figure 3.2. At a **planning level**, the component actions in the prepared sequence are

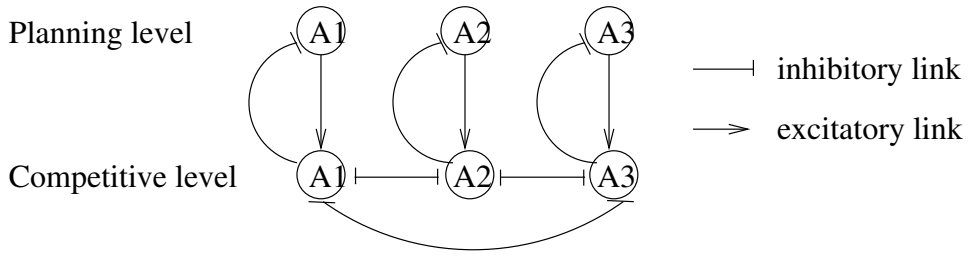


Figure 3.2: A competitive queueing model of action sequencing (after Houghton and Hartley, 1995)

active in parallel. At a second **competitive level**, the same actions are represented, and each receives activation from its associated action in the first level. However, in the second level, a winner-take-all scheme is in place, which selects a single action as the action to be executed next. Whichever action is selected as the winner at this level subsequently inhibits its corresponding action at the planning level, so that the next-most-active action at this level becomes the next winner and is executed in turn. For instance, in Figure 3.2, assume that the most active action in the planning level is A1, followed by A2 and A3. To beging with, A1 will dominate at the competitive level, and will be executed. After a delay, A1 will inhibit itself at the planning level. At this point, A2 will dominate at the competitive level, and thus will be executed next. Similar reasoning causes A3 to be executed third.

Note that the competitive queueing model is very similar to the model of serial visual search by inhibition-of-return proposed by Itti and Koch (see Section 2.4.5). In their

computational model, there are two representations of salient regions of the visual scene: in one, multiple regions are active simultaneously (the planning layer), and in the other, the most salient of these is selected through lateral inhibition (the competition layer) and then inhibits its associated region in the planning layer.

3.2.1.1 Evidence for competitive queueing mechanisms in dorsolateral PFC

There is good direct evidence for the competitive queueing model of sequencing. To begin with, at a behavioural level, the model nicely predicts a type of error which is common in the execution of action sequences, namely **reversal** of the serial order of two adjacent actions. For instance, when typing a word, it is common to reverse the order of adjacent letters (typing 'fro' instead of 'for', for instance). If the order of actions depends on the initial activation levels of units representing them, then reversals will result from a situation where one action has higher initial activation than it should (perhaps due to noise, or other external factors).

There is also firm evidence for competitive queueing from neurophysiological studies. In a very elegant study, Averbeck *et al.* (2002) trained monkeys to draw simple polygons (triangles and squares). After training, they identified groups of cells in the dorsolateral PFC which encoded different individual movements in shape-drawing sequences. They found that cells encoding all the individual movements for a given shape were active when the monkey prepared to draw that shape—and moreover, that the relative activation level of cell groups predicted the order in which movements were generated.

3.2.1.2 Adding context to a competitive queueing model

Competitive queuing models have difficulty accounting for planned sequences which involve repeated actions. Each action type is only represented once at the planning and competitive levels, so the initial gradient in the planning level cannot represent two occurrences of the action. To allow for repeated actions, competitive queueing models are often augmented with an external 'context signal' which evolves independently in time. An extra 'context layer' is added to the network to hold this signal. Links between the context layer and the planning layer allow actions to become active at different moments in time. Houghton and Hartley (1995) show that a combination of competitive queueing and time-activated action preparation can model a rich variety of prepared actions.

3.2.1.3 Tonically active representations in competitive queueing models

In the competitive queueing model, when an action sequence is executed, its representation at the planning level is destructively updated. However, there are many situations where

the sequence needs to be *retrieved* after it is executed. For instance, in Averbeck *et al.*'s experimental paradigm, the monkey is rewarded for performing the same sequence several times. The monkey must thus remember the correct sequence from one trial to the next.

A competitive queueing model must provide an additional mechanism for a sequence to be retained in this way. One common suggestion is that this additional mechanism is working memory in Baddeley's sense—i.e. a declarative memory for a sequence of items. In Rhodes *et al*'s (2004) influential competitive queueing model, a representation of the activation gradient associated with a particular sequence is stored separately in working memory, and transferred to the planning level each time it needs to be executed. In their model, this additional working memory storage is not just a stipulation to allow for repeated sequences. The mechanism for loading a new sequence into the planning level also allows the model to reproduce some of the timing characteristics found in human planned action sequences—in particular, the fact that longer sequences take longer to initiate.[1]

Rhodes *et al.* suggest that the working memory of an action sequence is held in dorsolateral PFC, based on the experiment of Averbeck *et al* (2002) cited earlier. But since the PFC signal identified in this experiment was modified as the monkey executed a planned sequence, it does not correspond well to the working memory signal; it appears much more like the signal in the planning layer. However, a more recent experiment by Averbeck and Lee (2007) finds evidence that dorsolateral PFC also contains a representation of a planned sequence which survives sequence execution. In this experiment, monkeys were trained to perform a succession of different three-movement motor sequences; each sequence was performed repeatedly for a block of trials. Recordings were made from PFC cells in the intervals between trials as well as during trials. It was found many dorsolateral PFC cells which encoded the upcoming sequence during the interval period. The key point for a competitive queueing model is that even just after a sequence was executed there remained a fairly strong signal in PFC encoding the prepared sequence. In other words, PFC holds a *tonic* representation of each action in a prepared sequence as well as the phasic representations which are inhibited when the sequence is executed. Averbeck and Lee consider the possibility that this tonic signal is transferred to PFC from another region each time it is prepared, but also suggest that the PFC might be responsible for holding the tonic signal itself.

[1] Another possibility is that the gradient of activation representing a sequence can be regenerated spontaneously, when neurons recover from inhibition and return to their previous level of activation (see Burgess and Hitch, 1999 for a model of this kind).

3.2.2 Associative chaining

Another popular model of action sequencing is known as **associative chaining**. The basic idea here is that the sensory consequences of executing the first action in the sequence trigger the execution of the next action. Arbib and Dominey's model of the sequencing of individual components of a grasp action is of this kind; for instance, the completion of the hand-opening action during a grasp preshape is what triggers the hand-closing command. However, it is also interesting to consider an associative chaining account of higher-level action sequences, such as the sort we are now considering. In this section I will present a version of associative chaining which draws on Miller and Cohen's model of PFC as a mechanism for delivering tonic biases to selected stimulus-response pathways.

The diagram presenting Miller and Cohen's model of PFC is repeated in Figure 3.3. To allow this model to encode prepared sequences, it suffices to imagine that some of the

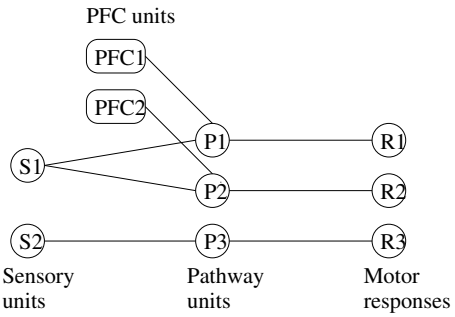


Figure 3.3: Miller and Cohen's (2001) model of PFC neurons

stimuli are reafferent consequences of earlier actions. When the first action in the sequence is executed, reafferent feedback about its successful completion will be generated. One of the primed stimulus-response pathways maps this reafferent signal onto the second action in the sequence, which results in the second action being generated, and so on. This model is illustrated in Figure 3.4. In this network, if PFC1 is active, it biases two separate stimulus-response pathways, one between stimulus S1 and response R1, and one betweeen stimulus S2 and response R2. If we assume that R1 generates S2 as reafferent sensory feedback, then this PFC unit will effectively trigger a sequence of actions: R1 followed by R2. (Naturally, it is more realistic to assume that *groups* of PFC cells encode sequences—we do not want to envisage a PFC cell for each sequence which could be learned. However, a localist model can illustrate the key idea of PFC activity biasing more than one pathway.)

There is good evidence for an associative chaining account along these lines. Barone and Joseph (1989) trained monkeys to attend to a visually presented sequence of objects, and

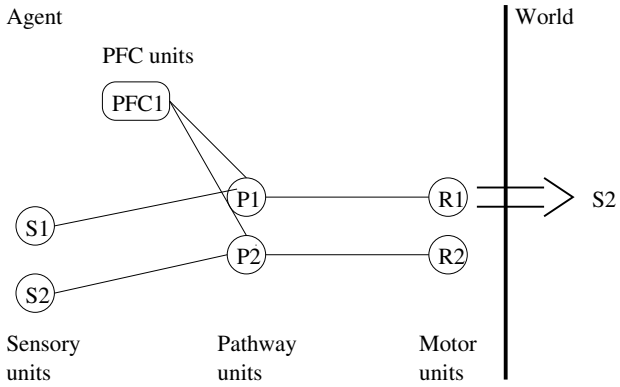


Figure 3.4: A PFC cell encoding a planned sequence of two actions (R1, R2)

then execute saccades to these objects in the same order after a delay period. Recording from the dorsolateral PFC during the delay period, they found neurons which responded only to specific sequences of targets. A neuron of this type can be interpreted as encoding a particular set of stimulus-response biases, which result in a sequence of actions being performed.

The idea that a sequence of actions can be prepared by selectively priming a set of stimulus-response pathways can be found in several models of sequencing. The model outlined above, in which multiple stimulus-response pathways are primed, is perhaps closest to a model by Shima and Tanji (2000). Shima and Tanji gave monkeys a task in which a sequence of three visual locations was presented; after a delay, the monkeys had to make saccades to the locations in the same sequence. In their model of this process, a perceptual representation of the sequence is first constructed, and this representation activates two separate elements, one linking the first and second actions, and one linking the second and third actions. However, these elements are not exactly stimulus response pathways, but rather links from one response to another; in addition, they are not tonically active throughout the prepared action, but only in sequence, during the preparation of individual actions. The model which I gave predicts that there will be elements representing whole prepared sequences which remain tonically active during the execution of the full sequence. For instance, in Averbeck *et al.*'s (2006) study, 9% of task-related PFC neurons studied showed a main effect relating to which sequence was being executed, but not to individual movements within a sequence (Averbeck, pc).[2]

[2]One potential objection to this model is that it is combinatorially expensive, requiring one 'pathway unit' for each possible stimulus-response pairing. However, the number of stimulus-response pathways

3.2.2.1 Adding context to the model

As with competitive queueing, the above model needs to be extended to allow the preparation of a sequence containing repeated movements. For instance, assume we want to plan the sequence R1, R2, R1, R3. We cannot just use reafferent feedback from the most recent action to trigger the next action, because the completion of R1 must trigger R2 in the first part of the sequence, but R3 in the second part. To solve this problem, the network needs to be able to maintain some internal state, keeping track of the history of recent events. The way to enable this is to add a representation of 'context' into the network, whose value at any point is a function of the current event *and of its own previous state*. A network which contains a layer updated by its own previous state is called a **recurrent network**.

The potential of recurrent networks for encoding sequences was first explored by Elman (1990) using simple artificial networks. Since that time, there have been several suggestions about how recurrent structures in real neural circuits might enable sequence encoding. Again, the focus has been on circuits involving the prefrontal cortex, because of its known role in sequence representation. In a model of action sequencing by Dominey *et al.* (1995), PFC is assumed to receive input both from itself (at a previous time point) and from an efferent copy of the current motor command. In a model of sequence encoding by Beiser and Houk (1998), PFC receives input directly from sensory stimuli, and is also reciprocally connected to structures in the basal ganglia; the loop involving PFC and the basal ganglia creates a recurrent source of input to PFC, and allows it to store sensory sequences. These models differ as to whether the PFC is updated by efferent copies of motor commands (Dominey *et al.*) or by perceptual stimuli (Beiser and Houk), but they share the idea that PFC participates in a recurrent circuit which allows it to encode sequences of recent events.

To add context to our model, I will assume that there are two separate areas of PFC: one for holding pathway-biasing units representing prepared actions (as in Miller and Cohen's model) and one for holding a representation of context encoding recent actions. The reason for assuming two areas is that the two types of unit have different basic properties: the context-encoding PFC units must be regularly updated by incoming events, while the pathway-biasing units, which hold the agent's current plan, must precisely *not* be subject to alteration by arbitrary current stimuli.

The model I propose is illustrated in Figure 3.5. In this model, each sensory stimulus updates the PFC context units. These units are also updated by a copy of their previous value (indicated by the recurrent loop), so that at each point they store a representation

which are actually used is likely to be much smaller than this. It may be that distributed representations of pathway units enable a smaller pool of units to encode the relevant stimulus-response associations. In addition, I will propose some methods of encoding attentional actions in Section 8.2.4.1 which strongly constrain the number of reafferent stimuli featuring in WM action representations.

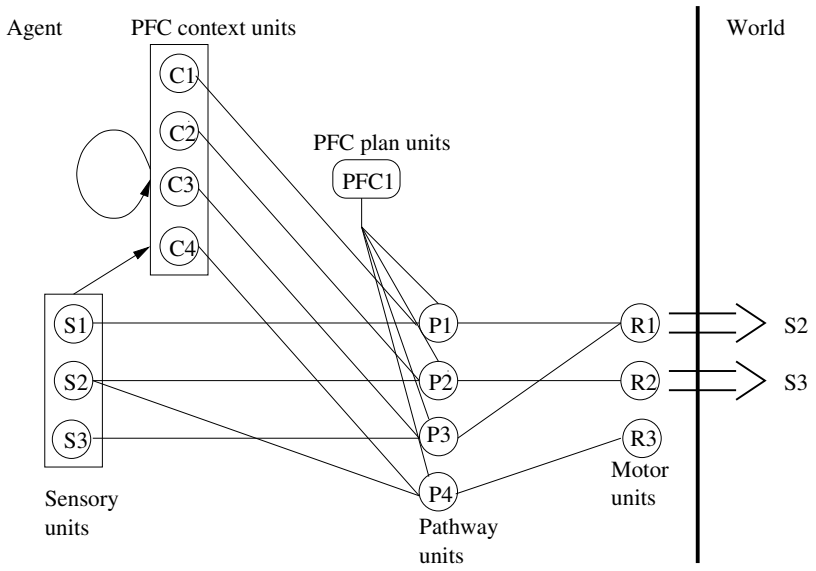


Figure 3.5: PFC sequencing network (extended with recurrent context representations) for encoding the sequence R1, R2, R1, R3

of the sequence of past stimuli, with an emphasis on the most recent. Instead of simple stimulus-response pathways, the model assumes pathways from current stimulus *and current context* to responses. As in the previous model, a PFC plan unit functions to bias the agent towards a particular set of these pathways, labelled as P1–P4. P1 is a pathway which maps an initial stimulus and context (S1 and C1) to response R1. P2 maps stimulus S2 and context C2 to response R2. P3 maps stimulus S3 and context C3 to response R1. P4 maps stimulus S2 and context C4 onto response R3. We assume that R1 produces S2 as sensory feedback, and R2 produces S3. We also assume that the sequence of sensory inputs S1, S2, S3, S2 causes the PFC context units to move through the sequence of states C1, C2, C3 and C4. With these assumptions, in the initial context S1-C1, the bias imposed by plan unit PFC1 will cause the agent to execute the sequence R1, R2, R1, R3.

3.2.3 PFC sequencing models and the reach-to-grasp action

It is likely that prepared sequences in working memory are represented using a combination of associative chaining and competitive queueing methods. Competitive queueing methods may dominate for short, well-learned movement sequences, where the timing and effects of individual movements are highly predictable—for instance, in the production of

sequences of phonemes during speaking, or of keypresses during typing. But for higher-level behavioural sequences, which take place over longer timescales and involve actions whose effects are less predictable, associative chaining models are more suitable.

Which sequencing model is most appropriate for our cup-grabbing sequence ('attend-to-self', 'attend-to-cup', 'activate-grab')? It is not easy to say. 'Attend-to-self' is presumably quite a reliable action, with predictable effects, so competitive queueing might work for the first two actions. But the effects of the second action, 'attend-to-cup', are less predictable—there may be no cup to attend to. So executing the grab action should probably wait until feedback about the success of this action has been received. Perhaps a transitive action like reaching-to-grasp is controlled using a mixture of competitive queueing and associative chaining mechanisms. However, note that even if reafferent sensory consequences of actions are not used to control sequence execution, they nonetheless still occur at the appropriate times when the sequence is performed.

3.2.3.1 Tonic and phasic sequence representations in PFC

It is important to note that both models of PFC sequencing require the existence of *tonically active* representations of the actions in a prepared sequence, both prior to the execution of the sequence, and during and even after its execution. The competitive queueing model requires a tonically maintained working memory of the activation gradient which defines a prepared sequence, which is not destructively updated when the sequence is performed, and which can be used to re-establish the sequence plan if the sequence is to be repeated (see Section 3.2.1.3). The associative chaining model also predicts the existence of tonically active sequence representations which survive sequence execution—at least, if it is formulated within Miller and Cohen's pathway-biasing account of PFC. In this model, PFC cells encode something like condition-action rules: these rules can remain active even after the sequence has been executed, because the conditions under which each rule triggers only occur once during the execution of the sequence.

3.2.3.2 Context representations in PFC

Note also that both models of sequence execution include a notion of context, which is updated as a sequence is executed. The context representation in the competitive queueing model evolves as a function of time, while the context representation in the pathway-biasing model reflects events occuring during sequence execution. Given that the cup-grabbing action is probably driven by a mixture of these two models, the context representation in PFC is likely to be a mixture as well.

3.2.3.3 A graphical notation for working memory sequence plans

The notion of a working memory sequence plan will play a central role in the model developed in this book, so it is useful to introduce a graphical notation for this structure. I will use the structure shown in Figure 3.6 to represent a sequence plan stored in working memory, at a level which is noncommittal about whether the sequence is encoded using competitive queueing or associative chaining or some combination of the two. This struc-

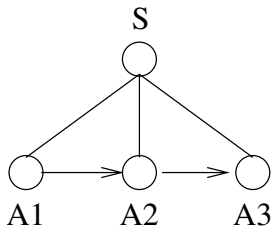


Figure 3.6: A graphical representation of a working memory plan to perform action sequence A1, A2, A3

ture represents the plan to perform the action sequence A1, A2 and A3. In a competitive queueing paradigm, the node S can be thought of as a PFC assembly which stores a gradient of activation over the three actions, such that actions earlier in the sequence are more active. In an associative chaining paradigm, S can be thought of as a PFC assembly which imposes a set of biases on the pathway from stimuli (and context representations) to responses, which primes the agent to perform A2 when A1 is complete, and to perform A3 when A2 is complete.

3.2.4 Reinforcement regimes for learning PFC sequence plans

We have now outlined two mechanisms for planning and executing a sequence of actions. However, we have not considered how it is that the agent comes to activate a particular sequence plan in PFC in a particular situation. A model of **reinforcement learning** is required to answer this question. The question of which plan to activate is ultimately a question about what is best for the agent. The tricky thing about reinforcement learning for action sequences is that the reward typically only arrives when the sequence is complete. There are several computational models of the process by which PFC learns to activate a plan in return for a reward which only occurs when the plan is fully executed (e.g. Braver and Cohen, 2000; O'Reilly and Frank, 2006). These models propose that plan learning is mediated by a class of neurons in the midbrain called **dopaminergic neurons**, which

appear able to *predict* rewards some time in the future. We do not need to go into the details of these models, but it is useful to know that there are reinforcement-based accounts of how sequence plans are acquired in PFC.

3.2.5 Summary

We now have an account of the working memory sequence representations in PFC which can function as a 'plan' to grab a cup. On this account, an agent who has a plan to grab a cup has an active assembly in PFC encoding a particular sequence of sensorimotor operations: 'attend-to-self', 'attend-to-cup', and 'grab'. (This assembly is active because the associated sequence has been found to be beneficial for the agent in similar situations in the past.) When the assembly is activated, 'attend-to-self' is performed, the agent moves into action-execution mode, and the agent receives reafferent feedback of this. Then 'attend-to-cup' is executed, and the agent receives reafferent feedback in the form of a representation of 'cup' (both in IT, as a category, and in parietal/premotor cortex as a set of affordances). Then 'grab' is activated. This imposes a strong top-down bias on the motor system to perform a particular kind of action on the cup. The PFC-imposed 'grab' action category then plans a sequence of biases on the movement vector, and a suitable sequence of hand preshapes, and a motor action is performed. Throughout this process, there remains a tonic representation of all three actions ('attend-to-self', 'attend-to-cup' and 'grab') in the working memory component of PFC.

3.3 Competition between PFC plan assemblies

When an agent explores a range of action sequences in a range of different situations, he will develop a number of different PFC plan representations, which are triggered by different stimuli. If we drop the localist assumption used in the illustrative networks above, each plan representation will consist of an assembly of cells. Cells within a given assembly should reinforce one another, so that activation of some elements of a plan leads to activation of the complete plan. Ultimately, cells in different assemblies should inhibit one another, to encourage the adoption of a single plan in any given situation. At some level, therefore, we expect a 'winning-assembly-take-all' structure in PFC.

However, it also makes sense to envisage an earlier level of plan representation in PFC, where several alternative plans can be simultaneously represented without competing with one another, similar to the planning layer in a competitive queueing model. At this level, multiple alternative plans are represented, and evidence accumulates for different plans from a range of different sources. At a later level, alternative plans compete with one

another, and the winning plan is selected to control the agent's behaviour. The idea of a working memory (PFC) medium in which alternative action plans compete with one another has been suggested several times, perhaps most notably in the contention scheduling model of Norman and Shallice (1986). In this section I review some evidence for plan competition in PFC, and discuss some possible mechanisms for plan selection and plan termination.

3.3.1 Evidence for multiple alternative plans in dorsolateral PFC

A good piece of evidence that dorsolateral PFC represents several alternative plans simultaneously comes from another study by Averbeck and colleagues (Averbeck *et al.*, 2006). In this study, monkeys were trained to perform two alternative sequences of eye movements, in response to two different cues. Each day, different cues were chosen to represent the two desired sequences. Halfway through the day, after the monkeys had learned the cue-sequence mapping very well, the cues were reversed, so that monkeys had to learn the opposite mapping of cues to sequences. During the period when the monkeys were relearning the mapping from cues to sequences, a cue triggered two ensembles of cells in dorsolateral PFC, one representing each planned sequence. Soon after reversal, the old, 'wrong' sequence encoding dominated, and as the monkey learned the remapping, the 'right' sequence encoding began to dominate. The relative strengths of the two sequence encodings closely predicted the monkey's performance in choosing the correct or incorrect task.

This finding is very analogous to Schall's (2001) account of how decisions about upcoming eye movements are taken in the frontal eye fields (see Section 2.4.1), and to Cisek and Kalaska's (2005) account of how decisions about upcoming reach movements are implemented in the premotor cortex (see Section 2.5.2). The difference is simply that the decisions are taking place in PFC, and the alternatives to choose between are complex abstract plans, rather than individual actions.

3.3.2 A possible role for posterior PFC and the SMA in plan selection

Where might PFC sequence plans compete with one another? One good candidate location is the supplementary motor area (the SMA; see Section 2.6.1.2). Prefrontal cortex projects to motor cortex indirectly, via the pre-supplementary motor area (pre-SMA) and then the SMA. There is good evidence that the SMA and pre-SMA both encode prepared sequences; see e.g. Shima and Tanji (2000). In addition, there is evidence from neuroimaging that

the SMA is particularly active in situations where a stimulus evokes two alternative and conflicting responses; see Rushworth *et al.* (2004) for a review.

Another area which is likely to be involved in plan competition is the posterior lateral PFC. Evidence for this idea comes mainly from studies of dual task performance in humans; see Marois and Ivanoff (2005) for a review. A particularly interesting study is that of Dux *et al.* (2006). There is good behavioural evidence that subjects must perform the 'response selection' component of two simultaneous tasks sequentially, rather than in parallel—i.e. that at the level of response selection, one task must be queued behind the other (see e.g. Pashler, 1998). Dux *et al.* found two frontal areas which showed queueing of response selection activity. One was the left posterior lateral PFC (Brodmann area 9); the other was centred on the SMA and pre-SMA. In their study, the former region was most strongly identified with task competition; they accord the SMA region a secondary role. However, it seems likely that both these regions are involved in competition between / selection of PFC plans.

3.3.3 Plan termination and the pre-SMA

PFC plan representations do not reflect transitory sensory events; they tend to persist until the agent changes plan (see e.g. Miller and Cohen, 2001). There are several models of how plan representations are actively maintained in PFC, including the reinforcement learning models of Braver and Cohen (2000) and O'Reilly and Frank (2006). However this is done, it means that there must be an active mechanism for *removing* a plan from PFC once it has been completed and is no longer needed.

There is good behavioural evidence for a mechanism by which a currently active plan is removed, from a phenomenon called **backward inhibition** (Mayr and Keele, 2000). Mayr and Keele devised an experiment where subjects had to perform a succession of different tasks on an array of visual stimuli. In one task, they had to respond to the colour of stimuli; in another, they had to respond to their orientation, and so on. The sequence of tasks was varied: in one condition subjects performed task A, then task B, then task C, while in another, they performed A, then B, *then A again*. It was found that subjects' response times were slowed when performing the repeated task A. Mayr and Keele's explanation was that switching tasks requires inhibition of the current task set; the slowed response to repeated task A was due to 'residual' inhibition associated with the prior inhibition of this task. The experiment is quite strong evidence that plans are removed from working memory by a process of self-inhibition.

The question then arises as to where this inhibitory process occurs. A study by Shima *et al.* (1996) provides some relevant evidence. Shima *et al.* found a population of cells in the pre-SMA of monkeys which are particularly active when the animal had to switch from

one plan to another. Monkeys performed a blocked task: in each block, they were required to learn a sequence of three motor movements using visual guidance, and then to perform this sequence several times. A signal was provided at the end of each block, indicating the beginning of the next block, in which a different sequence was to be learned. The signal used the same visual stimuli as were used during visual training. A population of pre-SMA cells was found which responded to this 'new block' indicating signal, but did not respond to the same visual stimuli when training with the same visual stimuli was under way. These cells thus appeared to be involved in the process of turning off the monkey's current motor plan, to allow another plan to be learned. Thus the pre-SMA appears to contain circuitry which allows the currently dominant plan to inhibit itself.[3] Rushworth *et al.* (2004) find similar activation in pre-SMA in imaging studies of humans.

Another possibility is that inhibition of plans happens in *right* prefrontal cortex, while activation of plans happens in *left* prefrontal cortex. This is supported by imaging evidence (see e.g. Aron *et al.*, 2004), and also by evidence from dysfunction (see e.g. Mayr *et al.*, 2006).

In the above paradigms, subjects were given explicit instructions about when to change their task. However, in other situations, agents have to decide for themselves when their current plan should be changed. One obvious place where this must occur is if the plan has just been successfully completed. The question of how an agent recognises that its current plan has been completed is not a trivial one; however, I will not consider it now.

3.4 PFC plan activation during action recognition

The preceding sections have described how PFC plan representations are activated when an agent is performing an action himself. We must also consider what happens when the agent watches a third party perform an action. As described in Section 2.7.5, during action recognition, the sensorimotor system exists in a different mode, with different connectivity in premotor cortex and PFC. During action execution, PFC exerts a causal influence on the action representations in premotor cortex, and these representations in turn result in physical movements. In action observation, observed movements create activity in STS, which causes premotor cortex representations, which in turn generate a PFC plan representation. In this case, the plan representation is inferred 'abductively', as an *explanation* of the observed action. Now that we have a more detailed model of plan representations in PFC, it is worth revisiting this account of its role in plan inference during action recog-

[3] In Section 5.6.2, when discussing much higher level communicative utterances, we will consider the possibility that humans can learn to execute this 'plan-inhibition' operation voluntarily in certain circumstances.

nition. In this section I will develop a model which draws mainly on the pathway-biasing account of sequence plans given in Section 3.2.2. However, it is also possible to develop a model within the competitive queueing paradigm; in fact, the network described in Rhodes *et al.* (2004) contains many of the elements needed for this model.

3.4.1 The attend-to-other operation

The first operation an observer must perform during action recognition is to attend to another agent. This is an operation which the observer might acquire through ordinary reinforcement learning; presumably it is often beneficial for the observer to attend to an event in the world rather than performing an action himself.

There are two consequences of the operation of attending to another agent. Firstly, a reafferent sensory state will be activated, evoking a representation of the agent attended to. As discussed in Section 2.7.6.2, this representation must have an intentional component; as of this point, the observer's PFC will be modelling the observed agent's predispositions to action, rather than those of the observer himself. Secondly, as argued in Section 2.8, the attend-to-other operation also puts the observer into action-recognition mode, in which action representations cause activity in PFC rather than vice versa. In the next section, I consider these causal processes.

3.4.2 Abductive inference of PFC states

Assume a scenario where the observer has attended to an agent A standing in a kitchen. The observer is in action-recognition mode; he is also in a new sensorimotor state (the state of attending-to-A). Before A does anything at all, the observer may well have prior expectations about what his plans are; for instance, he might assume A's two most likely current plans are to sit down and to grab a cup. These two plans will be represented in the observer's PFC, and will compete with each other as described in Section 3.3—but in action recognition, the winner is determined primarily by representations generated from visual processing (see Section 2.7.6.1). The first operation in the 'sit-down' plan is (we assume) 'attend-to-chair'; the first operation in the 'grab-cup' plan is 'attend-to-cup'. Once the observer has entered action recognition mode, he establishes joint attention with A, so whatever attentional action A executes, the observer will also execute. In other words, as soon as the observer activates an attentional action, he can begin to infer A's plan.

A method for implementing this inference is illustrated in Figure 3.7. Note that in this figure, activation flows from action representations to pathway units, and from pathway units to PFC units. We assume that a pathway unit can only be activated if it receives input both from its associated stimulus and from its associated action. (Recall a similar

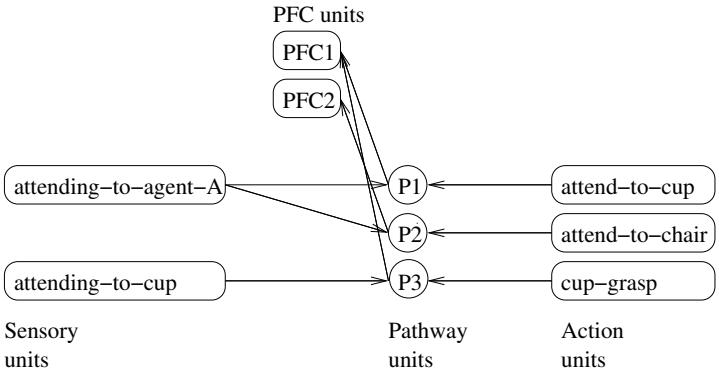


Figure 3.7: A mechanism for abductive inference from an action representation to a PFC state

requirement in action execution mode, where a pathway unit can only become active if it receives input both from the right stimulus and from the right PFC unit.) In our action-recognition example, since A attends to the cup (causing the observer to attend to it too), $P1$ will activate rather than $P2$. The active pathway unit in turn activates its associated PFC plan unit, in our case $PFC1$. Note that this plan unit is also involved in biasing another stimulus-response pathway, from the reafferent sensory state 'attending-to-cup' to the motor action 'cup-grasp'. In activating $PFC1$, therefore, the observer is effectively *anticipating* the remainder of A's action. The $PFC1$ cell thus behaves very much like the PF/PFG cells discovered by Gallese *et al.* (2002) and Fogassi *et al.* (2005) described in Section 2.7.4. Recall that these cells fire when a monkey observes a sequence of two actions, after the first action has occurred but in anticipation of the second action.

3.4.3 Training the abductive network

How might the connections in Figure 3.7 be learned? One possibility is that they are copied from those which the observer uses when acting as an agent himself. To make this work, the 'attending-to-agent-A' stimulus created when entering action-observation mode must essentially be the same as the 'attending-to-self' stimulus. However, this does not allow for the representation of agents with different predispositions to act than the observer himself. It may be, for instance, that the observed agent A has a different plan about what to do with a cup when entering a kitchen, and executes 'cup-smash' rather than 'cup-grasp'. The observer must be able to learn to infer this plan, even though it is inconsistent with his own, as discussed in Section 2.7.6.2.

One plausible scenario is that the 'attending-to-agent-A' stimulus is similar to 'attending-to-self' (so that in the absence of any information to the contrary the observer will use his own PFC connectivity during action observation) but nonetheless different in some respects, to allow new connections to be learned which are specific for this observed agent. Recall from Section 2.7.5 that new learning of this kind cannot be driven by reinforcement, since the observer of an action does not receive a reward for its successful accomplishment, but must rather be driven by predictions about the agent's forthcoming actions, encouraging representations which minimise the errors in these predictions. Returning to Figure 3.7, note that activating $PFC1$ creates a prediction that 'attending-to-cup' will result in 'cup-grasp'. Without going into detail, we can envisage that if this prediction is not fulfilled, the PFC connections responsible for generating it could be weakened, to allow the development of alternative connections specific for this observed agent.

3.4.4 The time-course of plan activation during action recognition

There are many situations in which an observer has to recognise an action using impoverished or inaccurate input data. For instance, it must be possible to recognise a grasp action from a still photograph: if the action is complete in the photograph, then the agent's earlier attentional actions must be inferred; if it is incomplete, then the final stages of the motor action must be inferred. Or again, an observer watching in real time may incorrectly identify one of the agent's actions, for instance by inaccurately anticipating the wrong target object and activating the wrong PFC plan. When the agent then grasps a different target object, the observer will have to backtrack, to seek a different interpretation of the wrongly classified action.

How might these mechanisms be implemented? Two points are worth noting. Firstly, recall from Section 3.3 that cells with a single PFC plan assembly mutually reinforce one another, while inhibiting cells in other assemblies. This means that PFC tends to relax to a stable state in which one complete plan is activated, even if sensory evidence only supports some components of this plan, as in the case of a static image. Secondly, it may be that prior to reaching this stable state, the agent can maintain several alternative candidate PFC plans, whose levels of activation reflect their current likelihood. A candidate plan assembly may start off with relatively high activation, but then be superseded by another candidate plan if this becomes more consistent with the action while it is in progress—again see Section 3.3.

However it is arrived at, by the end of action recognition, the observer will have a single active PFC representation of a planned sequence of three actions: attention to agent,

attention to cup, and cup-grasp. It is important to note that the activation of this PFC representation is not tightly synchronised with the occurrence of the actions themselves. It may occur in advance of one or more actions, since it is essentially predictive, or it may be established retrospectively, after they are all complete. The key point is that the plan representation settled on at the end of the process can now take on a causal role in an account of the agent's actions: the observer can assume the agent activated the plan in advance of initiating any action, and that it was maintained in PFC without change while the sequence of actions occurred.

3.5 'Replaying' PFC plans: simulation mode

To sum up Sections 3.2–3.4: I have outlined a model of the role of the PFC in executing the action of grabbing a cup, and in recognising this action when performed by another agent. During action execution, a PFC plan is activated by a sensory stimulus and then actively maintained; while it is being maintained, it supports the execution of a planned sequence of three actions (attend-to-self, attend-to-cup, cup-grasp), and upon successful completion of this sequence, it is removed from PFC. During action observation, the observer allows his sensorimotor system to be driven by the observed action, and as a consequence experiences a sequence of sensorimotor states; in response to these, he generates a PFC state which represents his best inference about the plan which generated the corresponding states in the observed agent. In both action execution and action generation, the PFC representation is essentially the same. According to the pathway-biasing account of sequence plans, it is a set of active units whose effect is to selectively enable three linked specific stimulus-response pathways. According to the competitive queueing model, it is a gradient of activation over the motor units which determines the order in which they fire.

The above account of PFC representations opens up the interesting possibility that an activated PFC plan can be 'played forward' by the agent in the absence of any overt motor action, either executed or observed. To begin with, consider the competitive queueing model. We already know from the existence of mirror neurons that the motor system can enter an 'action rehearsal' mode in which premotor action representations can fire without triggering explicit motor movements. Imagine that an agent has a prepared action sequence in working memory. If the agent enters this 'rehearsal' mode and then transfers the sequence in working memory to the planning layer, a sequence of premotor activations will result, with no overt motor reflexes.

This form of motor rehearsal is incomplete—it does not include the reafferent sensory consequences of the motor actions being rehearsed. However, we can also envisage a way whereby these consequences can be rehearsed. We just need to specify that rehearsal mode

also includes a condition which ensures that the activation of a premotor representation automatically triggers the reafferent consequences of its successful completion. It is quite easy to see how this condition could be implemented, given our account of how PFC exerts a top-down influence on attentional and motor actions. Recall from Section 2.4.2 that a top-down attentional action is simply a bias on the visual system to enter a certain state—for instance, a bias to find an object of a certain type in the visual scene. Normally, this bias can only activate the desired state if there is also bottom-up evidence for that state—otherwise we would hallucinate the objects we wanted to find. However, it is quite possible to imagine a mode in which this requirement is dropped, in which case a top-down bias would function as a self-fulfilling prophecy about the next sensorimotor state. A similar argument can be made for motor actions. As discussed in Sections 2.5.2 and 2.5.3, a top-down bias for a motor action normally only functions if there is bottom-up evidence that this action is afforded. Again, it is easy to imagine this constraint being dropped. In fact, some theorists suggest that motor programmes are cognitively represented by their intended reafferent consequences; see in particular Hommel *et al.* (2001). The idea of simulating the reafferent consequences of a motor action sits especially comfortably with this proposal.

We can therefore envisage a special 'simulation mode', in which an agent holding a planned sequence of sensorimotor operations in PFC internally rehearses this sequence, along with the reafferent sensory consequence of each operation. Note that a similar story can be told for a pathway-biasing model of planned sequences. In that case, the self-generated reafferent consequences have a more important role to play, since they are required to trigger successive actions in the sequence. However, the effect will be just the same: a sequence of rehearsed actions, interleaved with their reafferent consequences.

Is there any evidence that agents do internally replay actions they execute or observe in this way? In fact, many theorists have suggested that 'remembering' or 'thinking about' an action involves a process of simulation of this kind (see e.g. Gallese and Goldman, 1998; Jeannerod, 2001; Grèzes and Decety, 2001; Barsalou *et al.*, 2003; Feldman and Narayanan, 2004). In Sections 3.6—3.8 I will review several studies which suggest that agents internally replay sequences initially stored in working memory. However, it is interesting to note that the model of working memory for sequences which we are developing is starting to bear some resemblence to Baddeley's model of working memory—in particular to his newly proposed 'episodic buffer'. It is a memory for 'episodes' (in the sense that an observed or executed action is an episode). It evokes sensorimotor representations in different modalities (the reafferent consequences of the prepared actions occur in a range of modalities). It is inherently sequential. And it can be internally rehearsed.

In what way might our account of working memory for events differ from the conception developed by Baddeley? The main issue relates to how long a 'working memory' can be

sustained. As Baddeley notes, a 'task set' can be maintained by an agent for quite some time, and does not seem to require rehearsal, while material held in working memory in his sense degrades rapidly if it is not rehearsed. Perhaps one way of reconciling the two accounts is to note that the task sets I am envisaging are unusually complex. For one thing, they are prepared *sequences*, rather than prepared individual actions. Some competitive queueing models do in fact require rehearsal of stored sequences—see especially the model of Burgess and Hitch (1999). Moreover, the first item in our prepared sequences has a special flavour: it is the action of attention to the *agent* of an action, which can be onesself or someone else. The sequences in question are thus encodings of complete episodes, rather than just of the agent's own prepared actions. While it might make sense for an agent to maintain a particular task set for long periods, it does not make sense for an agent to maintain a particular episode representation for long periods. (Though doing so may be symptomatic of a clinical condition, perhaps obsessive-compulsive disorder.) In summary: it does not seem unreasonable to identify our mechanism for holding prepared sensorimotor sequences with Baddeley's episodic buffer. In Section 3.5.1 I will state this idea more explicitly, and preview the linguistic role I envisage for working memory episodes.

3.5.1 Working memory episodes

The proposal just made is that we can think about dorsolateral PFC sequence plans as *episode representations*, rather than just as action planning devices. The question of how agents represent episodes in working memory is a very important one. It is of great relevance to models of language processing, because cognitive representations of 'sentence semantics' are probably working memory episode representations (see the discussion in Section 6.1.4.2). But representations of episodes in working memory are also required in accounts of reasoning and long-term memory. One crucial question is how to distinguish the different semantic roles which objects or individuals can play in episodes—for instance how to represent the fact that the man is the agent of the 'grab' action while the cup is the patient. The suggestion I have just made is that the order in which the agent and patient are attended to when the episode is experienced can be used to distinguish them in working memory: since they are attended to in a characteristic sequence, a working memory representation of this sequence will contain the necessary information about agents and patients.

The basic idea that memory stores not only 'representations' but also the perceptual or attentional operations which activated these representations is not a new one—it is found in several models of memory (e.g. Kolers and Roediger, 1984; Kent and Lamberts, 2008) and of perception (e.g. Noton and Stark, 1971; Ballard *et al.*, 1997). My main new proposal is that experience of an episode has a characteristic sequential structure, and that

this sequential structure allows the representation of episodes in memory as sensorimotor routines.

I will make use of this idea extensively in the remainder of the book. I will use the term **working memory episode**, or **WM episode**, to refer to a PFC sequence plan which stores the sensorimotor sequence associated with experience of an action. Later I will extend the term to cover stored sensorimotor sequences associated with experience of events in general, and of states.

To connect with a different literature, I intend the notion of a WM episode to play a role somewhat akin to 'schemas' in certain other models of working memory. In Norman and Shallice's (1986) model of action planning (see also Cooper and Shallice, 2000), a schema is a working memory representation of an action, which can have a sequential structure. Schemas compete with other schemas to determine what the agent does next; schemas can be hierarchically organised, so that one schema can participate as an action in a higher-level schema. Arbib's conception of schemas gives them a similar role in the planning and control of actions. He also foresees a role for schemas in assembling the results of perceptual and perceptual operations in working memory (see e.g Arbib, 1998). The word 'schema' is used by many theorists, in many different senses; for clarity, I will use the term 'WM episode' to refer to the particular sequence-based working memory construct I have introduced in the current chapter.

To summarise: after an agent grabs a cup, or watches someone else grab a cup, he is left with a sequence plan representation in PFC, which can be used to internally replay all the steps in the executed or observed action, together with their reafferent consequences. I will refer to this plan representation as a WM episode: a working memory representation of the experienced episode.

Looking ahead for a moment, in Chapter 5 I will argue that the sensorimotor characterisation of a natural language sentence makes reference to this process of replaying a sensorimotor sequence. To put it simply: if a native speaker of a language performs or observes a 'grab' action, the reason why she can immediately utter a sentence describing this action is that she has in her PFC a working memory representation of the planned sequence of sensorimotor operations which make it up, which she can use to replay the sequence. (The replay happens in another special mode, where sensorimotor operations have verbal reflexes, which I will describe in Chapter 6.)

3.6 Episodic memory and the hippocampal system

Before moving to linguistic matters, I will conclude the present chapter by turning to the topic of **episodic memory** (long-term memory for events). PFC representations are

relatively short-lived; in order to describe an action which occurred some time ago, a speaker must retrieve the action from episodic memory.

To incorporate an account of episodic memory into the sensorimotor model given so far, I will make two proposals: first, that events are stored in episodic memory in the form of sensorimotor sequences, and second, that when these sequences are retrieved from episodic memory, they create working memory representations very similar to those created for just-experienced events. I begin in this section by giving some background about episodic memory, and outlining the key role played by the hippocampus and related regions. In Section 3.7 I survey the evidence that the hippocampus has a special role in encoding sequences, and describe a model of how it does this. In Section 3.8 I describe the role of working memory in encoding and retrieving material from episodic memory, again with an emphasis on the encoding and retrieval of sequences.

Episodic memory is long-term memory for significant events in an agent's life. It should be distinguished from other forms of long-term memory; in particular from **semantic memory**, which is long-term memory for 'facts about objects' (see Section 9.2), and from long-term knowledge in the sensorimotor system (which has already been discussed in detail in Chapter 2). It should also be distinguished from working memory. In Baddeley's model, working memory for events lasts on the order of seconds or (with the aid of rehearsal) perhaps minutes. An agent's ability to recall (or 'relive') events dating back for longer periods is normally attributed to a quite separate type of memory, with its own neural substrate, labelled episodic memory (Tulving, 1983; 2002).

While working memories are assumed to be maintained predominantly in PFC, episodic memory is primarily associated with the **hippocampal system**, a network of brain areas in the medial temporal lobe (see e.g. Gluck and Myers, 2001 for a review). A classic illustration of the role of the hippocampal system in episodic memory is the case of H.M., a patient who underwent a rare form of surgery in which a large portion of the hippocampal region was removed in both hemispheres, together with portions of the temporal lobe (Scoville and Milner, 1957). After this surgery, H.M. developed a very pure form of **anterograde amnesia**; he was essentially unable to remember any events which happened to him post-surgery for more than a few minutes. He also showed a degree of **retrograde amnesia**, with impaired memory for events happening up to two years before the surgery. The fact that H.M.'s memory for events which occurred a long time before surgery is relatively intact suggests that the hippocampal system is not the place where long-term memories are permanently stored. Rather, it suggests that the hippocampal system is necessary for the creation of *new* episodic memories. The presence of retrograde amnesia for events occurring shortly before surgery has been seen as evidence that the hippocampal system provides a temporary store for newly-created memories, which are then **consolidated** in other cortical regions over an extended period of time. Reasoning from H.M.'s

case alone, we could conclude that it takes up to two years for memories to move from temorary hippocampal/temporal storage to more permanent storage elsewhere in cortex. In computational models of memory, an influential suggestion (see e.g. Marr, 1971; McClelland *et al.*, 1995) is that the hippocampus and the cortex are complementary memory stores: the hippocampus is specialised for the creation of new memories, while the cortex is specialised for storing longer-term memories, and for extracting generalisations which hold between individual long-term memories. (Generalisations are probably stored in semantic memory, as will be described in Section 9.2.1; thus cortex is involved in both semantic and episodic memory.) Identifying generalisations requires slow interleaved learning from a large set of individual episodes; the suggestion is that the hippocampus provides the set of individual episodes from which the cortex can gradually learn generalisations.

However, there is still considerable debate about how to interpret the role of the hippocampal system in consolidating memories. The pattern of retrograde amnesia following hippocampal damage depends on several factors, including the degree of hippocampal damage, the species being studied, the degree of damage to other areas of temporal cortex, and the type of information being stored (see e.g. Nadel and Bohbot, 2001). There also appear to be two different types of consolidation, which occur over different timescales: a 'cellular' consolidation process which is internal to the hippocampus, and takes less than a day, and a 'systems-level' consolidation process which involves transfer to neocortex, and takes on the order of weeks in rodents and years in humans (see e.g. Nader, 2003). Other studies in rats and monkeys suggest that removal of the hippocampus impairs distant memories as well as recent ones (Rolls, 1996). There is also debate about the role of the hippocampus proper in the wider temporal region involved in episodic memory encoding. Some theorists hold that the hippocampus proper is mainly involved in spatial computations, and contributes only peripherally to the wider memory system (see e.g. Holscher, 2003). Other theorists give the hippocampus a central role in encoding both spatial and nonspatial memories (see e.g. Rolls, 1996 and subsequent work). I will adopt the latter position in this chapter, and will therefore focus on hippocampal data and models. However, it should be borne in mind that the hippocampus is only one component of the temporal region which initially encodes episodic memories, and that as time passes, episodic memories become consolidated in other areas of cortex as well.

How does the hippocampal region function as a memory for events? As shown in Chapter 2, witnessing an event involves activating representations in a range of different cortical regions. The basic idea is that the hippocampus serves to link these activated representations together, so that at a later date they can be reactivated together, re-evoking the original experience of the event. It is well established that the hippocampal region receives (indirect) inputs from a wide range of cortical areas, including the inferotemporal, parietal and prefrontal cortices (see e.g. Van Hoesen, 1982; Suzuki and Amaral, 1994); it

also sends back (indirect) projections to all of these regions. There is also considerable evidence that retrieving material from episodic memory reactivates representations in a diverse range of sensorimotor areas. In this section I will introduce a model of how the hippocampus stores links between representations, and of the special role of spatiotemporal representations in episodic memories.

3.6.1 The hippocampus as an autoassociative network

An influential model of episodic memory in the hippocampus is that of Edmund Rolls (see Rolls, 1996 for a summary). In Rolls' model, a region of the hippocampus called CA3 functions as an **autoassociative network**. Cells in CA3 receive inputs derived from specific cortical regions, but they are also highly (and homogeneously) interconnected. Moreover, many of these connections are Hebbian in nature: when two connected cells are active simultaneously, the connection between them is strengthened. Rolls and colleagues have developed a computational model of CA3 to explore the characterics of a network of cells with these properties; see e.g. Rolls *et al.* (1997b). Inputs to the network are patterns of activation deriving from different areas of cortex. When a pattern is presented, a subset of units becomes active, and the connections between these units are strengthened. Subsequently, to recall one of the input patterns, it suffices to re-activate a subset of the cells involved; the connections between them then cause the remainder of the group to become active.

To begin with a simplistic example (which we will progressively refine), imagine that a pattern in the CA3 network encodes a combination of three concepts: 'man', 'cup', and 'grab'. If this pattern is presented to the network, the connections between these concepts are strengthened, and the combination is stored. If we then try to retrieve this pattern by activating some of its component elements—say those corresponding to 'man' and 'cup'—the 'grab' node will become active, completing the pattern originally presented.

The neuronal mechanism by which Hebbian strengthening occurs in the hippocampus is called **long-term potentiation** (**LTP**; see e.g. Abraham *et al.*, 2002; Cooke and Bliss, 2006). Very briefly: if two connected cells fire simultaneously (or within around 100ms of one another), the synapse which connects them is strengthened. A complementary mechanism of **long-term depression** (**LTD**) also occurs across the hippocampus, which is assumed to counteract this process and keep average synaptic strengths constant throughout the network.

CA3 is just one region of the hippocampus. Cortical projections arrive in CA3 via another region called the dentate gyrus, and projections back to cortex pass through two other regions, CA1 and the subiculum. Rolls' model includes hypotheses about the role of these way-stations in pre-processing the input received by CA3, and in post-processing its

outputs. However, I will not discuss these in any detail.

3.6.2 Episodic memory and context representations

3.6.2.1 Place cells

The hippocampal region is not only involved in episodic memory; it also has a well-established role in spatial cognition. Early single-cell recordings in rats found hippocampal cells dubbed **place cells**, which responded maximally whenever the animal moved through a particular point in the environment (O'Keefe and Nadel, 1978). These cells appear to encode the rat's location in an environment-centred frame of reference—no mean feat, given that its sensory inputs arrive in a frame of reference centred on the animal itself. In primates, the hippocampus appears to provide a richer representation of space, including the location of other objects and agents. Many cells in the primate hippocampus fire when the animal attends to a particular location, or when a particular object is found at a particular location (Rolls *et al.*, 1997a; Rolls, 1999). These latter cells have been termed 'view cells'. Representations of location are again allocentric, and are invariant to changes in the animal's own location in the environment. Even in rats, the hippocampus does not *just* record spatial information. Hippocampal cells have been found which respond to specific tasks or perceptual stimuli independently of location, or which respond to particular combinations of tasks/stimuli and locations (Wood *et al.*, 1999).

Recent single-cell recordings in humans (Ekstrom *et al*, 2003) have found both place cells and view cells in the human hippocampus.[4] In humans, tasks involving spatial cognition appear to preferentially activate the right hippocampal region (see e.g. Abrahams *et al.*, 1997 and Maguire *et al.*, 1997 for evidence from lesion studies and neuroimaging).

3.6.2.2 Spatial context representations

Hippocampal place cells only provide a representation of a localised spatial environment. There is evidence that when a rat moves from one relatively coherent environment to another, its place cells are entirely remapped (see e.g. Muller and Kubie, 1987). It is not clear how a rat's spatial world is decomposed into 'environments', but it seems that environments tend to be reasonably closed, reasonably homogeneous at the level of ambient stimuli like floor and wall colour, and linked together by well defined access routes. For

[4]Ekstrom *et al.* also found cells they dubbed 'goal cells', which appear to encode the location of a landmark which the agent is currently trying to reach in a navigation task. I will consider these in Section 10.9.3.2.

instance, two areas of a run linked only by a small door would be likely to be represented as two separate environments.

It appears that the individual environments which a rat knows are encoded as distinct **spatial contexts**. When a rat is placed in a familiar environment, this triggers *recognition* of the environment, implemented as activation of the associated spatial context. The information provided by place cells must be decoded with reference to the currently active spatial context representation, which is an integral part of the rat's representation of space.

Where in the hippocampal system are individual spatial contexts stored? There is evidence from a range of studies that representations of spatial contexts are found in the **parahippocampal cortex** (**PHc**)—see Eichenbaum *et al.* (2007), Diana *et al.* (2007) for recent reviews. Particularly interesting is an area in PHc called the **parahippocampal place area** (or **PPA**): fMRI experiments show this area responding preferentially to visual stimuli depicting visual stimuli with 'topographical structure', as opposed to other types of visual stimuli (Epstein and Kanwisher, 1998). Moreover, the PPA responds more strongly to familiar locations than to unfamiliar ones (Epstein *et al.*, 2007). In another fMRI study, visual objects with strong associations to particular spatial contexts preferentially activated a wider area of PHc, together with areas in an area in the cingulate cortex cortex, the retrosplenial cortex (Bar and Aminoff, 2003). The retrosplenial cortex is also active during episodic memory retrieval (see Burgess *et al.*, 2001).[5] I will thus assume that individual spatial contexts are represented in PHc and in retrosplenial cortex. A more extended discussion of spatial context representations is given in Section 10.3.1.

3.6.2.3 Spatiotemporal context representations and the structure of episodic memory

Why would the hippocampal system be specialised for both spatial representations and episodic memory? A well-established suggestion is that events are individuated in episodic memory by linking their content (i.e. a representation of what happened) with a representation of a particular time and place (O'Keefe and Nadel, 1978). Thus if a given action (say 'man grabs cup') occurs twice in the agent's experience, the different occurrences can be stored as separate episodic memories by creating an association between the action and two distinct spatiotemporal contexts; see Figure 3.8. I noted in Section 3.6.2.2 that representations of spatial context could be contributed by the parahippocampal cortex (PHc). In fact, several authors have suggested that this region contributes more abstract representations of context as well as purely spatial representations—see e.g. Bar and Aminoff (2003)—in particular, that it contributes representations of temporal contexts, which can

[5]Burgess *et al.* have a different account of this memory-related activation, which I discuss in Section 10.3.3.3.

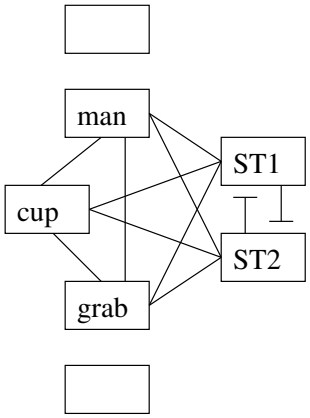


Figure 3.8: An encoding of two occurrences of 'man grab cup' at two spatiotemporal contexts. The contexts are represented by units ST1 and ST2 (which mutually inhibit one another). Only strong positive connections are shown. When either context unit is activated, the network will relax into a state in which 'man', 'grab' and 'cup' are active.

serve to distinguish two similar episodes which occur at different times. In fact, episodic memory is typically characterised as having a strong sequential or narrative structure, holding sequences of episodes in the order they occurred (Tulving, 1972; 1983).

There are several models of how a succession of temporal contexts is created in the hippocampus. In most recent models, there is an element of recurrency in the evolution of temporal context representations, with the new context being a function of the current context plus the most recent episode (see e.g. Howard and Kahana, 2002). This explains why there is a 'forward bias' in recall of episodic memories, such that it is easier to remember episodes in the order they were experienced.

I will assume, again as a first approximation, that PHc contributes representations of individual spatiotemporal contexts, which are used to individuate episodes, and encode their sequential order.

3.6.3 The hippocampus as a convergence zone

Note that while the hippocampal region stores associations between sensorimotor representations, it does not necessarily have to hold the representations themselves—at least, not in any detail. It just has to be able to activate these representations. I have already proposed that context representations are computed in parahippocampal cortex and action representations are computed in premotor and prefrontal cortex. In order for the

hippocampal region to store associations between these representations, it is not necessary for it to maintain its own copies of them—rather, it can simply act as a way-station whose neurons are connected to both regions.

This role of the hippocampus in creating indirect links between cortical representations is present in Rolls' model, in that cell assemblies in CA3 are very compressed encodings of the cortical representations they derive from. The important thing is that these compressed encodings can re-activate the same cortical representations which led to their own activation.

A similar idea is proposed by Damasio and Damasio (1994). In their account, the hippocampus is a **convergence zone**, which holds links between representations of objects, situations and actions stored elsewhere in cortex. If the hippocampus is damaged, representations of individual objects, situations and situations are preserved, but the relations which encode particular episodes are lost.

3.6.4 Representation of individuals in long-term memory

The hippocampus is a convergence zone, storing events as associations between context, action and object representations held in separate areas of cortex. In this section I will look in more detail at what kind of object representations are involved in these associations. We have already discussed sensory representations of objects—in particular, the category representations derived in inferotemporal cortex. However, we need to make sure that the representations brought togther in an episodic memory are of *individual* objects, rather than just object categories. An episodic memory is about a particular man and a particular cup, just as it is about a particular time and place. Any individual object may participate in many episodic memories. An object persists through time, and can be involved in many events. We need a way of representing individual objects in long-term memory, separately from the episodes in which they participate, so that an individual can be associated with many stored episodes. I will call these individual object representations **LTM individuals**. How are LTM individuals stored in the hippocampal system, and how are they linked to episodes in episodic memory?

3.6.4.1 Evidence for LTM individuals in the perirhinal cortex

Memory researchers attempt to study 'memory-for-individuals' by devising tasks which access this form of memory without relying on memory for episodes. They do this by comparing tasks which require **recognition** of previously presented objects with tasks which require **recollection** of events in which objects participate. The distinction between pure object recognition and event recollection has a fairly clear phenomenal correlate:

recognition is accompanied by a 'feeling of familiarity', which can be strong even if there is no recollection at all of any episodes involving the object. This dissociation gives us an empirical window onto memory for individuals.

What might a long-term memory representation of an individual object look like? There are several properties which it must have by definition. First, it must be a representation which is linked to sensory characteristics of the object, so that when the object is presented, the individual representation is activated. This is what 'recognising the object' must consist of. Note that this sensory representation must be a lot richer than just a single category representation in IT. I must be able to tell the difference between *this* cup and *this other cup*. I must be able to tell the difference between *the man who grabbed the cup* and all the other men I know. I will assume that when a particular cup is presented to the senses, IT (and the visual cortex more widely) register a complex conjunction of categories, which jointly associated with an LTM individual. Damasio and Damasio (1994) use their notion of convergence zones here too. For them, an LTM individual is a convergence zone associating a complex set of sensory manifestations of the individual, not just in vision, but in all modalities.

A second definitional property of an LTM individual is that it can participate in one or more episodic memories. Damasio and Damasio propose that the hippocampus is a 'second-order' convergence zone, which represents events as associations between LTM individuals (which are themselves first-order convergence zones).

We now have a reasonable idea of the necessary properties of LTM individuals, and of the experimental designs via which they can be studied. Can these designs tell us where LTM individuals are stored? Damasio and Damasio (1994) claim that the first-order convergence zones which represent individuals are found in the anterior temporal cortices. If this region is damaged bilaterally, a patient can no longer distinguish individuals, but crucially can still show knowledge of the component representations from which individuals are formed (for instance, the ability to categorise objects at the basic level). More recently, a consensus is emerging from single-cell studies in rodents (Eichenbaum *et al.*, 2007) and imaging work in humans (Diana *et al.*, 2007) that LTM representations of individuals are stored in an adjacent, slightly posterior area, the **perirhinal cortex**—an area of medial temporal cortex bordering the hippocampus. Perhaps most dramatically, there is recent evidence from single-cell recordings in humans being treated for epilepsy that individual cells in the medial temporal cortex encode particular individuals. Quiroga *et al.* (2005) showed subjects pictures of several well-known people and landmarks while recording the activity of medial temporal neurons. They found individual neurons with very specific responses to individuals, in the parahippocampal gyrus, entorhinal cortex and amygdala, and in the hippocampus itself. For instance, in one patient, one such neuron responded strongly to images of the actress Halle Berry, from a variety of different angles, and in

different costumes, but hardly at all to any other image. It also responded strongly to a line drawing of Halle Berry, and (perhaps most impressively) to the written word *Halle Berry*. Quiroga *et al.* argue that these neurons form part of assemblies which encode long-term memory representations of individuals.

3.6.4.2 Memory for object locations

In the model proposed so far, the hippocampus stores events, the parahippocampal cortex stores individual spatiotemporal contexts, and the perirhinal cortex stores individual objects that persist across time. One final issue to consider is how the *location* of individuals is represented in memory. Objects move around over time, so we need to have a way of storing the location of objects at particular times. How is this done? There is evidence that certain regions of the hippocampus (CA3 and the dentate gyrus) are particularly involved in storing the associations between individuals and spatial contexts (see Manns and Eichenbaum, 2006 for a review of evidence in several species). I will sketch a model of object location memory here, which will be elaborated on in Section 10.3.

Recall from Section 3.6.2.1 that the hippocampus provides a cognitive map of locations within the agent's current environment and is able to store combinations of objects and locations within this environment. If we also recall from Section 3.6.2.2 that the information provided within this map cannot be interpreted without reference to a representation of the agent's current spatial context, and recall from Section 3.6.2.3 that memories can be retrieved by reactivating an earlier spatiotemporal context, it follows that reactivating an earlier spatiotemporal context during memory regenerates a representation of the locations of objects at this earlier context in the cognitive map. Thus activation of an LTM individual in this earlier context will reactivate the location in the cognitive map occupied by this individual at the remembered time. Likewise, activation of the cognitive map location in this context will activate the LTM individual.

This model predicts that object location memory can be impaired in two ways; firstly, by damage to the medium which represents the cognitive map (i.e. the hippocampus), and secondly by damage to the media which represent individual spatiotemporal contexts (i.e. parahippocampal cortex and retrosplenial cortex). Both of these predictions are borne out; see e.g. Stepankova *et al.* (2004); Milner *et al.* (1997).

3.7 Hippocampal episode representations as sequences

One obvious problem with the account of episodic memory given thus far is that a grasp episode is stored simply as a collection of representations, without any internal structure.

For instance, the representations of the grasp episode in Figure 3.8 do not encode which object is the agent of the grasp action and which is the patient. Several solutions to this problem have been proposed. Most models make use of explicit representations of participant roles (e.g. 'agent' and 'patient'), which are then bound to different object representations in a variety of ways (see e.g. Shastri, 2001; 2002; Plate, 2003; van der Velde and de Kamps, 2006). However, the account of action execution and recognition developed in Chapter 2 opens up an alternative method for identifying thematic roles within event representations. In Chapter 2 I suggested that a reach-to-grasp action is associated with a characteristic *sequence* of sensorimotor operations, in which the agent and patient can be identified by their serial position: the first operation is attention to the agent, the next is attention to the patient, and the last is activation of the reach/grasp motor programme. If the sensorimotor experience of an action involves a characteristic sequence, and the working memory representation of the action encodes this same sequence, as argued in the first part of this chapter, then maybe this characteristic sequence is also used to store the action in longer term episodic memory.

In one sense, the idea that episodic memory is memory for sequences is a very old one. In fact, as noted in Section 3.6.2.3, the sequential character of episodic memory is taken by Tulving (1972; 1983) as being one of its defining characteristics. But recall that Tulving is thinking about sequences of whole episodes: each episode is individuated in memory by its sequential relationship to other episodes. In the account I sketched in Section 3.6.2.3, episodes were individuated by their association with spatiotemporal context representations in parahippocampal cortex; evoking a sequence of episodes would thus involve evoking a sequence of spatiotemporal context representations. In the model I am now considering, each episode is itself stored as a sequence. The model thus requires that the hippocampus is able to store fine-grained sequences of sensorimotor stimuli, as well as coarser-grained sequences of whole episodes.

There is no direct evidence that the hippocampus stores individual episodes as sensorimotor sequences. However, several theorists have recently proposed that the hippocampus is naturally configured for storing sequential patterns, rather than just conjunctions of stimuli, and thus that its primary function is to store sequences; see e.g. Wallenstein *et al.* (1998), Eichenbaum (2004) for general reviews. If experience of a reach-to-grasp episode has a characteristic sequential structure, which is preserved in working memory representations, and the primary function of the hippocampus is to store sequences, we might expect that such episodes are stored as sequences in the hippocampus.

There is good evidence that the hippocampus can store fine-grained sequences of stimuli without updating parahippocampal context representations. There is also evidence that individual elements in these fine-grained sequences correlate with specific representations in a range of cortical areas. I will review evidence for these fine-grained sequences and their

cortical correlates in Sections 3.7.1 and 3.7.2, and I will discuss some models of how the hippocampus stores such sequences in Section 3.7.3. In Section 3.7.4 I will give a worked example of how the hippocampus might store two successive episodes as sequences.

3.7.1 Storage of fine-grained temporal sequences in the hippocampus

As discussed in Section 3.6.2.1, the rat hippocampus contains place cells which encode particular locations in the animal's environment. Recently, it has been discovered that the hippocampus can also store sequences of locations. I will first review this evidence, and then consider evidence that the hippocampus is more generally involved in storing sequences, both in rats and in humans.

One indication that the hippocampus can encode sequences comes from cells which appear to encode specific trajectories in an environment. For instance, Frank *et al.* (2000) recorded from the hippocampus and entorhinal cortex of rats navigating a T-shaped maze, allowing a left or right turn at the end of a central corridoor. Some cells fired when the rat was running the central corridoor, but only when the rat subsequently turned left; other cells showed a similar preference for a subsequent right turn. These cells appear to be prospectively encoding trajectories—i.e. sequences of locations. But it is also possible that they are recording some static component of the animal's cognitive set.

More recently, there have been many studies which show that the hippocampus can actively *rehearse* sequences of locations. Striking evidence of this kind comes from studies of the rat hippocampus during sleep (see e.g. Lee and Wilson, 2002;, Nádasdy *et al.*, 1999; Buzsàki and Chrobak, 1995; Jensen and Lisman, 2005; Lisman, 2005). For instance, Nádasdy *et al.* (1999) found that sequential patterns of hippocampal place cell activity which occur when a rat navigates an environment during waking experience tend to be replayed when the rat is asleep. Interestingly, the reactivated sequences are speeded compared to the original sequences, occurring during bursts of hippocampal activity called **sharp wave ripples**. These findings provide quite convincing support for the suggestion that the hippocampus is able to store and re-evoke sequences. Sharp wave ripples also occur during waking experience, typically while the animal is resting rather than acting. It has recently been found that the hippocampus can replay experienced place cell sequences in both forward and reverse directions in sharp wave ripples in awake rats, immediately after they are experienced (Foster and Wilson, 2006; Diba and Buzsàki, 2007). Foster and Wilson suggest that reverse replay may have a general role in reinforcement learning of sequential behaviours. Finally, Diba and Buzsàki (2007) found that the sequence of place cells activated when a rat navigates a maze is also played in advance of the rat's

behaviour, again during sharp wave ripples. This suggests that the animals are able to rehearse a planned action sequence, as well as one which has already taken place.

Recall from Section 3.6.2.3 that the activity of place cells is dependent on higher-level representations of spatial (and perhaps temporal) context in parahippocampal cortex: when the animal's environment changes, place cells 'remap'. However, in the above experiments, the animal's environment is fairly small; the sequence of place cells denotes a sequence of places within a single environment rather than a sequence of environments. Place cell sequences are therefore evidence that the hippocampus can encode fine-grained sequential structure within a single spatiotemporal context. In a recent experiment, Davidson *et al.* (2009) found that rats running paths in a larger environment showed extended replay of hippocampal place-encoding assemblies, spanning *sequences* of sharp wave ripples occurring when the animals paused during exploration. They also found replay of hippocampal assemblies encoding relatively remote locations in the environment. They suggest that a single sharp wave ripple event may encode a fine-grained sequence of locations, and that sharp-wave ripples may themselves be organised into sequences, encoding a higher level organisation of memory. They also suggest that individual sharp wave ripples may be associated with specific hippocampal or parahippocampal context representations, and that sequences of ripple events might be stored using recurrent circuitry which updates these representations. On this scheme, we could think of individual ripples as holding the sequential structure of individual 'episodes', and sequences of ripples as holding sequences of episodes. While this idea is still very speculative, the existence of sequential structure both within sharp wave ripples and between them certainly suggests the hippocampal system can encode sequences at more than one level of hierarchy.

Since sequences of locations constitute trajectories, they are rather a special kind of sequence. Recently it has been found that sequences of nonspatial stimuli are also stored in the rat hippocampus (Fortin *et al.*, 2002; Kesner *et al.*, 2002). The sequences in both these experiments involved odours. For instance, in Fortin *et al.*'s study, rats were first presented with a sequence of different odours. After an interval they were presented with two odours from the sequence, and rewarded for choosing the one which occurred earlier. Normal rats could learn this task after some training, but rats with hippocampal lesions could not. Interestingly, lesioned rats could still distinguish between odours which had occurred in the sequence and new odours that had not been presented; in other words they retained an ability to encode the 'familiarity' of odours. Their impairment was thus selective for the sequence in which odours had occurred.

In humans, there is also evidence that the hippocampus is involved in remembering the sequential order of stimuli. For instance, patients with selective hippocampal damage have impaired memory for sequences of words (Shimamura *et al.*, 1990) and sequences of faces (Holdstock *et al.*, 2005). Other evidence comes from fMRI studies. For instance,

Kumaran and Maguire (2006) gave subjects two sequences of pictures of faces to learn. In one condition, some faces occured in both sequences. Thus one sequence might involve faces A, B and C, while the other might involve faces D, B and E. In the other the two sequences were separate (e.g. faces A, B and C in one sequence and faces D, E and F in the other). The former condition is particularly hard: since face B occurs in two places, the sequences cannot be encoded by remembering the transitions from one face to the next. Subjects must somehow differentiate between the two *contexts* in which face B can occur. Kumaran and Maguire recorded the hippocampal activity of subjects as they learned the pair of sequences. They found that in the 'hard' condition with overlapping sequences, subjects' hippocampal activity during learning correlated with their success in a subsequent recall test: the more hippocampal activity during learning, the better the sequences were recalled. For the easier condition, there was no such correlation. This is good evidence that the hippocampus has a specific role in encoding the serial order of stimuli in humans. Note that all these experiments involve nonspatial stimuli, indicating that the role of the human hippocampus also extends beyond spatial memory.

It is not clear whether these nonspatial sequences are sequences of 'episodes', each associated with a distinct spatiotemporal context, or sequences of stimuli within a single episode/context. In fact, it is probably misleading to make a clearcut distinction between these possibilities. It is likely that the hippocampal system makes use of a very rich system of context representations, with some dimensions of context representing physical features of the environment and others representing task demands (Smith and Mizumori, 2006). Some components of a context representation update rapidly, and others update more slowly. Smith and Mizumori suggest that hippocampal neurons contribute to these rich representations of context. Thus in some sense, a place cell is a representation of a particular 'context'—though it is a dimension of context which happens to update very rapidly.

In summary: the idea that the hippocampal system represents an individual episode in episodic memory as a 'fine-grained' sequence of sensorimotor states is reasonably consistent with what we know about how the hippocampus represents sequences.

3.7.2 Cortical associations of hippocampal sequences

In Section 3.6.3 I reviewed the idea that the hippocampus is a convergence zone, storing associations between representations distributed elsewhere in cortex. If the hippocampus naturally encodes sequences, the notion of a convergence zone must be somewhat extended. We can think of a single hippocampal assembly as storing static associations between cortical representations, but we can also think of a stored sequence of hippocampal assemblies as orchestrating the activation of a sequence of distributed cortical representations.

There is indeed evidence that hippocampal sequences trigger sequences of cortical representations. Again, this evidence comes from sharp wave ripples occurring during sleep. For instance, Ji and Wilson (2007) found that there are cells in the rat's visual cortex which are correlated with the activity of hippocampal place cells during waking experience (presumably because the rat's retinocentric view of the environment correlates with its current location). When sequences of place cells were replayed in sharp wave ripples during sleep, the corresponding visual cells also replayed in sequence. Euston *et al.* (2008) and Peyrache *et al.* (2009) have found a similar effect with prefrontal cortex cells. In Euston *et al.*'s study, the activity of some PFC cells was correlated with hippocampal place cells during waking experience (presumably because different locations were associated with different motivational states). When sequences of place cells were replayed during sleep, the corresponding PFC cells also replayed in sequence. So there is some evidence that hippocampal sequences are associated with sequences of cortical representations, in at least two cortical areas.

3.7.3 A model of sequence encoding in the hippocampus

There are several computational models of how Hebbian learning in the hippocampal system can enable the storage and recall of temporal sequences. I will focus on models of the CA3 region of the hippocampus, which is the area where most sharp wave ripples appear to originate (see e.g. Behrens *et al.*, 2005). Most models make use of some form of **asymmetric plasticity**, whereby connections are formed between neurons which fire in a particular order, which subsequently allow this order to be internally reproduced. Some models invoke a mechanism involving groups of CA3 neurons which are asymmetrically connected (see e.g. Levy, 1996; Wallenstein *et al.*, 1998); others invoke a mechanism at the level of single synapses (Bi and Poo, 1998; Jensen and Lisman, 2005). I will outline Wallenstein *et al.*'s model below, but either type of model is possible.

Wallenstein *et al.*'s model draws on two properties of CA3 neurons. One is that activity in CA3 neurons takes time to propagate to other CA3 neurons via recurrent synapses. The other is that the firing of CA3 neurons is temporally extended: a firing neuron tends to remain active for a brief period before it returns to resting potential. Both of these properties make for a certain amount of overlap in the firing of neurons in CA3. Imagine that CA3 is presented with a simple sequence of two stimuli, which activates neurons N1 and N2 in strict succession. Imagine also that N1 and N2 are each connected sparsely and randomly to other CA3 neurons. When N1 fires, it will activate a pool of other CA3 neurons. Since activity in these other neurons takes time to develop, they will still be active when N2 fires. Hebbian learning can then take place, creating associations between this pool of neurons and N2. If the network is repeatedly exposed to the sequence N1, N2,

these secondary associations cause the sequence to be replayed when the first item (N1) is presented by itself.

Wallenstein *et al.* have implemented a network which functions in this way. After a sequence has been presented several times to the network, they note that secondary neurons (those not involved in representing the original input items) begin to fire at particular places in the sequence. In fact, secondary neurons tend to remain active for sub-intervals within the sequence, rather than just at single points; this phenomenon originates because of the temporal extension of CA3 neuron firing, and is amplified by the recurrent connections in the network. Wallenstein *et al.* refer to these extended subsequence-denoting cells as having **context fields**. A population of cells with overlapping context fields allows a sequence of items to be recalled, even if the temporal separation between the items in the sequence is an order of magnitude larger than the time window within which LTP operates.

The notion of a 'context representation' has now occurred in two places in our account of episodic memory. In the model of hippocampal sequence encoding just given, CA3 cells with context fields are used to help store the sequence of sensorimotor items associated with a single episode in memory. But recall that context representations are also invoked in our account of how the hippocampus keeps individual events separate in episodic memory in Section 3.6.2. In that section, it was suggested that episodes are individuated by being linked to separate parahippocampal spatiotemporal representations. For clarity, I will call the former context representations **micro-contexts**, and the latter ones **macro-contexts**—though it should be borne in mind that context representations probably exist on a continuum, as suggested by Smith and Mizumori (2006).

3.7.4 An example: storing two successive episodes in the hippocampal system

It is useful to give an example. Say I want to store a sequence of two episodes: first John grabs Cup 1, then he grabs Cup 2. I assume that each episode is individuated by a unique macro-context representation in parahippocampal cortex. Call the macro-context associated with the first episode $C1$. Activating $C1$ as a hippocampal memory cue should trigger a hippocampal sharp-wave ripple, in which the hippocampus steps through a sequence of states which activate cortical representations. The first state activates the LTM individual which represents John, the second activates the LTM individual which represents Cup 1, and the third activates the premotor 'grab' schema. I assume that the macro-context $C1$ remains active throughout this elicited sequence, and indeed helps to generate it. (I also assume that presenting the sequence can reciprocally activate the macro-context $C1$.) The sharp wave ripple, plus the active macro-context $C1$, combine

to trigger an update of the macro-context via a wider recurrent circuit, to create a new macro-context $C2$. $C2$ in turn triggers a sharp wave ripple representing the second event, which results in activation of the LTM individual representing John, then activation of the LTM indivudal representing Cup 2, and finally activation of the 'grab' schema.

On this scheme, epsiodic memories in the hippocampal system are stored as sequences, at two levels of granularity. Individual episodes are stored as patterns of synaptic weights which generate sharp wave ripples in the hippocampus, which in turn activate sequences of cortical representations. Sequences of episodes are stored as patterns of weights in a recurrent circuit which supports update operations in a parahippocampal area representing spatiotemporal contexts.

3.8 Cortical mechanisms for encoding and retrieval of episodic memories

Section 3.7 gives a model of the *format* used by the hippocampus to represent episodic memories. However, it is also important to consider the processes whereby episodic memories are created and accessed. The storage of an episode which occurs in an agent's life does not happen automatically: storage is a cognitive operation which can be performed selectively, and to different degrees. The process of retrieving information from the hippocampus is likewise a cognitive operation in its own right. These cognitive operations involve both the hippocampus and the cortex, and the PFC is heavily involved in both of them. In this section I will review what is known about these operations, considering encoding in Section 3.8.1 and retrieval in Section 3.8.2. In each case I will attempt to relate results to the model of sequence storage just outlined.

3.8.1 Cortical operations involved in encoding episodic memories

3.8.1.1 Evidence for PFC involvement in episodic memory encoding

Our ability to recall an episode is influenced not only by events occurring at the time of recall, but also by events occurring at the time it was experienced. By changing the way a subject first encounters and processes a stimulus, we can manipulate how successfully it is later recalled (see classically Craik and Lockhart, 1972). Much of the initial processing which determines how well a stimulus is encoded takes place in the prefrontal cortex; for reviews of the considerable evidence supporting this claim, see Fletcher and Henson, 2001, Paller and Wagner (2002), Simons and Spiers (2003), Blumenfeld and Ranganath (2007). To take a couple of examples: ERP studies have shown that the level of PFC activity

when an individual word is presented is a good predictor of whether that word will later be successfully remembered (see e.g. Rugg, 1995); similar evidence has been provided using fMRI techniques (see e.g. Rotte *et al.*, 2000). There is a strong consensus that PFC-based working memory representations play an important role in the storage of stimuli in hippocampal episodic memory.

3.8.1.2 PFC-hippocampal interactions during working memory retention

Of course, the hippocampal system must also be involved in the process of encoding a stimulus in episodic memory, since this is where the stimulus is ultimately stored. In fact, there is a great deal of evidence that the PFC and the hippocampal system are jointly active when a stimulus is held in working memory, and that there is some exchange of information between them during this time; again see Simons and Spiers (2003) for a review. There seem to be several components to this exchange. In this section I will distinguish some of these, so that I can focus on one in particular in the next section.

Firstly, it now appears that some parts of the hippocampal system play an integral role in activity-based working memory. If the stimuli to be encoded are 'new', then regions in the parahippocampal system appear to be active during the retention period in the same way that PFC cells are active during delay periods (see e.g. Ranganath and Blumenfeld, 2005 for a review of evidence from both dysfunction and imaging). These parahippocampal regions appear to be involved in working memory representations of individual objects, especially novel ones. Thus there is a component of the hippocampal system (perhaps especially those areas which store LTM individuals) which is essentially an extension of PFC-based working memory.

Secondly, there is evidence in rats that PFC and the hippocampus communicate extensively during behaviour. For instance, Siapas *et al.* (2005) found that the firing of many PFC cells in behaving rats was phase-locked to the hippocampal theta rhythm. PFC cells fired slightly later than hippocampal cells, suggesting that synchronisation between the two regions is controlled by the hippocampus, and perhaps that information flows from the hippocampus to the PFC during this process. Note that this is not strictly speaking a working memory interaction, since it happens during real-time sensorimotor experience. Siapas *et al.* suggest that the interaction is involved in the consolidation of hippocampal memories in cortex. However, since it occurs during experience rather than offline, it may also be part of the process whereby hippocampal memories are originally encoded. For instance, it may be that PFC stores working memory representations of experienced hippocampal sequences, in the same way that it stores working memory representations of other stimulus sequences.

Thirdly, it must be that at least some of the hippocampal activity occurring during

working memory retention is influenced or driven by processes in PFC. As just noted in Section 3.8.1.1, there are many experiments showing that PFC-based working memory processes have an influence on how episodic memories are encoded. If episodic memories are stored in the hippocampal system, it must be that working memory representations in PFC somehow influence this storage process. It is this aspect of the interaction between PFC and the hippocampus which I will focus on in the next section.

3.8.1.3 A role for working memory replay in episodic memory encoding?

If we assume that a cup-grasp episode is stored in working memory as a sequence (Section 3.5.1), and that the episode is stored in the hippocampal episodic memory as a sequence (as just proposed in Section 3.7), what might the role of working memory be in encoding the episode in the hippocampus? One influential suggestion is that sequences to be stored in the hippocampus are first buffered in working memory, and then *replayed* from working memory to the hippocampus. We have already looked at models of how PFC can support the internal replay of sequences held in working memory—see Section 3.5. If we assume that the hippocampus stores sequences, and that processes in PFC are involved in the encoding of these sequences, it is very natural to suggest that a PFC-based replay operation is involved in encoding memories in the hippocampus.

In fact, there are very good reasons why an experienced episode must be buffered in working memory before it is encoded in the hippocampus. It may take some time for an episode to be experienced—perhaps several seconds or even minutes. But recall from Section 3.6.1 that the hippocampus creates associations using LTP, which requires associated items to be active within around 100ms of each other. Even with the bridging context representations discussed in Section 3.7.3, it is implausible that sequences of items occurring several seconds apart can be directly encoded in the hippocampus. The idea that working memory must act as a way-station for episodic memory is one of Baddeley's main arguments for the concept of an episodic buffer, as alluded to in Section 3.1. Baddeley (2000) suggests the episodic buffer may store sequences, which can be rehearsed through a process of 'sequential attention'. He does not explicitly propose that these sequences are replayed to the hippocampus, but this idea has been proposed (and implemented) by others—see especially Jensen and Lisman (2005; 1996) and Lisman and Otmakhova (2001).

Is there any evidence that encoding sequences in the hippocampus involves a process of working memory replay? I will discuss several supportive, though not conclusive, lines of evidence.

To begin with, there is considerable evidence for the serial replay of recently experienced sensorimotor events. We have already discussed replay of recent hippocampal place-cell sequences in sharp wave ripples during pauses in behaviour (see Section 3.7.1). There is

evidence that these waking replay events are involved in learning (see e.g. Axmacher *et al.*, 2008; Ego-Stengel and Wilson, 2009), though it is not clear whether replay is driven by the hippocampus or by some other site. There are also studies which show that sequences of widely distributed cortical representations are replayed during waking experience shortly after they occur. For instance, Hoffman and McNaughton (2002) gave monkeys a sequential reaching task, and recorded from several cortical areas during task performance and during a subsequent ten-minute rest period. They found that cells in a range of sensorimotor areas (posterior parietal, motor and somatosensory cortex) which were active together during behaviour also tended to be active together during the rest period—and moreover that sequences of cell complexes which occurred during behaviour tended to be repeated during the rest period. They did not record from the hippocampus, so it is not clear whether these replay events correlate with hippocampal replay events. But they did record from PFC; interestingly, the PFC does not appear to participate in these sequences. But recall that in the model given in Section 3.2.3.1, many components of sequence plans in PFC are tonically active during rehearsal of a sequence. On this model, we might expect PFC to be involved in replay, even if the transitory representations which are replayed are mainly in sensorimotor areas outside PFC.

In another line of evidence, there are several studies which suggest that the communication between working memory areas and the hippocampal system during memory encoding has a cyclical, phasic character. For instance, Weiss and Rappelsburger (2000) recorded EEG signals in subjects encoding words, and measured the synchronisation between signals from different neural regions. They found an increased synchronisation in several frequency bands between frontal and temporoparietal areas during encoding, and moreover found that this synchronisation was higher for words which were subsequently recalled than for those which were not recalled. A similar result was found by Fell *et al.* (2003) using intracranial EEG on subjects memorising lists of words. There was more synchronisation between hippocampal and cortical EEG signals during encoding of stimuli which were successfully recalled than during encoding of stimuli which were not recalled. The cortical area studied in this experiment was a parahippocampal area rather than PFC. But recall from Section 3.8.1.2 that the parahippocampal cortex is also implicated in activity-based working memory. Some theorists propose that the working memories of place cell sequences are replayed from parahippocampal cortex to the hippocampus during the period when these two regions are synchronised (see especially Jensen and Lisman, 2005). Of course, the observed synchronisation may not necessarily be due to a serial replay process; there are several other accounts of neural communication which invoke the idea of temporal synchrony. But it is one possibility.

There are other accounts of the interface between working memory and long-term memory which more explicitly indicate a role for the rehearsal of sequences in working memory.

A particularly interesting account is that of Burgess and Hitch (2005), which centres on a phenomenon called the **Hebb repetition effect**. Hebb (1961) gave subjects a series of digit sequences in an immediate serial recall task. Some of these digit sequences occurred many times, interspersed with other random sequences. Hebb found that these repeated sequences were *gradually* learned: immediate recall was better after several presentations than for the initial presentations. In other words, sequences rehearsed in working memory are gradually encoded in long-term memory. It was later found by Milner (1971) that patients with localised damage to the hippocampus did not show the Hebb repetition effect, though their performance on immediate recall was otherwise normal. This is good evidence that sequences held in working memory are gradually consolidated in hippocampal memory. Burgess and Hitch review evidence which suggests that the working memory structure which interfaces with hippocampal storage may be Baddeley's episodic buffer. They note that the Hebb repetition effect is unaffected by manipulations of phonological similarity in the lists to be recalled, and by articulatory suppression—so the working memory buffer which interfaces with long-term memory does not appear to be the phonological loop. However, the Hebb effect is weakened if the repeated items are presented with different rhythms, causing them to be chunked differently. Given that the episodic buffer is defined as the structure involved in chunking (c.f. Section 3.1.3), this finding is evidence that memories enter long-term storage through the episodic buffer. This conclusion is also consistent with the finding that long term memory of verbal material recovers its gist rather than its precise phonological structure (see again Section 3.1.3).

Burgess and Hitch (2005) note that there are some strong parallels between models of sequence storage in working memory and in episodic memory, in that both types of model make use of an evolving context signal to store successive items. (I have also discussed the use of context signals in both forms of memory; see e.g. Sections 3.6.2.3 and 3.2.3.2.) Burgess and Hitch suggest that context representations form a point of contact between working memory and episodic memory—specifically, that the context representations used to store sequences in working memory also participate in long-term storage of sequences. They argue that this overlap is what gives rise to the Hebb repetition effect, and what permits the transfer of sequences from working memory to long-term memory.

In summary: there is some evidence supporting the idea that encoding a sequentially structured stimulus in hippocampal memory involves an operation of replay from working memory. The evidence is not very clearcut. But there certainly appear to be many sequential replay operations in the period shortly after a sequential stimulus is experienced. Where these replay operations originate is not yet clear. It is perhaps most likely that there are reciprocal interactions between prefrontal and hippocampal areas during memory encoding, and that replay operations are jointly orchestrated between these areas.

3.8.2 Cortical processes involved in access of episodic memories

The process of retrieving a memory has several dissociable components. Firstly, a **memory cue** needs to be produced. In some cases a memory cue is delivered to the agent directly, in the form of a fairly complete sensory stimulus. The memory task in this case is termed **recognition**. In other cases, creation of a memory cue is more deliberate, and involves an active process in which an initially weak memory cue is elaborated or strengthened. The memory task in this case is termed **recall**. However it is formed, the memory cue needs to be presented to the hippocampus, to initiate the process of retrieval. In Sections 3.8.2.1–3.8.2.3 I will discuss memory cues, and the processes involved in presenting a cue to the hippocampus. In Sections 3.8.2.4–3.8.2.6 I will discuss the processes which are triggered when the hippocampus responds to a memory cue and retrieves some material from episodic memory.

3.8.2.1 Creation of memory cues

There is considerable evidence suggesting that the operation of generating or refining a cue stimulus is implemented in frontal cortex; see e.g. Wheeler *et al.* (1997), Buckner and Wheeler (2001) for reviews. For instance, a fairly coarse-grained neuropsychological finding is that patients with frontal lesions are often more impaired in recall tasks than in recognition tasks (see e.g. Gershberg and Shimamura, 1995), suggesting that frontal cortex is particularly important for recall. In more detailed fMRI studies, Buckner *et al.* (1998) found that activity in posterior frontal cortex during recall correlates with retrieval *effort* rather than retrieval success. Subjects were presented with verbal stimuli under conditions of 'deep' or 'shallow' semantic processing, the assumption being that shallowly encoded stimuli would require more effort to retrieve at recall time. Indeed, at recall time, there was less activity in posterior frontal cortex for deeply-encoded stimuli than for shallowly-encoded stimuli, even after controlling for recall success. This suggests that posterior frontal cortex is involved in the process of creating a cue stimulus which is strong enough to retrieve material from episodic memory.

3.8.2.2 The form of cues and responses in episodic memory

Say we are interested in retrieving an episode from the hippocampal system. We might imagine several types of memory cue which could be presented to the hippocampus.

One type of cue is a single stimulus, which happens to be associated with a particular episode. For instance, say a subject is shown a certain novel object in a particular situation. When he later re-encounters this object, it may evoke for him the episode in which he first encountered it (which is often referred to in memory experiments as the 'study episode').

In other situations, an agent can search his memory for a particular episode. In this case, the memory cue is an episode representation. In one case, the cue is a complete episode specification, and the agent simply wants to know if the episode happened. The action of querying memory with a complete episode representation is analogous to the action of asking a 'yes-no' question—for instance, *Did the man grab the cup?*. In other cases, the episode representation used as a memory cue can contain uninstantiated variables. Querying memory with a cue of this kind is analogous to asking a so-called *wh*-question—for instance, *Who grabbed the cup?*, *What did the man grab?*, and so on.

Note that answering a question does not necessarily transport an agent back in time to relive a 'study episode'. Sometimes it just involves accessing facts about the world. I will argue later that *wh*-questions are best thought of as queries on *semantic memory* rather than episodic memory. Semantic memory is a separate form of memory altogether, which I will discuss in detail in Section 9.2. So for the moment I will just discuss situations where a fully specified episode is presented as a memory cue to the hippocampus. These situations can include yes/no questions. But they can also include situations where the agent's purpose is not to answer a question, strictly speaking, but just to re-evoke the context associated with a particular episode. In language, the closest analogue to this type of cue is a subordinate clause introduced by *when*—for instance, *When the man grabbed the cup, (...)*. These memory operations have the effect of transporting the agent to a particular past situation, and allowing him to re-experience what occurred in that situation. Their main purpose is to put the episodic memory system into a certain state, which allows the cue episode, and associated episodes, to be relived. These operations are erhaps most consistent with Tulving's (2002) conception of episodic memory as 'mental time travel'.

If episodes are stored in working memory and in the hippocampus as sequences, it is most likely that an episode-sized cue to episodic memory also takes the form of a sequence, replayed from working memory to the hippocampal system. Of course there must be a difference between presenting an episode-denoting sequence as a query to the hippocampus and presenting it as a new fact to be encoded, which I will discuss in the next section. But if it is presented as a query, what will be the response? I suggest that if there is an episode which matches the query sequence, the hippocampus responds by activating the (macro) context representation associated with this episode. This operation is what initiates 'mental time travel'. The reactivated context allows the event to be replayed, possibly producing a richer and more detailed sequence of sensorimotor representations. It also provides a means of activating other stored episodes which are associated with this context. In particular, it allows activation of the episode/episodes which happened *next*.

3.8.2.3 Cue presentation mode

However a memory cue is represented in the hippocampus, it is important that the hippocampus can recognise it as a cue, rather than as the occurence of a new episode. A cue (e.g. *Which man grabbed the cup?*) is very different from a new episode (e.g. *A man grabbed the cup*), even though their content can be very similar. When a new episode is presented to the hippocampus, learning must take place, so it is encoded in memory. When a cue is presented to the hippocampus, learning must be disabled; instead, a retrieval process must be initiated, which results in information from the hippocampus being returned to working memory (as discussed in the next section).

How are queries distinguished from new events? Several differences have been found. Encoding and retrieval appear to involve different areas of prefrontal cortex; for instance, encoding appears to involve the orbitofrontal cortex (Frey and Petrides, 2002). They also preferentially activate different areas of the hippocampus; in an fMRI study by Eldridge *et al.* (2005), the dentate gyrus and CA2 and CA3 regions were more active during encoding, while the subiculum was more active during retrieval. Network models of CA3 often assume that CA3 inputs are clamped to high activation levels during encoding, and presented at lower activation levels during retrieval, to allow recurrent afferents to dominate and to minimise Hebbian learning (e.g. Treves and Rolls, 1992). Note that several neurotransmitters appear to modulate the degree of plasticity of the hippocampus. For instance, the neurotransmitter dopamine appears to have a role in modulating how strongly the hippocampus is modified by information presented to it (see e.g. Wittman *et al.*, 2005). These neurotransmitters could perhaps configure the hippocampus for encoding or retrieval.

In summary, we can say that input can be provided to the hippocampus in two different modes: either **episode encoding mode** (in which a new episodic memory is created), or **cue presentation mode** (in which a memory cue is presented). Not much is known about the origin of the control signal which places the hippocampus in one or other of these modes. But given the above evidence that prefrontal cortex is involved in creating memory cues, I will assume that the control signal comes from prefrontal cortex—perhaps from orbitofrontal cortex.

Note that if we assume that an episode-denoting memory cue is a sequence played to the hippocampus, we must also assume that the control signal which puts the hippocampus into the appropriate mode must occur *before* the sequence is played. The logic is similar to the logic which demands that the mirror system be put into action execution or action recognition mode before representations are evoked in it. In fact, we can note another ordering constraint to add to the collection in Chapter 2:

Ordering constraint 14 *When a new episode is encoded in episodic memory, the hippocampus must be placed in episode encoding mode before the episode is presented.*

Ordering constraint 15 *When a query is made to episodic memory, the hippocampus must be placed in query presentation mode before the query is presented.*

I have already proposed that episodes are encoded in the hippocampus as replayed sensorimotor sequences (Section 3.8.1.3), and that queries are presented to the hippocampus in a similar form (Section 3.8.2.2). The above constraints indicate that there is an extra operation at the start of the episode-denoting sequence, both for encoding and for query presentation. I will return to this issue in Section 5.6.3, when I consider the syntax of questions.

3.8.2.4 A model of retrieval as rehearsal

We have now discussed what it is to present a cue stimulus to the hippocampus. I now turn to the processes which occur when the hippocampus returns a response to a query.

What is it to retrieve a memory from the hippocampus? We can start with Tulving's proposal that retrieving an episodic memory involves reliving or replaying past experiences. Tulving (1983; 2002) suggests that retrieval happens in a special cognitive mode, which is distinct from ordinary sensory experience. In **experience mode**, the agent's high-level sensorimotor representations are activated in the regular way, by perception and his own motor actions. In **retrieval mode**, on the other hand, these same sensorimotor representations are activated by the episodic memory system (in our scenario, by the hippocampus). The idea is that retrieving the memory of an episode re-evokes the same sensorimotor representations in cortex as were activated when the episode was originally experienced. Note that experience mode and retrieval mode are modes of the whole sensorimotor system, not modes of hippocampal operation. So they are quite different from the episode-encoding and cue-presentation modes discussed in Section 3.8.2.3. They are also distinct from the two modes of operation of the motor/premotor system, action-execution mode and action-perception mode (see Section 2.8.3). In fact, they presumably subsume these latter modes, because an agent can recall actions performed by himself or by someone else. But they are similar to modes of the motor/premotor system, in that they are defined by the establishment of a particular set of interfaces between different neural regions.

There is considerable evidence that recall from episodic memory reactivates sensorimotor representations in cortex; for a recent review, see Kent and Lamberts (2008).[6] For instance, Burgess *et al.* (2001) asked human subjects to remember events embedded in a rich spatial environment. When subjects recalled these events, a network of brain areas was activated, including temporal areas associated with the hippocampus (as might be expected) but also including parietal cortex, which (as reviewed in Section 2.3) holds sensory representations related to action execution. Burgess *et al.* suggest that these parietal activations account for the richness of the visual and spatial representations which are retrieved from episodic memory.

There is also evidence that hippocampal representations reactivated during sleep activate associated sensorimotor representations, which I have already presented in Section 3.7.2. As discussed there, replayed hippocampal place cell sequences appear to reactivate associated sequences in visual cortex (Ji and Wilson, 2007) and in prefrontal cortex (Euston *et al.*, 2008). These reactivated hippocampal activations are assumed to be involved in the consolidation of hippocampal memories in the cortex, rather than in the kind of explicit memory retrieval operations I am considering in this section. However, they do provide evidence that sensorimotor or set-related representations in the cortex can be activated by hippocampal replay as well as by 'the real world'. In this sense, they corroborate Tulving's proposal about 'retrieval mode'.

The above model of retrieval allows us to be a little more precise in defining the retrieval operation and the sensorimotor processes which it triggers. I will define the **retrieval operation** as occurring after a memory cue has been presented to the hippocampus. It has two components: firstly there is an operation which puts the sensorimotor system into retrieval mode, so that it receives its input from the hippocampus rather than from the senses; and secondly there is activity in the hippocampal system in response to the memory cue (which because of retrieval mode will evoke activity in the sensorimotor system). I will define **post-retrieval processes** as the cognitive processes (largely prefrontal) which are involved in the *interpretation* of this re-evoked sensorimotor activity, which reassemble the sensorimotor activity into a working memory episode representation. I will argue that these processes are quite similar to those involved in storing directly experienced episodes in working memory.

3.8.2.5 The retrieval operation

The retrieval operation generates a response in the hippocampus to a presented memory cue, and configures the cortex to process this response. There is good evidence that the

[6]Retrieval from *semantic* memory also reactivates sensorimotor representations, as I will discuss in Section 9.1. But for now I am just thinking about episodic memory.

hippocampus is active during successful retrieval of episodic memories—see e.g. Eldridge *et al.* (2000). This study found hippocampal activity when a study episode was remembered, but not if subjects simply reported a feeling of familiarity with the memory cue. As already mentioned, Eldridge *et al.* (2005) found a particular area of the hippocampus, the subiculum, particularly active during retrieval as opposed to encoding.

The retrieval operation also appears to involve areas of the brain other than the hippocampus. One line of evidence for this comes from ERP studies of the recall process—see especially Allan *et al.* (1998), Johansson and Mecklinger (2003). These researchers have consistently found an early short burst of activity in left parietal cortex which they suggest is a neural reflex of the retrieval operation.

The extra-hippocampal components of the retrieval operation are probably involved in establishing retrieval mode, the special mode in which the cortex is activated by the hippocampus rather than directly by the senses. Establishing this mode presumably involves activating certain interfaces within the cortex and hippocampus and deactivating others.

It is important to distinguish between the neural operation which *establishes* retrieval mode, and the neural correlates of retrieval mode itself. The operation occurs early, and at a fairly localised point in time, while the mode which it imposes lasts for a more protracted period, while the retrieved representations are generated and interpreted. I will discuss the neural states associated with of retrieval mode here, though it should be noted that these states probably persist throughout the 'post-retrieval processes' discussed below.

There are several suggestions about what the neural correlates of retrieval mode are. Buckner and Wheeler (2001) identify a sustained parietal/frontal signal which they suggest indicates that the sensorimotor representation is due to a memory process, rather than to some other source. This is reminiscent of the 'match' signals associating motor representations with onesself or with an observed agent, as discussed in Section 2.8.2. Other researchers have suggested a particular area of prefrontal cortex called **rostral prefrontal cortex** is responsible for establishing this mode (see e.g. Buckner, 1996). A recent proposal by Burgess *et al.* (2007) is that rostral PFC has a more general function, which is to select whether cognitive processing is to be driven by perceptual stimuli or by stimuli generated internally. (Memory retrieval tasks involve 'internally-generated' stimuli, but Burgess *et al.* suggest that there are other types of stimulus-independent processing which also recruit circuits in rostral PFC.) Like my account of action-perception and action-execution modes in Section 2.8.3, Burgess *et al.*'s proposal is expressed using the idea of gating and competition. In 'stimulus-oriented' mode, pathways from perceptual modalities to central representations are gated open, while in 'stimulus-independent' mode they are gated shut, and top-down pathways activating central representations from internal sources are gated open. Burgess *et al.* suggest that internally generated stimuli compete with sensory stimuli for expression in higher-level sensorimotor areas (and thus for higher-level cognitive pro-

cessing), and that rostral PFC is involved in the pathways which can bias the competition towards internal stimuli.

3.8.2.6 Post-retrieval processes

Many theorists have proposed that material retrieved from episodic memory must be processed in some way by working memory. One suggestion is that retrieved material must be maintained or integrated in working memory, in somewhat the same way as incoming sensorimotor experiences are maintained and assembled into working memory episode representations. Another suggestion is that material retrieved from memory must be compared against the memory cue which retrieved it, to see if it matches: if it does not, the cue may perhaps be refined and retried, or an explicit 'fail' response can be generated. On this proposal it is very hard to distinguish post-retrieval processes from cue creation and presentation processes, because retrieval may trigger subsequent rounds of cue creation/presentation. Another suggestion is that there are some components of post-retrieval processing which relate to decision-making in a variety of domains, and are not specific to episodic memory retrieval at all. These suggestions are not necessarily exclusive—and in fact, there is evidence for all of them, though the picture is still somewhat confused. In this section I will review some attemts to localise the different frontal mechanisms which are active during retrieval. There are two broad distinctions which emerge in the literature: one between left and right hemispheres, and one between different prefrontal regions. I will consider these in turn.

Left and right frontal contributions to retrieval Several theorists have suggested that the right frontal cortex has a specific role in post-retrieval processes. For instance, the ERP studies of Allan *et al.* (1998) and Johansson and Mecklinger (2003) identify an extended burst of activity in right frontal cortex during recall which they identify with post-retrieval processing. They class it as 'post-retrieval' because it comes after the short burst of parietal activity which they associate with the retrieval operation itself. Allan *et al* suggest these prefrontal processes 'generate or maintain a representation of the study episode' from the raw data obtained from the hippocampus. Johansson and Mecklinger's experiment suggests that they are particularly involved in reconstructing memories involving conjunctions of stimuli.

This finding is partially echoed by imaging studies. One common finding from neuroimaging is that right prefrontal cortex is preferentially activated in situations where subjects successfully retrieve the 'study episode' associated with a cue stimulus, rather than just experiencing a feeling of familiarity with the stimulus (see Fletcher and Henson, 2001 for a review). Presumably, there is more retrieved material in the former case than

the latter; this increased right PFC activity has been taken as evidence that this area is involved in processing the retrieved material.

However, there are also some suggestions that this right PFC activity may not be *specific* to processing of material retrieved from episodic memory. Fleck *et al.* (2006) compared right PFC activity during episodic memory tasks with activity in the same area during a perceptual task. They found that the amount of right dorsolateral PFC activity was correlated more with subjects' confidence in their judgements, rather than with the task, leading them to suggest that right dorsloateral PFC is involved in 'evaluating decisions', both in a memory and a perceptual context.

There are also imaging studies which implicate a role for left PFC during retrieval. Henson *et al.* (1999) found that left PFC is activated more when subjects successfully recall the study episode than when they identify a stimulus as new. In fact Henson *et al.* also found increased activity in this same left PFC region when a cue stimulus is identified as familiar. Henson *et al.* suggest that left PFC may be involved in maintaining retrieved material in working memory, both for episodic memory retrieval and for familiarity judgements. But another imaging study (Dobbins *et al.*, 2002) found increased activity in left frontopolar and dorsolateral PFC when subjects were asked to retrieve the study episode rather than simply to identify the cue as familiar or unfamiliar, regardless of retrieval success. In summary, it is fair to say that there is still considerable uncertainty as to how to interpret data about the lateralisation of frontal mechanisms involved in episodic retrieval.

Anterior, posterior and dorsolateral PFC contributions to retrieval Another suggestion which is often voiced is that different areas of prefrontal cortex play different roles in retrieval (see Buckner, 2003 for a review). Posterior areas tend to respond differentially to the modality of retrieved stimuli—for instance, whether they are pictures or verbal stimuli. (I will discuss the areas responsive to verbal stimuli in some detail in Section 6.1.) Anterior and dorsolateral areas appear to be involved in aspects of retrieval which require some form of 'cognitive control'. Different theorists explain the relevant notion of control in different ways: Buckner suggests that anterior PFC contributes to the 'dynamic selection of representations' during memory retrieval, where specific representations must be 'momentarily constructed to solve a task goal'. On this view, anterior PFC is involved in the 'selection' and 'construction' of representations. Dobbins *et al.* (2002) have a similar analysis of the left frontopolar and dorsolateral PFC regions which were activated more by episodic retrieval than by familiarity, suggesting that these regions are involved in 'control processes that guide the monitoring or evaluation of episodic recollections'. Again, there is no reason to think these operations are specific to episodic retrieval. In fact, left anterior PFC is often given a role in 'selection' in language-processing tasks (see e.g. Wagner *et*

al., 2001, which I will discuss in Section 6.1.4.2). And the dorsolateral PFC is involved in constructing working memory representations of sequences and (in my account) of episodes (see Section 3.2 and Section 3.5.1). And in the left hemisphere, both anterior and dorsolateral PFC are involved in representing the semantics of verbs (as I will also discuss in Section 6.1.4.2).

A suggestion about one role of PFC in post-retrieval processing While it is still quite unclear what role PFC plays during post-retrieval processing, I suggest that one of its functions is to *reconstruct a working memory episode representation* from the stimuli evoked by the reactivated hippocampal trace. This proposal draws on the idea discussed above, that many functions of the PFC during episodic retrieval are not specific to retrieval, but also feature in linguistic or perceptual tasks. It is also consistent with the proposal that the hippocampus 'replays' a sequence of representations in sensorimotor cortical areas during episodic recall which can be monitored by PFC (see Section 3.8.2.4). Finally, it is consistent with the suggestion that areas of PFC (especially left anterior and dorsolateral PFC) are involved in 'constructing' representations from the material retrieved from episodic memory.

The idea that PFC recreates working memory episode representations during episodic recall makes sense given the other roles I have proposed for WM epsiodes in episodic memory. I have already suggested that WM episodes buffer the material which is sent to the hippocampus, during both encoding (see Section 3.8.1.3) and retrieval (see Section 3.8.2.2). On these assumptions it is quite natural to assume that retrieval regenerates a WM episode representation, so that retrieval and encoding are inverses of one another, and so that memory cues have the same form as the material they retrieve.

There are also some independent reasons for thinking that the PFC recreates WM episode representations during post-retrieval processing. One of these turns on the schema-like nature of WM episodes, which was briefly touched on in Section 3.5.1. It is well known that retrieval of information from episodic memory is subject to top-down influences and distortions due to 'world knowledge'—see Kellogg (2003, Ch 7) for a review. This general finding is often expressed using the theoretical framework of 'schemas' (see classically Bartlett, 1932) or 'scripts' (Schank, 1977). Schemas/scripts are structures which encode conventional situations, tasks, and patterns of action, often at quite high levels of representation spanning multiple episodes. These structures set up expectations which influence the creation of semantic representations in the same way in a range of different tasks, including perception, action, language interpretation, and also recall from episodic memory. In this framework, material retrieved from episodic memory must end up in a format which can interface with schemas or scripts. The idea that PFC creates WM episode representations

during post-retrieval processing is one way of explaining how this can happen.

Note that once a WM episode representation has been recreated in PFC from a retrieved hippocampal sequence, the episode can again be internally rehearsed as a sensorimotor process, just as it was when it was first experienced. While the hippocampus stores the objects which participate in an episode as LTM individuals (see e.g. Section 3.7.4), the PFC stores objects as 'sensorimotor operations'—specifically as attentional operations—which have reafferent sensory side-effects. When a WM episode is rehearsed, each sensorimotor operation has a reafferent side-effect. Moreover, as described in Section 2.11, the agent and the patient each feature twice as reafferent representations during rehearsal. In some ways, therefore, the working memory representation of the episode enriches the representation retrieved from the hippocampus. The enrichment here is not so much related to the influence of world knowledge, as with recreating the *experience* of the retrieved episode. Again, this idea is in line with the suggestion that episodic memories are relived experiences, rather than just retrieved facts.

3.9 Summary: cognitive processes occurring during the replay of a grasp episode

In Chapter 2 I developed a model of the sensorimotor processes which occur when an agent executes a cup-grabbing action, or perceives another agent performing this action. In the present chapter, I extended this model to cover how the perceived cup-grabbing event is stored in working memory and in episodic memory. I suggested in Chapter 2 that the sensorimotor processes associated with the cup-grabbing action have a characteristic sequential structure. In the present chapter I argued that an agent's working memory and episodic memory representations of the action preserve this sequential structure. I argued that the sequence held in working memory can be replayed to longer-term hippocampal storage. I also suggested that an episode recalled from hippocampal storage reproduces a working memory representation very similar to the one created by the agent when the episode was first experienced.

In the sensorimotor characterisation of syntax which I develop in Chapter 5, the process of internally replaying a sequential event representation held in working memory will have a central role: I will attempt to characterise the syntactic structure of the transitive sentence 'The man grabbed a cup' as a description of this replay process. It is useful to conclude the current chapter by describing the replay process in detail. To make it easier to refer back to the process, I will provide labels for the different elements which it comprises.

The basic idea is that there are two types of activity during the replay of a grasp episode,

whether it originates from an observed or executed action, or from a recalled episode in memory. One type of activity takes the form of a sustained signal which is maintained throughout the replay process. The other is a sequence of transitory activations associated with the execution of the different operations in the sequence.

The sustained signal is the PFC activity which holds the tonically active representation of the sequence being rehearsed. (In the competitive queueing model of sequence preparation, this is the 'working memory' representation of the activation gradient associated with the planned sequence, which survives the inhibition of actions in the sequence following their execution. In the pathway-biasing model, it is the ensemble of tonically active pathway-biasing units.) In either case, there are three components to the signal: the planned action of attention to the agent (which we will denote $plan_{attend\_agent}$), the planned action of attention to the cup ($plan_{attend\_cup}$) and the planned grasp action ($plan_{grasp}$).

The key transient sensorimotor signals are those associated with 'the current action being executed', and those associated with the reafferent sensory consequences of this action. First there is an action of attention to the agent, either in action execution or action perception mode (*attend_agent*), and its sensory consequence, an evocation of the agent and of the associated mode (*attending_to_agent*). Then there is an action of attention to the cup (*attend_cup*) and its sensory consequence, a representation of the cup (*attending_to_cup*). Then there is the 'grasp' motor action (*grasp*) and its reafferent consequence, a re-evocation of the agent (*attending_to_agent*). When the grasp action is completed, there is an extra piece of reafferent feedback: a re-evocation of the cup via the haptic interface (*attending_to_cup*).

A third transient signal is the context representation which helps control the execution of the planned sequence. The initial context representation (C_1) helps to trigger the initial action (*attend_agent*). The passage of time, and/or the reafferent feedback from this action generates the next context representation (C_2), which helps to trigger the second action (*attend_cup*). More time and/or reafferent feedback generates the next context representation (C_3), which helps to trigger the third action (*grasp*). There are no more actions to trigger; the final context representation (C_4) is associated with the endpoint of the action: the point at which haptic attention to the cup has been established.

A summary of the replay process is given in Table 3.9. This figure shows the representations which are active at each point during the process. Note that the replay process is very regular; it consists of four successive stages, each of which has the same internal structure. I will refer to each stage of the replay operation as a **cycle**. The first three cycles each consist of three sub-stages: the evocation of a context, followed by an action, followed by its reafferent consequence, which triggers an update to the next context. The final cycle is somewhat different, as it does not involve an action, and the final context coincides with the haptic evocation of the cup.

Sustained signals	Transient signals		
	Context signals	Action signals	Reafferent sensory signals
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_1	*attend_agent*	*attending_to_agent*
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_2	*attend_cup*	*attending_to_cup*
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_3	*grasp*	*attending_to_agent*
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_4		*attending_to_cup*

Figure 3.9: The time course of signals occurring during the replay of the cup-grabbing episode in working memory

The leftmost column shows the signal which is tonically active throughout the rehearsed sequence. There are three components of this signal, associated with the three actions being prepared. The figure duplicates these signals for each cycle in the process, so we can consider any cycle in isolation and see what signals are present. (The arrows are intended to show that the signal is sustained for each of the substages of each cycle.)

3.10 An assessment of the sensorimotor model

Chapters 2 and 3 jointly provide a model of what it means to actively entertain an event representation in working memory. (From now on I will refer to this model as 'the sensorimotor model'.) The model draws on a wide range of experiments and models in neuroscience and psychology. It attempts to synthesise these in a coherent way; however the synthesis goes well beyond the individual models on which it is based. There are many other models which have been synthesised from the huge literature on the performance, perception and memory of reach-to-grasp actions, and no doubt many more which could coherently be synthesised. The model must therefore be evaluated in its own right.

What is the best method for evaluating the sensorimotor model? What testable predictions does it make? One approach is to look for specific predictions deriving from the

particular way it combines existing models. For instance, the suggestion that executing a grasp action involves attention to the agent and then to the patient followed by a *re*-attention to agent and patient generates predictions which could be tested. My account of the role of 'mode-setting' operations in the self-other distinction also contains new ideas which need to be tested. It predicts, for instance, that executed actions will involve an early neural operation (perhaps the cortical readiness potential) which is also found when the agent remembers performing an action himself, but not found when the agent remembers an action performed by someone else. My suggestion that Baddeley's episodic buffer takes the form of a prepared sensorimotor sequence is another synthesis of two models which generates predictions—for instance, we expect action preparation and working memory for events to activate the same areas in imaging experiments. However, these predictions do not test the model *as a whole*. In fact it is very hard to think of predictions which do this, since it consists of many different components derived from many different sources of data. This is one of the difficulties in producing a model of high-level cognitive processes—even when these processes are grounded in concrete sensorimotor experience, as in my case.

My approach will be to leave the cognitive model just outlined as a hypothesis, and turn to the question of its relationship with language. If the model permits a neat account of the grounding of syntactic structures in sensorimotor representations, this may provide some way of assessing it as a whole, rather than piece by piece.

Chapter 4

A syntactic framework: Minimalism

In this chapter, I introduce a syntactic account of transitive sentences. As with the sensorimotor model, I want this account to be uncontroversial as a syntactic model in its own right—i.e. to draw as much as possible on current syntactic theory. Again, it is impossible to avoid some measure of controversy; there are several different paradigms for syntax, and it is necessary to choose one of them. The key contenders are the Chomskyan paradigm of **generative grammar** on the one hand (see e.g. Chomsky, 1995) and the more 'empiricist' paradigms of **construction grammar** (c.f. e.g. Goldberg, 1995; Jackendoff, 2002) and **unification-based grammar** (see e.g. Pollard and Sag, 1994) on the other. Construction grammars and unification-based grammars are in fact quite closely related, so at the highest level, the choice of a grammatical paradigm is basically a dichotomy.

In my case, the choice of syntactic paradigm is quite strongly constrained by the claim I want to argue for, namely that the structure of a transitive sentence is an encoding of the sensorimotor and memory processes outlined in the previous chapters. This makes a strong assumption about universal grammar: at some level of abstraction, the linguistic structure of a transitive sentence in any language must be the same. While there are treatments of linguistic universals in all three syntactic paradigms, it is only within the tradition of generative grammar that a strong claim about universal underlying syntactic structures is made (and even then, only within some versions of generative grammar). Accordingly, it makes sense to start by looking for a syntactic model within the framework of generative grammar. In this chapter, I will describe a recent incarnation of generative grammar called **Minimalism** (Chomsky, 1995). In fact I also want to incorporate elements from empiricist models of syntax, but I will defer this discussion until Chapter 6.

The goal of the present chapter is to introduce the concepts from Minimalism which are required in order to understand and motivate an analysis of a transitive sentence—specifically, our example sentence *The man grabbed a cup*. To understand Mimimalism,

it is often helpful to make reference to the theory which preceded it, which was called the **Government-and-Binding** (**GB**) theory (Chomsky, 1981). I will often introduce concepts by presenting arguments which relate to versions of GB theory, but the framework I eventually adopt will be a version of Minimalism. The ideas which I outline here are introduced more thoroughly in several textbooks; some good ones are Haegeman (1991), Radford (2004) and Carnie (2007).

4.1 What is a syntactic analysis?

Before I describe the technical details of GB and Minimalism, I will briefly describe what the objective of a syntactic theory is. The basic objective is to *describe a human language*—to catalogue the range of sentences in the language, and to specify what each of these sentences means. Thus a syntactic theory of a given language should provide two things: firstly a method which identifies the set of syntactically well-formed sentences in that language, and secondly a method for identifying the meaning of each of these well-formed sentences.

Given that there is effectively an infinite number of well-formed sentences in any natural language, the methods in question need to involve the application of general principles, rather than simply an enumeration of all possible sentences. These general principles state that a sentence can be broken up into different constituents or **phrases**, whose individual well-formedness can be defined relatively independently. Phrases can in turn be decomposed into smaller phrases, and ultimately into words (or further, into the morphological components of words). The principles which specify how sentences and phrases are composed can be thought of mathematically as describing, or **generating**, a set of well-formed sentences. The term 'generate' is used in a special sense in syntax. The aim of a syntactic theory (as just defined) is not to describe how sentences are actually produced by human speakers in particular communicative situations, but rather to describe the set of well-formed sentences in a language. Given that these cannot be enumerated, they have to be constructed by the application of general principles.

Once a set of general principles has been specified, we can indicate for any given well-formed sentence how the principles permit the generation of this sentence. Since the principles break the sentence recursively into phrases, an account of how the sentence is generated can be given by presenting a tree structure, in which the terminal nodes are the atomic elements of the sentence, and the nonterminals represent applications of particular principles. To take a simple example, consider a context-free grammar, such as that given in Figure 4.1. The general principles in this case are simple rules which decompose a sentence into smaller phrases; for instance `S-->NP,VP` states that a sentence can be formed

by a noun phrase followed by a verb phrase. The analysis of a given sentence can be given as a tree structure, in which each nonterminal is a phrase, whose children are sanctioned by the application of a rule.

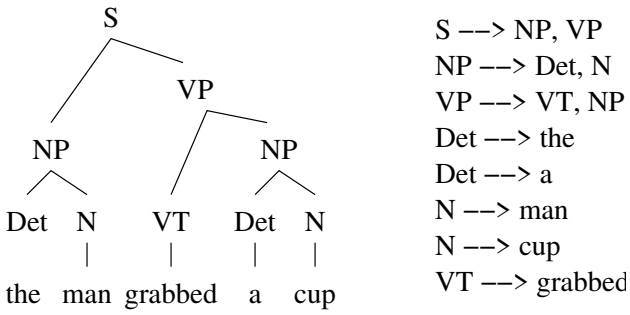


Figure 4.1: A simple syntactic tree

Simple context-free rules such as those shown in Figure 4.1 are not sufficient to describe the set of sentences in a natural language. However, all syntactic theories involve the postulation of general principles decomposing sentences recursively into intermediate phrases, and as a consequence, the use of tree diagrams to illustrate the decomposition of particular sentences. Basically, a tree diagram of a given sentence is a description of how the mechanism which generates *all* the sentences in a language generates *that particular* sentence.

4.2 Phonetic form and logical form

Our aim is to present an account of syntax grounded in sensorimotor cognition. We will assume that individual differences between people's sensorimotor apparatus are relatively small, and thus that the same sensorimotor theory will do for all people. However, the languages people speak vary widely. If we want to ground language in sensorimotor cognition, we cannot assert any *direct* mapping between sentence structures and sensorimotor processes. A way round the problem is offered by theories like GB and Minimalism. These theories hold that an adequate linguistic model of sentences requires their analysis not just at a surface level, but at an underlying level at which many of the surface differences between languages disappear. This underlying level of linguistic representation is thus a promising candidate for characterisation in sensorimotor terms.

GB posited four levels of linguistic structure: **phonetic form** (**PF**), **surface structure** (**SS**), **deep structure** (**DS**) and **logical form** (**LF**). In Minimalism, only two levels of

structure are invoked: PF and LF. I will adopt the Minimalist position, and assume just LF and PF in this chapter, and in the remainder of the book. PF is the representation of the surface form of the sentence, including such features as word order, intontation and phonology. LF is defined as the syntactic representation of a sentence which 'interfaces with meaning'. If the 'meaning' of a sentence has something to do with sensorimotor simulation, as I am suggesting, then we expect LF to tell us something about how linguistic representations interface with sensorimotor ones.

The Minimalist theory comprises two components: a declarative representation of **phrases** (elements of hierarchical syntactic structure) and a **generative mechanism** for forming and altering phrases. Structurally, phrases are quite simple: the general form of phrases will be described in Section 4.3. Most of the complexity of the theory resides in a specification of the generative mechanism for building and manipulating phrases. Again, the generative mechanism is not to be understood as a 'processing model' of how sentences are actually interpreted or produced. Rather, it is an abstract theoretical device which is used to characterise the full range of well-formed sentences in a language. The goal of a linguist studying a given language is to define a mechanism which produces all (and only) the well-formed sentences in the language under investigation.

The generative mechanism produces an LF structure and an associated PF structure. The stages involved in the process, as envisioned within the Minimalist model, are illustrated in Figure 4.2. The generative mechanism begins with a series of **phrase-formation**

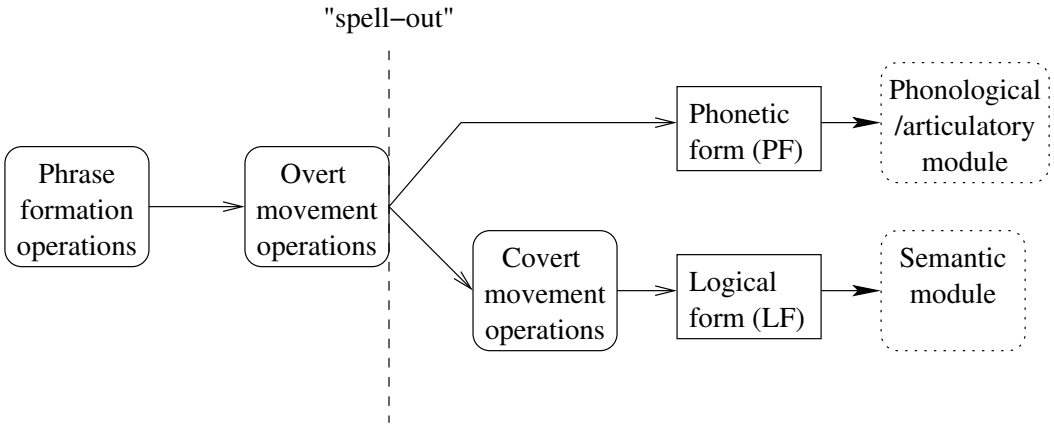


Figure 4.2: Flow chart for derivations in Minimalist syntax. (Syntactic operations in solid rounded boxes; syntactic representations in solid square boxes.)

operations, to create a hierarchical structure of phrases. A series of **movement opera-**

tions is then executed on this basic structure. This series is divided into two subsequences: a sequence of **overt movement** operations, followed by a sequence of **covert movement** operations. The structure generated after the overt operations is the one which interfaces with the phonological and articulatory processes in the model. Further movement operations are then applied to generate the LF of the sentence. These are called covert, because they do not show up in the phonological representation of the sentence. The full sequence of operations involved in generating the LF of a sentence is called a **derivation**. The point in the derivation at which PF is read off is called **spell-out**.

4.3 X-bar theory

The hierarchical phrase structure of a sentence can be very complex. But an interesting proposal in generative grammar is that phrase structures are all built from the same basic building blocks. The building block is called an **X-bar schema**. X-bar theory, introduced by Jackendoff (1977), is one of the most enduring components of generative grammar—in fact, it features in many other grammatical theories as well. There are two key ideas.

The first idea is that knowledge about phrase structure is stored in the lexicon, distributed amongst individual lexical entries, rather than in a set of disembodied phrase structure rules. According to this idea, each word in the lexicon comes with a piece of syntactic structure specifying the local syntactic context in which it can appear. In a generative model, the word can be thought of as 'creating', or **projecting**, its own local syntactic environment. The phrase structure of a whole sentence can then be broken up into components, each of which is projected by one of its constituent words. The structure of the sentence is thus formed by plugging together the local structures associated with the individual words.

Words of different grammatical categories will of course project different grammatical structures. A noun (N) projects a noun phrase (NP), a verb (V) projects a verb phrase (VP), and so on. The second idea in X-bar theory is that each word in the lexicon projects a syntactic structure with the same basic form, regardless of its grammatical category. The basic form is shown in Figure 4.3. This figure shows an X-bar schema, or **XP**, which is the unit of phrase structure projected by a word of category X. Within the schema, there are three distinct syntactic positions. The word itself appears at the **head** (**X**) position. The other two positions are places where other XPs can appear. One is called the **specifier**, and the other is called the **complement**. The XP unit itself is sometimes called the 'maximal projection' of the X head: it represents the complete piece of phrase structure contributed by the head.

Complements are best defined semantically. A head word can be thought of as con-

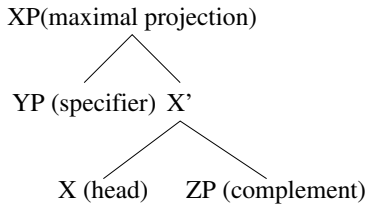


Figure 4.3: The X-bar schema

tributing a component of meaning to a sentence. But the meaning must in many cases be further specified. For instance, a verb like 'grab' presumably contributes a representation of a 'grab' action. But in order to properly specify this action, we must describe what object was grabbed. The complement of an XP provides a position for another phrase contributing the additional information. A phrase which contributes some of the semantic information required by a head is called an **argument** of the head. Thus the complement XP of a head is the position at which one of its arguments should appear.

In Figure 4.3, the complement of the X head is labelled 'ZP'. ZP should be understood as a placeholder for another maximal projection, contributed by some other word. In other words, the X-bar schema is recursive: one maximal projection has slots in it which are to be filled by other maximal projections. Figure 4.4 shows two X-bar schemas joined together, one appearing as the complement of the other. Note that these are separate 'applications'

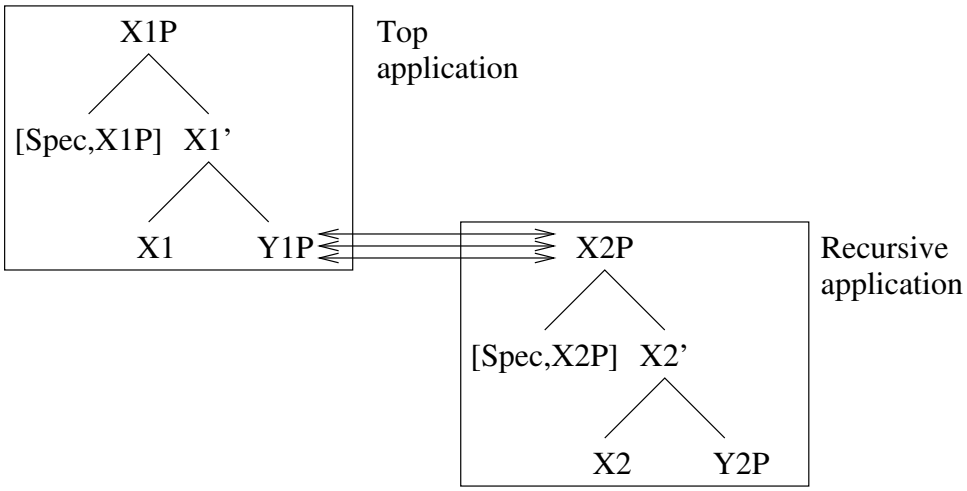


Figure 4.4: A recursive application of the X-bar schema

of the same general schema. They might have different heads: for instance, X1P might be a VP, and X2P might be an NP.

The X-bar schema also contains a second slot which can be recursively filled by an XP: the **Specifier** (or **[Spec,XP]**) position. The specifier is projected at the highest level of structure in the schema. While the distinction between specifiers and complements is very important in Minimalism, it is surprisingly difficult to define specifiers in a pre-theoretical way. However, in one sense, specifiers function very much like complements, providing a syntactic position where an argument of the head can appear. A head may require several arguments, of different kinds. For instance, the verb *grab* describes an episode involving an agent performing an action on a target object: the VP projected by this verb must therefore contain syntactic positions for both the agent and the target. A simple way of thinking about the distinction between specifiers and complements is that they are positions for different types of argument. In the case of verbs, arguments are distinguished using labels which refer to the different semantic roles which they play in the described episode. These roles are called **thematic roles** or **theta roles**. A verb like *grab* requires reference to an 'agent' and a 'patient'.[1] A common proposal is that the agent and patient of a transitive verb like *grab* appear respectively at the specifier and complement positions of the VP it projects. (At least, in an active sentence.)

There have been three main alterations to the X-bar schema since it was first proposed. The first concerns the type of grammatical element which can appear at the head of an XP. In Jackendoff's original conception, the head of an XP was always a word. But later it was proposed that a number of other primitive syntactic elements can project their own XPs (see e.g. Stowell, 1981; Abney, 1987). For instance, the inflections on verbs project 'inflection phrases', as will be discussed in Section 4.5. XPs headed by words are termed **lexical projections**, and those headed by other syntactic elements are termed **functional projections**.

The second alteration concerns the position occupied by **modifiers**. Modifiers also introduce whole XPs. But semantically, they provide additional, parenthetical information about the object contributed by the head, rather than obligatory information. For instance, the modifier of a verb head might be an adverbial phrase, describing the manner of an action, or a prepositional phrase describing where it took place. Traditionally, the X-bar schema also contains a position in between Spec and YP at which modifiers can be

[1] Of course, different verbs require reference to different roles. There are many suggestions about what might constitute a complete set of theta roles: the set often contains roles like 'location', 'path', 'stimulus', 'experiencer', etc. Since I am focussing on the verb *grab*, I will make do with just two theta roles, 'agent' and patient, for now. In fact, there is an interesting proposal by Dowty (1991) that there are just two basic types of theta role, whose definitions centre around the concepts of agent and patient.

recursively attached. More recent analyses of modifiers allow them also to adjoin at the very highest level to XP, and I will assume this analysis in Sections 4.5 and 4.8.

The final alteration to X-bar theory concerns assumptions about the linear order of elements of the schema. Originally, the X-bar schema specified the hierarchical relationship between heads, specifiers and modifiers, but it was assumed that their linear order was unconstrained, and was in fact a dimension along which languages varied. In Minimalism, since Kayne (1994), the assumption is that the X-bar schema imposes a fixed order on its elements, with XP modifiers preceding specifiers, specifiers preceding heads and heads preceding complements. Again, I will adopt this assumption in Sections 4.5 and 4.8.

4.4 The structure of a transitive clause at LF: Overview

The LF for the English transitive sentence *The man grabbed a cup* is given in Figure 4.5. Two things must be noted immediately. To begin with, notice that the figure is annotated

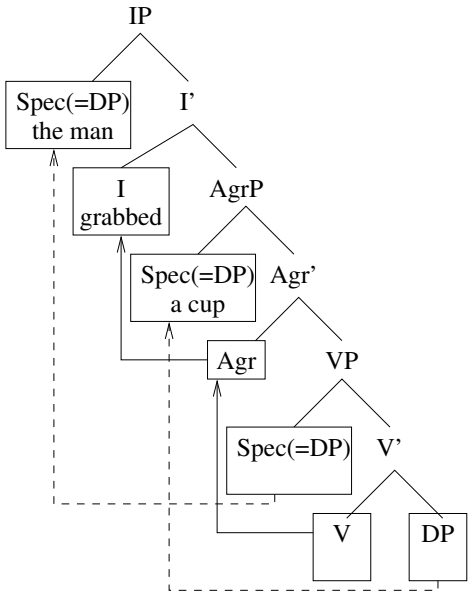


Figure 4.5: X-bar structure of a transitive clause at LF

with a number of arrows. These arrows indicate the movement operations which have occurred during the derivation of LF. There are two kinds of movement; one concerns the verb, and is shown with solid arrows; the other concerns noun phrases, and is shown

with dashed arrows. Most of this movement is covert, occurring after spell-out. The only overt movement is that of the subject *The man*. Note also that the XPs involved in the LF structure are mostly unfamiliar: there is an 'IP', an 'AgrP', and some 'DP's. In fact the only familiar one is likely to be VP. I will briefly outline all these innovations in this section, and motivate them in more detail in Sections 4.5—4.8.

Firstly, in GB and Minimalist syntax, a verb (V) appears at a separate position in LF from the inflection (I) on the verb which agrees with its subject. In fact, the whole clause is seen as a projection of the inflection of its main verb. (**Inflection phrase** or **IP** is the generative grammar term for 'sentence'.) Note that the specifier of the IP projection ([Spec,IP]) is a position associated with the subject of the sentence. Note also that the main verb originates at the head of VP, but moves (through an intermediate head) to the head of IP. It is through this mechanism that the verb is related to its inflection, as I will explain in Section 4.5.

Secondly, note that the phrases *the man* and *a cup* have been labelled **DPs** or **determiner phrases**, rather than 'noun phrases'. There is considerable agreement amongst syntacticians that the heads of 'referential' expressions like *the man* and *a cup*, as well as of 'quantifying' expressions like *every man* and *many cups*, should be determiners (*the*, *a*, *every*, *many* etc) rather than nouns. I will not discuss this proposal in any detail until Chapter 7.2, when I consider the internal syntax of such phrases. But it is so commonly accepted, both in Minimalism and in other syntactic theories, that I will use the term 'DP' to refer to what are pretheoretically called 'noun phrases' from now on.

Thirdly, following an argument by Koopman and Sportiche (1991), the subject DP appears in two positions. It originates at [Spec,VP], and raises to [Spec,IP]. I will discuss DP raising in general in Section 4.6, and the VP-internal subject hypothesis in particular in Section 4.7.

Fourthly, there is an extra XP intervening between IP and VP called **AgrP** or **agreement phrase**. While IP is associated with the subject, AgrP is associated with the object: its head (Agr) is the 'home' of any morphology on the verb agreeing with the object, and its specifier is a position to which the object raises. The AgrP projection will be discussed in Section 4.8.

4.5 The IP projection

IP is a functional projection, whose head is—or at least can be—an inflection. An inflection is a morphological affix on a word, rather than a word in its own right. It can carry various simple types of information; for instance, in English the inflection on a verb can carry information about tense, or about its subject. For instance, in Example 4.1, the verb *walks*

is decomposed into a **verb stem** *walk*, and the inflection *-s*, which carries information about tense, and about the grammatical person and number of the subject *John*.

(4.1) John walks.

A sentence containing an inflected main verb is termed **finite**. In simple finite sentences, like the one above, or like *The man grabbed a cup*, the head of IP is occupied by the inflection of the main verb—i.e. by *-s* or *-ed*.

An argument is obviously needed to explain why the inflection of the main verb of a sentence should appear separately from—and apparently on the wrong side of!—the verb itself. This argument involves data from several languages, and I will summarise it below.

To begin with, note that in English, verbs do not always need to be inflected. But when a verb is not inflected, it must be introduced by another syntactic item called an **auxiliary verb**—for instance, *can*, *will* and so on.

(4.2) *John walk.

(4.3) John can walk.

Note also that if an auxiliary verb appears, the main verb *cannot* be inflected: if it is, the sentence is ill-formed:

(4.4) *John can walks.

Inflections and verbs seem to be in complementary distribution: at least one of them must be present, but they cannot both be present. This suggests that auxiliary verbs in some sense fulfil the same role as inflections. The GB claim is that auxiliary verbs and inflections actually occupy the same position at LF—the head of a projection called IP, which dominates VP, as shown in Figure 4.6. In this scheme, the main verb originates at the V head and inflections and auxiliaries originate at the I head. (The subject of the sentence is in [Spec,IP], for reasons which I will discuss in Section 4.6.) If the head of I is an auxiliary, the V and I heads stay in these positions. But in finite clauses, the inflected verb is produced by **head-to-head movement** bringing the inflection and the main verb together. Following work by Emonds (1978), the kind of movement involved is assumed to be different in different languages. Emonds proposed that in English, the inflection moves down to the verb, while in 'Romance' languages like French and Italian, the verb moves up to the inflection. This proposal is supported by many pieces of evidence—I will discuss three which are commonly used to introduce the idea.

Firstly, the ordering of verbs and adverbs is different in English and Romance. In English, adverbs come before verbs, while in Romance, they come after:

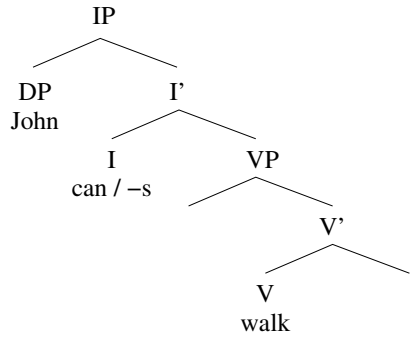


Figure 4.6: IP and VP. At LF, the head of IP can contain either an auxiliary or an inflection.

(4.5) John often drinks tea.

(4.6) *John drinks often tea.

(4.7)
Jean	boit	souvent	du	thé.
John	drinks	often	of	tea.

'John often drinks tea'

(4.8)
*Jean	souvent	boit	du	thé.
John	often	drinks	of	tea.

If we assume that adverbials originate immediately above the VP, then the proposal of I-to-V lowering in English and V-to-I raising in Romance explains this ordering data, as shown in Figure 4.7.

Secondly, note that English and Romance also differ as regards how negations are formed. In English, negative polarity sentences use an auxiliary verb (such as *do*) and an uninflected main verb, with the negative polarity marker *not* positioned in between:

(4.9) John does not walk.

In Romance, they use an inflected verb, positioned to the left of the negative polarity marker *pas*:[2]

(4.10) Jean ne marche pas.
John *ne* walks not.

[2]The word *ne*, also necessary in negative polarity sentences, is assumed to adjoin to the verb, for reasons I will not discuss.

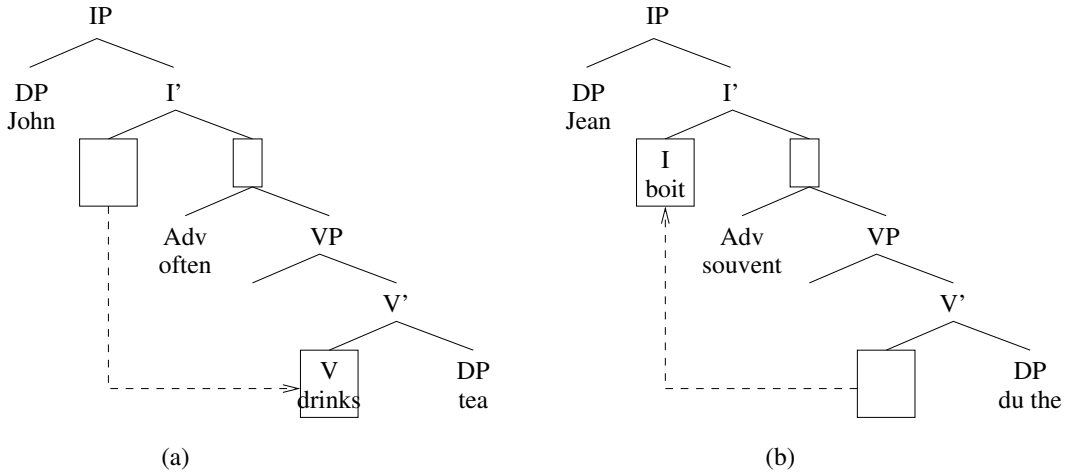


Figure 4.7: I-V lowering in English and V-I raising in French

If we assume that the negative polarity item occurs in a position similar to adverbs, then Example 4.9 is what we expect if there is no movement, and Example 4.10 is what we expect if V raises to I.

Finally, the idea that V raises to I in French but not in English receives additional support from some data on a separate topic: question formation. To begin with, note that in English questions, the auxiliary verb appears before the subject:

(4.11) Can John walk?

The GB story is that questions require movement of the I head to the head of a higher projection, the **complementiser phrase** or **CP**, as shown in Figure 4.8. This movement is another instance of head-to-head movement. (It incidentally provides a simple precedent for the idea that the head of one projection can move to the head of a higher projection.)

In question-formation in Romance languages such as French, it is not just the auxiliary which moves to the left of the subject; the whole inflected verb moves.

(4.12) Marche t-il?
Walks he?

(4.13) Aimez-vous Brahms?
Like you Brahms?

Emonds (1978) proposed that question formation in Romance involves *two* instances of head-to-head movement, as shown in Figure 4.9. First, the uninflected V moves up to

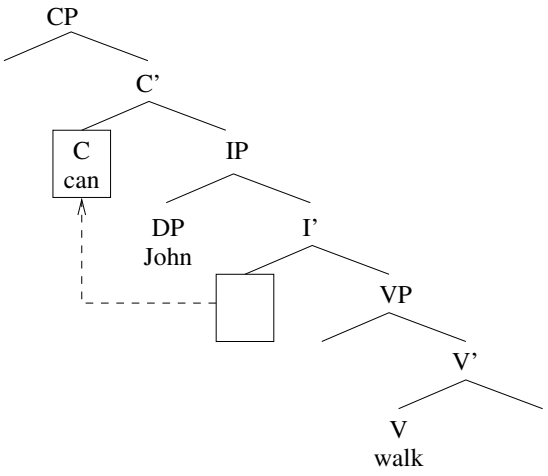


Figure 4.8: Movement from I to C in English question formation

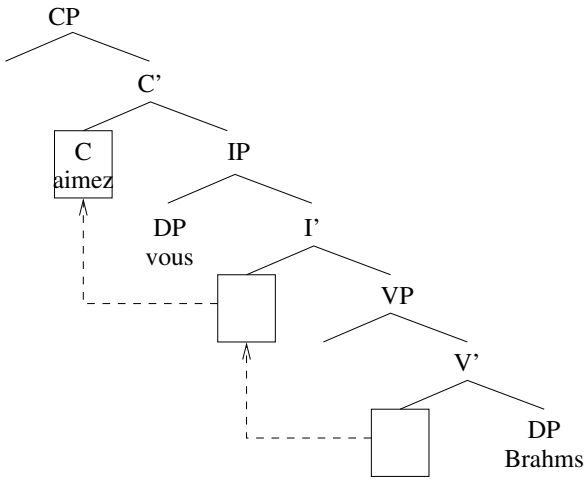


Figure 4.9: Movement from I to C in French question formation

the I head, to join with the inflection. Then, the whole inflected verb moves up to the preposed C head position. In general, there is an assumption that head-to-head raising works **cyclically**, by a series of movements from one head to the head immediately above. English questions (e.g. Example 4.11) always involve preposing of auxiliaries, leaving the uninflected verb in its original position.

In summary, the postulate of movement between V and I positions allows a neat explanation of three differences between English and Romance languages, in question formation, adverb placement and negation placement.

It is important to say a little about the mechanism which allows V to raise and I to lower, because it is very different in GB and in Minimalism. In GB, a rather stipulative rule of 'affix-lowering' is proposed by which I moves down to V in English. In Minimalism, it is assumed that verbs appear fully inflected in the V position, but need to **check** the semantic features contributed by these inflections by raising to higher head positions. Recall that in Minimalism, some movement occurs before spell-out (and is therefore reflected in a sentence's phonological form), while other movement occurs 'covertly', after spell-out. The Minimalist suggestion is that V-to-I movement happens prior to spell-out in French, but after spell-out in English, which is why English inflected verbs appear at V, while French verbs appear at I. This difference between English and French is attributed to the differing **strengths** of English and French agreement feaures: French verbs have strong agreement features, and English verbs have weak ones. The assumption is that only words with strong agreement features can move to I before spell-out. The strength of verb agreement features in turn correlates with overt morphology: French verbs have overt inflections carrying agreement information (person, number, gender) while English verbs (except *have* and *be*) do not. The idea that verbs raise to I if they carry overt agreement morphology is supported for a range of different languages; see Rohrbacher (1994); Vikner (1995). The Minimalist analysis of verb movement is less stipulative and more general, and I will adopt it henceforth. However, no matter how I-lowering is construed, the general idea of movement between V and I positions has considerable explanatory power.

4.6 DP-movement and Case assignment

In GB and Minimalist theory it is not just heads which can move from one position to another during a derivation. Whole DPs can move too, for different reasons, and subject to different constraints. The phenomenon I will focus on is called **DP-raising**.

To begin with a concrete example, consider the following sentence:

(4.14) It seems John walks.

There are two verbs in this sentence: *seem* and *walk*. Semantically, *seems* is a predicate which applies to a whole proposition, and thus takes a sentential complement which provides this proposition. The subject of *seems* (which is *it*) is analysed as an **expletive DP**, which is present for syntactic reasons (in English) but does not contribute semantically to the sentence. *Walks*, an intransitive verb, is a verb with one argument, which is assigned the thematic role 'agent'.

Now consider a variant of Example 4.14:

(4.15) John seems to walk.

This sentence means roughly the same as Example 4.14. However, on the face of it, the subject of *seems* is now the DP *John*, and the verb *walk* appears without an obvious subject, in the VP complement of *seems*. Of course, the meaning of the sentence still requires us to identify John as the agent of the walk action. We somehow have to explain how it is that *walk* assigns its 'agent' theta role to *John*, even though *John* appears to be the subject of *seems*.

In GB and Minimalism, the proposal is that Examples 4.14 and 4.15 in fact derive from similar underlying LF representations, where *seems* introduces a complement clause whose subject is *John* and whose main verb is *walk*. In Example 4.14, this LF structure is preserved at PF. But in Example 4.15, the DP *John* moves from the subject position of the embedded clause to the subject position of the top-level clause, as shown in Figure 4.10. (In Minimalism, we must add that this movement is *overt*.) When it moves, it creates what is called an **A-chain** linking its original position with its new position. All elements in an A-chain are associated with the same thematic role. The thematic role 'agent' which is assigned to the lower position of the A-chain is thus transmitted to the higher position, which ensures that the agent of *walk* is contributed by the DP *John*.

The reason why *John* can move to the subject position of the top-level clause is that this position is empty; as we have seen, it does not contribute anything to the semantic representation of the sentence. However, note that DP-raising in the example just discussed is not just possible; it is obligatory. If the DP *John* stays as the subject of *walk*, the result is ungrammatical:

(4.16) *It seems John to walk.

In GB and Minimalism, the story about why DP-movement is obligatory in this context involves reference to an abstract concept known as **Case**. (The capital is used to refer to a technical conception of case which is related to but distinct from the case inflections needed on nominal elements in several languages.) The claim is that every DP must be 'assigned Case' by some other element in the syntactic structure of the sentence. Exactly what Case

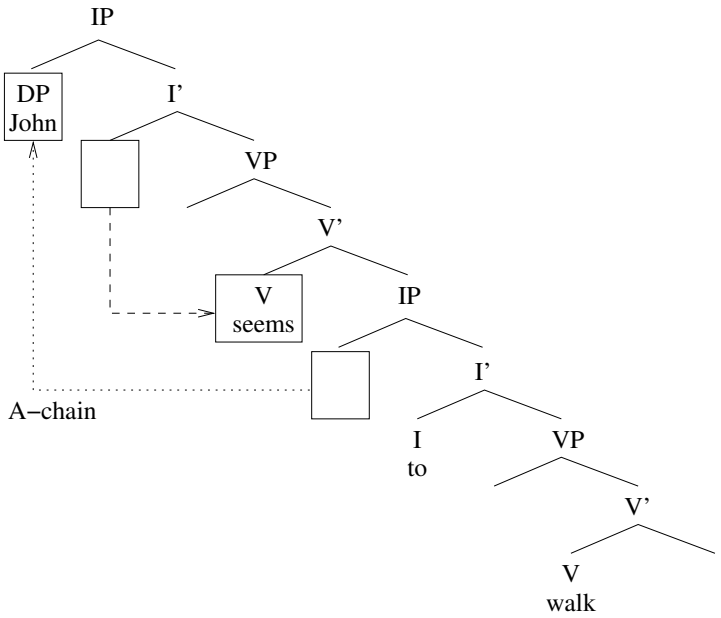


Figure 4.10: DP movement

is, and why DPs must be assigned Case, is not spelled out in the theory; Case and Case assignment are simply postulates motivated by their usefulness in a parsimonious theory of sentence syntax. GB stipulates that Case can only be assigned by head constituents, and that they can only assign Case to constituents which they **govern**. Simplifying greatly: a head X governs a DP if the DP appears as its specifier [Spec,XP], or its complement YP. The object DP of a verb can thus be assigned (accusative) Case by a verb head. However, the subject DP, which appears as [Spec,IP], is assigned (nominative) Case by the head of IP, namely an inflection. To explain the ill-formedness of Example 4.16, GB stipulates in addition that nominative Case can only be assigned by a *finite* I head. The DP *John* cannot receive Case as the specifier of the nonfinite I in the embedded clause. It therefore moves to a free position where it *can* be assigned Case, by the finite I associated with the higher verb *seem*. The principle which allows this movement in GB is called **move-alpha**, and there is a similar principle in Minimalism. It is somewhat dramatic: it says that any constituent can move to any position, provided this movement does not violate any of the other constraints given in the theory. As with head-to-head movement, the relationship between DP-movement and actual mechanisms for producing and interpreting sentences is extremely distant; GB simply does not address this question. But it is a question which has to be addressed at some point.

In summary, to explain sentences such as those given in Examples 4.14 and 4.15, GB makes use of a number of abstract theoretical constructs: nominative and accusative Case, Case assignment mechanisms (including the notion of government), A-chains created by movement, and the assignment of thematic roles to A-chains. (Minimalism also invokes all these constructs, except for government.) While these constructs are extremely abstract, and the relationship between the move-alpha device and actual sentence generation/interpretation mechanisms is very unclear, there is again considerable explanatory power in the idea that a DP requires both a thematic role and Case, and can sometimes obtain these from heads at different positions in a sentence. In addition, the notion of (finite) I as a case assigner meshes well with the general account of IP given above in Section 4.5.

4.7 The VP-internal subject hypothesis

In later incarnations of GB, and in Minimalism, the notion of DP-raising plays a role not just in sentences featuring clausal complements, but in simple sentences too. Koopman and Sportiche (1991) propose that the subject DP of a sentence originates in the specifier of VP [Spec,VP], and raises to [Spec,IP] to get Case. This model basically assumes that I is a raising verb, which behaves much like *seem*. There are a number of reasons for this suggestion, of which I will note three.

Firstly, consider an auxiliary verb like *can*, which as argued in Section 4.5 appears as the head of an IP projection. Example 4.3 is repeated below:

(4.17) John can walk.

The verb *can* in fact has a lot in common with a verb like *seem*. Semantically, it can be analysed as contributing a predicate whose argument is a whole proposition; *can* would typically be analysed as a modal operator over propositions. But syntactically, the subject of this proposition is distanced from the predicate: it appears in the specifier of IP, while the predicate appears in the complement of IP. Moreover, auxiliary verbs like *can* allow an expletive subject, which is not assigned an external theta role:

(4.18) It can rain hard in these parts.

Koopman and Sportiche therefore argue that the subject DP of a sentence like Example 4.17 in fact originates within the VP—specifically, in [Spec,VP] position. To complete the story, they then suggest that [Spec,VP] is not a position in which the subject DP can receive Case, either from the V head or the I head. Movement to [Spec,IP] is then obligatory, as shown in Figure 4.11.

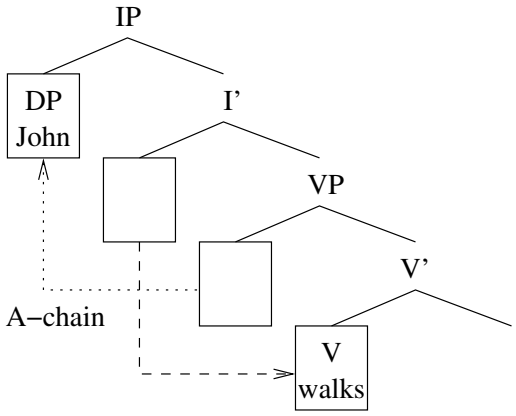


Figure 4.11: DP raising from [Spec,VP] to [Spec,IP]

Another way of motivating the above account of I as a raising category is to note that it simplifies the story that needs to be told about how the thematic roles associated with a verb are structurally assigned. Recall from Section 4.3 that a verb projects a VP structure, containing empty specifier and complement positions. Each verb has a particular set of thematic roles, which must be assigned to different structural positions: for 'grab', the subject must be assigned the 'agent' role, and the complement must be assigned 'patient'. It is most natural to assume that the verb assigns its theta roles locally, to positions within the VP it projects. Then, as already proposed in Section 4.3, 'agent' can be assigned to the specifier position, and 'patient' with the complement. But in languages like English and French, we also need a subject position outside of the VP, because in these languages, the subject appears to the left of the verb in surface sentences. So again, it is attractive to assume that the subject originates at [Spec,VP], where it receives its theta role from the verb, but then moves to a higher position in IP (to receive Case). The alternative to this idea, that the subject originates at [Spec,IP], requires that the verb is able to assign its agent theta role in a special way, to a position outside its maximal projection. This was indeed the proposal in GB, but Koopman and Sportiche's proposal is a more economical one.

A final motivation for the VP-internal subject hypothesis comes from cross-linguistic considerations. There are many languages where the subjects, verbs and objects of transitive sentences appear in a canonical order, allowing subject and object DPs to be identified by their serial position in a sentence.[3] Different languages have different canonical order-

[3]In other languages, the ordering of constituents in a sentence is unconstrained. These are languages

ings: for instance, English and French are SVO (subject-object-verb), Korean is SOV, Welsh and Polynesian languages are VSO. If we assume a model where subjects originate at [Spec,IP], SVO order is easy to explain. SOV and VSO order are more difficult. I will consider SOV order in Section 4.8; for the moment I will just consider VSO order. The problem here is that if the subject originates at [Spec,IP], there is no obvious way the verb can move beyond it: [Spec,IP] is the highest position in the sentence. On the other hand, if the subject originates at [Spec,VP], then we can easily explain VSO languages by assuming that in these languages the subject is pronounced at this lower position, while V is pronounced after it has raised to I, as shown in Figure 4.12. We have already seen evidence in Romance languages that V can raise to I, as discussed in Section 4.5, so this account of VSO order is quite attractive. (And much useful work on VSO languages has

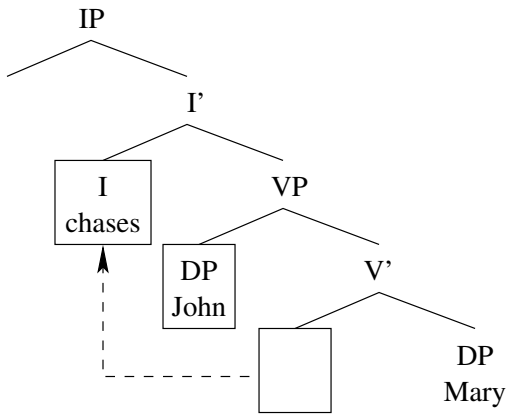


Figure 4.12: Generating VSO order by raising V to I and keeping subject at [Spec,VP]

been done using this basic analysis; see e.g. Pearce, 1997.) Note that we must still assume that the subject raises to [Spec,IP] to get Case; it just does so covertly in a VSO language. In SVO languages, of course, the subject moves overtly to this higher position.

In summary, there are several arguments which converge on the hypothesis that there are two subject positions in a transitive sentence. One has to do with commonalities between auxiliaries and raising verbs; one has to do with the mechanism by which a verb assigns a thematic role to its subject; and one relates to an analysis of VSO languages. Note that these arguments mesh well with the hypothesis that inflections have their own IP projection (discussed in Section 4.5) and with the idea of DP movement (discussed in

where subjects and objects can be distinguished by overt (morphological) case marking. I will not discuss free word-order languages here.

Section 4.6).

4.8 The AgrP projection

A final ingredient in the clause syntax sketched in Figure 4.5 is the AgrP projection. In late GB models and in Minimalist models, it is commonly proposed that objects as well as subjects need to raise to a higher projection to get Case (or in Minimalism, to check their Case features). According to this proposal, just as subjects raise to the specifier of IP to get nominative Case, objects raise to the specifier of a projection which intervenes between IP and VP, called AgrP, to get accusative Case.

The idea that the object of a verb can raise to a position in between IP and VP can be motivated in different ways. I will give two arguments.

4.8.1 Motivating AgrP: an argument from SOV word order

As already mentioned, there are many languages in which the object can appear in between the subject and the verb, to give 'SOV' order. For instance, here are examples from a southern Irish infinitival complement clause and a German subordinate clause, discussed by Carnie (2007).[4]

	Ba mhaith liom	[Seán	an abairt	a scríobh	]
(4.19)	Good with	[John	the sentence$_{ACC}$	write	]
	'I want John to write the sentence'				

	Weil	[ich	die Katze	nicht	streichle	]
(4.20)	Because	[I	the cat	not	petted	]
	'Because I did not pet the cat'					

If the object is generated as the complement of VP at LF, then how can we explain an SOV ordering, as in the above examples? In GB, it was assumed that the ordering of the head, specifier and complement of an XP could be different in different languages. SOV order would then result from the head of VP being positioned after its complement at LF. But Minimalism rests on an influential proposal by Kayne (1994), that the hierarchical position of constituents within an XP determines their linear order at LF, so that the head of an XP always comes after its specifier, but before its complement. In this scenario, to

[4]In fact there are probably several different types of SOV construction, which must be given different syntactic analyses. In particular, there is a difference between an 'object shift' construction and a 'scrambling' construction, which I will not go into here (but see Thráinsson, 2001 for a recent account).

allow SOV order, an object must move to a position in between IP and VP. We must thus postulate an additional XP sitting in between IP and VP.

While positing an intermediate XP clearly complicates the phrase structure of the sentence, it also enables a more economical account of how DPs are assigned Case, and of the agreement inflections on verbs. In the account of DP movement given in Section 4.6, we had to postulate two mechanisms for assigning Case: accusative Case was assigned by a V head to its complement, and nominative Case was assigned by an I head to its specifier. If objects raise to the Spec of an intermediate XP, then we can tell a uniform story, where all Case is assigned by a head to its specifier: the head of IP assigns nominative Case, and the head of the new XP assigns accusative Case. In the account of subject-verb agreement given in Section 4.5, we proposed that verbs raise from the V head to successively higher heads. There are also languages where verbs agree with their objects (see Baker, 2008 for an extensive review). Postulating an intermediate XP allows us to use the same account of verb movement to explain these inflections as well. The head of this XP is assumed to be the 'home' of the agreement features contributed by object-agreement inflections on the verb. (Consequently, this XP is called Agr_OP, or 'object agreement phrase', which I am abbreviating to AgrP.) The inflected verb first raises to the head of AgrP, to check its object agreement features, and then to the head of IP, to check its subject agreement features. The new account is more complicated in positing an extra XP, but also more economical, in that the *mechanisms* associated with Case and verb inflections are now exactly the same for the subject and the object.

4.8.2 Pollock's argument for AgrP

The idea of an intermediate projection between IP and VP was first proposed by Pollock (1989), for reasons to do with verb movement rather than to do with the position and Case of objects. His argument is quite complex, but it is worth rehearsing in some detail.

Pollock's starting point is Emonds' (1978) discussion of verb and adverb ordering in English and French, which was touched on in Section 4.5. To summarise that discussion: in English, auxiliary verbs appear on the left of adverbs and negation, while inflected verbs appear to the right of adverbs and are not present in negative polarity sentences, whereas in French, both auxiliaries and inflected verbs appear to the left of both adverbs and negation. This data was used to argue for an IP projection separate from the verb, and for V-to-I raising in French inflected verbs and I-to-V lowering in English inflected verbs. Pollock focusses on a slightly different distinction: rather than auxiliary and inflected verbs, he distinguishes between the verbs *have* and *be* (*avoir* and *être* in French), which can function as auxiliaries, and the class of 'ordinary' **lexical** verbs, which cannot function as auxiliaries. English *have* and *be* appear to the left of adverbs and negation,

just like auxiliaries (Examples 4.21 and 4.22) and unlike other English lexical verbs (e.g. Example 4.23).

(4.21) John has often smiled. / John has not smiled.

(4.22) John is often happy. / John is not happy.

(4.23) * John drinks often tea. / * John drinks not tea.

Pollock notes that while there is a distinction in English between lexical verbs and *have/be* which is not present in French, a similar distinction does resurface in French *infinitive* clauses, in which lexical verbs can raise past adverbs but not past negation. The relevant data is summarised in Table 4.1. This table describes how English and French verbs

		V Movement beyond Neg	V Movement beyond Adv	No V movement
Fin	Fr. be/have	√J n'est pas heureux	√J est souvent heureux	* J souvent est heureux
	Fr. lex vbs	√J n'embrasse pas M	√J embrasse souvent M	* J souvent embrasse M
	Eng be/have	√John is not happy	√John is often happy	√John often is happy
	Eng lex vbs	* John kisses not Mary	* John kisses often Mary	√John often kisses Mary
Inf	Fr. be/have	√N'être pas heureux	√Etre souvent heureux	√Souvent être heureux
	Fr. lex vbs	* N'embrasser pas M	√Embrasser souvent M	√Souvent embrasser M
	Eng be/have	√To be not happy	√To be often happy	√To often be happy
	Eng lex vbs	* To kiss not Mary	* To kiss often Mary	√To often kiss Mary

Table 4.1: Summary of data from Pollock (1989)

behave when moved to the left of negation markers (*not/pas*), to the left of adverbials (*often/souvent*) and left in situ to the right of adverbials. The table shows the behaviour of lexical verbs (such as *kiss/embrasse*) and *be/have* (with *be/être* being the example shown). It also shows the behaviour of verbs in both finite (fin) and nonfinite (inf) contexts.

Note first that in both finite and nonfinite clauses, English *be/have* can raise beyond both adverbs and negation, while English lexical verbs cannot do either. There is thus a distinction between lexical verbs and *be/have* in English. This distinction is not present for French finite clauses: in finite clauses, both lexical verbs and *be/have* can raise beyond both adverbs and negation (and in fact must do so). However, the distinction between lexical verbs and *be/have* reappears in French *nonfinite* clauses. Here, while both kinds of verb can raise beyond adverbs, only *be/have* can raise beyond negation.

Pollock identifies four generalisations from the data given in Table 4.1.

1. Moving past negation is harder than moving past adverbs. If a given verb can move past negation, it can always move past adverbs, but the reverse is not true.

2. All French verbs, whether lexical or *be/have*, can move past adverbs. But in English, only *be/have* can do this.

3. *be/have* can move past negation (and therefore past adverbs, by Generalisation 1) in both English and French.

4. Moving past negation is easier in finite clauses than in nonfinite ones (because French lexical verbs can do so in finite clauses, but not in nonfinite ones).

From these generalisations, Pollock argues that verb movement beyond negation actually happens in two stages: the verb first moves from V to the head of an intermediate projection immediately dominating VP, and then from here to I. This is illustrated in Figure 4.13. He

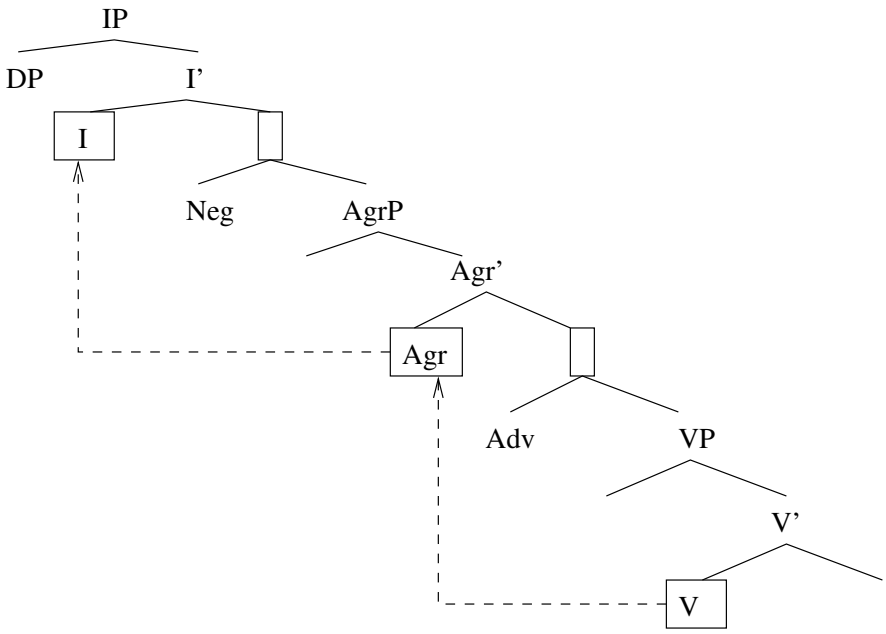


Figure 4.13: Pollock's proposal for movement from V to I via Agr

calls the intermediate projection AgrP. The key thing about the Agr head is that it is to the left of adverbs, but to the right of negation elements. (Again, the figure remains neutral about what kinds of projection introduce negations and adverbs.) We can now make use of the idea motivated in Section 4.5, that head-to-head movement is always cyclic; in other words, that to move V to I, we must first move V to Agr, and then move Agr to I. This counts as an explanation of Generalisation 1 above. To capture the data in Table 4.1, we can then formulate separate constraints on these two movement operations, as follows.

- Constraint 1: French lexical verbs can move from V to Agr, but English lexical verbs cannot. (This principle is in turn derived from the idea that French has a richer inflectional system for verb agreement.) This explains Generalisation 2.
- Constraint 2: The verbs *be* and *have*, whether in English or French, can always move from V to Agr and from Agr to I. This explains Generalisation 3.
- Constraint 3: In finite clauses, lexical verbs can move from Agr to I, but in nonfinite clauses, they cannot.

We can now note a corollary of these three constraints:

- In French finite clauses, where all verbs can move to Agr (either by Constraint 1 or Constraint 2), all verbs can move on to I (by Constraint 3). But in English finite clauses, where only *be*/*have* can get to Agr in the first place (by Constraint 2), only *be*/*have* can move on to I (by Constraint 3). This explains Generalisation 4.

It is assumed (for reasons we won't go into) that negative polarity items like *not*/*pas* head their own **NegP** projections, which dominate the VP projection. So in negative polarity sentences, there is an extra projection in between the I and V heads, and I has further to travel to adjoin to V (or V has further to travel to raise to I).

What is the function of the Agr projection? We already saw one proposal, in Section 4.8.1, that an intermediary projection is needed to explain SOV word order and object-agreement inflections on the verb. In fact, Pollock's original idea was that his new intermediate projection contributed the *subject* agreement morphology on the verb, while IP contributed *tense* agreement morphology. He suggested renaming IP 'TP', denoting 'tense phrase'. However, there are good reasons for arguing that the projection contributing subject agreement sits *above* the projection contributing tense. These stem from the fact that verbs pick up agreement morphemes before they pick up tense morphemes (Belletti (1990). For instance in Italian, tense morphemes attach directly to the verb, with agreement morphemes attaching onto the complexes thus created. E.g. in the inflected verb *parl-av-ano* (*they were speaking*), the tense morpheme *-av* ('past') attaches first to the verb stem *parl-* ('speak'), and the agreement morpheme *-ano* (3rd person plural) attaches second. Thus Belletti agrees with Pollock that IP should be split into TP and AgrP, but places her AgrP above TP. The idea that Pollock's intermediate projection is associated with object agreement is originally due to Chomsky (1995). He suggested that a clause contains *two* Agr projections, as well as TP: there is a subject Agr projection above TP, as proposed by Belletti, and an object Agr projection below TP, which provides the intermediate Agr projection hypothesised by Pollock. Chomsky termed Belletti's Agr projection $\mathbf{Agr}_S\mathbf{P}$, and Pollock's Agr projection $\mathbf{Agr}_O\mathbf{P}$. Since I am not considering tense

projections for the moment, I will continue to use the term 'IP' to describe the subject agreement projection, and the unsubscripted term 'AgrP' to describe the object agreement projection.

4.9 Summary: strengths and weaknesses of the Minimalist model

We have now arrived at the basic structure of a clause originally presented in Figure 4.5. Naturally, we have only motivated this structure in broad outline. However, a considerable weight of argumentation converges on the model I have presented here.

Current work within Minimalism is mainly concerned with extending the model, in two directions. Firstly, it is possible to motivate a great many additional syntactic projections, using linguistic argumentation similar in style to that already presented. Most of these new projections are functional projections, whose heads have little or no phonological content, but which play an explanatory role by providing landing sites for various different types of movement. Examples of these functional projections are AgrSP (Belletti, 1990), ForceP (Rizzi, 1997) and ZP (Stowell, 1996). While all of these projections can be motivated linguistically, there is still considerable debate about each of them.

The main difficulty with Minimalist theory is that the two main tools it employs—functional projections and accounts of constituent movement—are both extremely powerful. Both of these tools can be very useful, as I hope to have shown in this chapter; there are good arguments that a theory of syntax should make use of these devices. However, accepting these devices presents a methodological problem: it becomes very hard to choose between different specific theories. Any particular theory is a combination of a complex proposal about a structure of XP projections and a complex proposal about constraints on movement of constituents within this structure. However, it is hard to look for evidence about a single element of this complex theory on its own. The empirical data of linguistics—judgements about the syntactic well-formedness of word sequences, or comparisons between the meanings of well-formed sentences—only provide direct evidence for a fully specified theory. And for a smallish set of relevant data, such as the example sentences I have given in this chapter, there are likely to be many such theories which can provide an explanation. (To take one relevant example: Iatridou (1990) has presented a theory which accounts for much of the data referred to by Pollock, without positing an AgrP projection.) To distinguish between competing fully-specified theories, linguists must formulate extremely complex arguments, drawing on a very large pool of linguistic data. While there is broad agreement within Minimalism on the kinds of principles and projections which

must feature in an adequate theory, there is considerable variation between specific theories presenting specific principles and projections. Making progress within the Minimalist programme is becoming increasingly difficult as the theory becomes more complex.

Chapter 5

The relationship between syntax and sensorimotor structure

If a syntactic theory provides a good model of human language, then it is tempting to look for cognitive (ultimately, neural) mechanisms and representations which correspond to the theoretical devices it employs. For instance, we might expect to find mechanisms or representations in the brain which correspond to elements of hierarchical phrase structure, such as verb phrases and noun phrases. Thinking specifically of the Minimalist model of syntax, we might expect to find cognitive correlates of mechanisms like verb movement or DP raising or Case assignment.

In fact, it is not necessarily the case that the components of a high-level 'symbolic' model like a formal grammar have clear neural correlates. One claim often voiced by neuroscientists and connectionist modellers (see e.g. many papers in Rumelhart *et al.*, 1986; McClelland *et al.*, 1986) is that symbolic models often misrepresent cognitive mechanisms by making clear distinctions between components which are in fact blended at the level of neural implementation. Conversely, proponents of symbolic accounts of cognitive phenomena often argue that descriptions of neural processes adopt the wrong 'level of representation' for these phenomena (see e.g. Broadbent, 1985; Fodor and Pylyshyn, 1988). If this is the case, there will be no point in looking for neural correlates of high-level syntactic constructs. But of course neither of these polemics should cause us to reject out of hand the possibility that there may be neural mechanisms which correspond to the components of a symbolic syntactic model. In fact, in the sensorimotor model which I presented in Chapters 2 and 3, there is some interesting high-level structure. It is an empirical question whether any of this structure can be identified with elements in a symbolic model of syntax.

In this chapter I will argue that there is indeed an interesting way of relating the sensorimotor model described earlier to the model of syntax outlined in Chapter 4—at

least for our example cup-grabbing episode. Specifically, I will suggest that the LF of the sentence *The man grabbed a cup* can be understood as a description of the sequence of sensorimotor processes involved in experiencing the cup-grabbing episode, as replayed in working memory. I make this claim on several grounds, but the main one is that there is an *isomorphism*—a strong formal similarity—between the LF representation motivated in Chapter 4 and the sensorimotor sequence argued for in Chapters 2 and 3. There is no particular reason to expect a similarity between the syntactic and sensorimotor models: they are motivated from completely different sets of empirical data, using completely different methods. If there are just a few similarities between the two models, it would be easy to dismiss these as a coincidence. But if the similarities run deep, as I will suggest they do, then they start to constitute evidence that the models are in fact related—in other words, that there may in fact be a close relationship between sensorimotor cognition and natural language syntax.

I will begin in Section 5.1 by summarising the model of sensorimotor processing and working memory introduced in Chapters 2 and 3. In Section 5.2 I will outline a basic proposal about how this model relates to the LF syntactic structure motivated in Chapter 4, and I will present the proposal in more detail in Sections 5.2–5.4. In Section 5.5 I will step back and consider some of the larger implications of characterising LF in sensorimotor terms. LF has a role in describing the interface between language and meaning, and a role in describing cross-linguistic syntactic generalisations, which must both be revisited if it is to be reconstrued in this way. Finally, in Section 5.6 I take a brief look at some other syntactic constructions mentioned in Chapter 4, to see if the sensorimotor characterisation of LF extends to these.

5.1 Summary of the sensorimotor model

In this section I will summarise the cognitive operations which will feature in the sensorimotor reinterpretation of LF, and I will introduce some terminology to refer to them.

Recall that we have focussed on a single concrete event: the action of a man grabbing a cup. In Chapter 2, I argued that the sensorimotor processes involved in executing or perceiving this action take the form of a strict sequence: first an action of attention to the agent (and its reafferent feedback, a representation of the agent), then an action of attention to the cup (and its reafferent feedback, a representation of the cup), then an evocation of the motor action (and its reafferent feedback, reattention to the agent), and finally haptic reattention to the cup. In Chapter 3, I argued that an agent can hold a representation of this grasp event in working memory, in the 'episodic buffer' proposed by Baddeley, and that this representation takes the form of a planned sequence of attentional and motor actions.

I further argued that the agent can 'internally replay' this planned sequence, for instance to store it in episodic memory. I concluded in Section 3.9 by reviewing the sequence of sensorimotor signals which occur during this replay operation. The schematic summary given in that section is repeated in Figure 5.1.

Sustained signals	Transient signals		
	Context signals	Action signals	Reafferent sensory signals
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_1	*attend_agent*	*attending_to_agent*
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_2	*attend_cup*	*attending_to_cup*
$plan_{attend\_agent/attend\_cup/grasp}$ ↓ ↓	C_3	*grasp*	*attending_to_agent*
$plan_{attend\_agent/attend\_cup/grasp}$	C_4		*attending_to_cup*

Figure 5.1: The time course of signals occurring during the replay of the cup-grabbing episode in working memory (repeated from Figure 3.9)

The replay process evokes a cyclic sequence of transient sensorimotor signals, mirroring the cyclic structure of the attentional and motor actions involved in perceiving or executing the cup-grabbing event. The general structure of an individual cycle is illustrated in Figure 5.2. Each cycle begins with the activation of an **initial context**, and with a tonically

action plan ↓ ↓ ↓	**initial context** **next context**	**sensorimotor action**	**reafferent sensory signal**

Figure 5.2: General schema for a single cycle of the replay process

active **action plan**. Together, the context and the action plan trigger a **sensorimotor action**, which in turn triggers a **reafferent sensory signal**. After a period of time, the action results in the establishment of the **next context**, and the cycle repeats. In the first

cycle in the replay process illustrated in Figure 5.1, the initial context is C_1, the planned action is $plan_{attend\_agent/attend\_cup/grasp}$, the action is *attend_agent*, the reafferent signal is *attending_to_agent* and the next context is C_2. To represent a sequence of cycles using this general schema, we can recursively define the 'next context' of one cycle as the 'initial context' of the following cycle. Thus the initial context for the second cycle in Figure 5.1 is C_2. I will refer in general to a sequence of cycles as a **sensorimotor sequence**, whether it is generated by a plan being internally replayed (as in the above example) or by some other mechanism.

Note that while successive cycles involve a succession of transitory context, action and reafferent representations, the action plan does not change from one cycle to the next. A single cycle involves the execution of a single action, but the plan representation which supports this is a plan to perform a sequence of actions. Thus while the 'context', 'action' and 'reafferent signal' elements of a cycle all describe operations local to this cycle, the 'action plan' element describes a representation which endures through several successive cycles in a sensorimotor sequence.

5.2 Sensorimotor interpretation of the LF of *The man grabbed a cup*: overview

My main proposal in this chapter is that the LF representation of the sentence *The man grabbed a cup* can be interpreted as a description of the replay process summarised in Section 5.1.

> **Principle 1** *The LF of the sentence 'The man grabbed a cup' is a description of the sensorimotor sequence which occurs in an agent when the agent internally replays a working memory representation of the episode of a man grabbing a cup.*

This proposal will be presented and motivated in detail below. But there are a few points which are worth making straight away.

Firstly, note that Principle 1 does not give a sensorimotor interpretation to the *surface* syntactic form of the sentence. It makes a claim about the underlying form of the sentence, LF—the form which is relatively invariant across languages, and which is understood as interfacing with semantic representations.

Secondly, note that Principle 1 does not constitute a 'processing model' of the sentence *The man grabbed a cup*. It is not concerned in any way with the processes which allow

a speaker to produce this sentence, or which allow a hearer to interpret it. The principle states that the LF of the sentence can be thought of as evoking a process, but this process is distinct from the process of generating or interpreting the sentence. (In Chapter 6, I will outline a procedural model of sentence generation, which will *exploit* the fact that LF is modelled as a process. Nonetheless, Principle 1 by itself does not say anything about sentence processing.)

Thirdly, it is worthwhile restating a general point made in Chapter 1. The proposal expresses a particular version of the 'shared mechanisms' hypothesis: it states that the linguistic structure of a sentence is somehow grounded in the sensorimotor mechanisms via which we perceive and interact with the world. There are many versions of this hypothesis. The basic appeal of the hypothesis has already been discussed in Section 1.1.1, and I will not reiterate the general motivation for it. As for the relationship of the present theory to other statements of the hypothesis in linguistics and cognitive science—this will be discussed in Section 13.1, after the present theory has been presented in detail.

5.3 A sensorimotor characterisation of the X-bar schema

I will elaborate the claim made in Principle 1 by proposing a general sensorimotor interpretation of the X-bar schema, which (as introduced in Section 4.3) is the basic unit of syntactic structure in Minimalism. The structure of the X-bar schema is reiterated in Figure 5.3. Every maximal projection XP has a **head** X and a **specifier** [Spec,XP], and

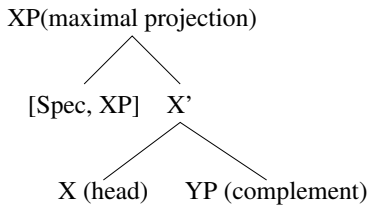


Figure 5.3: The X-bar schema (repeated from Figure 4.3)

introduces a **complement**, YP.

I propose that a single application of the X-bar schema describes a single cycle in a replayed sensorimotor sequence.

Principle 2 *A single application of the X-bar schema in LF can be understood as a description of the signals which are active during a single cycle of a replayed sensorimotor sequence, in which:*

- **XP** *denotes the* **initial context***;*
- *the* **X** *head denotes the* **sensorimotor action***;*
- **[Spec,XP]** *denotes the* **reafferent sensory signal***;*
- *the* **YP** *complement of X denotes the* **next context***.*

These proposed sensorimotor interpretations of the different components of an X-bar schema are illustrated (in blue) in Figure 5.4.

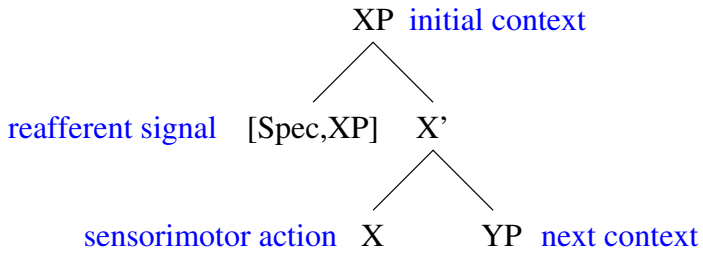


Figure 5.4: Sensorimotor interpretation of an X-bar schema

Principle 2 states a correspondence between a syntactic construct (an LF X-bar schema) and a cognitive process (a cycle in the replay of an episode representation in working memory). An X-bar schema application is taken to be a description of the 'signals which are active' during one cycle of the replay process.

I now consider how to give a sensorimotor interpretation of multiple X-bar schemas linked together. Recall that the X-bar schema is recursive: the complement of the X head is itself an instance of the X-bar schema. When we consider two applications of the X-bar schema, as in Figure 5.5 (repeated from Section 4.3), we can see how the YP of the higher application is identified with the XP of the lower application.

Now recall from Section 5.1 that cycles in a sensorimotor sequence are also defined recursively: the 'next context' of one cycle can be recursively specified as the 'start context' of the next cycle. This means that there is a natural characterisation of a right-branching recursive X-bar structure as a sequence of cycles.

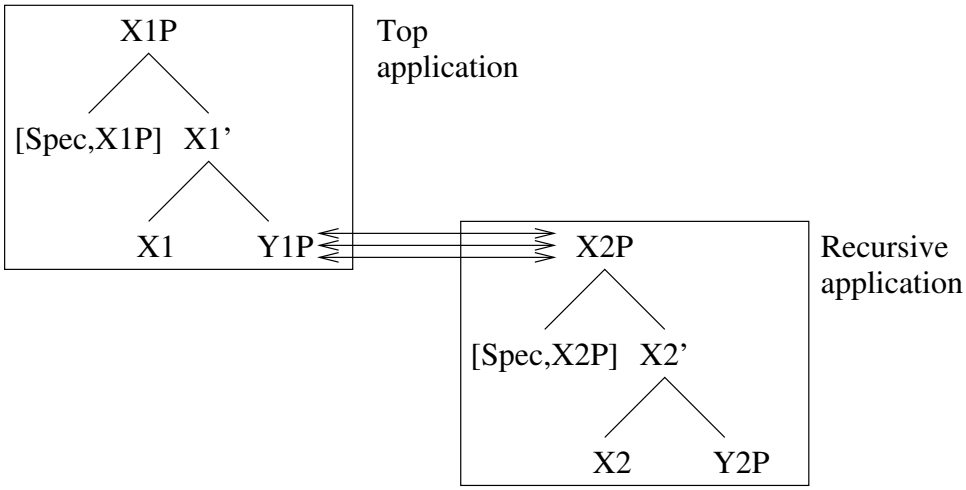


Figure 5.5: A recursive application of the X-bar schema (repeated from Figure 4.4)

Principle 3 *A sequence of n right-branching X-bar schema applications in the LF of a sentence can be understood as a description of n consecutive cycles in a sensorimotor sequence replayed from working memory.*

(This principle is actually a corollary of Principle 2, given the recursive definitions of both X-bar schemas and cycles.) To illustrate Principle 3, Figure 5.6 shows a generic right-branching structure of three X-bar schemas, and its sensorimotor interpretation (in blue) as a description of three successive cycles.[1]

Note that this sensorimotor characterisation of X-bar syntax is consistent with a key constraint in Kayne's (1994) account of the ordering of elements in an X-bar schema, namely that heads always precede their complements at LF (see Section 4.3). Kayne suggests that there is a constraint on LF representations which is closely analogous to the constraint on linear ordering of words in a surface sentence. This constraint is principally motivated from linguistic argumentation, but he also suggests (pp36–38) that there is a temporal component to LF representations. The present characterisation of an X-bar schema as a cycle in a sensorimotor sequence can be seen as an explanation for the fixed order of heads and complements in Kayne's model of X-bar structure.

[1]The last cycle is a special case, as it is where recursion terminates.

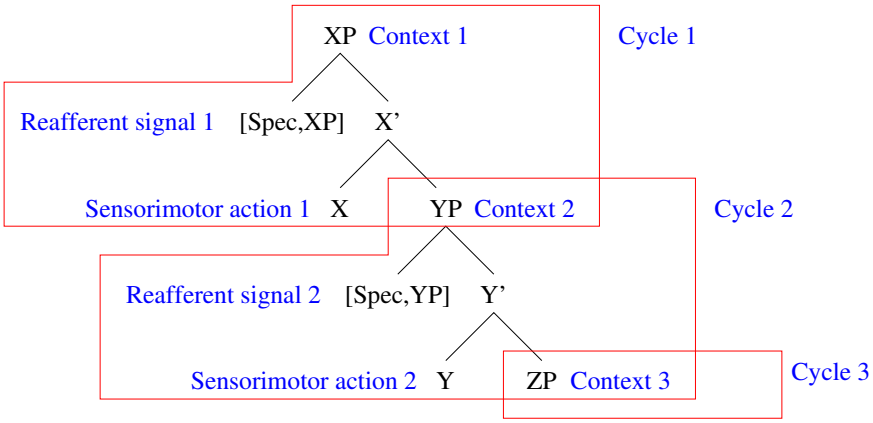


Figure 5.6: Sensorimotor interpretation of a right-branching recursive application of the X-bar schema

5.4 Sensorimotor interpretation of the LF of *The man grabbed a cup*

The above sensorimotor interpretations of linguistic structure are very general, because they are cast in terms of X-bar theory. In Minimalism, right-branching structures of X-bar schemas are common currency in LF representations: these structures emerge naturally as a result of the operations through which LF representations are 'derived'. Even if we limit ourselves to 'concrete' sentences, describing events or states which can be directly experienced by sensorimotor processing, this interpretation makes some very strong claims. Before even considering how to justify these claims, I will return to our example sentence, to show how they are instantiated for this sentence, and to introduce some additional principles to structure a discussion of their wider applicability.

Recall from Chapter 4 that the LF of *The man grabbed a cup* consists of four right-branching X-bar schemas, as illustrated in Figure 5.7. Principle 3 commits us to the claim that this LF structure describes a replayed sensorimotor sequence comprising four cycles. Now refer back to the summary of the sensorimotor model given in Section 5.1: the process of experiencing a cup-grabbing action involves four operations, which can be stored in, and replayed from, working memory. The signals evoked during replay are structured in four cycles, as shown in Figure 5.1. For this particular sentence, therefore, Principle 3 commits us to the following proposals:

- The IP projection describes the first cycle: i.e. the signals evoked during the replayed

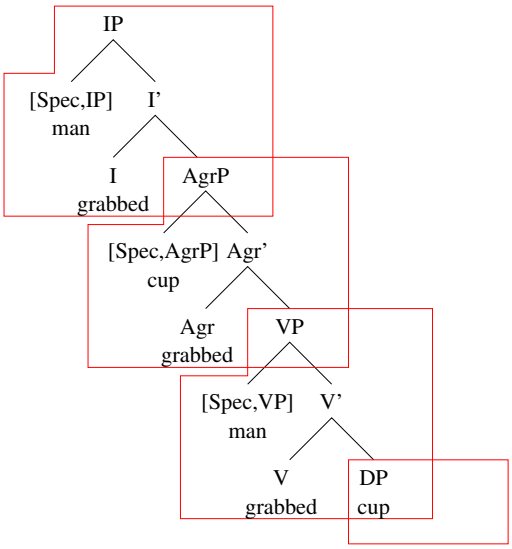


Figure 5.7: The LF structure of *The man grabbed a cup*. (The arrows denoting movement relations have been dropped from the figure, and the constituents which move have been explicitly repeated.)

action of attention to the agent.

- The AgrP projection describes the second cycle, i.e. the signals evoked during the replayed action of attention to the cup.
- The VP projection describes the third cycle, i.e. the signals evoked during the replayed 'grasp' motor programme.
- The DP projection introduced as complement of the V head describes the fourth (terminating) cycle, i.e. the signals evoked during the replayed haptic establishment of the cup.

To summarise: my proposal is that the LF of *The man grabbed a cup* can be understood as describing the sensorimotor experience of watching (or being) a man grabbing a cup—as this experience is replayed from working memory, one sensorimotor action at a time.

Now recall that Principle 2 provides a general sensorimotor interpretation for *each constituent* in an XP. The head of each XP describes one particular sensorimotor action. The XP itself describes the initial context in which the action occurs. The specifier describes the reafferent signal generated by the action. And the complement YP describes the new

context which the operation brings about. We can therefore label each constituent in the LF representation with a specific sensorimotor interpretation, as shown in Figure 5.8.

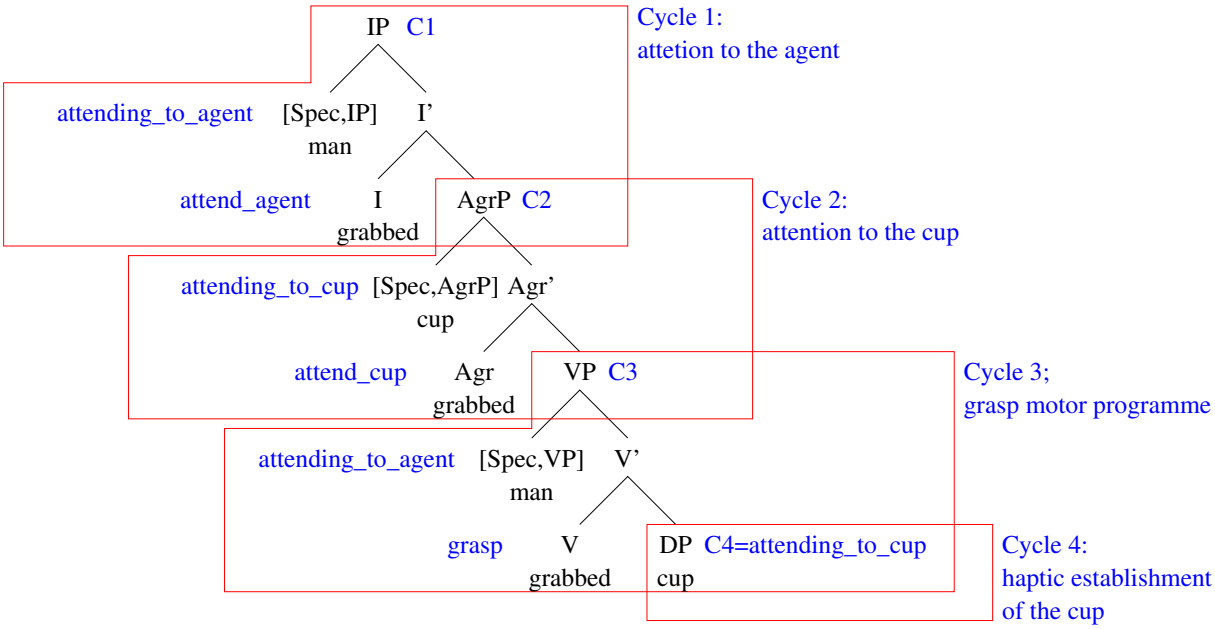


Figure 5.8: Sensorimotor interpretation of the LF structure of *The man grabbed a cup.* Sensorimotor interpretations of each constituent are given in blue, and each cycle in the replayed sequence is given in red.

To summarise the figure: the LF of *The man grabbed a cup* provides a description of the replayed cup-grabbing sequence, by specifying information about each context, about each sensorimotor action, and about each reafferent sensory signal. The functional head I describes the action of attention to the agent, and the DP which sits at its specifier describes the reafferent sensory state of attending to the agent. The functional head Agr describes the action of attention to the cup, and the DP which sits at its specifier describes the reafferent state of attending to the cup. The lexical head V describes the 'grasp' motor programme, and the DP which sits at its specifier describes the reafferent consequence of this action: reattention to the agent (but this time, as discussed in Section 2.10, as an animate agent rather than just as an object.) Finally, the DP which sits at the complement of V describes the state brought about by the completed grasp action, of reattention to the cup (but this time, as discussed in Section 2.5.4 in the haptic modality). Prima facie, the cycles in the sensorimotor sequence match up very well with the available positions in the LF structure.

IP, the projection where the subject is assigned Case, describes an action of attention to the agent. AgrP, the projection where the object is assigned Case, describes an action of attention to the cup. VP, where the verb and its arguments are introduced, describes the execution/monitoring of a motor action. Notice also that the Minimalist idea of DP movement finds a correlate in the sensorimotor idea of *reattention* to agent and patient. The subject originates at [Spec,VP] and raises to [Spec,IP]: these two positions both denote the state 'attending_to_agent'. The object originates at the VP complement and raises to [Spec,AgrP]: these two positions denote the state 'attending_to_patient'. The interpretation of X-bar structure thus shows some interesting isomorphism between the LF of our example sentence and the structure of the associated sensorimotor model. But of course we must look much deeper than these simple correspondences to give substance to a sensorimotor interpretation of LF. What does it mean to say that the I and Agr heads denote 'actions of attention'? And what sensorimotor interpretation can be given to syntactic operations like DP-movement, and head movement? I will consider these questions in the remainder of this section. More fundamentally, how does our sensorimotor interpretation of LF square with the conception of LF within syntactic theory? In Minimalism, an LF structure is created by a generative process, whose specification essentially determines the range of possible LF structures. Is there a sensorimotor interpretation of this process as well? If so, what is it? Furthermore, in Minimalism an LF structure is understood as the level of syntactic representation which 'interfaces with meaning' in the cognitive system. If we understand LF as a description of a rehearsed sensorimotor sequence, does this oblige us to adopt a particular model of sentence meaning? If so, is it a defensible one? I will consider these questions in Section 5.5.

5.4.1 I and Agr as attentional actions

In this section, I will consider in more detail what it means to say that the I and Agr heads are 'descriptions of attentional actions'. But before I begin, I will make one observation about the way our sensorimotor interpretation of the X-bar schema is being applied to the LF of our example sentence, because it makes some important assumptions about how sensorimotor information can be linguistically signalled.

According to Principle 2, the head of each X-bar schema describes a sensorimotor operation, and its specifier describes its reafferent consequence. In the interpretation illustrated in Figure 5.8, the head of VP conveys a 'sensorimotor operation' quite directly, in an open-class verb. But the heads of IP and AgrP do not carry rich open-class information. If they 'describe attentional actions', they only do so in a rather impoverished way. On the other hand, there is rich information about attentional actions in the *specifiers* of IP and AgrP—which hold DPs—while in the specifier of VP, there is no information at all about the

motor action. Applying our sensorimotor interpretation of X-bar structure to our example sentence obliges us to posit some redundancy in how the sensorimotor action in a given cycle is conveyed linguistically: it can either be conveyed directly (as is done within the VP), or indirectly via its reafferent consequences (as is done within IP and AgrP). In other words, motor actions are signalled linguistically as actions, while attentional actions are signalled linguistically by the sensory states which they result in. There is thus a division of labour within each XP as to how its associated sensorimotor operation is described, and different XPs divide the labour up in different ways.

I now turn to the characterisation of the I and Agr heads as 'descriptions of attentional actions'. The idea that the head of an XP describes a sensorimotor action is reasonably intuitive for a VP: a verb like *grab* can be naturally understood as describing an action. But the idea that 'I and Agr describe attentional actions' is not intuitive; I and Agr have very little overt phonological content, and are motivated to a large extent by their role in the operations of DP movement and head-raising, which are themselves invoked to explain a complex array of cross-linguistic distributional data. To give substance to the idea that I and Agr describe attentional actions, we have two options. We can argue that what little phonological content is possessed by I or Agr can be understood as a signal of an attentional action, within the sensorimotor model we are using. Or we can argue that interpreting I and Agr as descriptions of actions of attention contributes to a convincing sensorimotor account of DP movement and verb raising. I will consider both arguments in turn.

The concept of an attentional action was introduced in Chapter 2, and extended in Chapter 3. An attentional action involves the direction of focal attention to a particular point in external space (see Section 2.4), or an internal operation of attention 'to onesself' (see Section 2.8). It can also involve a 'top-down' activation of the object one expects or wants to see (see Section 2.2.2). An attentional action can be stored in working memory by associating a location with a top-down object representation, and it can be 'replayed' by entering a mode where top-down object representations automatically evoke their bottom-up counterparts (see Section 3.5). To motivate the claim that the I and Agr heads describe attentional operations, we must show that their associated phonological signals can be understood as reflecting operations of internal or external focal attention, together with an associated top-down object representation. Given the weakness of these phonological signals, this test is more of a reality check than a motivation. Nonetheless, the kind of information conveyed by agreement morphology does appear to convey some components of an attentional action quite well. In particular, it can signal grammatical 'person' information, indicating whether the speaker refers to himself, to his interlocutor, or to a third party or object. These differences are quite plausibly attentional in origin. Agreement morphology can also signal grammatical number. I will argue in Chapter 7 that the

distinction between 'singular' and 'plural' is also attentional in origin.[2] In summary, the information conveyed by agreement morphology is *consistent* with the proposal that I and Agr signal attentional actions. But in order to positively motivate this proposal, we need to focus on the syntactic roles of I and Agr, and show how our sensorimotor interpretations of I and Agr contribute to a convincing sensorimotor account of DP movement and verb raising.

Syntactically, I and Agr have two important roles. Firstly they assign Case to their specifiers, and are therefore central to the Minimalist account of the dual positions of subject and object. Secondly they are the positions to which the verb must raise to 'check' its subject and object agreement morphology—so they are also central to the Minimalist account of agreement inflections on verbs, and of the placement of inflected verbs. The question is, then: does interpreting I and Agr as descriptions of actions of attention contribute to a convincing sensorimotor account of DP movement and/or verb raising? I will answer 'yes' in both cases.

5.4.2 A sensorimotor account of DP movement and Case

In Minimalism, both subject and object originate in positions within the VP, and then raise to the specifiers of projections above VP. The higher and lower positions have different functions. VP-internal positions are used to assign thematic roles; the verb assigns 'agent' role to its specifier, and 'patient' role to its complement. The VP-external specifier positions are where DPs are assigned Case; the subject raises to [Spec,IP] to be assigned nominative Case by the I head, and the object raises to [Spec,AgrP] to be assigned accusative Case by the Agr head. A general principle is proposed: that 'DPs must raise to get Case'.

Is there a sensorimotor interpretation of this principles? I propose that there is—and moreover, that it makes reference to the characterisation of I and Agr as 'attentional operations'. My proposal is as follows:

Principle 4 *The principle that 'DPs must raise to get Case' can be understood as an expression of the constraint that an object must be attended to before it can participate in cognitive routines.*

[2]The case of gender is somewhat different, because it is at least partly a matter of convention. I will also briefly consider gender in Chapter 7.

The idea that objects must be attended to before they can participate in cognitive routines is at the heart of Ballard *et al.*'s (1997) deictic model of cognition. The 'cognitive routines' in our cup-grabbing example are to do with action monitoring: as was argued in Chapter 2, we cannot execute (or even classify) a grasp action until we have attended to the agent, and until we have attended to the intended target. The actions of attention which establish the parameters necessary for the cognitive routine to operate must be quite separate from the cognitive routine itself, and must occur before it is initiated. The proposal in Principle 4 is that Case-assigning heads denote the special attentional operations which must precede the initiation of a cognitive routine. The reafferent consequences of these preliminary attentional operations are sensory states in which object representations are evoked. According to our sensorimotor interpretation of of the X-bar schema (Principle 2), these states are signalled by the specifiers of Case-assigning heads, containing subject and object DPs. However, the attentional operations also update the agent's context, enabling new sensorimotor routines to operate. Again by our interpretation of the X-bar schema, these new routines are signalled by complement XPs.

If the higher positions of DPs ([Spec,IP] and [Spec,AgrP]) are understood as denoting the reafferent consequences of initial attentional actions, how should we characterise the VP-internal DP positions (i.e. [Spec,VP] and the complement of V)? In the Minimalist model, these are the positions at which the agent and patient are 'assigned thematic roles' by the verb. In the sensorimotor account, these positions denote states of *reattending* to the agent and patient. Is there anything in these states of reattention analogous to the assignment of thematic roles? I would argue that there is. Firstly, note that the operation of reattention to the agent is quite different from the operation of reattention to the patient. Reattention to the agent happens as a side-effect of an ongoing action monitoring process, while reattention to the patient happens when action monitoring is complete. Reattention to the agent is a special form of attention, constrained to apply to an animate, acting entity: this type of attention seems well suited to provide a sensorimotor basis for the mechanism which assigns the 'agent' thematic role. Secondly, note that the nature of the attentional processes which happen during and after action monitoring are *specific to the action*, unlike the processes described by IP and AgrP. This is particularly so for the patient: the manner in which the target is reattended to by the establishment of a stable grasp is quite specific to the grasp routine. Obviously it would be premature to frame a generalising sensorimotor interpretation of thematic roles after only having considered one action. But we can say somewhat tentatively:

Principle 5 *The principle that 'a verb assigns thematic roles to its specifier and complement' can be understood as a statement of the fact that cognitive routines involve reattention to objects in new (and routine-specific) modalities.*

Finally, is there a sensorimotor interpretation of DP movement itself? Recall from Chapter 4 that the movement of a DP creates an object called an 'A-chain', which spans all the positions involved in the movement operation, and that Case and thematic roles are in fact assigned to A-chains rather than to individual positions. My suggestion here is that the mechanism of DP movement captures something about the role of action monitoring in the formation of multimodal object representations. As discussed in Chapter 2, during experience of a reach-to-grasp action the agent and the patient are both attended to twice, in different modalities. Thus the attentional states denoted by the two DP positions in an A-chain are representations of a single object in two different modalities. In our analysis, [Spec,IP] holds a representation of the agent 'as an object', while [Spec,VP] holds a representation of the agent 'as an agent'. The former representation might primarily involve shape, colour, size and so on, while the latter might primarily involve motion, as well as intentional representations in PFC. As discussed in Section 2.7.6.2, our representation of 'animate agents' must incorporate elements from both these modalities. Likewise, [Spec,AgrP] holds a visual representation of the cup, while the complement of V holds a representation of the cup as a goal motor state, i.e. of the affordances of the cup. As discussed in Section 2.5.4, our representations of 'manipulable objects' must incorporate visual form, and also motor affordances. A successful grasp action provides some special opportunities for *learning* the required crossmodal representations. When we are monitoring an action, we are generating a representation of an animate agent which can be understood *axiomatically* as the same object we established (as an object) in our first action of attention. When we have a stable grasp, we are generating a haptic representation of an object which is *axiomatically* the same object we established in our second action of attention. Again it is too early to speculate in general about a sensorimotor interpretation of DP-movement. But we might tentatively suggest:

Principle 6 *The movement of a DP from a VP-internal position to the specifier of a Case-assigning head can be understood as a reflection of a process which creates a multi-modal object representation.*

I have stated Principles 4–6 quite generally, to allow for their extension beyond the current reach-to-grasp example. Of course, the principles have only been motivated for this one example; whether they extend to others is a matter for further investigation. This will be one of the questions I will return to in the chapters which follow.

5.4.3 A sensorimotor interpretation of head movement

In the Minimalist account of LF derivation, a fully inflected verb is generated at the V head, and it must move to the Agr and I heads in turn to 'check' the semantic features associated with these inflections. Can we give a sensorimotor interpretation of this movement? And if so, does it make reference to the idea that the I and Agr heads are attentional operations?

At first sight, this type of movement is very puzzling. We are assuming that I and Agr denote actions of attention to the agent and patient, and that V denotes a motor action. Why can the inflections conveying these actions of attention appear together with the verb at the head of VP? And why can the verb (plus its inflections) appear at the head of IP? These types of movement involve sensorimotor actions being reported *out of sequence*. If V 'raises to' I, then information about the motor action is being conveyed 'prospectively', in anticipation of the time it actually occurs. If the inflections on the verb are present at V, then information about the attentional actions is being conveyed 'retrospectively', after they actually occur.

In fact, there is a good interpretation of these out-of-sequence appearances of I, Agr and V. And it does indeed make reference to the fact that all these heads denote sensorimotor actions. We simply need to assume that the linguistic associations of sensorimotor actions are with *planned* actions, rather than with transitory action signals. Recall from Section 5.1 that the cup-grabbing event is stored as a planned sensorimotor sequence in PFC. The PFC-based plan representation features representations of each of the component actions; they are active in parallel, and remain tonically active throughout rehearsal of the sequence. This further assumption can be stated as a refinement of Principle 2:

Principle 7 *The head of an XP denotes a* planned *sensorimotor action, rather than a transitory action signal. It is a reference to the currently active plan representation in PFC.*

The basic idea is that linguistic information about sensorimotor actions is 'read from' plan representations in the PFC, rather than from action signals as they appear in sequence. In Chapter 6, I will present some good evidence for this proposal: it does seem that inflected

verbs are associated with PFC rather than with brain regions evoking direct attentional or motor signals. The details are given in Section 6.1.4.2.

We can understand Principle 7 as suggesting that evolution happened to find a means of generating linguistic signals of sensorimotor actions by signalling actions as they are planned, rather than as they occur. Of course this introduces some further indirection into the way linguistic expressions signal sensorimotor operations. Each reference to an action is in fact a reference to *all* the actions in the sensorimotor sequence currently being rehearsed—so information about the order of these actions is lost. But note also that there are other sources of information about the order of these actions. Although the actions themselves do not have direct linguistic reflexes, their reafferent consequences do—at least in the case of attentional actions (as we saw in Section 5.4.1). In addition, there are strong general constraints on the ordering of motor and attentional actions: motor actions cannot occur without appropriate attentional actions. These sources of information can compensate for the ambiguity introduced by using tonically active plan representations to convey sensorimotor actions.

Principle 7 requires us to revise the sensorimotor interpretation of our example sentence. The head of each XP now denotes a *set* of sensorimotor actions, comprising each of the actions in the action sequence currently being rehearsed, as shown in Figure 5.9. Now note

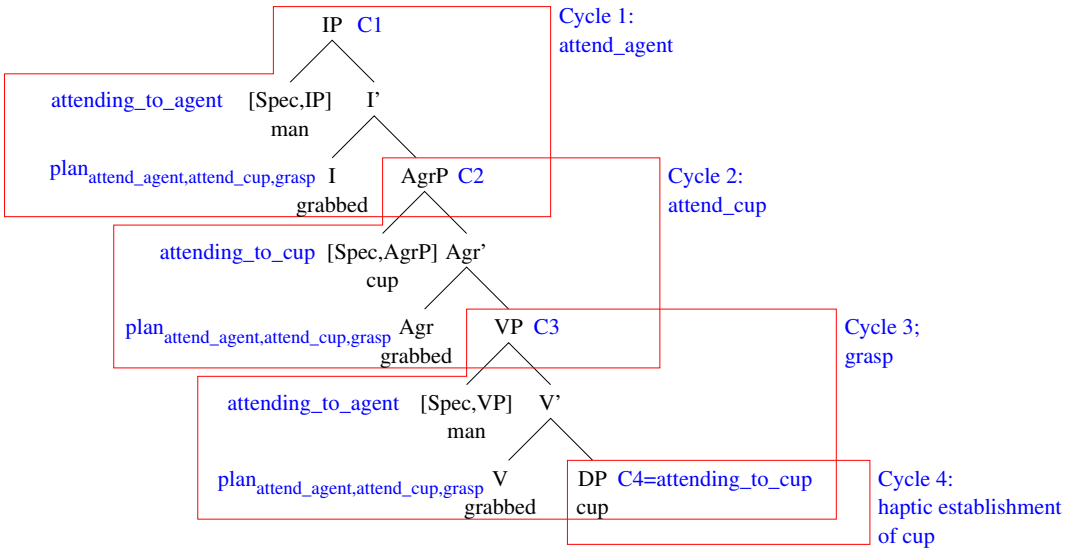


Figure 5.9: Revised sensorimotor interpretation of the LF structure of *The man grabbed a cup*, with XP heads denoting planned sensorimotor actions.

that this revised interpretation of LF provides a promising basis for a sensorimotor account of head movement. A linguistic signal combining information about the motor action and the two attentional actions can be read out at each iteration of the replayed sensorimotor sequence. A discussion of how an agent decides *when* to read out this information will be given in Chapter 6, which gives an account of how the mapping from LF to 'phonetic form' is learned. (The details are in Section 6.4.3.) In the meantime, we can give a partial sensorimotor interpretation of head raising:

Principle 8 *The non-locality of information which is captured by the mechanism of head raising is a reflection of the fact that the heads denote sensorimotor actions indirectly, by reference to the currently planned sequence of actions being rehearsed.*

There are some interesting additional syntactic corollaries of this principle. For one thing, recall that head raising is always 'successive cyclic': a head always raises to the immediately higher head; intermediate heads are never skipped. This falls out directly from the fact that a planned sensorimotor action sequence is active throughout the rehearsal of the sequence. In addition, note that the domain over which a head raises can be understood as a reflection of the temporal period during which a particular sensorimotor sequence plan is active and is rehearsed. In complex sentences, there are often 'barriers' to head raising—for instance, heads do not appear to raise across clause boundaries. Principle 8 may force us to posit mechanisms for changing the *sequence plan* from which a sensorimotor sequence is replayed, so that different domains of a sentence can correspond to times during which different plans are being replayed. I will consider a few such mechanisms in Section 5.6.

5.5 The role of LF revisited

When proposing a sensorimotor interpretation of LF, it is obviously important to keep in mind what the role of LF is within syntactic theory, and to ask whether LF still fulfils this role under the sensorimotor interpretation. In Minimalism, there are two ways of thinking about LF: one is as a product of the 'generative process', which produces all (and only) the well-formed sentences in a given language; and one is as 'the syntactic representation which interfaces with semantics'. In this section, I will consider whether the sensorimotor interpretation of LF is suitable to allow it to play these roles.

5.5.1 A sensorimotor interpretation of the generative process

To begin with: what is the sensorimotor interpretation of 'the process which generates an LF representation'? A description of the generative process is central to the Minimalist account: constraints on possible LF structures are typically stated as constraints governing the construction of an LF, rather than as static constraints on completed LF structures. At the same time, the generative process is not to be understood as a description of sentence processing. The end goal is simply to specify a way of dividing an infinite set of possible word sequences into well-formed and ill-formed sentences. If we are looking for a sensorimotor interpretation of LF, we must also give an interpretation of 'the generative mechanism' which creates LF representations—and we must explain why we expect this mechanism to generate all (and only) the LF representations of well-formed sentences.

If LF describes a replayed sensorimotor sequence, then 'the generative mechanism' must be in part the mechanism which produces sensorimotor sequences, and in part the mechanism which stores and replays them. Describing 'the mechanism which produces sensorimotor sequences' is obviously a tall order: it involves describing all the possible ways in which we can interact with the world, which is partly a function of our sensorimotor capabilities, and partly a function of the way the world is. But of course we do not need to provide a complete description: what we must do is to express the *constraints* on the structure of our sensorimotor experience of the world which manifest themselves in language. In the model I am envisaging, the primary constraint is that we must experience the world sequentially. But there are then many additional constraints on the structure of sensorimotor experiences. One general one is the idea expressed in Principle 4, that objects have to be attended to before they can participate in cognitive routines, is one such constraint. A more specific one relates to the cup-grabbing scenario: the fact that the observer has to attend to the agent before he can attend to the patient. I suggest that the mechanism which generates LF structures can be thought of in large part as a specification of the constraints on the ordering of sensorimotor operations—expressed at the level of granularity at which these operations are signalled in natural language, naturally.

However, there is another component to the generative mechanism: it must also be thought of as a description of the cognitive routines involved in storing sensorimotor sequences in working memory and in replaying them. Several aspects of these routines might be relevant. For instance, as proposed in Principle 8, the way LF describes a replayed sensorimotor sequence appears to reflect the fact that the actions in a prepared sequence are individually identifiable and active in parallel. But there are many other aspects we have not yet considered. In particular, there are issues to do with the *units* in which sensorimotor sequences are stored and replayed. The points at which an agent chooses to interrupt sensorimotor experience, replay the sequence currently buffered in working mem-

ory and then reset the buffer, do not seem at all arbitrary. In language, they correspond to sentence boundaries—or to clause boundaries, or other major phrase boundaries. Later in this chapter (in Section 5.6) I will consider some examples of complex sentences where the question of how replayed sensorimotor sequences are organised into units is an important one. The question will also arise in the account of noun phrases I give in Chapters 7 and 8. In the meantime, my basic proposal about the generative mechanism can be summed up as follows:

Principle 9 *The mechanism which generates LF structures can be understood as a specification of the constraints on the order in which sensorimotor operations can be applied, and a description of the working memory mechanism which stores and replays sensorimotor sequences.*

This reinterpretation of the generative mechanism is a draconian one. In Minimalism, the generative mechanism is a tightly defined algorithmic process which creates LF representations *from the bottom up*, starting with a VP and progressively adding the projections which introduce this phrase. It is within this framework that 'movement' operations are defined. My interpretation of the generative mechanism does not have this right-to-left order. (If anything, constraints on the order of sensorimotor operations are better understood as applying forwards in time, with one operation setting up the context for the operations which occur next.) Neither, consequently, does my model have an explicit notion of 'movement'. But it must be remembered that the generative mechanism is a formal device for generating a space of LF representations. Constructing LFs bottom-up and having elements raise to higher positions is a way of accounting elegantly for the nonlocal syntactic relations in a sentence. But these relations can also be interpreted as reflections of operations of reattention, and of the presence of planned sequence representations in working memory, as described in Principles 1–8, with the generative mechanism understood as described in Principle 9. So while there is no sensorimotor interpretation of Minimalism's bottom-up generative mechanism *as a procedural mechanism*, I want to preserve the basic insights which are expressed within Minimalism's procedural framework. To be quite clear: I do not want to 'deny' that there are operations like DP-movement or head raising—I just want to *reinterpret* these operations. Of course, it remains to be seen whether all of the principles which are formulated within Minimalism's procedural generative framework can be reinterpreted; it is likely that some of them will not survive. On the other hand, my reinterpretation of the generative process has the advantage that it makes reference to an actual *cognitive process*—the process of internally rehearsing an episode-denoting sensorimotor sequence. This means there is scope for it to feature in a psychological account of

sentence processing, as well as in a declarative model of grammatical competence, as I will discuss in detail in Chapter 6.

5.5.2 LF as a representation of sentence meaning

As mentioned in Section 5.5.1, LF is also accorded a semantic role in the Minimalist model: it is the syntactic representation of a sentence which conveys its meaning to the nonlinguistic cognitive system. If we are to give a sensorimotor interpretation of LF, we must also explain how it is that replayed sensorimotor sequences can convey meanings.

At least in relation to concrete sentences, we are in a good position here. Take a concrete sentence S, which describes an event or state that can be apprehended through sensorimotor experience. My proposal is the LF of S can be understood essentially as a *simulation* of this experience. This simulation will activate rich sensorimotor representations of objects, actions and intentions, which interface to the cognitive system in the same ways that the sensorimotor experience itself does. The account of sentence meaning which emerges from a sensorimotor interpretation of LF is exactly what is demanded by theories of 'embodied cognition', which protest about the impoverishedness of logical formulae as representations of meanings, and the difficulty of expressing world knowledge in logical axioms (see e.g. Harnad, 1990; Brooks, 1991; Clark, 1997; Barsalou, 2008), and also by theories which argue that semantic representation involves simulation (see e.g. Gallese and Goldman, 1998; Jeannerod, 2001; Grèzes and Decety, 2001; Barsalou *et al.*, 2003; Feldman and Narayanan, 2004).

Note also that interpreting LF as a description of sensorimotor processing forces a *dynamic* conception of sentence meaning. There is a strong movement in natural language semantics arguing that utterances should be understood as describing the epistemic operations of an agent discovering information about the world, rather than as direct descriptions of the world itself (see in particular Heim, 1982; Kamp and Reyle, 1993). A sensorimotor interpretation of LF yields a conception of sentence meaning which is quite similar. In fact, I will examine these similarities in some detail in Chapters 7 and 8.

The question of what account of *abstract* sentences is afforded by a sensorimotor interpretation of LF is, of course, unanswered. As discussed at the start of the book in Section 1.2.1, I will only be considering concrete sentences in this book, in the hope that a sensorimotor interpretation of these will later be useful in suggesting hypotheses about analogous mechanisms underlying abstract sentences.

5.6 Predictions of the sensorimotor account of LF: looking at some other syntactic constructions

In this chapter, I have given some arguments to support a sensorimotor interpretation of LF as a description of sequencing structures in the PFC. So far, only simple finite transitive clauses have been considered. In subsequent chapters I will consider how the account can be extended to a range of other syntactic constructions, but I will conclude the current chapter by looking briefly at a few constructions which the sensorimotor interpretation appears to make some testable predictions about. In Sections 5.6.1 and 5.6.2 I will consider some predictions about the syntax of nonfinite and finite clausal complements. (These sections will also provide some useful background for the model of the LF-PF interface to be developed in Chapter 6.) In Section 5.6.3 I will consider some predictions about the syntax of questions. Semantically, all of these constructions go some way beyond simply reporting sensorimotor experiences: to account for them, we need to draw on the models of working memory and long-term memory which were outlined in Chapter 3. Since the sensorimotor interpretation of LF also makes reference to these models, it generates predictions about the syntax of the associated constructions which we can square against the facts.

5.6.1 Control constructions

The account of DP-movement which I gave in Section 4.6 referred to a sentence with an infinitive clausal complement:

(5.1) John seems [to walk].

I argued that at LF, the subject of this sentence (*John*) appears in the complement clause as the subject of *walk*, but has to move to the higher subject position to get Case. Movement is possible, because the higher verb *seem* has an empty subject.

There are in fact several types of infinitive clausal complement. In this section I will focus on a different kind, exemplified in Example 5.2.

(5.2) The man wants [to grab the cup].

Here again, there are reasons to suggest that the subject (*the man*) should appear in the complement clause at LF, as the subject of *grab*, because the man is the agent of the verb *grab*. However, in this case, the man is also the agent of the higher verb *want*, so it does not seem right to speak of the lower subject 'moving' up to an unoccupied position. To illustrate, note that only Example 5.1 has an alternation involving an expletive subject:

(5.3) It seems John walks.

(5.4) *It wants the man grabs a cup.

In generative grammar, it is proposed that verbs like *want* introduce an infinitive clause whose subject, rather than moving to a higher position, is simply not pronounced. The unpronounced subject is designated by the symbol 'pro'. Its appearance is governed by some special constraints: pro can only appear in subject position in a complement clause introduced by a certain type of verb, and it has to corefer with a noun phrase in the higher clause. The verbs which license clausal complements with pro subjects are termed **control verbs**, and the higher coreferential noun phrase is said to **control** the pro subject. The relationship between the two noun phrases is conventionally indicated by coindexing, as shown below.

(5.5) The man$_i$ wants [pro$_i$ to grab the cup].

Note that pro can only occur as the subject of an infinitive clause:

(5.6) *The man$_i$ wants [?$_i$ grabs the cup].

So it is likely there is a special circumstance which leads both to nonfiniteness and to the use of a pro subject.

Is it possible to give a sensorimotor interpretation of control constructions? Control verbs introduce clausal complements describing actions which are in some sense 'unrealised'. For instance, *want* introduces a clause describing an action which is intended, but not yet achieved. Intentions feature prominently in my sensorimotor account, so there would seem to be scope for a discussion of such sentences. In this section I will outline a suggestion about the intentional representations and operations which might be 'described' by the LF of a control construction.

In my account of planned action sequences, a plan to perform a sequence of actions is active in PFC in its entirety while the individual actions in the sequence are carried out. In our example cup-grasping scenario, the actions are in fact all carried out. However, I will now consider a situation where an agent *intends* to grasp a cup, but for some reason has not yet executed this action.

I have already suggested that there is a special cognitive operation which takes a PFC sequence plan and internally rehearses it (see the account of 'simulation mode' in Section 3.5). In that earlier section, the operation was invoked automatically, after a successfully completed action sequence (and functioned among other things to create a trace of the sequence in episodic memory). However, it may be that the rehearsal operation can also

be initiated *overtly*, as a matter of choice, to achieve the effect of 'reading out' the as-yet-unrealised intentions of an agent being attended to. There may be several reasons why we might want to make an observed agent's intentions explicit, the most obvious being in order to express them linguistically.

Before an observer can represent an agent's intentions in PFC, he must attend to the agent. Recall from Section 2.7.6.2 that after an agent has been attended to, the observer's PFC will evoke a space of plans reflecting the predispositions of that agent. (If the observer 'attends to himself', i.e. prepares to act, his PFC will evoke his own plans.) Recall also that PFC represents a number of different alternative plans, which compete with one another, so that a single dominant plan is selected (as discussed in Section 3.3). So in order to rehearse an agent's dominant intention, the observer must first attend to the agent, and then overtly initiate a rehearsal operation, to simulate the dominant plan before it has actually been executed.

There is a complication to the above process. If the operation of 'attending to an agent and then rehearsing his dominant plan' is an overt one, it must *itself* be planned by the observer. Of course the observer does not want to rehearse this 'meta-level' plan! So once the plan-rehearsal operation has been performed, the observer's currently dominant meta-level plan must be *inhibited*, to allow the observed agent's substantive plan to become dominant. After this operation, the plan which the observer actually rehearses will be the observed agent's dominant plan, rather than his own meta-level plan. The idea of plan inhibition was motivated in some detail in Section 3.3.3, in connection with the phenomenon of backward inhibition (Mayr and Keele, 2000; Shima *et al.*, 1996; Rushworth *et al.*, 2004). I now suggest that plan inhibition plays an important role in the cognitive mechanism which allows an observer to overtly simulate an observed agent's intentions.[3] If this is the case, we expect that the LF representation of a sentence like Example 5.5 will make reference to a plan-inhibition operation.

I suggest that a mental state verb like *want* has a special sensorimotor denotation: rather than denoting a motor programme, like *grab*, it denotes an internal cognitive operation, which has the effect of inhibiting the currently dominant plan, and moving deliberately into simulation mode.

The interesting thing about this proposal is that it also appears to explain the *syntactic* idiosyncracies of sentences featuring the verb *want*—i.e. of control constructions. I suggest

[3]In support of this idea, it is interesting to note that two of the areas asssociated with plan inhibition and plan competition, the SMA region and the lateral PFC region (Section 3.3) are also frequently activated in tasks where subjects represent the mental states of other agents—see e.g. the review in Carrington and Bailey, 2009. I will make use of the idea of plan-inhibition to account for several operations evoking mental states—see Section 5.6.2 for a (brief) account of beliefs, and Section 6.3.3 for a (more detailed) account of communicative actions.

that the complement clause introduced by *want* (or similar control verbs) has the form it does because of the special conditions under which rehearsal occurs when it is initiated in this way.

There are two things to explain: the first is that the clausal complement of *want* is nonfinite in form (i.e. its verb is uninflected—*to grab* rather than *grabs*); the second is that it has a *pro* subject. Both of these properties have natural explanations on the 'plan-inhibition' model just outlined.

Firstly, consider the fact that the verb of the complement clause is uninflected—specifically, it is missing an inflection agreeing with the subject.[4] On my analysis, the subject agreement inflection on a verb appears because the verb and its inflection are read from the plan representation which drives the replay of a sensorimotor sequence, which is tonically active throughout replay (see Section 5.4.3). But note that in the scenario for explicitly rehearsing an agent's intention which I have just described, the observer must *inhibit* the plan which allowed attention to the agent before he can rehearse the intention. In this scenario, the planned action of attention to the agent which normally produces subject-agreement morphology on the verb is not active when an agent's intention is rehearsed. We therefore predict that the verb in the complement clause will have no subject agreement inflection.[5]

Now consider the fact that the complement clause has no overt subject. Again this is predicted by the meta-plan-inhibition scenario just presented. When the observer attends to the agent, he establishes a particular cognitive context, in which he is attending to an agent, and evoking this agent's plans in PFC. This is the context in which he will rehearse the agent's dominant plan. But note that it is also the reafferent state which results from executing the first operation in the plan to be rehearsed. So when the rehearsal process is initiated, it will start from the second step, skipping the operation of attention to the agent. We therefore predict that the complement clause will have no overt subject.

In fact, the plan-based account of 'want' sentences gives a good explanation of the special dependence of the unpronounced pro subject on a noun phrase from the higher clause in control constructions. The reafferent state which obtains at the start of the

[4]It also lacks a tense inflection, but I am just considering agreement inflections for now.

[5]Note that there is no corresponding prediction about object agreement inflections. In the above account there is nothing to suggest that object inflections will be lost in the complements of control verbs. And indeed, in languages with object inflections, object agreement can be retained in 'nonfinite' clausal complements of control verbs. Urdu is a case in point. In Urdu, verbs carry object agreement morphology in finite sentences and also in complement clauses (see e.g. Butt, 1995):

	laRkooN-nee	[	pro	yah	kitaab	paRh-nii	] caah-ii
(5.7)	boys			this	book.fem.sg	read-inf.fem.sg	want-pf-fem.sg
	'The boys wanted to read this book'						

rehearsed plan is brought about by an earlier action of attention, that determines which agent is attended to, and thus which plans are evoked in PFC. The pro subject is thus structurally constrained to co-refer with a 'controlling' DP.

The above suggestions about control constructions are fairly preliminary, but they seem quite promising as an account of the intentional representations and mechanisms which underlie control constructions. My reason for mentioning them here is that they corroborate the idea that plan-rehearsal operations underlie the syntactic phenomenon of verb raising. Control constructions are interesting because their *semantics* requires reference to an agent's planning or intentional representations. I proposed a 'semantic' account of how an observer can gain access to an observed agent's plans, which is based on the sensorimotor model introduced in Chapters 2 and 3, involving notions of planned sensorimotor sequences, plan rehearsal, plan competition, and plan inhibition. This semantic account is quite separate from the syntactic account I gave of the role of plan rehearsal in verb movement and verb inflections in simple finite sentences. But the two accounts together make exactly the right predictions about the syntax of control constructions.

5.6.2 Finite clausal complements

Want is not the only type of 'mental state' verb. There are also mental state verbs which have finite clausal complements—for instance *believe* and *think*. A finite clausal complement is illustrated in Example 5.8.

(5.8) The man thinks [that Bill grabbed the cup].

I suggest that this type of mental state verb also involves an explicitly scheduled plan-inhibition operation. The observer must first attend to an agent—in the above case, 'the man'—and then perform an operation which inhibits the currently active plan, and explicitly enters rehearsal mode, to allow another plan-based representation to be read out. In this case, the second 'plan representation' is not literally one of the attended agent's plans. We do not want to evoke the state of the agent's PFC when he is planning an action, but rather when he is entertaining some particular fact or episode in working memory. The observer must therefore perform some additional cognitive operation which causes his PFC to simulate the contents of the observed agent's working memory, and select the WM episode which is most plausibly the dominant one. This operation might take different forms for different verbs. For a verb of perception like *see* or *notice*, it could involve establishment of joint attention with the agent, followed by regular episode perception. For *think*, it is likely to involve some form of inference, as well as perceptual or memory-related processes. I will not consider the operation in any detail—I just want

to make a simple point about the syntax of finite clausal complements introduced by verbs of this kind.

Note that in a finite clausal complement, the clause boundary, roughly coincident with the word *that*, constitutes a barrier for verb raising: the verb in the embedded clause does not raise beyond this boundary. This is a syntactic observation. But again, it is predicted by the plan-based model of how mental states are evoked. According to the model just presented, the observer executes a meta-level plan to attend to an agent and then evoke (and rehearse) the WM episode the agent is most likely to be entertaining. The meta-level plan must be inhibited before the new WM episode can be evoked. If the account of head-raising given in Section 5.4.3 is correct, we therefore predict that there are two domains of verb movement in the full sentence describing the complete sequence of operations executed by the observer. There is a domain in the matrix clause, where the higher verb (*think*) raises to agree with the higher subject (*the man*), and a domain in the complement clause, where the lower verb (*grab*) raises to agree with the lower subject (*Bill*). But there will be no raising of heads across the clause boundary. This is indeed what the sentence shows.

Note also that the set of 'mental state' verbs taking a finite clausal complement includes the verb *say*. I will make further use of the idea of a self-inhibit meta-plan in an account of how communicative actions are represented, and of what role they play in language acquisition, in Section 6.3.3.

5.6.3 Questions and V-to-C raising

Recall from Section 4.5 that the LF of a question involves raising of material from the head of I into the head of a higher projection C. In English, auxiliaries appear at I in an indicative sentence:

(5.9) John has grabbed the cup.

The raising of I to C explains why auxiliaries switch position with the subject in questions:

(5.10) Has John grabbed the cup?

In French, the inflected V raises overtly to I in indicative sentences:

(5.11) Jean prend la tasse. (Jean takes the cup)

In questions, V-to-C raising causes the full inflected verb to move beyond the subject:

(5.12) Prend Jean la tasse? (Takes John the cup?)

Can we find an interpretation for V-to-C raising in the sensorimotor model? The account of questions as involving V-to-C raising is an important part of the justification for the existence of an IP projection, and for the general Minimalist idea that there is an operation which allows material to 'raise' from one head to the immediately higher head. If head-raising is to be explained by thinking of heads as references to planning representations tonically activated during a replayed sensorimotor sequence, then we expect an account of the differences between propositions and questions to make reference to differences in the way sensorimotor sequences are rehearsed in these two cases.

To begin with: if LF is to be understood as a description of a cognitive mechanism, what kind of mechanism is being described by the LF of a question? Recall that our example sentence *The man grabbed a cup*, which is indicative, is thought of as describing a replayed piece of sensorimotor experience. The original, pre-linguistic function of this replay operation is to encode an experienced episode in hippocampal episodic memory (see Section 3.8.1). But note that there are also special cognitive operations involved in the *querying* of episodic memory, as discussed in Section 3.8.2. I assume that the LF of a question describes a cognitive operation which queries long-term memory, in the same way that the LF of an indicative sentence describes a cognitive operation storing something in long-term memory. I further assume that the fact that I 'raises to C' in questions but not in indicative sentences reflects differences in these two memory-access operations. On this assumption, I-to-C raising is not an arbitary operation, somehow encoded in a universal language faculty or present by convention in certain languages, but a reflection of differences in the mechanisms involved in storing and in accessing information from memory.

There are two relevant aspects of the model of memory retrieval given in Section 3.8.2. In Section 3.8.2.2 I suggested that a full cue to episodic memory (of the kind that is involved in 'recognition', rather than recall) takes the form of a sensorimotor sequence—its form is basically the same as a WM episode. And in Section 3.8.2.3 I suggested that the sequence must be presented to the hippocampus in a special mode, in which it functions as a pattern to be matched, rather than as the representation of a new event to be learned. I suggested that the hippocampus can be accessed in two modes: 'cue presentation mode' (the mode in which a query sequence is presented) and 'episode encoding mode' (the mode in which a new episode is recorded). I now suggest that the CP which dominates the LF of a question describes the operation which puts the hippocampus into cue presentation mode. Specifically: the head of CP describes this operation, while its specifier describes the reafferent side-effect of the operation, and its complement IP describes the new cognitive context which the operation brings about.

Now consider in what kind of context the operation which establishes cue presentation mode is likely to arise. When we ask a question, it seems unlikely that we first decide

to ask a question and only afterwards make a decision about the content of the question. It seems more likely that we entertain a proposition and then decide to find out whether this proposition is contained in our memory. Questions are planned actions, just like motor actions. An agent must decide to ask a question, in the same way that he decides to perform an action. (Presumably the decision to ask a question of a certain form is one which is beneficial to the agent in certain circumstances, and is reinforced when it is made in these circumstances.) Now note that a planned question will be executed as a planned sequence of sensorimotor operations, the first of which will be the operation which establishes cue presentation mode, and the remainder of which will be the operations which generate the cue sequence itself. In this case, we expect information about the operations in the cue sequence to be present in the planning representation which supports execution of the query sequence, *even at the time the first mode-setting operation is ececuted*. In the linguistic interface, therefore, we expect the information which is manifest at the V, Agr and I heads of the complement IP to also be manifest at C. In other words, we allow the possibility that information associated with I, Agr and V can be pronounced at C. (In French, it is all pronounced here; in English, only some of it is.) Thus if we accept the sensorimotor interpretation of head-raising, the raising of I to C in questions falls out quite neatly as a direct consequence of the way queries to episodic memory are formed.

Note that things are different for indicative propositions like *The man grabbed the cup*. Syntacticians often assume there is a 'silent' CP introducing IP in this case too. We might imagine that this CP denotes the operation of putting the hippocampus into 'episode presentation mode'. (This would permit a simple general account of the kinds of operation which can be described at CP: they are all to do with operations configuring the mode in which a sequence is to be presented to long-term memory.) Of course, an agent can pursue a planned action to achieve a particular episode, or can decide to observe an episode and have strong expectations about what it will contain. However, it does not make sense for him to *update episodic memory* in these increments, because things may not turn out the way he hopes or expects. Instead he must first decide to interface with the perceptual world (and move into action exectution or observation mode), and only some time afterwards replay an event representation into long-term memory. At the time the former decision is made, the agent has effectively decided to encode a new episode in long-term memory—but he does not yet know what its content is. We do not therefore expect a planned sensorimotor sequence representing the event to be present at the time this first decision is made. We therefore predict that heads within the IP of an indicative clause do not raise to the C head—as is in fact the case.

5.7 Summary

In this section, I have introduced the main proposal in the book: that the LF of our example concrete sentence *The man grabbed a cup* can be understood as a description of the sensorimotor sequence involved in experiencing a cup-grabbing episode, as replayed from working memory (Principle 1). The proposal was introduced as an instance of some very general hypotheses about how syntactic structures and operations can be interpreted. There was a general proposal that an XP schema describes a single action in a replayed sensorimotor sequence (Principle 2), a proposal that a right-branching structure of XP schemas describes a sequence of consecutive actions in a replayed sequence (Principle 3), a proposal that the raising of DPs to Case-assigning positions reflects the fact that objects must be attended to before they can participate in cognitive routines (Principle 4), a proposal that DP-movement reflects the construction of multimodal object representations (Principle 6) and a proposal that head raising reflects the units in which a replayed sequence is stored as a plan (Principles 7 and 8). Each of these proposals was supported by arguments relating to our chosen example episode, pointing to similarities between the LF of the sentence reporting the episode and the structure of sensorimotor processes involved in experiencing it and representing it in memory. Together, these reveal a fairly deep isomorphism between the LF structure and the associated sensorimotor processes.

Of course, the proposals make very general claims, which must be tested by looking in detail at the roles of LF within Minimalism, to see whether the sensorimotor interpretation allows it to play these roles, and by looking at a large range of other linguistic examples, to see if the proposed sensorimotor interpretation holds up in these cases. A discussion of the former issues was begun in Section 5.5, and a few predictions of the sensorimotor interpretation were examined in Section 5.6. In each case, there were some preliminary indications supporting the sensorimotor interpretation of LF. However, the main purpose of this chapter has simply been to present the interpretation. Much of the rest of the book will be given over to examining how well it extends to other constructions.

Before I turn to this question, however, I want to elaborate on the sensorimotor account of LF presented in the current chapter, by considering what implications it has for an account of 'surface' linguistic representations—i.e. of phonetic form. In Chapter 6 I will consider the sensorimotor interpretation of LF in the light of current ideas about how language is implemented in the brain, and put forward an account of how an infant learns to 'read out' a sequence of words from a sensorimotor sequence.

Chapter 6

A model of surface language, and of the LF-PF mapping

I have not yet said anything about how specifically linguistic representations are implemented in the brain. This question has of course been intensively studied, using a variety of empirical techniques. Experiments call on human subjects to produce or interpret language—perhaps single words, perhaps larger units—and investigate the mental processes involved in a variety of ways. These might involve analyses of subjects' behaviour during controlled tasks, or imaging of brain activity, or studies of neurological patients with language impairments. Experiments can also investigate changes in linguistic behaviour over time, particularly in children as they acquire language. Data gained from these empirical paradigms are used to build models of the mental processes and representations involved in producing or generating language. I will focus on computational models—in particular, connectionist models.

In this chapter I will first survey what is known about where and how the brain represents 'surface language'—i.e. words and word sequences. I will then outline some of the well-known models of how language develops in infants. Many of these adopt an **empiricist** standpoint, arguing that language is acquired through general-purpose learning mechanisms, rather than via the innate language-specific mechanisms favoured by Minimalists. I will summarise the debate between nativists and empiricists, and describe the models of language which are proposed by empiricists. In the second half of the chapter I will present a computational model of how children learn the meanings of words, and a computational model of how children learn syntactic constructions. My intention is that this latter model can be understood by Minimalist linguists as a model of how children learn 'the LF-PF mapping' in a particular language—i.e. how they learn conventions about how word sequences should be be 'read off' LF representations in that language—but that it is

also recognisable as an empiricist model of language processing and language development.

Of course, a Minimalist model of the LF-PF mapping (in any given language) is not a 'processing model'. It is a model of the conventions which a mature speaker of that language has learned about how to read phonetic representations off LF structures. However, if we model the *mechanism* through which this learning happens, rather than just the knowledge which is learned, we must be concerned with processing. Learning happens when a child registers utterances made by mature speakers, and manifests itself in a child's ability to generate sentences conforming to the learned conventions. The processing which is done by a child will initially be very unlike that of a mature speaker, of course, but as the child develops, it will increasingly resemble that of a mature speaker. The same is true for the connectionist system I will present: its ability to generate well-formed sentences will develop gradually as it is exposed to training data. Once the system's learning is complete, it should be possible to analyse its behaviour and give a declarative description of 'what it has learned', abstracting away from processing details. This description should be understandable by Minimalist linguists as a description of a set of conventions about how to read phonetic representations out from a logical form.

In Section 6.1 I will discuss the neural substrates of phonological representations, of words, and of 'syntactic processing'. In Section 6.2, I will discuss the process of language development in children, reviewing some of the important models of the acquisition of words and of syntactic structures. This section will also contain an introduction to the empiricist tradition in linguistics, including overviews of the main theoretical tools associated with this tradition: statistical language models, construction grammar formalisms and Elman's simple recurrent network architecture. In the remainder of the chapter, I will present a model of language development which interfaces with the model of sensorimotor cognition introduced in Chapters 2 and 3, and with the sensorimotor interpretation of syntax just given in Chapter 5. In Section 6.3 I outline a model of how children learn the meanings of individual (concrete) words. And in Section 6.4 I present a model of how children learn to associate whole working-memory episode representations with *sequences* of words. The latter model is basically a connectionist model of sentence generation. It incorporates some familiar empiricist devices, in particular an Elman network. Its main novel feature is that the semantic representation which provides the input to the generation process is structured as a sequence, just like the phonetic representation which is output by the generation process. I argue that thinking of an episode representation as a sequence (as I am proposing in this book), rather than as a static representation (as most connectionist linguists do) considerably simplifies the computational task of mapping episode representations to surface utterances. Thus the idea that 'the LF of a sentence describes a sequence' not only enables a useful sensorimotor interpretation of declarative syntactic structures (as discussed in Chapter 5), but also enables some interesting new

connectionist architectures for sentence processing.

The core of the new model I propose is given in Section 6.4.3. There I describe a network which takes a sensorimotor sequence as input and learns to selectively 'read out' some of the signals in this sequence as words. This network is actually fairly simple—readers who want to cut to the chase can go straight to this section. However, it is obviously important to situate this network in relation to what is already known about how linguistic information is processed and represented in the brain, and about how infants acquire language. The earlier parts of the chapter are there to provide this context. Sections 6.1 and 6.2 are fairly straight reviews of the literature on the neural substrates of language and on language development in infants. Sections 6.3 and 6.4 introduce a new model of language processing, but it is only in Section 6.4.3 that this model relates directly to the sensorimotor interpretation of LF presented in Chapter 5.

6.1 Neural substrates of language

In this section, I will give an overview of what is known about the neural mechanisms involved in processing language, and about the neural representations which subserve these mechanisms. In Section 6.1.1 I will discuss the neural representation of phonemes, and of sequences of phonemes. In Section 6.1.2 I will discuss how the brain represents the 'meanings' of concrete nouns and action verbs. (In our embodied approach to semantics, these words have sensorimotor denotations: concrete nouns denote object categories and action verbs denote motor programmes. Thus this section is largely a matter of recapping aspects of the sensorimotor model.) In Section 6.1.3 I provide a definition of 'words' as neural assemblies: following Pulvermuller (2001), words are viewed as assemblies which map phonological representations to semantic representations. I also discuss how these assemblies function during production and generation of language, and discuss the location of these assemblies for nouns and verbs. Finally in Section 6.1.4 I discuss the neural locus of 'syntactic processing'. There are a number of well-known arguments from language disorders—Broca's and Wernicke's aphasia—suggesting there is a localised neural region specialising in syntactic processing. These arguments are controversial, but I will summarise some basic findings. I will also review some more recent studies which shed light on the neural mechanisms involved in syntactic processing.

6.1.1 The neural locus of phonological representations

'Surface language' can take many forms: it can be a sequence of visual patterns (in written language), or a sequence of gestures (in sign language). But the primary medium for

surface language is in speech, and I will focus on speech in this section. A speech signal is a sequence of articulated sounds, or **phonemes**. How are phonemes represented and generated in the brain? And what is the relationship between the phonemes which we hear (which are auditory signals) and those which we produce (which are articulatory motor programmes)?

6.1.1.1 A mirror system for phonemes

A consensus is emerging that the auditory perception of phonemes involves the identification of the articulatory gestures which led to the heard stimuli—in other words, that there is a 'mirror system' for phoneme perception, just as there is for the perception of limb actions. In fact, a mirror system hypothesis was proposed for articulation some years before it was first proposed for hand/arm actions.

Evidence for a mirror system for articulatory actions comes from two distinct sources. Firstly, there is evidence from the linguistic discipline of **phonology**, which studies the sound systems of language, and how these contribute to meaning. One of the tasks of phonology is to build models of how acoustic signals can be mapped onto phoneme sequences. This is a difficult task because of the phenomenon of **coarticulation**: the pronunciation of one phoneme is often interfered with by the preparation of the next, so that phonemes are often only partially produced (which is what makes automated speech recognition a hard task). Several theorists have argued that formulating an adequate model requires us to posit phonological units which are in effect gestures—see in particular Liberman and Mattingly (1985); Browman and Goldstein (1995). These theories posit that hearers can recognise a speaker's *articulatory goals* prior to their achievement, just as observers of a reach action can recognise its intended target before the hand gets there. This means that phonemes can be recognised even if they are not fully articulated.

Other evidence for the existence of a mirror system for phonemes is directly experimental. For instance, in an fMRI imaging study by Wilson *et al.* (2004), subjects were scanned performing two tasks: listening to a nonsense syllable, and pronouncing a nonsense syllable. A brain region in the ventral premotor cortex was found which was activated by both these tasks. (Crucially, since the tasks involve nonsense syllables, activation of this region must be due in each case to their phonological properties rather than to any associated meanings.) Other experiments have used transcranial magnetic stimulation over the motor cortex to amplify motor representations generated while listening to speech (see e.g. Fadiga *et al*, 2002). Subjects heard two types of auditory stimuli: one type were words or nonwords whose pronunciation requires a large tongue movement (e.g. *birra*, *berro*); the other type were words or nonwords whose pronunciation is less reliant on the tongue (e.g. *baffo*, *biffo*). They found that TMS activated subjects' tongue muscles more for the former

stimuli than for the latter, which suggests that hearing words causes their associated motor programmes to be activated. Perhaps the most direct evidence for a mirror system for articulation comes from a study by Ito *et al.*, 2009. These researchers used robotic devices to stretch the skin around a subject's mouth, to mimic the somatosensory consequences of producing particular phonemes—specifically, the lateral stretches involved in producing the vowels in *head* and *had*. The stretch signals were synchronised with the delivery of audio word stimuli, which subjects had to identify. It was found that subjects' perceptions of words were biased in the direction of the motor movements, suggesting that the neural pathways involved in actively producing phonemes (which rely on somatosensory feedback from the vocal tract) are also involved in auditory processing of phonemes. In summary, there is good evidence that auditory verbal stimuli are encoded as motor programmes.[1]

The phenomenon of coarticulation suggests that the mapping from auditory signals to articulatory gestures is learned for specific sequences or clusters of phonemes, rather than at the level of individual phonemes. Work in phonology (see Blevins, 1995 for a review) and in psycholinguistics (see e.g. Treiman, 1986) suggests that phoneme sequences are represented at several levels of hierarchical structure. A key unit is the **syllable**, a unit which is most commonly defined in terms of 'sonority'. A stream of speech alternates cyclically between high-sonority sounds (vowels) and low-sonority ones (consonants); each cycle is associated with a single syllable. A syllable has its own hierarchical structure, consisting of an **onset** and a **rhyme**. The onset and rhyme are both sequences of phonemes. Phonemes can be **vowels** or **consonants**. Simplifying considerably, the onset is a short (possibly empty) sequence of consonants, and the rhyme is a vowel followed by another short (possibly empty) sequence of consonants. Connectionist models of phonological structure tend to use localist encodings to represent individual phonemes, and also to represent higher-level phonological units such as specific onsets, codas and syllables. These higher-level units activate groups of phonemes in parallel, often in conjunction with specialised phoneme-sequencing machinery. A particularly elegant model is that of Hartley and Houghton (1996), which combines the idea of parallel activation of individual phonemes and syllables with the idea of central pattern generators, responsible for the cyclic rhythms of syllabic speech. In this model, each possible syllable in the language is encoded as a single unit. Syllable sequences are stored using a competitive queueing mechanism (supplemented with

[1]For a dissenting opinion, see Lotto *et al.* (2009), who argue that the motor theory of speech perception is only superficially consistent with the claim that there is a mirror system for phonemes. They argue that if this claim is taken as asserting that the auditory and articulatory systems interact, it is uncontroversial, and if it is taken as asserting that we fully understand the relationship between speech perception and speech production, it is a considerable overstatement. Obviously I do not want to endorse this latter claim. But I do think that the hypothesis of a mirror system for phonemes has proved, and is still proving, useful in generating interesting research into the relationship between hearing and speaking.

a temporally evolving context signal, to support repeated syllables). The competitive queueing mechanism means that several syllables are active simultaneously, allowing for syllable inversions which are often found in generated speech. At the same time, each syllable links to separate 'onset' and 'rhyme' units. When a syllable is activated, its onset and rhyme units are activated sequentially, one at a time. The onset and rhyme units are each connected in parallel to a set of phoneme units. Phoneme units are activated partly from the currently active onset or rhyme unit, and also partly from an independent pattern generator which cycles through the different phases in a syllable. The parallel activation of phonemes allows the possibility of Spoonerisms, where the onset of one syllable is combined with the rhyme of another. The decomposition of syllables into onsets and rhymes, together with the operation of the pattern generator, ensure that the output stream of phonemes conforms to the phonological regularities of the language, even in the presence of errors. I will assume that the mapping which children learn from auditory to articulatory representations is learned at the level of whole syllables, rather than of individual phonemes or lower-level phonological units. This is a simplification, but it should not affect the model of higher-level language processing which I develop later.

How do we learn to map auditory syllable representations onto their associated articulatory gestures? A common suggestion is that children begin by exploring random articulatory movements ('babbling'), and come to associate these movements with the reafferent auditory stimuli which they produce, through regular Hebbian learning (see e.g. Westermann and Miranda, 2004). This model is analogous to the Iacoboni (2001) / Keysers and Perrett (2004) model of the origin of mirror neurons for motor actions. The circuit I assume is shown in Figure 6.1. The basic idea is that an infant generates a random se-

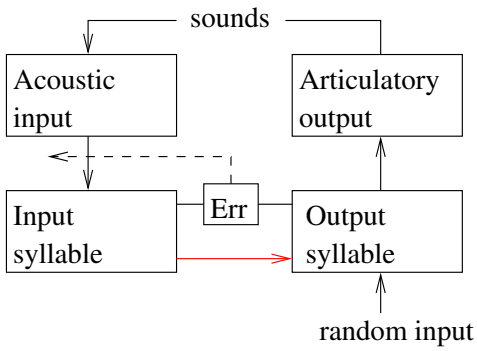


Figure 6.1: A circuit for learning to map acoustic representations onto articulatory syllables

quence of syllable-sized articulatory units (the Output syllable box), which are converted

into motor gestures (the Articulatory output box). The reafferent sounds generate acoustic representations (the Acoustic input box), and a function converts these representations back into articulatory gestures (the Input syllable box). This function is trained (dashed line) using an error term representing the difference between the syllable gesture it actually computed and the one it should have computed. When the function is fully trained and the error term is low, the system enters a new mode where input syllable representations are mapped directly onto output syllable representations (red arrow). During training, representations in the output system influence representations in the input system, but after training is complete, input representations are copied into the output system. In this latter mode, the infant can reproduce a syllable spoken by a different speaker.

What are the neural pathways involved in mapping auditory signals to motor gestures? The pathway obviously begins in **auditory cortex**, which is in the dorsal superior temporal gyrus, and ends in the areas of premotor (and finally, motor) cortex which control the vocal tract. An influential model by Hickok and Poeppel (2007) proposes that auditory signals are converted into phonological representations in the auditory cortex, and also middle and posterior portions of the superior temporal sulcus (STS). Hickok and Poeppel propose that there are two neural pathways for processing the phonological representations computed in STS. One pathway runs through the left temporoparietal junction in inferior parietal cortex (an area comprising the angular and supramarginal gyri, which has recently been dubbed 'Geschwind's territory'), and then directly to articulatory areas in the premotor cortex and inferior frontal cortex. This pathway turns input phonological representations directly into motor gestures, without accessing 'words'; it is this pathway which is implicated in the development of the mirror system for phonemes. A possible anatomical basis for this pathway has recently been found; there are well-defined projections from the auditory cortex to Geschwind's territory, and from there to **Broca's area** in left inferior frontal cortex (Catani *et al.*, 2005).[2] The other pathway runs through posterior and anterior temporal cortices, and maps phonological representations in STS onto words. This pathway will be further discussed in Section 6.1.3.2.

Note that Hickok and Poeppel's direct pathway for mapping input phoneme representations onto articulated phonemes is very similar to the pathway implicated in the mirror system hypothesis for hand/arm actions, introduced in Section 2.7. The direct phoneme pathway for phonemes runs from auditory cortex through STS and inferior parietal cortex to premotor and motor cortices. The pathway which maps visually perceived hand/arm actions onto motor representations runs from visual cortex through STS and inferior pari-

[2]An alternative route is via the **arcuate fasciculus**, which directly connects the dorsal superior temporal gyrus with Broca's area. But the angular gyrus has an important role to play in phonological working memory, as I will discuss in Section 6.1.1.2, so I will assume a sensorimotor route through this area, following Hickok and Poeppel.

etal cortex to premotor and motor cortices. The modality of the input areas (auditory vs visual) and output areas (articulatory vs hand/arm) are different, but the intermediate areas are the same. In fact, it seems likely that the intermediate areas play a role in cross-modal sensory integration during speech perception. It is well known that watching the movements of a speaker's mouth can influence auditory speech interpretation (McGurk and MacDonald, 1976). Sekiyama *et al.* (2003) have recently found fMRI and PET evidence that left STS is activated by speech recognition tasks which require a combination of audition and lip-reading. Recall from Section 2.7.2 that STS encodes visually perceived motion stimuli, and is sensitive to faces as well as to motor gestures. It seems likely that STS is where auditory and visual information about an observed speech signal are combined.

6.1.1.2 The phonological loop

Typically a hearer perceives not just an isolated syllable, but a sequence of syllables closely following one another. If syllables are represented as articulatory actions, the hearer will in effect be receiving information about a sequence of actions. What does the hearer do with this information?

It is well established that short sequences of syllables can be retained in working memory, and repeated later. This ability is attributed to a form of working memory called the 'phonological loop', which we have already briefly introduced in Section 3.1.2. As discussed there, the phonological loop stores phonological information rather than semantic information, it is sequentially structured, and it requires rehearsal to be maintained for more than a few seconds.

What is the phonological loop for? Baddeley *et al.* (1998) argue persuasively that it plays a role in language acquisition. For instance, it has been found that there is a correlation between phonological working memory capacity and vocabulary size in children (Gathercole and Baddelely, 1990). There are several computational models of vocabulary acquisition which foresee a role for the phonological loop. Its function is typically to deal with the asynchrony of word-meaning pairings in the training data to which a language learner is exposed. A child must learn the meaning of the word *dog* in a situation where he sees a dog and hears the word being pronounced; however, there is often some delay between the onset of word and meaning representations. The phonological loop might enable words to be buffered, so that each word can be associated with a range of candidate meanings, improving the efficiency of word meaning learning. I will discuss this proposal in more detail in Section 6.2.2.1. There may also be a role for the phonological loop in the acquisition of syntax. A study of preschool children found that their phonological working memory capacity correlated with the syntactic complexity of the sentences they produced in spontaneous speech (Adams and Gathercole, 1995). This may be because

learning grammatical constructions requires short-term storage of extended word sequences (as argued by Baddeley *et al.*, 1998). But it may also be to deal with synchrony issues in the training data available to children. The training data for learning syntax consists of pairs of event representations and word sequences. Again, these pairs may not always arrive simultaneously; the phonological buffer may provide a method for synchronising the presentation of a word sequence with the onset of its associated event, so that a detailed mapping between them can be learned. These ideas will be discussed further in Sections 6.3 and 6.4.

The role of the phonological loop in mature sentence processing is more controversial. There are many case studies of adults with severely impaired phonological working memory, but apparently intact abilities to generate speech (see e.g. Shallice and Butterworth, 1977) or to interpret it (see e.g. Vallar and Shallice, 1990). This condition is termed **repetition conduction aphasia**. At the same time, there is good evidence that acquired deficits in phonological short-term memory tend to co-occur with language deficits, which suggests that the phonological loop is involved in normal language processing (see e.g. Martin and Saffran, 1997). One way of reconciling these findings is to note that there is good evidence for two distinct phonological buffers: a **phonological input buffer** which temporarily stores an incoming verbal signal as a sequence of phonemes, and a **phonological output buffer**, which supports the motor production of a sequence of phonemes. The phonological input buffer is closely related to the auditory system; it holds a memory of phonological material which has recently been heard, which decays quite rapidly as time passes. The phonological output buffer, on the other hand, involves premotor articulatory planning. Representations which are added to this buffer must be presumed to increase in activation over time, until they reach a threshold where they are executed.

The most direct evidence for a phonological output buffer comes from studies of errors made during spontaneous speech—for instance Spoonerisms, where two phonemes in the speech stream are swapped with one another (as when a speaker says *the darn bore* rather than *the barn door*). These errors suggest that a speaker internally represents the sequence of phonemes which are about to be pronounced, in a format which allows for transpositions of phonemes to occur prior to the sequence actually being spoken. A competitive queueing model, where the phonemes in a prepared sequence are active in parallel, seems broadly appropriate for modelling such errors, just as it is for modelling the phonological loop.

Evidence for a distinction between phonological input and output buffers comes primarily from studies of dysfunction. For instance, Shallice *et al.* (2000) describe a patient with a condition known as **reproduction conduction aphasia**, who showed a similar pattern of phonological errors when asked to repeat a spoken word, to read a written word, or to name an object. Since these tasks are quite diverse, but all ultimately involve spoken output,

it can be assumed that the patient's deficit was in the phonological output buffer.[3] Crucially, the patient also had an intact digit span, suggesting the existence of a separate (and undamaged) phonological input store where an auditory phonological sequence could be maintained. Patients with repetition conduction aphasia, who have impaired phonological STM but are unimpaired at generating or interpreting speech (Shallice and Butterworth, 1977; Vallar and Shallice, 1990), can be assumed to have the converse problem: damage to their phonological input buffer, and an intact phonological output buffer. (This argument rests on the assumption that spontaneous speech generation involves sending phonemes to a phonological output buffer, but does not involve the phonological input buffer, and that ordinary speech interpretation can take place in real time, without buffering auditory input.)

If the phonological input and output buffers are distinct, we should expect to see dissociations between the neural areas where they are stored. If the input buffer is a sensory working memory, and the output buffer holds articulatory representations, we might expect to see the input buffer close to auditory cortex, and the output buffer in the prefrontal or premotor system. Some early imaging studies on normal subjects suggested just this. For instance, Paulesu *et al.* (1993) localised the 'articulatory rehearsal' system in left Brodmann area 44, the posterior part of Broca's area in inferior PFC, and distinguished this from the 'phonological store', which they localised in the left supramarginal gyrus (which is on Hickok and Poepel's direct sublexical route from input to output phonological representations). An imaging study by Henson *et al.* (2000) is roughly consistent with this picture, localising the phonological store in left supramarginal gyrus (but also in the inferior frontal gyrus), and the articulatory rehearsal system in left lateral premotor cortex.

However, things are not this straightforward. It seems likely that reproduction conduction aphasia results from damage to an area encompassing the auditory cortex in the left hemisphere, called the **left posterior superior temporal gyrus** (**lpSTG**), and also known as **Wernicke's area** (Damasio and Damasio, 1980; Anderson *et al.* 1999). This area is associated with speech interpretation, as already mentioned in Section 6.1.1.1. If it is damaged, there are frequently serious difficulties understanding speech, as I will discuss in Section 6.1.3.2. So it is somewhat surprising to find it implicated in output phonological processing, which we might expect to situate in frontal or premotor areas. But Hickok *et al.* (2000) present fMRI evidence that the lpSTG is indeed active during speech production, even when the subject cannot hear the sound of his own voice.

It may seem hard to reconcile the idea of distinct phonological input and output buffers

[3]It is interesting to note that patients with reproduction conduction aphasia typically have more problems repeating (pronounceable) nonwords than with repeating words (Caplan and Waters, 1992). This can be understood as another lexical bias effect: phonemes which are grouped into words will tend to be strongly associated with one another, and will thus be more resilient to damage.

with the apparent involvement of the lpSTG in both input and output phonological processing. But in fact a close physical proximity between phonological input and output areas is predicted by the hypothesis that there is a mirror system for phonemes. As noted in Section 6.1.1.1, if acoustic signals are turned into premotor articulatory representations during phonological input processing, the function which learns to do this must have access to 'genuine' articulatory representations on which to train. So we expect there to be phonological output representations in the close vicinity of the phonolgical input system. The presence of both input and output phonological representations within the lpSTG is perhaps analogous to the presence of both canonical and mirror neurons in macaque premotor cortex (see Section 2.5.3 and Section 2.7.5). Whether damage to the lpSTG results in phonological input difficulties or phonological output difficulties or both may well relate to fine-grained functional organisation within this area which is hard to observe with lesion-mapping or imaging techniques, and may also be subject to large individual differences from one speaker to another.

In summary: the phonological input buffer is probably implemented in the supramarginal gyrus, and in circuits connecting this area with the phonological input representation areas in lpSTG (and with associated phonological representations in STS, as mentioned in Section 6.1.1.1). The phonological output buffer is probably distributed between Broca's area, articulatory premotor cortex and output representation areas in the lpSTG.

Phonological short-term memory normally involves 'rehearsal' of phonological material, as we have already discussed. What are the neural circuits involved in this process? Most models assume that phonological items can be 'played out' from the phonological input buffer directly to an articulatory process, the output of which can in turn refresh the phonological input buffer. (Note that if we assume a mirror system for phonemes, it is likely that the phonological input buffer can hold articulatory representations, in which case it would be quite appropriate to refresh it in this way.) One of the most successful models of the phonological loop is that of Burgess and Hitch (1999), which uses competitive queueing as its central mechanism, as well as a temporally updating context signal, which provides a means for representing sequences containing repeated elements. The model is basically very similar to the competitive queueing models of prepared action sequences which we discussed in Chapter 3.2. A somewhat adapted version of Burgess and Hitch's model is shown in Figure 6.2. (The figure builds on the mirror system circuit for syllables given in Figure 6.1; connections only involved in the phonological loop, and labels associated with models of the phonological loop, are shown in red.) The core device in the model is a temporally evolving context signal, which can be set to progress through a deterministic sequence of states as a sequence of syllables is presented to the system. Short-term weights are created between context representations and individual syllables in

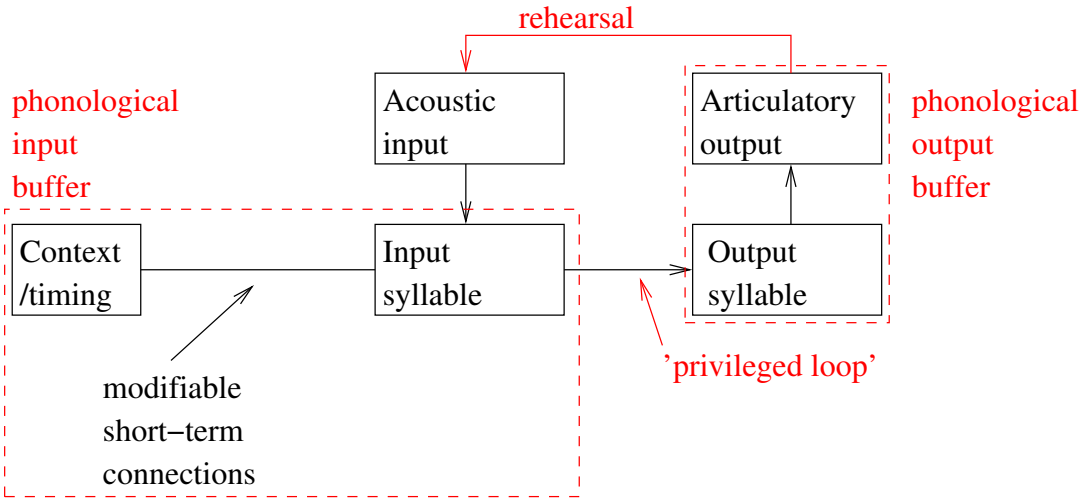


Figure 6.2: A model of short-term memory for syllable sequences

the sequence. These weights decay rapidly, but before they decay it is possible to replay the context signal and read out the associated syllables. Input syllables are mapped directly to output syllables—a mechanism which Burgess and Hitch call the 'privileged loop'. Output syllables generate articulatory representations. To model rehearsal of a stored syllable sequence, these articulatory representations can be passed back into the input system, where they can refresh the decaying weights between context and syllable representations, and allow the operation to be repeated.[4]

The model in Figure 6.2 actually differs from Burgess and Hitch's original model in several ways. For one thing, their model uses phonemes as articulatory units, rather than syllables. For another thing, their model includes phonological representations of words. (I will discuss this extra level of complexity in Section 6.1.3.1.) Finally, their model does not distinguish between phonological input and output buffers. (In this sense, the model I have presented is more like that of Jacquemot and Scott, 2006 in which phonemes are cycled from the input buffer to the output buffer, and then back again.)

Note that according to this model, the connections involved in learning the mirror system for phonemes overlap with those which implement the phonological buffer. The

[4]We must assume that the articulatory representations here are still *premotor* representations, albeit low-level ones, if we want to be able to model subvocal rehearsal. In fact, since these premotor representations must be mapped back to the input stream, we must probably assume a separate mirror system for these lower-level units—for instance, a system which learns to represent incoming phonemes as articulatory units. However, I will not extend the model below the level of single syllables.

information travelling along these connections is the same for both tasks, even though the way it is used, and the direction in which it flows, are somewhat different.

In Sections 6.3 and 6.4, I will pursue the idea that representations in the phonological STM system are involved in other aspects of language learning. If the phonological input system represents the speech of a mature speaker whom the child is listening to, and the phonological output system represents the output of the child's language production network, configured to attempt to reproduce this speech from a semantic input, the differences between the two systems can be used as an error term to train the production network.

6.1.2 Neural representations of the semantics of concrete nouns and verbs

Where in the brain are the meanings of words represented? There are many different conceptions of word meaning; the answer depends on which conception we pick. For instance, logicians and formal semanticists tend to adopt an 'extensional' account of word meanings (see classically Tarski, 1936/1983), where meanings are defined with reference to a notional database containing the set of all known objects and exhaustively enumerating all the facts which are true of each object. The meaning of a common noun like *cup* is defined as the set of all objects which are cups. The meaning of a transitive verb like *grab* is the set of all ordered pairs of objects $\langle O_1, O_2 \rangle$ such that O_1 grabbed O_2. This conception of meaning stresses 'encyclopaedic' knowledge: knowing the meaning of a word is knowing all the facts about the world in which it participates. For instance, we know something about the meaning of the word *grab* if we know which pairs of objects have been related by a 'grab' event; and we know something about the meaning of the word *cup* if we know all the objects in the world which are cups. Using this conception of meaning, we expect some component of word meanings to reside in the structure of long-term memory for objects and their associated properties and events. In Section 9.2 I will argue that a form of long-term memory called **semantic memory** carries this sort of information, and I will consider the contribution of encyclopaedic knowledge to word meanings in that section.

Another way of modelling the meaning of a word is to represent how it is used by speakers. This conception of meaning is pragmatic; it it sometimes associated with Wittgenstein's (1953/2001) slogan 'meaning is use'. It is a conception which is often adopted by empiricist linguists, who model language as a disparate collection of learned conventions (as will be discussed in Section 6.2.5.2). The 'situations of usage' of a word can be modelled in different ways. In computational linguistics, the situations in which a word is used are often modelled by making reference to the words with which it commonly co-occurs in natural texts (see e.g. Harris, 1954; Pantel and Lin, 2002). I will discuss this conception

of word meaning in Section 6.2.5.4. In psychology, 'situations of usage' are often defined in relation to the speaker's intentions. This definition is particularly important in an account of children's early use of words as holophrases. I will discuss this conception of word meaning in Section 6.2.3.

The extensional and the distributional conceptions of meaning both emphasise the interrelated nature of word meanings. Word meanings are massively interrelated, in a variety of different ways. Each fact that we represent about the world is a relation between several semantic elements: the extensional account of word meaning sees words as pointers into this complex web of relations. In surface language, words appear in complex distributional patterns with other words: the usage-based account of word meaning emphasises this type of interrelatedness between words.

While both of the above conceptions identify components of word meaning, I want to focus on another conception of word meaning, which can be termed the 'embodied' conception. According to this conception, it is necessary to ground the web of interrelated word meanings in sensorimotor experience (see e.g. Harnad, 1990). The meanings of *concrete* words provide a means for doing this. Concrete words are words which describe the properties or types of physical objects or actions. Like all words, these words participate in the web of meanings defined by the extensional or usage-based accounts just described. But they also have an 'embodied' or 'sensorimotor' meaning, because they are directly associated with particular sensorimotor concepts. The embodied meaning of a concrete noun like *man* or *cup* consists of the representations which are evoked in the sensorimotor system when an observer attends to an object of the relevant type. The embodied meaning of a concrete action verb like *grab* is the motor programme which is activated when an observer performs the associated action, or observes another agent performing this action. In this section, I will consider what these embodied meanings might comprise, and what neural representations they might correspond to.

Several researchers have found evidence that processing concrete words involves activation of associated sensorimotor areas. I will consider three studies here. Damasio *et al.* (1996) investigated the production of concrete nouns. They asked a large group of patients with a range of focal brain lesions to name objects from several different categories. Two of the interesting categories were 'tools' and 'animals'. They found several patients with selective difficulties in naming objects from one category or the other, even when there was good evidence the objects had been successfully categorised. Moreover, these difficulties were correlated with the position of the patients' lesions: difficulties in naming animals were associated with damage to left anterior IT cortex, and difficulties in naming tools were associated with damage to posterolateral IT, along with junctions to neighbouring temporal, occipital and parietal cortices. Damasio *et al.* suggested that representations of object categories in IT have to be 'enriched' in order to trigger associations with specific lexical

items, and that damage to certain areas of IT can prevent this enrichment. Damasio *et al.* also conducted a PET imaging study on normal subjects naming objects from the same categories. They found that naming animals and naming tools activated different regions in IT, and that these regions corresponded approximately to the regions associated with the naming deficits in the patient population. The important thing about these results is that IT is not a 'language' region. As we saw in Section 2.2, it is involved in the visual classification of objects: a task which does not implicate language at all. The activity in IT required for the naming task is likely to be in the creation of visual representations of object types which are rich enough, or strong enough, to trigger associated words.

More recently, a series of experiments about the sensorimotor associations of concrete words has been conducted by Friedemann Pulvermüller and colleagues. Pulvermüller *et al.* (1999) used a number of imaging techniques (ERP, EEG and MEG) to investigate neural activity found in subjects presented with written words in a lexical decision task. The words included nouns with strong visual associations but weak motor associations, and verbs with the converse pattern of associations. Nouns tended to evoke activity in visual cortices, while verbs evoked activity in motor, premotor and prefrontal cortices. In other studies, more detailed distinctions within the class of verbs were found. Hauk *et al.* (2004) used an fMRI paradigm to present subjects with verbs describing actions of the mouth, arms or legs (*lick*, *pick* and *kick*). They found that these verbs elicited activity in the corresponding mouth, hand and foot areas of the motor and premotor cortices. Pulvermüller *et al.* (2005) found similar results in an MEG paradigm, showing additionally that the motor associations of verbs are activated very fast, on the order of 150ms after word onset. Again, there is good evidence that concrete words are strongly associated with specific sensorimotor denotations.

Finally, Lu *et al.* (2002) found similar evidence for sensorimotor word denotations from a study of brain dysfunction. They investigated the naming abilities of patients who had had their left anterior temporal lobe surgically removed. This procedure cuts the uncinate fasciculus, which links the temporal lobe with frontal areas. It was found that these patients had a selective deficit in naming actions and tools, in comparison to control patients who had undergone the same procedure in the right hemisphere. Given that frontal cortex is where motor concepts are represented, this is additional evidence that the semantics of concrete verbs, and of nouns denoting tools, are primarily given in motor terms.

The activation of *prefrontal* assemblies by action verbs (see especially Pulvermüller *et al.*, 1999) is particularly interesting given our hypothesis about head movement, as set out in the previous chapter. Recall from Section 5.4.3 that head movement is hypothesised to be due to the fact that heads—including verb heads—are signalled linguistically from planning representations in prefrontal cortex rather than from transitory action signals in premotor or motor cortex. Pulvermüller's finding that action verbs activate prefrontal

representations is consistent with this hypothesis. In fact, there is also evidence that there are (partly non-overlapping) areas of prefrontal cortex which are involved in the processing of nouns and object representations. I will introduce this evidence when I discuss Broca's area in Section 6.1.4.2, and return to it in Chapter 7.

6.1.3 The neural representation of words

We have discussed where phonological sequences are represented, and where sensorimotor denotations of concrete words are represented. Where then are words themselves represented? There are two ways of defining words, which need to be considered in the right order. Firstly, words are phonological entities: commonly-occurring sequences of syllables. Secondly, words are associations *between* well-defined syllable sequences and semantic representations—or in our simplified case, sensorimotor representations. I will consider these two conceptions of words in the following two sections.

6.1.3.1 Words as phonological entities

We have already identified two areas where phonological representations are maintained: the phonological input system (in STS, parts of the left posterior STG and the supramarginal gyrus), and the phonological output buffer (distributed across posterior Broca's area, articulatory premotor cortex and other parts of the left posterior STG). As argued in Sections 6.1.1.1 and 6.1.1.2, these areas represent phonemes as prepared articulatory actions, and syllables as higher-level articulatory actions involving multiple phonemes. Extending this conception, it is natural to think of phonological word representations as still higher-level articulatory action plans, to produce sequences of syllables. As discussed in Section 2.5, motor representations tend to have a rich hierarchical structure; there is no reason why articulatory motor representations should be an exception. Phonological word representations can be thought of as special assemblies in the phonological input and output systems, supporting the interpretation and generation of frequently occurring complex articulatory actions.

Phonologically, an individual word is a sequence of syllables.[5] We have already proposed that syllables (and their constituent phonological units) are represented in both input and phonological systems. The same appears to be true for word-sized phonological word representations. In fact, the distinction between a **phonological input lexicon** and a **phonological output lexicon** is a venerable one in cognitive neuropsychology. One

[5] A phonological word is also associated with a stress pattern, or 'prosodic structure', which has its own influence on word production and retrieval processes (see e.g. Lindfield *et al.*, 1999). I will not consider word prosody in this review.

condition which is often adduced as evidence for this distinction is 'pure word deafness'. To illustrate, Auerbach *et al.* (1982) describe a patient who can produce single words normally, and can also read words normally, as well as repeat spoken words. However, the patient cannot understand spoken words. In this patient, the phonological input buffer is not completely destroyed, because he can accurately repeat phonological material; it is just that input phonological representations are not able to access word-sized semantic representations in the normal way. In fact, there are two possible reasons for this deficit. It may simply be because the connection from phonological input representations to semantics is lost (as I will discuss in Section 6.1.3.2). But it may also be because of damage within the phonological input buffer, preventing the emergence of word-sized phonological assemblies sufficiently well defined to activate specific semantic representations. (This hypothesis mirrors Damasio *et al.*'s hypothesis about how damage local to IT can prevent object naming, even when object recognition is intact; see Section 6.1.2.) But note that in either case, it is clear that words as phonological *output* groupings are still intact. So in either case, it is legitimate to talk about a 'phonological input lexicon', whether it is stored in structures within the phonological input buffer or in links from this buffer to semantic representations. There are other patients with a converse pattern. We have already discussed reproduction conduction aphasia (Shallice *et al.*, 2000), which involves damage to the phonological output system. The damage appears to be to output word representations; patients with this condition have a particularly hard time producing complex words, whether they are repeated or spontaneously produced, and their errors in these words tend to be phonemic in nature. These patients can nonetheless understand the words they cannot produce. In fact, normal speakers also make phonological errors quite regularly. There is a lexical bias on these errors; they take the form of words more often than would be expected by chance (see e.g. Baars *et al.*, 1975; Dell and Reich, 1981). This suggests that the buffer of phonemes 'competing to be spoken' is hierarchically organised to reflect groupings found in actual lexical items.

I will assume that there are phonological word representations both in the input and output phonological systems, as shown in Figure 6.3. Again, the figure is a modification of the model of Burgess and Hitch (1999); the main difference is that I distinguish between input and output word representations. The figure builds on the model of phonological short-term memory given in Figure 6.2. Notice there are no direct connections between input and output phonological word representations. As Hickok and Poepel propose, the direct mapping from phonological input to phonological output representations is 'sublexical'.

Input phonological word representations are part of the phonological input buffer cir-

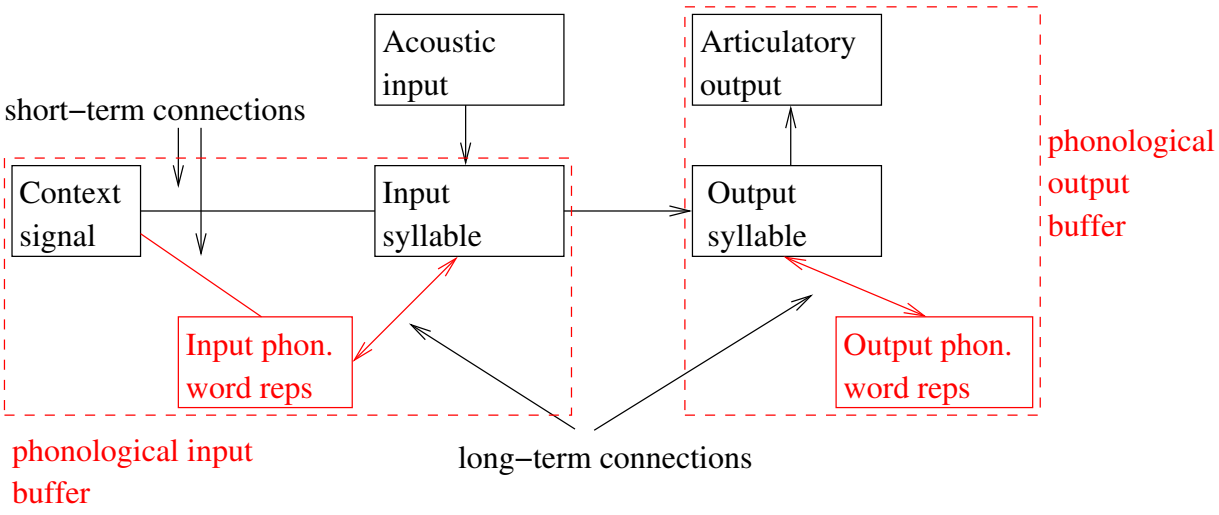


Figure 6.3: Input and output phonological word representations

cuit. I assume they are mainly in STS and left posterior STG[6] rather than the supramarginal gyrus, since this latter structure is the start of the 'sublexical' route from input to output phonological representations. They are connected bidirectionally to input syllable representations. They should be thought of as representations which become active when particular sequences of syllables are presented to the system, but which can also trigger the 'replay' of their associated syllable sequence. Burgess and Hitch model the role of word representations in phonological short-term memory by positing short-term connections between word representations and the evolving context signal. Thus if a syllable sequence in the input evokes a word representation, this word representation is linked to a particular temporal context. This circuitry reproduces the finding that sequences of words are easier to recall than sequences of nonwords.[7] Additional 'lexical effects' in short-term memory are modelled by assuming that word representations compete with one another, and that they activate their associated syllables in parallel. These assumptions allow the

[6]Some good evidence that lpSTG is involved in storing phonological word representations comes from a recent study by Graves *et al.* (2008), who show that word-sized phonological priming effects seen during a nonword repetition task correlate with activity in this region. Whether the phonological word representations in this area are in the input or output system or in both is hard to judge, for the reasons already discussed in Section 6.1.1.2.

[7]In fact in Burgess and Hitch's model, the context signal links *only* to word representations, and not to sublexical items. But as they note, to model memory for nonword syllable sequences, we must envisage links from the context signal to both levels of representation.

system to make recall errors in which whole words are transposed, and also produce a tendency for errors in recalled syllable sequences to take the form of words.

Output phonological word representations encode sequential regularities in the input speech stream, just like input phonological word representations. However, they encode these regularities in the motor domain: rather than passively storing and replaying commonly-occurring syllable sequences, they are best thought of as autonomous action plans which can actively *produce* such sequences. The suggestion that a sequence of heard phonemes can activate a word-sized articulatory plan can be thought of as an instance of 'intention recognition', similar to that described for hand/arm actions in Sections 2.7.6 and 3.4. In this case, the premotor representations are syllable-sized articulatory representations, and the right sequence of premotor representations 'abductively' activates the plan which would produce this sequence if executed by the observer. Again I will not consider the details of the circuit which learns these representations, but I will assume that output phonological words compete with each other, that they activate their constituent syllables in parallel, and are activated in turn by these syllables, just like input words. These assumptions allow us to model various types of error in spontaneous speech, including syllable transposition errors, and the tendency for errors to take the form of words. The 'phonological output buffer' now comprises all of the right-hand side of the figure; we can envisage multiple words active simultaneously and competing with one another, each activating its set of component syllables, which in turn activate their component phonemes. The most active words and syllables dominate the buffer, but the parallelism allows for output errors to occur.

6.1.3.2 Words as form-meaning associations

A word can also be understood as an association between a phonological representation and a semantic one. One of the best-developed accounts of words in this sense comes from Pulvermüller (2001). He suggests that words are implemented in the brain as large, distributed cell assemblies, which combine phonological or articulatory representations with semantic representations (or sensorimotor representations, in our concrete case). The phonological parts of a word assembly are principally in Broca's area, Wernicke's area and STS, as already discussed, while the semantic parts are in different areas for different word types. For concrete nouns they will include representations in the temporal cortices, where object categories are found, while for action verbs they will include representations in the premotor and prefrontal cortices. The areas involved in mapping phonological to semantic representations will be distributed between the phonological regions and the specialised semantic areas. Pulvermüller suggests that these associations are learned through regular Hebbian processes. We will consider the learning mechanisms which may be involved in

Section 6.2.2. But the basic idea is that associations are created between phonological word representations and semantic representations which frequently co-occur together.

Figure 6.4 shows (very schematically) the links between phonological and semantic representations. I am emphasising sensorimotor meanings here, and in particular the plan-

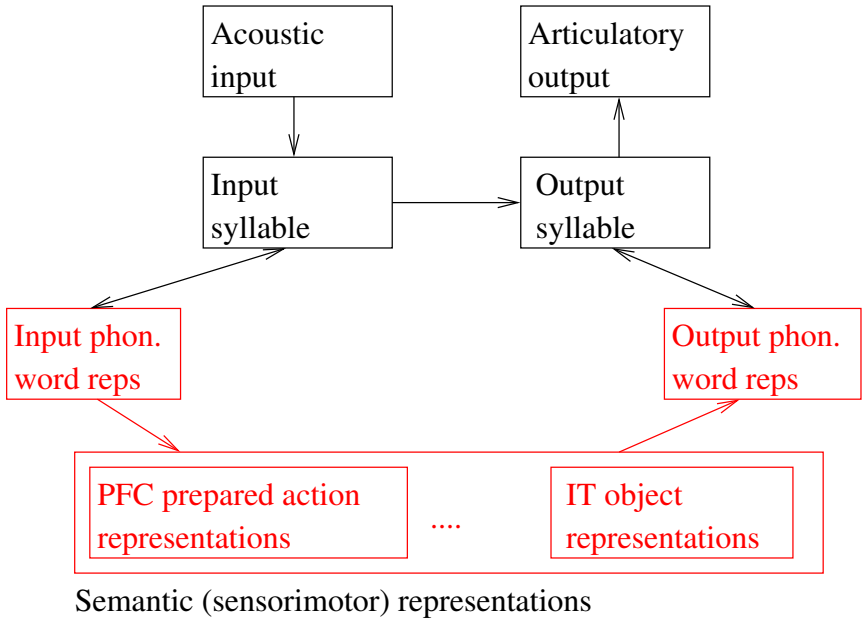


Figure 6.4: Associations between semantic and input/output phonological representations

ning representations in PFC and the object representations in IT, as just discussed in Section 6.1.2. Note that there are two unidirectional mappings: one from input word representations to semantic representations, and one from semantic representations to output word representations. These two mappings create another indirect route from input to output phonological representations, via semantic representations. If this route is intact, but the direct mapping from input to output syllables is damaged, we will be able to use semantic representations to reproduce the 'gist' of acoustically presented words. However, we will have a tendency to produce semantically related words instead, even if they bear no phonological resemblance to the input word (for instance, echoing 'tulip' as 'crocus'). And we will be particularly bad at repeating nonword stimuli. This pattern of errors is indeed attested, in a condition called **deep dysphasia** (see e.g. Butterworth and Warrington, 1995). It is also possible to selectively damage the connections from the input lexicon to semantic representations, or from semantic representations to the output lexicon. For

instance, Kay and Ellis (1987) describe a patient who can understand concrete words, but has severe difficulties generating the same words (for instance in naming pictures or objects). Importantly, this patient is relatively good at repeating words—even those words he cannot spontaneously produce. This suggests that his output word representations are intact. Given that his semantic representations of concrete words are also intact, his word-production deficit appears to be due to damage to the connections from semantic representations to the output lexicon.

'Wernicke's area' appears to be particularly important in storing mappings between phonology and meaning, in both directions. The exact boundaries of Wernicke's area are not at all precisely defined; the core area is the area surrounding the primary auditory cortex in the left hemisphere—the left posterior STG area which has already been discussed. But some definitions also include parts of the superior temporal sulcus (STS) or the temporoparietal junction (see e.g. Wise, 2001). Classically, damage within these areas is associated with difficulties in both understanding and producing language. Speech production in fact tends to be fluent, and grammatical (a fact which I will discuss in Section 6.1.4.1), but it lacks meaning. Speech comprehension is very poor. The traditional explanation is that Wernicke's area is involved in mapping word meanings onto phonological forms and vice versa. This idea is still broadly accepted, but there are some important qualifications to note. Firstly, mappings from word meanings to phonological forms almost certainly involve regions outside Wernicke's area as well. For instance, as already discussed in Section 6.1.2, Damasio *et al* (1996) found that damage to higher-order temporal cortices can result in difficulties in naming objects, even when these objects can be successfully identified, and Lu *et al.* (2002) found that damage to anterior temporal cortex caused selective difficulties in naming actions. Lesions to temporal cortex can also give rise to difficulties in understanding single words which do not appear to be phonological in origin (see e.g. Hickok and Poeppel, 2004). Secondly, Wernicke's area (and associated temporal areas) are not the only areas where words as phonological entities are stored. As already discussed in Section 6.1.3.1, word-sized articulatory units are also formed in the phonological output system, which is implemented across left prefrontal and premotor areas as well as in (parts of) Wernicke's area. Since Wernicke's aphasics generate phonological words comparatively well, we must assume that these prefrontal/premotor areas are sufficient for producing word-sized units. The role of Wernicke's area during generation must be in selecting the phonological units which convey an intended meaning. Finally, not all of Wernicke's area is involved in implementing a mapping between phonological representations and meanings. Posterior parts of 'Wernicke's area' (in particular the temporoparietal junction) appear to be involved in the direct neural pathway from auditory to articulatory representations, which is involved in establishing the mirror system for phonemes (see Section 6.1.1.1) and in the phonological loop (see Section 6.1.1.2), which bypasses lexical

representations altogether. But note that while this pathway is not directly involved in representing words, it nonetheless plays a crucial role in word learning. A child learns word-meaning associations by *hearing* words and pairing them with concepts, but must be able to exploit this learning by *producing* the words associated with appropriate concepts. The mirror system for phonemes allows the phonology of heard words to be represented as articulatory gestures, and thus ensures that phonology-meaning associations which the child learns passively, through hearing words spoken by mature speakers, can subsequently be used to actively generate words.

6.1.3.3 'Speaking' and 'hearing' modes

As was discussed in Section 2.8, while action perception and action execution make use of the same premotor and intentional representations, the circuitry active during action perception has to be very different from that active during execution. I suggested that action representations can be activated in two modes: 'execution mode' and 'perception mode'. In execution mode, circuits are enabled which allow intentional representations in PFC to activate premotor representations, and allow these premotor representations to generate overt motor actions. In perception mode, circuits are enabled which cause premotor representations to be activated by perceptual representations, which gate shut the link from premotor to motor representations, and which allow 'abductive' inference from premotor action representations to the intentional representations in PFC which best explain the actions being observed. In execution mode the intentional representations in PFC precede premotor representations, while in perception mode the direction is reversed: premotor representations trigger intentional representations in PFC.

A similar concept of 'modes' may be needed in the neural system which maps between phonological and semantic word representations. If the agent has established perception mode, and is attending to another person who is talking, the direction of activation should be from phonological word representations to concepts, while if the agent is talking, the direction should be from concepts to words. We might thus need to postulate two different modes of phonological processing, **speaking mode** and **hearing mode**, and envisage special internal operations which allow an agent to 'enter' one or other of these modes. (When a child learns language by attempting to reproduce the utterances of other speakers, it may be that both modes must be engaged simultaneously, as I will discuss in Sections 6.3 and 6.4.)

6.1.4 The neural locus of syntactic processing

I will now review some general ideas about where 'syntactic processing' is carried out in the brain. Studies which investigate this question are hampered by the fact that no-one knows what 'syntactic processing' *is*—but obviously, we expect that experimental studies can help shed light on this question. I will begin in Section 6.1.4.1 by reviewing some basic patterns of language dysfunction which appear to suggest that syntactic processing happens in an area of prefrontal cortex called Broca's area, and surrounding prefrontal regions. In Section 6.1.4.2 I will discuss imaging evidence which bears out this claim, and suggests that different parts of Broca's area and associated prefrontal regions have different roles in producing syntactic competence. In the remaining sections, I will discuss some proposals about the neural mechanisms in these areas which allow them to play these roles. And in Section 6.1.4.7 I will discuss more recent evidence that anterior temporal cortex has a role in syntactic processing during sentence interpretation.

6.1.4.1 Broca's aphasia

Broca's area occupies the left inferior frontal gyrus. It is classically defined as comprising Brodmann areas 44 and 45 (respectively, the pars opercularis and pars triangularis of the third frontal convolution). It is part of prefrontal cortex, though its posterior part neighbours premotor cortex. Traditionally, damage to Broca's area is associated with a condition called 'Broca's aphasia', which I will discuss in this section. However, it is important to point out that the pattern of symptoms considered typical of Broca's aphasia is now associated with an area extending well beyond Broca's area, encompassing other adjacent prefrontal regions (see e.g. Alexander *et al.*, 1990; Bookheimer, 2002). When additional prefrontal regions are involved, aphasia becomes more chronic, and more pronounced.

The classical symptoms of Broca's aphasia are nonfluent delivery of sentences, generation of agrammatical sentences, and abnormal repetition of words, but with relatively preserved ability to generate individual words (see e.g. Saffran, 2000). Ability to comprehend speech is also relatively preserved. However, it has been known for some time that damage to Broca's area also causes difficulties in correctly interpreting sentences with complex or non-canonical word orders (see e.g. Zurif *et al.*, 1972; Hickok *et al.*, 1993). The problems here are often in assigning appropriate thematic roles to entities in a sentence, rather than in recovering the meanings of individual words. The preservation of word meanings, together with difficulties in generating or interpreting syntactic structure, suggest that Broca's area plays a role in encoding 'specifically syntactic' knowledge or mechanisms. What these are has been a matter of considerable debate—partly because

the pattern of deficits in Broca's aphasia is very variable (much more so than is suggested by my summary above), and partly because there is still much disagreement as to what syntax is, as already mentioned.

The pattern of deficits found in Broca's aphasia is very different from that associated with damage to Wernicke's area. As already discussed in Section 6.1.3.2, in Wernicke's aphasia, speech tends to be fluent, and grammatical, but it lacks meaning. There are also more serious deficits in sentence comprehension than in Broca's aphasia. Broca's aphasics can often identify the meanings of individual words, but Wernicke's aphasics are typically unable to do this. (See Brookshire, 1997 for a summary of these deficits.) Wernicke's area, and associated areas in posterior temporal cortex, appear to be involved in implementing our knowledge of words as mappings from phonogical/articulatory representations to semantic representations. The important thing to note here is that the apparent preservation of syntax in Wernicke's aphasia supports the view that syntax is a neurally autonomous mechanism, which does not rely on knowledge of the meanings of words.

6.1.4.2 The components of syntactic competence

In order to perform 'syntactic processing', whether during sentence generation or sentence interpretation, a number of distinct capabilities are required. The agent must be able to maintain complex semantic representations—for instance episode representations—in working memory. (During generation there only needs to be one of these; during interpretation, there are proabably several alternative candidates.) He must be able to maintain a sequence of words in phonological working memory. (During generation, words are probably queued in the phonological output buffer; during interpretation, the phonological input buffer is probably involved as well.) And he must be able to build 'syntactic structures', which map between these phonological and semantic working memory representations. Given the syntactic deficits associated with Broca's aphasia, we expect to find that some or all of these functions involve Broca's area or associated prefrontal regions. We have already seen evidence that working memory episode representations are held in dorsolateral prefrontal cortex (see e.g. Section 3.5.1). We have also seen evidence that the phonological output buffer is implemented in Broca's area and adjacent articulatory premotor cortex (see Section 6.1.1.2). In this section I will discuss some neuroimaging experiments on sentence processing which corroborate and extend this evidence, and help to isolate the role played by Broca's area in syntactic processing.

BA 44 and syntactic processing The hypothesis that Broca's area is involved in syntactic processing has been tested in imaging experiments in which subjects are presented with sentences of varying degrees of syntactic complexity. For instance, Stromswold *et*

al. (1996) found that sentences featuring centre-embedded relative clauses (e.g. *The juice that the child spilled stained the rug*) activate Broca's area more than sentences featuring right-branching relative clauses (e.g. *The child spilled the juice that stained the rug*). The area involved is BA 44, the pars opercularis of Broca's area, a finding replicated by Caplan *et al.* (1998). Note that this manipulation of 'syntactic complexity' is also a manipulation in working memory requirements. As discussed in Section 6.2.5.4, processing a centre-embedded construction requires more working memory than an equally deeply nested right-branching construction. Fiebach *et al.* (2001) conducted a sentence interpretation experiment in which working memory demands were varied independently of sentence complexity; they found that BA 44 varied according to syntactic working memory rather than according to syntactic complexity.

BA 44 is also involved in the generation of syntactic structures. In an elegant experiment by Indefrey *et al.* (2001), subjects were asked to describe a simple spatial scene using a syntactically complete sentence (e.g. *The red square launches the blue ellipse*), a fixed pattern of noun phrases (*red square, blue ellipse, launch*) or a fixed pattern of individual words (*square, red; ellipse, blue; launch*). An area posterior to, and partly overlapping with, BA 44 responded differently to these conditions. It was most active during generation of complete sentences, least active during generation of a pattern of single words, and active at an intermediate level during generation of a pattern involving whole noun phrases. These results suggest that this area is involved in the creation of syntactic structures—and moreover, that it is involved in the creation of local structures as well as of complete sentences. This posterior area of BA 44 seems distinct from the central region involved in syntactic working memory. We might tentatively conclude that there are separate neural substrates for the mechanisms which construct syntactic phrases and the working memory mechanisms involved in processing embedded clauses.

BA 45/46/47 and representations of nouns and objects Even if syntactic knowledge is relatively distinct from knowledge of word meanings, syntactic processing must still involve the manipulation of word meanings. As discussed in Section 4.1, a model of syntax must specify how the meanings of individual words in a phrase are combined together to create the meaning of the whole phrase. We therefore expect to find neural areas involved in syntactic processing whose function is to evoke or access the semantic representations of individual words. There do indeed appear to be such areas in prefrontal cortex. Moreover, the areas involved in representing noun semantics and verb semantics seem to be somewhat different.

Interesting evidence for the representation of noun semantics in PFC comes from a study by Demb *et al.* (1995). Subjects were asked to study nouns in two conditions: one

involving semantic processing (a decision about whether the noun was concrete or abstract) and the other involving 'superficial' processing (a decision about whether the noun was presented in upper or lower case). The former task differentially activated several PFC areas: Brodmann's areas 45, 46 and 47. Moreover, these same areas showed a sensitivity to the phenomenon of semantic priming. Subjects performed the semantic task faster for words they had already seen (in the semantic task). Activation of areas 45–47 was lower for repeated words in the semantic task than for words presented for the first time, suggesting that these areas are involved in producing the priming effect. Several similar findings are summarised in Bookheimer (2002). BA 45 is the anterior part of Broca's area. There is fairly good consensus that BA 47, which is anterior to BA 45, has a role in language processing, and should be included in the network of language-related frontal areas (see e.g. Hagoort, 2005). BA 46 is another prefrontal area which is implicated in syntactic processing (see e.g. the neuropsychological survey in Dronkers *et al.*, 2004).

Additional evidence that these PFC areas are involved in working memory object representations comes from Ungerleider *et al.* (1998). These researchers note that the inferotemporal cortex (representing object categories) projects to BA 45, while parietal cortex (representing object locations) projects to BA 46, and show fMRI evidence that working memory representations of object category and object location are preferentally maintained in BA 45 and BA 46 respectively.

Note that these these prefrontal areas appear to hold semantic representations, rather than purely phonological ones. Wagner *et al.* (2001) gave subjects lists of three nouns to maintain in working memory. In one condition, they simply had to covertly rehearse the words as a phonological sequence. In the other condition, called 'elaborative rehearsal', they had to reorder the words from least to most desirable. This latter task involved accessing the semantics of the words, and selection of particular words in order to reorder the list. Both conditions activated posterior parts of Broca's area, extending into premotor cortex; these are the areas involved in the phonological output buffer. But elaborative rehearsal preferentially activated BA 45, 46 and 47. Gough *et al.* (2005) found corroborating evidence for a distinction between phonological and semantic processing in Broca's area, in a transcranial magnetic stimulation experiment. They found that TMS over the posterior part of Broca's area selectively impaired performance of a phonological task (judging whether two written words sound the same) while TMS over anterior Broca's area and the PFC areas anterior to this selectively impaired performance of a semantic task (judging whether two written words have the same meaning).

The question of whether these prefrontal areas hold information about objects or information about nouns is more controversial. But there is certainly some evidence that morphological processing of nouns happens in these areas. Shapiro *et al.* (2000) describe a patient with damage to these areas (as well as to Broca's area quite generally) who

has more difficulty producing nouns than verbs, and has more difficulty producing plural inflections on nouns than on verbs. Interestingly, the patient had difficulty producing inflections on pseudowords if they were presented as nouns (e.g. *a wug* vs *two wugs*), but less difficulty producing inflections on these same pseudowords if they were presented as verbs (e.g. *you wug* vs *he wugs*). In summary, there is some evidence that BA 45, 46 and 47 are particularly involved in evoking the semantic representations of nouns and in generating the grammatical inflections of nouns, though they also have a non-linguistic function in evoking working memory representations of objects. Of course, areas in inferotemporal cortex are also activated when processing concrete nouns denoting objects, as already discussed in Section 6.1.2. But there also appear to be areas in prefrontal cortex specialised for the processing of nouns as syntactic objects.

Dorsolateral PFC, verbs and inflections There are also areas of PFC which seem particularly involved in evoking the semantics of verbs. Generating verbs appears to involve a large region of left anterior PFC, extending over BA 45/46, but also extending more dorsally to dorsolateral PFC, BA 9 (see e.g. Perani *et al.*, 1999 and the papers cited in Tranel *et al.*, 2001). We have already seen evidence that dorsolateral PFC is involved in storing working memory episode representations (see Section 3.5.1). Given that verbs play a central role in creating the syntactic structures which portray episodes, we might expect verbs and working memory episodes to be represented in the same neural area. In fact, this is an explicit prediction of my sensorimotor interpretation of LF structure. A verb can have inflections, agreeing with its subject and its object. In this sense, the syntactic domain of a verb extends over a whole clause. In Minimalism, as discussed in Section 4.5, a fully inflected verb is generated at the head of VP, but it must raise to the heads of AgrP and IP to 'check' its inflections. In Section 5.4.3 I suggested that verb movement might reflect the fact that verbs are read out from working memory episode representations which are tonically active while an episode is being rehearsed. This hypothesis predicts that the neural areas which hold working memory episode representations will also be involved in accessing inflected verbs.

This does indeed seem to be the case. There is interesting evidence from neuropsychology that BA 9 is selectively involved in producing morphologically complex verbs. Shapiro *et al.* (2003) reported a patient with damage to left prefrontal cortex, who had more difficulty generating morphological inflections of verbs than of nouns. This selective deficit extended to nonwords (thus the inflection on *he wugs* was harder to produce than the inflection on *the wugs*), and was also found for words which function both as nouns and verbs (thus the inflection on *he judges* was harder to produce than that on *the judges*). This is the opposite pattern of performance from the patient reported in Shapiro *et al.*

(2000). The patient with a selective impairment for verb morphology had damage to the dorsolateral PFC, Brodmann area 9, which was preserved in the patient with a selective impairment for noun morphology. Shapiro *et al.* (2001) also found evidence for category-specific morphological processes in PFC using a TMS paradigm on normal subjects. They found that TMS centred on an area of PFC anterior and superior to Broca's area (including portions of BA 9) disrupts the generation of inflections on verbs, but not the generation of inflections on nouns. The effect is found for pseudowords as well as actual words. In summary, there is evidence that a network of anterior prefrontal regions, in particular BA 9, is responsible for holding not just working memory episode representations, but is also involved in the generation of verbs and their inflections.

Summary Syntactic competence involves a range of different abilities. An agent must be able to hold phonological sequences in working memory, to hold episode representations in working memory, to evoke object and action representations and map these onto appropriately inflected nouns and verbs, to construct phrases, and to represent the kind of syntactic dependencies found in sentences with embedded clauses. The evidence I have just reviewed suggests that these different abilities are to some extent associated with different areas in left prefrontal cortex. These areas include the classically defined Broca's area, but also extend well beyond this area.

What are the neural mechanisms implemented in these prefrontal areas? In the remainder of this section, I will consider some of the hypotheses which have received most attention, which are mostly proposals about what happens in Broca's area.

6.1.4.3 Broca's area and syntactic movement

One idea which derives from syntactic theory—in fact from generative grammar (circa Chomsky, 1981)—is that Broca's area is involved in the processing of syntactic structures which require 'movement' of constituents (see Grodzinsky and Santi, 2008 for a review). The key evidence comes from the claim that sentences in which the noun phrases have been moved, so that they appear in a non-canonical order, are harder for Broca's aphasics to comprehend. Examples of movement include passives (e.g. *The cat$_i$ was chased t_i by the dog*), where the subject is analysed as moving from an underlying object position) or relative clauses (e.g. *The cat$_i$ that the dog chased t_i*), where the head noun is analysed as moving from an underlying object position.[8] The suggestion that Broca's aphasics have particular problems with these sentences is interesting, though the data are still controversial (see e.g. Caramazza *et al.*, 2005). However, the issue of what mechanisms in Broca's

[8]Note that these forms of movement are different from the case-assigning DP movement we have been focussing on. I will consider them later in the book.

area might be responsible for this specific deficit is still very unclear. One difficulty is that syntactic movement as conceived within generative syntax is not a processing operation: it is simply one of the formal operations defining the 'generative mechanism' which produces the space of well-formed sentences in a language. If a processing account of movement can be provided, this may help to assess Grodzinsky's proposal about the function of Broca's area.

6.1.4.4 Broca's area and general serial cognition

Broca's area is part of prefrontal cortex, a brain region which is associated with general working memory, and with working memory for sequences in particular, both in humans and primates (see Section 3.2).

Several theorists have suggested that our syntactic ability is in some measure an ability to process sequences of items in working memory, and that the syntactic operations performed by Broca's area are sequencing operations. A straightforward identification of syntactic competence with sequence-learning ability is obviously too simplistic—and indeed it has been shown by Goschke *et al.* (2001) that Broca's aphasics are capable of learning simple nonlinguistic sequences. However, several variants on the basic sequence-learning hypothesis have been proposed. Conway and Christiansen (2001) distinguish between simple sequence-learning mechanisms and more complex mechanisms which can learn hierarchical structure in sequences. (We will consider such mechanisms in Section 6.2.5.4.) They suggest that non-human animals only show the former type of sequence learning, and that the modification in Broca's area which allows language is one which allows learning of sequences with hierarchical structure. Dominey *et al.* (2003) suggest that a key ability in syntactic competence is the ability to sequence abstract symbols, which stand for arbitrary objects, and thus permit the recognition of sequences with particular abstract forms, as well as sequences of tokens. They find evidence that Broca's aphasics are impaired on this more complex sequencing task, though not on simple token-sequencing tasks. Hoen *et al.* (2006) found fMRI evidence that the complex sequencing task activates BA 44, but not BA 45, and consequently Dominey *et al.* (2006) present a model of sentence interpretation in which BA 44 is involved in representing abstract sequential patterns using a recurrent network, while BA 45 is involved in selecting the lexical items which feature in these patterns.

6.1.4.5 Broca's area and combinatorial sensorimotor mechanisms

There are some theorists who suggest that the syntactic capacity in Broca's area derives from its original use in the control and perception of motor actions. Rizzolatti and Arbib

(1998; see also Arbib, 2005) argue that Broca's area is the human homologue of area F5 in the macaque, the area where mirror neurons are found. They suggest that the mechanism for combining object and action representations in F5 may be the origin of a more general mechanism for composition of concepts, as required by natural language syntax. In humans, there is some evidence that Broca's area and associated frontal areas have a role in general (i.e. non-linguistic) motor cognition. For instance, Müller and Basho (2004) found that left prefrontal cortex is activated by nonlinguistic sensorimotor and working memory processes as well as by linguistic processes. There is also some interesting evidence from dysfunction which bears quite directly on Rizzolatti and Arbib's hypothesis. Broca's aphasia commonly co-occurs with **ideomotor apraxia**, a condition where patients are unable to accurately imitate hand gestures observed in other agents, and unable to 'pantomime' common hand actions involving tools in the absence of these tools (see e.g. Papagno *et al.*, 1993). Imitation is very naturally analysed in terms of the mirror system framework; prima facie, a correlation between imitation deficits and syntactic deficits appears to support for Rizzolatti and Arbib's proposal. In fact, there are clear cases ideomotor apraxia occurs without aphasia, and vice versa (see again Papagno *et al.*, 1993). This suggests that the associated skills involve separate but adjacent frontal regions, rather than a single region. However, the neuroanatomical proximity of the relevant regions still provides some support for a model in which syntactic mechanisms derive from the motor mirror system through copying of circuitry, rather than straightforward co-opting (Arbib, 2006).

Rizzolatti and Arbib's proposal focusses on the compositional properties of natural language syntax, rather than on hierarchical structures or sequencing. But it might nonetheless relate to the general sequencing capabilities described in Section 6.1.4.4. Action representations in F5 are strongly sequential—for instance in Fagg and Arbib's (1998) model, motor schemas in F5 and antecedent regions hold prepared motor sequences.

In a similar vein to Rizzolatti and Arbib, Ullman (2004) suggests that the frontal areas which include Broca's area are specialised for representing 'procedural' memories as opposed to declarative ones. Procedural memory holds information about how to perform cognitive or motor tasks. This information tends to be acquired through practice, and tends to be implicit, i.e. hard to verbalise. Again, this characterisation of Broca's area is likely to overlap with the characterisation which emphasises sequential mechanisms, since procedural memory often involves the encoding of sequences of cognitive or motor actions.

6.1.4.6 Broca's area and cognitive control

As discussed in Section 2.6.2, prefrontal cortex is known to be involved in representing an agent's current task set, and in switching between different task sets. In Miller and Cohen's

2001 model of PFC, PFC delivers a modulatory bias on the pathways linking stimuli to responses, allowing stimuli to be responded to in some circumstances and ignored in others. Some theorists have suggested that the syntactic mechanisms in Broca's area are involved in high-level control of syntactic processing. In particular, Novick *et al.* (2005) suggest that Broca's area is involved in initiating 'backtracking' during sentence interpretation, in situations where the expected syntactic analysis of an initial sequence of words is found to be inconsistent with its continuation. The basic idea that Broca's area is involved in the overriding of habitual response behaviours is supported by several studies. Particularly interesting is an fMRI study by Thomson-Schill *et al.* (1997) which suggests that Broca's area is involved in the 'selection' of semantic representations, in situations where several alternative representations are evoked. In Miller and Cohen's model, PFC was involved in 'selecting' an explicit motor action appropriate for the circumstances. It may be that Broca's area is involved in selecting appropriate internal cognitive actions, or cognitive representations, during language processing.

6.1.4.7 Syntactic processing in anterior temporal cortex during sentence interpretation

While syntactic processing is traditionally associated with Broca's area, recent work has found that there are other areas involved, in particular during interpretation of sentences. As mentioned in Section 6.1.4.1, Broca's aphasics are more impaired in generating sentences than in interpreting them. They have difficulties interpreting syntactically complex sentences, but for simple sentences, they appear able to use quite effective heuristics, which have their origin in world knowledge and knowledge of the canonical word orderings in the language (see e.g. Hickok *et al.*, 1993). We must therefore assume the existence of a 'shallow' method for interpreting sentences based on surface word orderings, which survives even after damage to Broca's area.

One interesting candidate for this area is anterior superior temporal cortex, a region of temporal cortex adjacent to Broca's area. This area has been found to be more activated by syntactically well-formed utterances than by lists of words (see e.g. Friederici *et al.*, 2000); Vandenberghe *et al.*, 2002). It was also found to be more activated by sentences reporting events than by nonlinguistic auditory stimuli depicting events (Humphries *et al.*, 2001). Damage to the area is also associated with syntactic interpretation deficits. In fact, in a large study of patients with brain injuries, Dronkers *et al.* (2004) found that syntactic interpretation deficits were more strongly associated with damage to this area than with damage to Broca's area. The model of syntactic processing which I present in Section 6.4 will be a model of generation rather than interpretation. However, I will include a brief discussion of sentence interpretation, and the role of the anterior superior temporal cortex,

in Section 6.4.5.

6.2 The basic stages of language development

In this section, I will review what is known about how children acquire language. It is important that a model of language can explain adult linguistic competence, but it must also be able to support an account of how this mature competence is acquired—a process which occurs gradually. To review this process without too much theoretical commitment, I will discuss five successive 'stages' of language development: a preliminary stage (Section 6.2.1); a stage where individual word meanings are acquired (Section 6.2.2), a stage where single-word utterances are produced (Section 6.2.3), a stage where simplified syntactic constructions are used (Section 6.2.4), and a stage where mature syntax is mastered (Section 6.2.5).

6.2.1 Preliminaries for word learning: phonological word representations and sensorimotor concepts

Recall that there are two ways of characterising words: they can be thought of as regularly occurring sequences of phonemes (see Section 6.1.3.1), or as mappings between such sequences and semantic concepts (see Section 6.1.3.2). In order to begin learning words in the latter sense, as associations between phonological forms and concepts, an infant must already have representations of words as phonological entities on the one hand, and concepts on the other.

There is good evidence that infants can form representations of words as phonological sequences by the time they are 8 months old. In a well-known study, Saffran *et al.* (1996) acclimatised 8-month-old infants to a stream of phonemes in which some subsequences (e.g. *ga, bi, ro* or *to, ba, di*) occurred much more often than chance. After two minutes of acclimatisation, infants were presented with a mixture of high-probability and low-probability phoneme sequences. They showed longer listening times for the low-probability sequences, indicating some measure of habituation to the high-probablility ones. This shows that even after short exposures, 8-month-olds can detect short, regularly occurring phonological sequences—i.e. that they are beginning to represent words as phonological entities at this age. Moreover, 8-month-olds can also generate short phonological sequences. 'Babbling' begins at around 6 months; at this age, infants can produce repeated sequences of a single syllable (e.g. *bababa*). By 8-9 months, infants start to produce short sequences of varied syllables (see e.g. Werker and Tees, 1999).

There is also good evidence that infants have developed a simple set of sensorimotor 'concepts' by the age of 8 months. By this age, infants can perform simple reach-to-grasp actions (see e.g. Dimitrijevi and Bjelakovi, 2004), and can also recognise such actions being performed in others, deriving quite sophisticated representations of the agent's intentions (see e.g. Woodward, 1998) and of the support and contact relations which are achieved by grasping (see e.g. Leslie, 1984; Needham and Baillargeon, 1993; Baillargeon and Wang, 2002). They also have fairly well developed concepts of physical objects; they have a good understanding of the spatiotemporal continuity of objects (see e.g. Spelke *et al.*, 1994; 1995), and can recognise objects by their shape in a variety of poses (see e.g. Ruff, 1978).

In summary: by the age of 8 months, infants have sufficiently well-developed representations of words as phonological entities, and of concrete object and action categories, to be able to begin learning associations between these representations—in other words, to begin acquiring words as mappings between phonological forms and meanings.

6.2.2 Learning the meanings of individual words

The rate at which infants learn the meanings of words depends on several factors. There are important differences between comprehension and production of words: production lags behind comprehension, especially to begin with. There are differences due to country and to socioeconomic status; there are also large differences between individual infants. The data given in Figure 6.5 comes from a well-known study of US infants from a representative range of backgrounds (Fenson *et al.*, 1994). The 8–16 month data relate to an inventory of 396 common words, and the 20–28 month data relate to an inventory of 680 common words. Word comprehension was only assessed at the younger age range; after 18 months passive vocabularies grow so fast that it becomes quite hard to assess their size using a standard inventory of words.

Between 8 and 12 months, infants acquire a small passive vocabulary. At 8 months, infants can typically understand between 5 and 50 words, and at 12 months, they can typically understand between 50 and 110 words. Their active vocabulary is very small during this period: from 0 to 2 words at 8 months, and from 3 to 18 words at 12 months. After 12 months, spoken vocabulary begins to grow sharply: at 18 months, infants are producing between 35 and 100 words, and by 28 months, children are producing between 315 and 570 words.

In the remainder of this section, I will review the dominant models of how word meanings are learned.

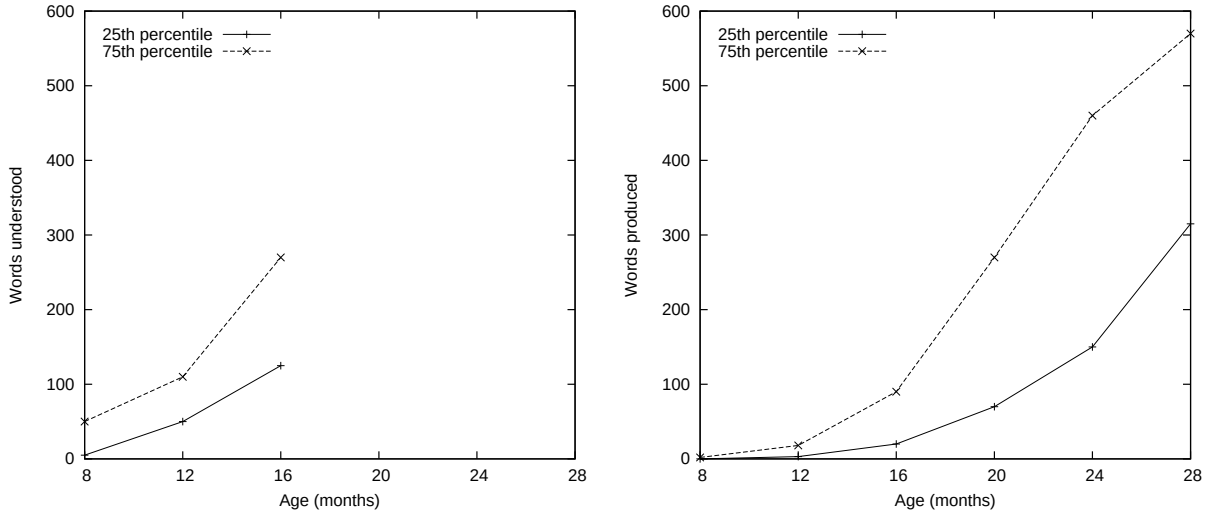


Figure 6.5: Vocabulary size (words understood and words produced) in infants from 8 to 24 months (data from Fenson *et al*, 1994).

6.2.2.1 Cross-situational learning

The simplest models of word-meaning learning are straight associationist accounts: children use general-purpose learning skills to learn mappings from concepts to phonological representations. Associative learning models trade on the fact that mature speakers often talk about perceptually accessible elements in their current environment. Since infants sample their perceptual environment, this means there are statistical correlations between the words that infants hear and the concepts they experience. Learning algorithms which acquire word-meaning mappings from these statistical correlations are called **cross-situational** algorithms; well-known computational models include those of Plunkett *et al.* (1992) and Siskind (1996). In fact, since words are very transitory perceptual stimuli, it is still rare for an infant to hear a word at exactly the time when its associated concept is being evoked, so learning directly from cross-situational correlations is extremely inefficient. As mentioned in Section 6.1.1.2, cross-situational algorithms tend to assume a phonological buffer of n words, and to strengthen associations between all of these words and the currently evoked semantic concept. This improves their efficiency considerably. But they are still very inefficient techniques.

It seems plausible to assume that an infant's initial passive vocabulary is acquired using this slow form of learning. However, a theory of word-learning must also account for the

fact that infants do not begin to *produce* words in any numbers until around 12 months, as well as for the significant increase in the rate at which children learn to speak words from 12 to 24 months. A straight associationist account has difficulties with both these findings. As noted by Tomasello (2003), infants demonstrate associative learning in many modalities from birth. By 8 months they have developed a repertoire of semantic concepts, and can represent words as phonological entities. They can also produce phonological words in 'babble'. A pure associationist account would predict that words start to be understood and produced at 8 months. But it is only at 12 months that infants begin to use words in any numbers. Many models of word learning therefore postulate that specialised mechanisms for acquiring words begin to come online around 12 months.

One of the most important proposals (see especially Tomasello, 2003) is that mature word-learning requires the development of a set of *social*, or *interpersonal* skills. I will discuss these skills, and their postulated role in vocabulary development, in the remainder of this section.

6.2.2.2 The role of joint attention in word learning

One interpersonal skill which has been accorded a role in word learning is the ability to establish joint attention—i.e. to attend to where an observed agent is looking. We have already discussed the important role of joint attention in action recognition; see Section 2.7.2. There is also good evidence that establishing joint attention is important for learning words. In a well known series of experiments (Baldwin, 1991; 1993; Baldwin *et al.*, 1996) it was found that at 16–19 months of age, infants will follow the gaze of an observed speaker to determine the referent of a new word the speaker is using. It has also recently been found that children's ability to establish joint attention at 12 and 18 months is predictive of their language ability at 24 months (Mundy et al., 2007). It makes sense for infants to establish joint attention with a speaker when associating the speaker's words with their own percepts; this tactic makes the mapping from words to percepts less noisy. (Of course there will still be noise: speakers do not always talk about what they are looking at, and even if they are, gaze direction does not unambiguously determine a single concept.) The capacity to achieve accurate joint attention develops gradually over the first year of life, becoming reliable in simple contexts at around 10 months (Scaife and Bruner, 1975), but in more complex contexts only around 18 months (see e.g. Butterworth and Jarrett, 1991). This might be one reason why word learning becomes more efficient during the second year of life.

6.2.2.3 The role of communicative intention recognition in word learning

The other key interpersonal skill Tomasello suggests is crucial for word learning is an ability to recognise the *communicative intentions* of other agents. We have already considered some general models of how an observed agent's intentions are inferred, in our discussion of action recogntion (see Section 2.7.6). In this section I will review how infants' ability to recognise intentions develops, and I will discuss the special case of communicative intentions, which Tomasello accords a special role in vocabulary development.

There is good evidence that infants have some awareness of the intended targets of observed reach-to-grasp actions from an early age, around 6 months. For instance, Woodward, 1998 showed 6-month-olds a video stimulus of a human hand reaching for one of two possible targets. In a habituation phase, infants were shown the hand repeatedly for the same target in the same position. The location of the targets was then swapped, and a new reach was shown. This reach could either be to the new target (i.e. to the same location as in previous reaches), or to the new location (i.e. to the same target as in previous reaches). Infants looked longer in the first of these cases, indicating that it was understood as a more novel event. Interestingly, this did not happen if the object moving towards the target in the habituation phase was a stick, rather than a human hand. These results suggest that 6-month-olds represent reach-to-grasp hand movements by encoding the identity of the target object more prominently than details of the movement trajectory.

Of course, the above experiment does not tell us that infants *explicitly* represent the intentions of an observed agent. Tomasello suggests that a more explicit representation of intentions is necessary in order to use language. He suggests that mature word learning only begins when an infant can recognise that an observed agent *wants to tell him something*—in other words can recognise communicative intentions in other agents. I will first outline Tomasello's proposal about what it means to explicitly represent a communicative intention, and then discuss the evidence he presents that communicative intention recognition is a prerequisite for word learning.

Tomasello suggests that representing the communicative intention behind an utterance involves representing relations between three entities in the world: the speaker, the intended hearer, and the thing in the world which the utterance will describe. In the case of early word learning, the utterance can be thought of as an individual word (perhaps extracted from a longer utterance), the 'thing in the world' will be an object or an action, and the speaker's communicative intent is to *direct the hearer's attention* to this object or action, so that both speaker and hearer are evoking the same concept. Tomasello notes that when an infant establishes joint attention with an adult, the infant and adult get into something like this state, in that they both end up evoking the same concept. However, a communicative scenario differs in the way this state is brought about. In communication,

the speaker *actively* directs the hearer's attention to a particular concept, i.e. brings about joint attention volitionally, rather than incidentally. And in communication, the jointly-attended concept is evoked through a mapping with the phonological properties of the utterance, rather than through joint attention to an object in the world.

Tomasello suggests that before infants can start learning the meanings of words, they must first learn that people use words to direct attention—i.e. that words have 'meanings' (or at least pragmatic effects). The general principle which an infant must learn might be expressed as follows: if an observed agent makes an utterance, this is a signal to move into a special cognitive mode where concepts should be evoked by words rather than by sensory experience. Note that if regular perceptual joint-attention situations are *accompanied* by utterances, they provide opportunities for infants to learn this attention-direction function of utterances. In summary: while the ability to recognise the special attention-directing function of communicative utterances goes some way beyond the simpler perceptual ability to establish joint attention, the simpler joint-attention mechanism is likely to play an important role in acquiring the more complex idea that utterances have communicative functions.

When do infants develop the concept of communicative actions? While infants begin to follow gaze quite early in their first year, they only begin following pointing gestures at around 11 or 12 months (Butterworth, 2006:224; Carpenter *et al.*, 1998). At around 14 months, there is evidence that infants are able to make use of the nonverbal communicative actions they see. In a study by Behne *et al.* (2005), infants were given the task of discovering where an object was hidden. They were helped by an experimenter making 'communicative gestures' indicating were to look; these gestures included pointing, with gaze alternating between the infant and the pointed-at object, and simple gaze alternation between the infant and the hiding place. Infants as young as 14 months were able to interpret these gestures, with performance improving at 18 months, and also (somewhat less) at 24 months. Gestures which were not overtly communicative—looking at the target with an 'absent-minded gaze', or 'distractedly pointing' while looking at the pointing hand—did not influence the infants' behaviour.

Interestingly, chimpanzees are not able to recognise overtly communicative gestures in a similar search paradigm, whether they involve pointing or looking (Hare and Tomasello, 2004). Chimpanzees are able to follow gaze (see e.g. Call *et al.*, 2004). They are also sensitive to the intentions of other agents. We saw in Section 2.7 that there are neural signals in macaques which encode the 'expected actions' of an observed agent. Great apes can use intention recognition to help decide how to act; for instance, if a human experimenter who is competing with an ape to find food *reaches* for a hiding place, the ape will reach to the same location (Call, 2004). Nonetheless, it seems that specifically *communicative* gestures—the kind which are executed cooperatively, rather than competitively—are not

understood. Tomasello suggests that an ability to recognise specifically communicative actions and their underlying intentions may be what sets humans apart from other primates, and proposes that this ability is responsible in fairly large measure for our special linguistic abilities.

6.2.2.4 Neural mechanisms underlying Tomasello's model of word learning: some open questions

Tomasello's model of word learning does not make reference to neural machinery. If we want to make contact with neural mechanisms subserving language, there are several important questions to address. I will consider two of these.

Firstly, how are communicative intentions represented in the brain? Tomasello describes informally what a communicative intention is—an intention to communicate some piece of content to another agent. But the issue of how to express this intention is still an open one in cognitive science. One key question we must consider before we even begin to think about neural representations is how to characterise the relationship between an utterance and its content; see Perner (1991) for a good introduction to the complex issues surrounding this question. An utterance is a physical action, made by an agent (the speaker) to another agent (the hearer). Its content is a representation of the world. To adopt a communicative intention is to adopt the intention that a certain agent in the world has a certain representation of the world. For instance, a speaker might intend that a certain agent *believes* a certain proposition.[9] Or, in our more simple referential scenario a speaker might intend that a certain agent *entertains* a certain concept. Expressing such intentions requires not just an ability to represent the world, but also an ability to attribute representations of the world to particular individuals. Another key question is how propositions are represented in the brain. This is a daunting enough question for 'simple' propositions which make no reference to mental states ('it is cold', 'I am happy' etc). But to implement Tomasello's model of word learning we must go beyond this, and formulate some idea about how propositions *about the mental states of others* are represented. Of course, 12-month-old infants' concepts of mental states are far from fully developed. (For instance, the ability to attribute counterfactual beliefs to other agents does not appear to emerge until around 4 years of age—see e.g. Wimmer and Perner, 1983). But we certainly need to make a suggestion about how an infant represents the special relationship between an utterance and its content.

Secondly, what is the neural mechanism via which an ability to represent communicative

[9]Or more elaborately: the speaker might intend that the agent believes *that the speaker believes* the proposition. There are many formulations of the necessary idea of mutual belief—see e.g. Cohen and Levesque (1990b; 1990a).

intentions allows an infant to efficiently acquire word-concept mappings? In Tomasello's account, the importance of communicative intention recognition for word learning is adduced experimentally, from the observations that infants learn to recognise communicative intentions at around the age that word learning gathers momentum, and that apes do not appear able to recognise communicative intentions at all. If communicative intention recognition is causally involved in acquiring word meanings, then recognising a communicative intention in a speaker must put an infant into a special state which encourages word-meaning mappings to be learned. Are there specialised neural mechanisms which support this process? Or does it result from domain-general intelligence and reasoning? In either case, further details must be given.

In Section 6.3 I will propose a model of word learning which gives speculative answers to both the above questions, drawing on the notion of intentions and working memory event representations developed in Chapters 2 and 3, and on the ideas about mental state representations briefly discussed in Section 5.6.

6.2.3 Infants' earliest single-word utterances

Once infants have learned to recognise communicative intentions and have begun to learn word meanings, they can begin to produce words themselves. In order to describe infants' active use of words, the concept of communicative intentions is again central, but now it is the communicative intentions of the infant which are at issue. As Tomasello (2003) notes, infants' first spoken words are uttered in the service of particular communicative goals. Effectively they are **holophrases**: single-word utterances which convey complete propositional messages.

Infants' earliest single-word utterances occur at around 12 months of age, as already shown in Figure 6.5. These utterances have a range of declarative and imperative functions. For instance, an infant might produce the word *car* to draw attention to a toy car, or to request the car. Single-word utterances can also be used to attribute properties to objects (*wet!*), to describe or request events (*up!*, *more!*) or simply to engage in social behaviour (*byebye!*).

In fact, infants' earliest overtly communicative actions are nonlinguistic. Pointing is a good example. At around 11 months, infants begin to point in order to communicate (see e.g. Butterworth, 2006:227). From 11 months, pointing gestures reliably have a communicative intepretation: either declarative ('look at that!') or imperative ('get me that!').[10] These nonlinguistic communicative actions provide a platform for the development of lin-

[10]The physical gesture of pointing is attested before 11 months, but these early uses of the gesture are more to do with the development of hand-eye coordination than with communication.

guistic communication. One of the skills required in order to use words communicatively is the ability to map a communicative intention onto a communicative action. This skill is already present in infants who can use pointing gestures for declarative or imperative purposes. Thus when infants learn that uttering words can achieve communicative goals, they are extending an existing repertoire of communicative behaviours, rather than discovering communication for the first time.

Of course, for linguistic communicative actions, the infant also needs to know the relevant word meaning. Producing a word involves making a decision 'to say something', but also a separate decision about which word to produce (or which sequence of words, for a mature speaker).[11] These two decisions rely on different forms of learning. The decision to say something relies on instrumental learning: learning that saying things can be effective in achieving communicative goals. The decision about about which word to produce relies on the declarative learning of an association between meanings and words. What binds these two decisions together is the infant's communicative goal itself. The communicative goal triggers the pragmatic decision to make an utterance. It also serves as the semantic representation from which words are generated.

How do infants represent communicative goals? While this is still an open question, as discussed in Section 6.2.2.4, it is reasonably uncontroversial to suppose that an infant's communicative goal has compositional structure, even if the associated utterance does not. For instance, suppose that an infant's goal is to make her interlocutor give her a cup which is out of reach. Representing this goal must involve representing the desired action, using an encoding in which the component concepts 'give', 'cup' and 'me' can be individually distinguished. Note that the holophrase produced in service of this goal is often an expression which in the adult language denotes one specific component of the goal representation. (For instance, the holophrase *cup!* uttered by the infant to acquire the cup, denotes the the desired object in the adult language.) Of course this is no surprise; it is the reason why the infant's intention can be recognised by the adult. But it does show that the declarative mappings between meanings and words which the infant acquires from adults are not *themselves* 'holophrastic'. As described in Section 6.2.2, these mappings are associations between individual sensorimotor concepts and individual words. What is holophrastic is an infant's *use* of single words to achieve goals. The stage is now set for the infant to learn more complex mappings from communicative goals to linguistic expressions, which more effectively achieve these goals—i.e. to learn syntax.

[11] Theories of adult utterance production (e.g. Levelt, 1989; Dale *et al.*, 1998) tend to assume a similar decomposition. The speaker's initial decision is how to achieve a given communicative intention—which in some cases, is to generate an utterance. Deciding what goes in the utterance is a separate question. Of course, both these decisions are greatly elaborated in adult speaking.

6.2.4 Learning syntax: early developmental stages

The question of how children acquire syntax is of course controversial. The controversy pits **nativists** against **empiricists**: broadly speaking, nativists believe that children are born with a substantial amount of grammatical knowledge, while empiricists believe that children acquire a substantial amount of grammatical knowledge through exposure to linguistic data. Naturally there are degrees of nativism and empiricism. Extreme versions of either position are untenable: humans must do *some* learning before they can speak or understand language, and conversely there must be *something* special about human brains which is responsible for the special complexity of human language. Nonetheless, the debate is one which seems to polarise linguists quite thoroughly. I will discuss the nativist-empiricist debate in some detail in Section 6.2.5. Before I do so, however, I will attempt to describe some of the early stages of syntactic development in a relatively theory-neutral way.

6.2.4.1 Simple word combinations

Children start producing multi-word utterances between 18 and 24 months (Tomasello, 2003). The communicative functions served by early multi-word utterances are essentially the same as those served by holophrases; all that changes is the amount of information which an utterance provides about its associated communicative goal. As just noted in Section 6.2.3, a communicative goal is likely to be expressed compositionally, in a way which activates several distinct semantic representations, and a holophrase often consists of a single word denoting one of these distinct representations. (For instance, to achieve the goal of getting a cup from an interlocutor, an infant might use the holophrase *cup!*.) At around 18 months, infants learn that utterances can consist of multiple words, which denote different elements of their associated goal representation, and which thus provide their hearer with more information about the goal. The earliest multi-word utterances tend to be two-word combinations. Thus an infant who wants a cup might say *my cup!* or *cup my!*.

6.2.4.2 Pivot schemas

It is not long before an infant's two-word utterances start to exhibit some simple regularities. Some regularities are to do with word sequencing. For instance, *my cup* might be more commonly produced than *cup my*. Others are to do with generalisations over words. For instance, the infant might produce *my cup*, and also *my bread* and *my milk*. The resulting linguistic patterns have been termed **pivot schemas** (Braine, 1976). Each schema is associated with a single word, which serves to structure the utterance by licensing another

word of a particular category, either to its left or to its right. The structuring word often indicates a particular speech act type as well. In the case of the schema *my* ___, the pivot word *my* licenses an object-denoting word to its right, and the whole utterance serves to express a request for this object. Tomasello *et al.* (1997) showed that pivot schemas are not just collections of memorised two-word utterances; children of 22 months were able to use an existing schema with a newly acquired word. Pivot schemas thus enable productive combinations of words. However, Tomasello argues that they do not encode sophisticated conventions about how the structure of an utterance relates to the structure of the message it conveys—i.e. that they do not encode properly 'syntactic' knowledge. He suggests that the word ordering regularities found in pivot schemas are just a reflection of surface regularities in the exposure language, and do not carry information about the structure of the underlying message.

6.2.4.3 Item-based syntactic constructions

Evidence of genuinely syntactic knowledge emerges in comprehension at around 18 months. If infants of this age are given a sentence of the form *A is hugging B*, and are shown a picture of this event as well as one of a confounding event in which B hugs A, they reliably look more at the described event than the confounding one (Hirsh-Pasek and Golinkoff, 1996). Production lags behind, as usual; children begin to produce transitive sentences with adult-like word-ordering conventions at around 24 months (Tomasello and Brooks, 1998). (Note that these ordering conventions cannot just be reflections of common surface word orders, because the words which denote the agent and the patient are frequently attested in both subject and object positions.) It is also around 24 months that children begin to use grammatical function words (e.g. determiners, auxiliary verbs) and morphological affixes (e.g. agreement inflections on nouns and verbs).

One important discovery about children's earliest syntactic constructions is that, like pivot schemas, they tend to be tied to specific words. For instance, Tomasello (2003) discusses a child who used two different syntactic constructions for introducing the instrument with which an action is performed. One construction made use of the preposition *with* (*Open it* ***with this***); the other used simple apposition (*He hit me* ***this***). The interesting thing is that the child systematically used different constructions with different verbs: *with* was always used with the verb *open*, and apposition was always used with the verb *hit*. More quantitively, Lieven *et al.* (1997) analysed the utterances of children between 1 and 3 years old, identifying for each child a set of templates capturing the form of utterances which were 'constructed' rather than holophrastic. Templates contained 'slots' which could be filled by a range of words from a particular syntactic or semantic class. (Utterances conforming to a particular template were assumed to have been 'constructed' from this

template by filling its slots.) The important finding was that templates also tended to contain specific open-class words. Thus the syntactic generalisations in the children's utterances tended to be incomplete; they were generalisations about the uses of particular words, rather than fully abstract syntactic rules.

As syntactic development progresses beyond item-based constructions, it becomes necessary to adopt a fully-fledged model of syntax in order to characterise children's utterances. At this point, it becomes hard to remain theoretically neutral, because there are two very different brands of syntactic model to choose from. In the next section I will review the two basic alternatives, and I will dwell in particular on the empiricist model, because this has not been discussed until now.

6.2.5 Learning syntax: nativist and empiricist models

The nativist-empiricist controversy is due in large part to two related disagreements. One is about the right way to model syntactic knowledge; the other is about how much syntactic information children can extract 'bottom-up' from the language which they hear. In Section 6.2.5.1 I will outline a nativist position on these issues, which is roughly that espoused by generative linguists (e.g. Chomsky, 1980; Pinker, 1994). This position rests on the model of syntax which I introduced in Chapter 4. In Section 6.2.5.2 I will outline an empiricist position, which is drawn mainly from Tomasello (2003), with some arguments taken from Jackendoff (2002). The empiricist position is associated with a class of syntactic models called **construction grammars**; in Section 6.2.5.3 I will introduce construction grammar, and discuss how it differs from the generative model of syntax presented in Chapter 4. Empiricism is also often associated with a particular neural network architecture, the **simple recurrent network**, which shows compellingly how several types of linguistic structure can be learned 'from scratch' by exposure to linguistic data. I will introduce this style of neural network in Section 6.2.5.4.

6.2.5.1 The nativist position

In the generative grammar tradition (e.g. GB and Minimalism), there is a strong focus on identifying syntactic principles which apply across languages. In order to make these principles as general as possible, they must be very abstract; in fact, as we saw in Chapter 4, they require the postulation of an underlying level of syntactic structure (which in the model I introduced was termed 'LF'). Generative linguists believe that positing covert structures allows a more economical statement of the grammars of individual languages: the idea is to express differences between languages as differences in the way covert structures

are expressed overtly in surface language. The extra machinery needed to model covert structures pays for itself by allowing a simpler account of the differences between languages.

Note that the structure of covert representations heavily constrains the range of overt structures which can be generated. To model a particular language, a relatively small number of 'parameters' can be specified, denoting choices about how different elements of covert structure are to expressed overtly. If the system works properly, the set of possible parameter values will describe the range of human languages which are actually found.

If variation between human languages can be well modelled by a small number of parameters, then it is interesting to compare the space of *actual* human languages to the space of *possible* human languages—i.e. to the space of languages which humans could conceivably adopt to communicate with one another. The space of actual human languages is likely to be a small subset of the space of possible human languages. If this is the case, the contingent similarities between human languages need to be explained. The Chomskyan explanation is that linguistic universals reflect aspects of language which are encoded innately. According to this view, a child is born with some 'grammatical knowledge'. (Obviously, this knowledge is implicit.) The knowledge manifests itself as a propensity to develop languages in the space defined by possible parameter settings, rather than outside this space. Chomsky used the term **universal grammar** to refer to this innate grammatical knowledge.

If children have some amount of innate grammatical knowledge, they do not have to learn the grammar of their native language from scratch. Instead, they only need to explore a fairly constrained set of hypotheses. The grammar of any natural language is a very complex thing: the task of acquiring a grammar without *any* prior knowledge seems a daunting one. Of course, there are many idiosyncracies in any language which a child has no option but to learn from the data. For one thing, all individual word meanings are idiosyncratic. (There are no universal generalisations about the meanings of words.) But an account which assumes that some syntactic principles are innate certainly reduces the difficulty of the learning task. Nativist developmental linguists argue that some of the syntactic knowledge which children acquire *could not* have been learned from the data. I will give two examples of this type of argument.

One type of argument focusses on a specific piece of syntactic knowledge. For instance, Lidz *et al.* (2003), study how children learn the generalisation that transitive verbs typically denote causative actions (actions where the agent brings about a change in the object). In the Indian language Kannada, causative verbs are linguistically signalled in two different ways: they are often transitive, and they often have a characteristic morphology. Distributionally, morphology is in fact a stronger signal of causitivity than transitivity. Yet children still use transitivity as an indicator of causativity, rather than morphology. Lidz *et al.* argue that they are innately predisposed towards this generalisation.

Another type of argument comes from situations where children grow up with impoverished linguistic input. A particularly interesting case is where children grow up in a community where a **pidgin** language is prominently spoken. A pidgin develops in an environment where people from different language communities come into contact with one another, and must invent a lingua franca in order to communicate. Pidgins tend to be syntactically very simple; they tend to consist of a vocabulary, and not much else. Words are typically uninflected, and there are few constraints on word ordering. But if a community of children grows up in an environment where a pidgin is spoken, there is some evidence that these children develop a much richer language called a **creole**, even in the space of a single generation. A creole is a proper natural language, with the same kinds of grammatical feature found in languages which have been spoken for centuries. Again this suggests that children have a propensity to find complex grammatical structure in their exposure language. This idea was first proposed by Bickerton (1981); for contemporary views on what creoles tell us about innate grammatical knowledge, see DeGraff (2001).

There are many potential holes in the above argument for innate syntactic knowledge. Firstly, and perhaps most importantly, note the argument is only as good as the syntactic account of principles and parameters which it rests on. Genuine linguistic universals are hard to find. If candidate universals are very concrete, there are likely to be exceptions. But if they are too abstract, they might be too powerful, and characterise too large a space of languages, which is also undesirable. Secondly, contingent similarities between human languages need not necessarily be the result of innate syntactic constraints. For instance, they may also be due to constraints on the learnability of languages. Some languages may be more easy to transmit from one generation to the next (see e.g. Kirby and Hurford, 2002; Christiansen and Kirby, 2003). Thirdly, the task of learning a complex grammar from exposure to data does not seem as daunting now as it did when the idea of innate grammar was first mooted by Chomsky. Powerful statistical techniques have been developed in the discipline of computational linguistics, which are able to extract all kinds of regularities from raw linguistic data. Pullum and Scholz (2002) consider several constructions which generative linguists have claimed cannot be learned from raw data, and demonstrate how they can in fact be learned without postulating any innate syntactic biases. (I will look at some examples of how linguistic structure can be extracted from raw data in the next section.) Given infants' proven abilities to extract statistical regularities in a linguistic signal (recall the experiment of Saffran *et al.*, 1996 described in Section 6.2.1), the possibility of a fully data-driven language acquisition mechanism does not seem as far-fetched as it once did. Finally, the argument from creoles as stated above is very simplistic. There is also good evidence that creoles tend to adopt syntactic structures from the languages which they blend (see e.g. Siegel, 2008). There are attested cases where this is not possible—for instance, a creole developed in a community of deaf children in

Nicaragua, where there was no mature sign language to adapt (Kegl *et al.*, 2001). But these cases are very rare, and data about how such creoles arise are quite scarce.

Another challenge for nativists is to give a neural account of innate syntactic knowledge. Where in the brain is an infant's 'innate syntactic knowledge'? Generative linguists tend not to consider this question in much detail. They do tend to concur on one point, however—namely that innate syntactic knowledge is *specifically linguistic*. Generative linguists tend to adopt some version of Fodor's (1983) modularity hypothesis (see Section 1.1), assuming that grammatical mechanisms and representations are the preserve of specialised neural structures. If this assumption is correct, we should find areas of the brain which specialise in maintaining syntactic representations and performing syntactic processing. As we saw in Section 6.1.4.1, there is some evidence that Broca's area and surrounding areas in frontal cortex have some specialisation for syntax, because damage to these areas results in specific syntactic deficits. If nativists speculate at all about neural mechanisms, they tend to suggest that Broca's area is the primary locus of innate syntactic knowledge. However, our understanding of what computations occur in Broca's area and surrounding areas is far from being good enough to attribute to it the kind of role it needs to play in a nativist theory of universal grammar. It is fair to say that there is no evidence for the kind of complex covert syntactic representations postulated by nativists, either in Broca's area or anywhere else in the brain. Indeed, given the impossibility of using invasive single-cell recording techniques in humans, it is doubtful whether such evidence could ever be gathered, even if nativists happen to be right.

It certainly seems that Broca's and associated areas are not *only* involved in linguistic processing. For instance, as already discussed in Section 6.1.4.1, these areas are also involved in nonlinguistic serial tasks (see e.g. Dominey *et al.*, 2003), in sensorimotor sequencing tasks (see e.g. Greenfield, 1991) and in nonlinguistic sensorimotor and working memory tasks (see e.g. Müller and Basho, 2004). Are these findings evidence against an innate grammatical capability? Not at all. In fact, the idea that linguistic processing 'shares mechanisms' with general sensorimotor cognition (see Section 1.1) sits very comfortably with the idea that some syntactic mechanisms are innate. It is unquestionable that much of our sensorimotor apparatus is 'innate' in just the right sense—i.e. that infants have a genetically determined predisposition to develop a particular structure of sensorimotor pathways. The central idea in this book, as described in Chapter 5, is that LF, the covert level of syntactic structure which nativists argue reflects innate linguistic knowledge, can be understood as a representation of nonlinguistic sensorimotor sequencing operations. According to this view, our 'innate linguistic knowledge' is really innate *sensorimotor* knowledge.[12]

[12]Jackendoff (2002) notes that Chomsky is uncommitted about the extent to which innate linguistic

In summary: I have outlined two possible arguments for innate syntactic knowledge. One is the classical argument advanced by generative linguists. This derives from a particular model of syntax, which assumes an underlying level of syntactic structure, and a paramaterisation of natural languages as regards how underlying structures are expressed in surface language. This argument has some appeal, but is far from watertight. Another way of arguing for innate syntactic knowledge is to claim that syntactic mechanisms supervene on some more general cognitive mechanism, which is itself innate in origin. Of course the effectiveness of this argument rests on how detailed the account of supervening is. Nonetheless, it emphasises that innate linguistic knowledge need not be *specifically* linguistic: it may also have a role in the development of of other non-linguistic capabilities.

6.2.5.2 The empiricist position

Empiricist linguists argue that children learn the syntax of their native language by extracting regularities from the linguistic data they are exposed to. The linguistic data takes the form of a set of utterances. Each utterance is a sequence of words, uttered by a speaker to achieve a particular communicative effect on a hearer.

The utterances that a child hears contain several sorts of regularity. There are regularities in the way words in utterances are sequentially organised: for instance, certain word sequences co-occur more frequently than others. There are more complex distributional regularities: for instance, there may be correlations between the appearance of words or word sequences in noncontiguous parts of an utterance. There are also—obviously—regularities in the relationship between word sequences and their communicative effects—which is why utterances have meanings. While nativists argue that children are born with some knowledge about the form of these regularities, empiricist linguists argue that children learn them mainly by exposure to data, using domain-general pattern-finding skills.

What pattern-finding skills are required? Tomasello (2003) mentions several kinds of skill, including an ability to find structure in sequentially organised stimuli, and an ability to find mappings between pairs of complex patterns. Both types of skill are well attested in non-linguistic cognition. For instance, in the domain of motor control, as we saw in Section 2.5.2.2, there are mechanisms which can learn a forward model of a motor effector,

knowledge is specifically linguistic. He quotes a footnote from Chomsky (1965): 'we do not, of course, imply that the functions of language acquisition are carried out by entirely separate components of the abstract mind or the physical brain. (...) In fact, it is an important problem for psychology to determine what other aspects of cognition share properties of language acquisition and language use.' However, it is pretty clear that Chomsky did not have overlaps with sensorimotor cognition in mind. He has famously argued against the idea that language emerges from the sensorimotor system—see in particular his debate with Piaget (Piattelli-Palmarini, 1980).

and can thus predict the future state of the effector from its current state and the current motor impulse being applied to it. In the domain of action recognition, the mirror system hypothesis postulates that agents learn to associate visual representations of actions with motor representations (see Section 2.7.5). The visual and motor patterns to be associated are complex; but agents are obviously able to find these associations.[13] In the domain of analogical reasoning, humans also show an ability to identify mappings between complex patterns which have 'structural similarity' (see e.g. Gentner and Markman, 1997). So there are good indications that these pattern-finding skills are found in domains other than language.

What sorts of pattern can be found by these mechanisms? They are different from the patterns encoded in traditional phrase-structure grammars, in two important ways.

Firstly, patterns are defined as *statistical tendencies*, rather than as universal generalisations about well-formed utterances. A traditional phrase-structure grammar divides the space of word sequences discretely into two categories, 'well-formed' and 'ill-formed'; there are no degrees of grammaticality. Any well-formed utterance will be assigned a set of possible syntactic analyses, each of which may have several different semantic interpretations. The learning mechanisms invoked by empiricist linguists are able to detect patterns of arbitrary statistical strength, varying on a continuum from 'impossible' to 'certain'. This is useful both in modelling sentence syntax and in modelling the mapping from sentences to meanings. In the former case it allows us to express *degrees of grammaticality*. A statistical language model assigns probabilities to word sequences, rather than simply labelling them as grammatical or ungrammatical. Such a model can reflect a language sample very directly, even if it includes the messy kinds of error found in real language use, such as incomplete sentences or repeated words. Traditionally, syntacticians have avoided trying to model the complexity of real usage data—a methodological decision enshrined in Chomsky's well-known distinction between syntactic **competence** (the abstract ability to generate all the sentences in a language) and syntactic **performance** (the way this ability manifests itself in agents with limited processing resources operating in real-time, real-world environments). Chomsky (1965) argued that it was legitimate for syntacticians to focus on competence, and therefore to study an idealised fragment of language, free

[13]It may not be a coincidence that the pattern-matching machinery in the mirror system is a good example of the kind of machinery required to learn mappings from words to meanings. As discussed in Section 6.1.4.5, Rizzolatti and Arbib's (1998) model of syntax turns on the claim that Broca's area is the human homologue of F5 (the mirror neuron area in macaque). They argue that the pattern-matching machinery which supports action recognition was co-opted by evolution for use in mapping sentences to meanings. (See also Arbib, 2005.) If this is so, then the machinery involved may not necessarily be 'general-purpose', as empiricists tend to claim. It may be quite specialised for a particular sensorimtor task.

from performance errors. But ultimately, of course, we would like to model real language, errors and all. In a statistical model, many types of 'performance errors' simply constitute noise, and are easily ignored. A statistical inference engine may thus be able to extract a model of syntactic competence directly from a sample of real language.

Statistical models are also very useful in explaining how linguistic expressions are mapped to meanings. Ambiguity is rife in natural language: many words have several meanings, and many word sequences have several syntactic interpretations. A traditional grammar with reasonable coverage assigns a very large number of possible interpretations to almost every well-formed sentence, rendering it almost useless without a separate disambiguation component (see Abney, 1996 for some very compelling examples). A statistical model can assign different probabilities to different interpretations; in this case the likelihoods are an integral part of the model, rather than being separately stipulated.

Secondly, pattern-finding mechanisms are geared to finding patterns in the *surface structure* of language. Traditional grammars analyse sentences as hierarchical objects, but pattern-finding mechanisms operate on surface sentences; the regularities they discover are sequential regularities in the speech stream. Among the first regularities which infants find are those which characterise words as frequently-occurring phoneme sequences; recall Saffran *et al.*'s (1996) experiments on word-segmentation reported in Section 6.2.1. Once a phonological conception of words has been acquired, then the same pattern-finding ability can identify commonly-occurring sequential patterns of words. Empiricists argue that many structures in language must be defined, at least in part, as surface patterns. Such patterns are often grouped under the rubric of 'idioms'. I will give several examples, all drawn from Jackendoff (2002).

A 'pure' idiom can be defined as an arbitrary sequence of words which collectively have an arbitrary semantic interpretation. An example of a pure idiom is *by and large*. The meaning of this phrase has nothing to do with the conventional meanings of the words *by*, *and* or *large*. The syntax of the phrase is also anomalous; there is no general phrase-formation rule which combines a preposition, a conjunction and an adjective. One way to think of the phrase is as a special kind of lexical item. It is an arbitrary mapping between a phonological sequence and a meaning—the phonological sequence just happens to comprise three words rather than one. Note that *by and large* is a commonly-occurring word sequence, which we can expect a statistical pattern-finding mechanism to detect.

Pure idioms are in fact relatively rare. Most idiomatic constructions exhibit some form of syntactic regularity. This can be of different kinds. In some cases, idioms have 'slots' which can be filled with phrases of a particular type, generated according to regular grammatical principles. For instance, *Far be it from NP to VP* is an idiom whose slots can

be filled by any (regular) NP and any (regular, nonfinite) VP.[14] In these cases, learning the meaning of an idiom involves learning an abstract mapping, which associates a particular *pattern* of surface words with a corresponding *pattern* of semantic representation. As already mentioned, we must probably assume some special pattern-recognition machinery in order to account for how such abstract mappings are discovered.

In other cases, idioms consist of syntactically regular sequences of words. For instance, consider *That's the way the cookie crumbles*, or *NP kicked the bucket*. The former idiom is a well-formed sentence; the latter is a well-formed VP. But note that the meaning of the former phrase has nothing to do with cookies or crumbling; the phrase is a conventional way of saying 'That's just the way things are'. Similarly, the meaning of the latter phrase has nothing to do with kicking or buckets. So even though these expressions have regular syntax, this syntax is not being used to derive their semantics, as it is in normal phrases.[15]

Syntactically regular idioms tend to show some limited degree of syntactic flexibility. For instance, if an idiom contains slots, it must often contain inflections signalling grammatical agreement with the material in the slot (e.g. *I* ***kick*** *the bucket*, *Bill* ***kicks*** *the bucket*). Often, the verbs in an idiom can also be inflected for tense or aspect, or modified for modality (e.g. *That* ***was*** *the way the cookie* ***crumbled*** *in those days*; *That's* ***probably not*** *the way the cookie* ***will crumble***). In such cases, the word sequence which makes up an idiom is not entirely frozen; there is some degree of variation in the direction of syntactic regularity. Sometimes, quite wholesale transformations are possible; for instance, *NP let the cat out of the bag* can be expressed in the passive (*The cat was let out of the bag [by NP]*). In other cases, such transformations are not possible (*The bucket was kicked by Bill* is a transformation too far). The degree of regularity which an idiom allows seems to depend on the degree to which the metaphor from which it derives is still accessible (see e.g. Nunberg *et al.*, 1994).

How prevalent are idioms in language? Empiricists argue that they are a core element of syntactic structure, and that language is full of them. Their central argument is that the syntactic structures in a language occupy a *continuum* on the dimension of idiomaticity, which extends from pure idioms at one end to fully productive constructions at the other. The 'semi-idiomatic' constructions we have just looked at occupy a point close to the idiomatic end of the spectrum. There are other constructions which occupy a more intermediate position. For instance, consider the class of **phrasal verbs**, such as *give up*

[14]Notice that the NP slot is often filled by the pronoun 'me'; this is a statistical tendency which can be well captured by a pattern-finding mechanism.

[15]Syntactically regular idioms tend to have their origins in metaphors. Understood metaphorically, it is often possible to derive the semantics of a syntactically regular idiom compositionally. What makes them idiomatic is the fact that their original metaphorical meanings tend to be hard to recover. Over time, in other words, they have acquired a conventionalised meaning rather than a metaphorical one.

or *pull over*. Phrasal verbs have special meanings; for instance, the meaning of *give up*—'concede'—cannot be derived from the conventional meanings of the words *give* and *up* (except through an extremely opaque metaphor). It is possible to represent phrasal verbs like *give up* within a fully compositional grammar, by making the verb do all the work. We can define a special sense of *give*, which has the meaning 'concede', and a special sense of *up*, which has no semantic content, and require that the special sense of *give* takes the special sense of *up* as an obligatory PP argument. But another way of representing phrasal verbs is as *simple idioms*. On this analysis, *give NP up* or *give up NP* is a surface pattern of words which has the conventionalised meaning 'to concede X', where X is contributed by the NP. Children learn such constructions in the way they learn normal idioms, as abstract mappings from surface patterns of words to semantic patterns.

The idiomatic analysis is supported by the fact that there are a large number of phrasal verb constructions, and that these constructions also appear to occupy a continuum of syntactic flexibility, just like prototypical idioms. Some phrasal verbs are very fixed; for instance *take NP to task*, or *put* NP_i *in* pro_i*'s place*. Others have some degree of flexibility. For instance, while *pull* in *pull over* does not have its standard meaning, the meanings of *pull over* and *pull across* do seem to be derived in some way from the words *over* and *across*. But the pattern does not generalise to *pull up*. At the most flexible end of the spectrum of phrasal verbs, we still observe collocational regularities. For instance, *jump* can take an arbitrary PP as an argument. But *jump up* is more common than *jump along*. An idiomatic or collocational analysis can thus be usefully extended right down to prototypically grammatical constructions.

The above argument for an empiricist model of linguistic structure is mainly from theoretical economy. A single notion of 'construction' is employed to analyse a wide range of linguistic phenomena, ranging from broad syntactic generalisations to statistically defined collocation patterns to fixed idioms and usage patterns. However, there are also more empirical arguments, which relate to observations about how children acquire syntactic constructions. One particularly telling argument relates to the presence of item-based constructions in children's early utterances (see Section 6.2.4.3). The fact that syntactic development proceeds through a stage of item-based constructions suggests that syntactic generalisations are formed by finding patterns in surface language, and then progressively abstracting away from these patterns. It can be interpreted as evidence for the continuity between item-based idiomatic constructions and more abstract syntactic generalisations.

In summary: the pattern-finding mechanisms which empiricists implicate in language learning are able to represent regularities of different statistical strengths, which can make reference to the surface structure of utterances. Notice that the ability to represent patterns of different strengths sits well with the analysis of idioms as occupying a continuum of degrees of fixedness. For empiricists, therefore, learning the syntax of a language consists

in becoming attuned to a very large number of statistically defined patterns in the language data. A child accomplishes this using general purpose learning mechanisms; an innate predisposition to acquire certain forms of grammar is not required.

A generative grammar like Minimalism has a very hard time accounting for idiomatic elements of language. Idioms are defined as statistical patterns, some components of which are patterns in surface language. There is no treatment of statistical patterns in Minimalism as it stands, and there is no treatment of surface patterns in their own right: surface sentences are products of a generative process which makes no reference to the sort of patterns found in idioms. It may be possible to revise the Minimalist model of the generative process to incorporate idiomatic constructions, but it would be a very radical revision; it is not clear what essential elements of Minimalism would remain.

On the other hand, the empiricist model of language has its own problems. In fact, these problems are the converse of those faced by a traditional grammar. The central problem for empiricist models is how to account for the elements of linguistic structure which *do* seem to be well described by conventional hierarchical grammatical rules. The basic story must involve an elaboration of the idea of a 'pattern' in surface structure. For instance, as discussed above, the idiom *Far be it from NP to VP* is a 'pattern'; it is a general template which matches lots of specific word sequences. Note that the material which appears in each 'slot' is unspecified both in terms of its lexical content and its length. Note also that there are much more general 'patterns' than this. For instance, the abstract pattern 'S→NP, VP' is a strong high-level pattern in English, which any grammatical model must represent, even if it is only understood as a statistical tendency.

In the next two sections, I will introduce two models which constructivists have used to express a notion of 'pattern' which is powerful enough to represent abstract syntactic structure as well as surface patterns in language, and which permits statistical generalisations to be expressed rather than absolute ones. The first, discussed in Section 6.2.5.3, is a style of syntactic model, which should be understood as a radical alternative to the generative model introduced in Chapter 4. The second, discussed in Section 6.2.5.4, is a class of neural network which can learn abstract regularities in sequential data.

6.2.5.3 Construction grammar and associated computational grammar formalisms

Construction grammars are founded on the empiricist assumption described above. Language is a collection of patterns of words: patterns can be of different degrees of abstractness, and of different strengths. Each pattern of words is associated with a semantic pattern. Children have to learn a large number of word patterns, as well as associations between these patterns and semantic patterns. Language is modelled as a collection of

memorised patterns, rather than as the product of a single unified generative mechanism. Constructions have to be individually learned, just like words. In fact, in many construction grammars, constructions are explicitly understood as residing in the lexicon. In this model, learning a new word involves learning the constructions in which it can participate as a 'head'. But even in models where constructions are stored separately from lexical items, constructions are still just memorised patterns, differing from lexical entries mainly in their complexity and abstractness.

The idea of using constructions as a core theoretical device in a model of syntax is due to several researchers—see in particular Lakoff (1987), Fillmore *et al.* (1988), Goldberg (1995). I will begin by introducing the notion of construction as it is presented in Jackendoff (2002). For Jackendoff, a model of a language must make reference to three separate domains: phonology, syntax and 'conceptual structure' (i.e. semantics). In each domain, there are mechanisms for encoding and learning patterns—in other words, for combining simple elements into more complex ones. The mechanisms are quite independent in the different domains; it is not possible to give a simple account of the combinatorial mechanism in one domain by reference to those in the other domains. However, there are also mechanisms for learning associations between specific patterns in different domains. Each of these associations is a construction.

The construction representing the partial idiom 'take NP to task' is shown (in a somewhat simplified form) in Figure 6.6. The head of the construction, indicated with the

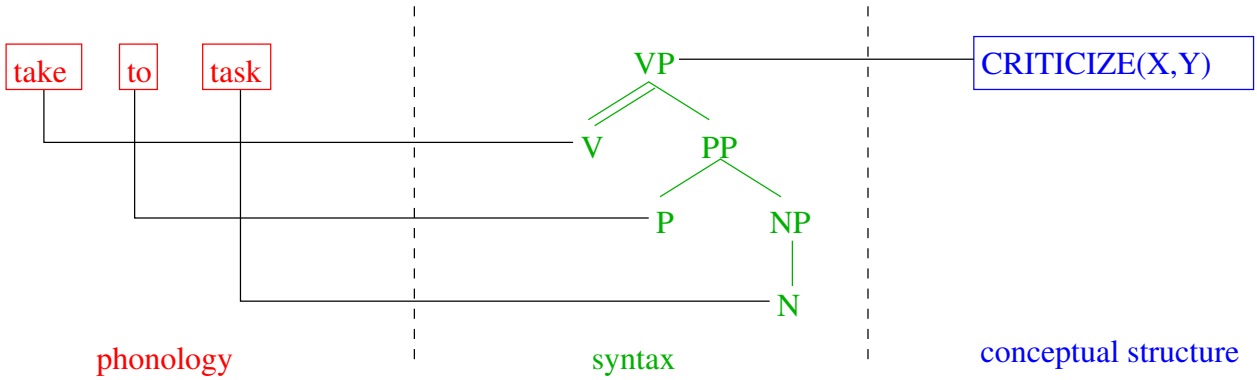


Figure 6.6: Jackendoff's (2002) representation of the partially idiomatic construction *take NP to task* (simplified)

double line, is the verb *take*: the construction can be thought of as describing one of the ways in which this word can be used. The notion of lexical heads is related to the notion of

heads in Minimalism, which was described in Section 4.3. In fact, construction grammars often make use of some form of X-bar theory. The main difference is that a lexical head can be associated with an arbitrarily large phonological pattern rather than just with a single word, an arbitrarily large syntactic structure, rather than just a single X-bar schema, and an arbitrarily large semantic template. In the above case, the phonological pattern (shown in red) is a sequence of phonological units. The syntactic pattern (shown in green) is a hierarchical syntactic structure. The semantic pattern is a predicate, denoting the action 'criticize', carrying two obligatory arguments. The black lines denote associations between these patterns. The single line linking the root of the tree (the VP node) to the semantic pattern encodes the fact that the meaning of the construction is conventional, rather than derived compositionally from its constituents. The lines linking individual phonological units to leaves in the syntactic tree encode the fact that the construction nonetheless possesses some syntactic structure. In virtue of this structure, the construction can be combined with NP arguments denoting the agent and patient of the 'criticize' action according to regular grammatical principles governing how the arguments of a transitive verb are syntactically realised. In Jackendoff's model, these principles include ordering constraints which allow the NP denoting the patient to be inserted in between the verb and the PP, giving the pattern *take NP to task*.

Jackendoff's scheme allows many of the important characteristics of constructions to be expressed. Firstly a construction can associate an arbitrarily large phonological structure with a meaning, allowing for large phrases to have conventionalised meanings. At the same time a construction can associate a phonological structure with a syntactic pattern, which allows a treatment of syntactic regularity within idioms. Finally Jackendoff's scheme provides a framework for modelling how children learn general grammatical principles by abstracting over item-based constructions. We can suggest that children first acquire direct associations between particular phonological patterns and particular semantic patterns, and then later identify generalisations over these associations which allow them to posit an intermediate level of syntactic structure. For regular syntactic constructions, associations via this intermediate level end up supplanting direct associations from phonology to semantics.

Some of the most interesting varieties of construction grammar have been developed within computational linguistics, a field in which models of language are expressed as computer programs. Several grammar formalisms have been designed or adapted to be expressed in this way; among the most widely used are **head-driven phrase structure grammar** (Pollard and Sag, 1994; Copestake, 2002), **combinatorial categorial grammar** (Steedman, 2000) and **lexicalised tree-adjoining grammar** (see e.g. Schabes, 1990; XTAG research group, 2001). Calling these formalisms 'construction grammars' is using the term rather loosely; however, they share some of the key properties of construc-

tion grammars.

Firstly, each of the above computational grammar formalisms is expressed as a collection of syntactic structures, tied to particular lexical heads. The formalisms also allow syntactic generalisations to be represented; there are *types* of words which can be used in similar contexts. But this typing system is secondary; syntactic generalisations ultimately describe patterns of language use. By contrast, in Minimalism and its antecedents, syntactic generalisations describe phrase-formation mechanisms which have their own existence independent of the lexicon. The lexicalist approach of computational grammar formalisms allows a mixture of very general and very idiosyncratic patterns in language to be captured. It also provides a framework for explaining how children acquire syntactic generalisations, by identifying patterns in the usage of certain specific words. In both of these ways, computational grammar formalisms pattern with construction grammar.

Secondly, computational grammar formalisms and construction grammar use the same basic device for allowing hierarchical patterns in language to be expressed: syntactic structures are allowed to contain 'slots', which can be filled by other syntactic structures satisfying certain constraints. In fact, all grammars provide a slot mechanism of some sort. But the possibilities afforded by slot-filling operations have been most thoroughly explored in computational grammar formalisms, where they are the central device for expressing all syntactic and semantic patterns. In computational formalisms, the syntactic objects which fill slots are modelled as complex entities, defined for a mixture of syntactic, semantic and phonological attributes. Constraints on the objects which fill slots can then refer selectively to particular attributes. In addition, if a structure contains several slots, constraints can be imposed on the relationship between the objects which fill these slots, requiring for instance that both objects have the same value for a particular attribute. These more powerful slot-filling operations (often termed **unification** operations) permit syntactic and semantic information to 'travel' through a syntactic structure. They therefore provide a means for modelling long-distance syntactic dependencies—for instance, subject-verb agreement—without postulating movement operations at a covert level of syntactic structure. Roughly speaking, what the Minimalist syntactic model accomplishes by positing 'movement at LF' is accomplished in computational grammar formalisms by defining complex slot-filling operations.[16]

Thirdly, computational grammar formalisms can be readily adapted to support a sta-

[16] An important part of the research programme for computational grammar formalisms is to define slot-filling operations which are *just powerful enough* to capture the types of structure found in natural languages. Interestingly, lexicalised tree-adjoining grammar, combinatory categorial grammar and a restricted version of head-driven phrase structure grammar have been shown to be weakly equivalent (Joshi *et al.*, 1991; Kasper *et al.*, 1995). They generate a set of languages which have been termed 'mildly context-sensitive'. Some have suggested that all human languages are also of this type.

tistical notion of 'patterns in language', as envisaged by empiricist linguists. The simplest statistical models of language are derived directly from raw text corpora. These models treat sentences as **Markov chains**, in which the probability of a given word at a given position in a sentence is conditional on the n previous words. Probabilities can be estimated from **n-gram counts** (counts of the frequency of each word sequence of length n) in a training corpus of naturally occurring text. If n is small, quite accurate probabilities can be estimated from n-grams; see e.g. Chen and Goodman (1998). Over short text spans, therefore, a Markovian model of language can approximate a syntactic theory all by itself; all other things being equal, it assigns low(er) probabilities to syntactically ill-formed word sequences, and high(er) probabilities to well-formed word sequences. However, if n is large, probabilities of word sequences of length n are hard to estimate from n-grams. Most of the word types in a naturally-occurring text corpus are very rare, so as n grows, data about the frequency of arbitrary n-grams quickly becomes extremely sparse. One way of remedying this problem is to introduce hierarchical structure into a probabilistic language model, so that probabilities are defined for combinations of phrases featuring certain words, rather than for surface sequences of words. For instance, to capture the statistical relationship between the verb *eat* and the noun *cake*, we could define the probability that a VP headed by the verb *eat* takes a NP complement headed by the noun *cake*. Expressing probabilities within a phrase-structure grammar allows us to capture statistical dependencies between noncontiguous words, which are harder to estimate from n-gram data. (Note that we are still interested in capturing dependencies between *words*, so it is important to use a lexicalised grammar formalism to express these statistical regularities.) The most accurate and sophisticated modern parsers all use statistical grammars. For instance Collins (1996) uses a probabilistic lexicalised context-free grammar; Hockenmaier and Steedman (2002) use probabilistic versions of combinatorial categorial grammar; Toutanova *et al.* (2005) use a probabilistic version of head-driven phrase-structure grammar. These grammars have a constructivist flavour, in the sense that they are induced from large corpora using general statistical techniques. (However, it is important to note that they are learned from corpora of parsed sentences, rather than from raw text.) Compared to Jackendoff's grammar formalism, they are not quite as well-suited for associating constructions 'holistically' with arbitrary phonological patterns. On the other hand, they develop the idea of constructions as statistically defined patterns much further.

All the varieties of construction grammar discussed above model hierarchical syntactic structure using the device of 'slots' in one form or another. In order to understand these grammars as models of cognitive processing, we need to explain how the slot-filling mechanism is neurally implemented. This is still an open question in cognitive neuroscience—in fact it is an instance of the 'binding problem' which we have already seen in various guises. In visual perception the issue is how to bind object representations to retinal locations (see

Section 2.4.3). In episode representation, the issue is how to bind object representations to thematic roles (see Section 3.7). To illustrate the problem as it applies to linguistic constructions, consider an example case, where a construction of type $C1$ is defined as having two slots $S1$ and $S2$, which can both be filled by constructions of type $C2$. To represent a token linguistic expression of type $C1$, we must first posit a neural representation of a particular instance of $C1$ (call it $c1$), containing slots $s1$ and $s2$. Then we must posit neural representations of two distinct instances of $C2$ (call them $c2_a$ and $c2_b$) to fill these two slots. Finally we must find a way of specifying which instance of $C2$ fills which slot. For instance we might specify that $c2_a$ fills $s1$ and $c2_b$ fills $s2$. Notoriously, simply activating $c1$, $c2_a$ and $c2_b$ as an assembly fails to achieve this: it does not establish links between specific slots and specific fillers.

As briefly mentioned in Section 3.7, there have been several proposals about how neural circuitry can associate specific slots with specific fillers. One proposal draws on the observation that neural populations often exhibit cyclic activity at a particular frequency, and that individual neurons often fire at particular phases of the population-level cycle. The proposal is that slots and fillers are only bound together if they fire at the same phase (see e.g. von der Malsburg and Buhmann, 1992; Shastri and Ajjanagadde, 1993; Shastri, 2002; Hummel and Holyoak, 2003). However, while binding by phase synchrony may be a solution to some binding problems, syntactic structures often require large numbers of variables to be bound. As Jackendoff (2002) notes, it seems unlikely that there are enough distinct phases in a single neural cycle to represent all the slot-filler relationships which need to be expressed in a moderately complex syntactic structure. Another proposal is that binding is implemented by a collection of dedicated units, each of which is fully connected to all slots of a given type and all items which can fill slots of this type (see e.g. van der Velde and de Kamps, 2006). In this proposal, one binding unit would be used to encode the link from $s1$ to $c2_a$, and another binding unit would encode the link from $s2$ to $c2_b$. This implementation of binding has been demonstrated in a computational model, and could certainly handle large numbers of variable bindings. However, there is as yet no evidence of dedicated binding units in the brain. (In fact, given that van der Velde and de Kamps' hypothesised binding units can represent arbitrary slots and fillers, finding neurophysiological evidence for them is going to be very hard.)

In summary, while construction grammars provide powerful and expressive formalisms for representing hierarchical and statistical patterns in language, there is as yet no recognised neural mechanism for implementing the basic slot-filling operation which they all heavily rely on. In the next section, I will introduce another model of hierarchical structure in language which is frequently adopted by empiricist linguists, which is much more easily associated with neural mechanisms.

6.2.5.4 The Elman network: a connectionist architecture for learning linguistic structure

The **simple recurrent network** (**SRN**), or **Elman network** (Elman, 1990) is a neural network which can learn patterns in sequential data. We have already seen networks which can learn sequences, in the accounts of sequence storage in working memory (see Section 3.2.2.1) and in the hippocampus (see Section 3.7.3). In a sense, an Elman network is another network which can learn sequences. What is distinctive about it is that it can learn structural regularities in sequences, as well as simple sequences of items.

The architecture of a SRN is shown in Figure 6.7. While an ordinary feedforward neural

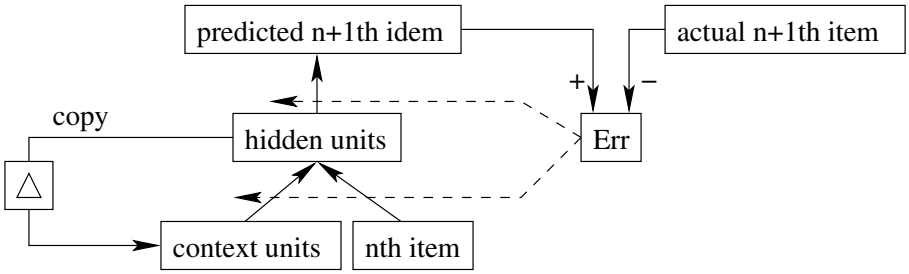


Figure 6.7: A simple recurrent network

network receives all its input at a single moment in time, and produces all its output at some later moment, a SRN's operation is extended in time. Time is represented discretely, as a series of **time points**. A SRN receives a sequence of patterns in its input layer, one at each time point. At each time point, a feedforward neural network generates a prediction about the pattern which will appear in its input layer at the next time point. This prediction appears on its output layer. The feedforward network makes use of a **hidden layer** of units. The pattern which appears in the hidden layer at a given time point is then copied into a layer of **context units** at the next time point. The context units provide information about the previous state of the network which can be helpful in predicting the next output.

To train the feedforward network, we create an error term, defined as the difference between the actual next input pattern and the pattern predicted by the network. This error term is used to update the weights in the network using a standard backpropagation scheme. In a single epoch of training, we initialise the pattern in the context units to some predetermined value, and then present a training sequence to the network, at each time point training the network to predict the next item in the sequence. After several iterations of this process, if we configure the system to copy its output at time point t_n to its input

at time t_{n+1}, and present the network at t_0 with the initial context pattern and the first input pattern, it will generate a sequence which reproduces some or all of the sequential structure in the training sequence.

Elman's original SRN was trained on a set of word sequences mimicking English sentence structure. These sequences were generated from some simple templates, producing structures which conformed to grammatical constraints, and also to semantic ones: for instance *man eat cookie*, *monster break glass*, *woman sleep*. This meant that the syntactic and semantic properties of any given word were reflected in the distribution of words which could appear before or after it in the training sequences. Sequences were given to the SRN one word at a time, and the SRN was trained to predict the next word in the sequence. Words in the input and output layer were represented using a 'one-hot' coding scheme: for a lexicon of n words, n units were used, with a different unit turned to 1 for each word and all other units turned to 0. Thus 'boy' might be encoded as $1000\ldots$, and 'girl' as $0100\ldots$. Since there were several words fitting each template category, the network could not be expected to make accurate predictions about the next word in a given sequence. However, the output vector could be interpreted as an estimated probability distribution for the next word, with higher activations for a given unit indicating higher probability for the word which it coded. When trained to predict the next word, the network learned very accurate probability distributions.

The way an Elman network learns to predict the next word in a sequence is by developing an encoding of each word in the hidden layer which reflects the words which typically precede and follow it. The recurrent connections of the hidden layer mean that this encoding can come to reflect not just the immediately preceding and following words, but a sequence of several words on either side. How long this sequence is depends partly on how complex the sequential patterns in the training data are (the network will not learn a more complex encoding than it needs to) and also on the size of the hidden layer and on the accuracy with which its unit activations are represented. Essentially, an Elman network ends up learning something very much like the n-gram model of language introduced in Section 6.2.5.3. It is trained to predict the next word in a sentence using the most recent n words as evidence, a regime which explicitly models sentences as Markov processes. The interesting thing is how much linguistic structure can be captured using this assumption.

Firstly, an Elman network can learn a taxononomy of basic parts of speech from scratch, without any prior knowledge. Elman examined the hidden-unit encodings of words which were developed when his network was trained to predict the next word. If there are n units in the hidden layer, then a trained network can be understood as mapping each word to a point in an n-dimensional space. Elman found that the encodings of all nouns clustered in one region of this space, while the encodings of all verbs clustered in another region; moreover, transitive verbs and intransitive verbs clustered in separate sub-regions

of the 'verb' region. In fact, the hidden layer encoded semantic word categories as well as syntactic ones. For instance, nouns denoting humans and nouns denoting breakable objects were encoded in separate sub-regions of the 'noun' region, as a result of their different collocations with specific verbs. Very similar methods for bootstrapping syntactic and semantic word categories have been developed in computational linguistics, using n-gram probabilities derived from text corpora; see e.g. Redington *et al.* (1998), Lin and Pantel (2001), Pantel and Lin (2002). In these studies, a small number (n) of 'representatitive words' are chosen, which are reasonably common and which cover a broad range of semantic and syntactic classes. Each word in the language is then mapped into an n-dimensional space encoding the frequency with which it co-occurs with each of the n representative words. Again, words from similar syntactic and semantic classes cluster into similar regions of this space. Recently there has been some very interesting work suggesting that this method of representing words captures something about the way words are encoded in the brain. Mitchell *et al.* (2008) represented 60 English concrete nouns in a 25-dimensional vector space, encoding for each word the frequency with which it co-occurred with 25 representative words in a trillion-word corpus of English text. They also recorded the fMRI activity evoked in human subjects for each of these 60 nouns. They then used a machine learning technique to learn a function mapping the collocation-based encodings of nouns to their associated fMRI patterns. Amazingly, the learned function was quite accurate, predicting the fMRI patterns for words held out of the training data at levels well above chance. Of course, we cannot infer from this result that humans represent words using vectors of word collocations. It may just be that there are correlations between these vectors and the distributed encodings of words which the brain actually uses. For instance, the most successful set of representative words found by Mitchell *et al.* was a set of concrete verbs: *taste*, *touch*, *smell*, *open*, *push*, etc. If concrete nouns are represented in the brain as distributed patterns of motor affordances, then we would expect word encodings based on collocations with concrete verbs to be effective in predicting fMRI activity even if words are not *explicitly* represented in the brain using collocations.[17] However, Mitchell *et al.* also found that many words with no obvious semantic or sensorimotor correlates (e.g. *seems*, *lots* and *various*) were effective as representative words. In summary, while some components of the fMRI activity evoked by concrete nouns probably reflect nonlinguistic representations, it is quite likely that other components really are distributed encodings of word collocations.

Secondly, an Elman network can learn a certain amount of *hierarchical structure* in

[17]In fact, the fact that noun encodings based on collocations with action verbs are particularly good at predicting fMRI activity can be interpreted as support for the 'embodied' conception of lexical semantics reviewed in Section 6.1.2. This is certainly the conclusion which Mitchell *et al.* reach.

language. For instance, Elman (1991) trained a SRN on sentences whose noun phrases could contain relative clauses, as illustrated below.[18]

(6.1) *Boys [who mary chases] see girls*

(6.2) *Cats [who chase dogs] walk*

Because of the embedded structure of relative clauses, some of the important syntactic dependencies in these sentences are non-local: for instance, there are important dependencies between the plural subject *boys* and its verb *chase* in Example 6.1, and between the singular subject *cats* and its verb *walk* in Example 6.2. Because of the recurrent connections in an Elman network, it can learn these nonlocal dependencies. For instance, after training, when given the following two initial sequences:

(6.3) *Boys [who mary chases ...*

(6.4) *Boy [who mary chases ...*

the network strongly predicted a plural verb (e.g. *see*, *chase*) as the next word for Example 6.3, and a singular verb (e.g. *sees*, *chases*) for Example 6.4. The network also learned to expect a missing noun inside a relative clause, but not inside a main clause. For instance, given Example 6.5 it expects a noun:

(6.5) *Mary chases ...*

while given Example 6.3, it expects a plural verb, as just described.

Note that the network does not simply memorise sets of specific word sequences. We have already noted how an Elman network maps words from similar syntactic and semantic classes into clusters in the vector space defined by its hidden layer units. These clusters allow the network to generalise, and make predictions about the distribution of words even in unseen syntactic contexts.

How does a SRN encode non-local syntactic dependencies? To explain, Elman notes that hidden-unit representations in the trained network encode words *in contexts* rather than words by themselves. For example, the context representations created by processing the closely related sequences in Examples 6.3 and 6.4 are subtly different. Training can help to accentuate this difference if it is important in predicting the next word. Of course, there is a limit to how much history an Elman network can encode. However, there is also a limit to the amount of nested syntactic structure which humans can process—a point

[18]The square brackets are included to make the nested structures clear for the reader—obviously they are not part of the actual sequence presented to the network.

made forcefully by Christiansen and Chater (1999). These authors distinguish between two types of recursive structures in language: **right-branching** structures, such as that illustrated in Example 6.6, and **centre-embedded** structures, such as that illustrated in Example 6.7.

(6.6) The dog chased [the man [who owned the cat [that caught the fish]]]

(6.7) The fish [that the cat [that the man [who the dog chased] owned] caught] swam

Right-branching recursive structures can be nested arbitrarily deep without the need to encode distant syntactic dependencies, because the embedded structure always occurs at the very end of the structure which embeds it. But centre-embedded structures require the encoding of increasingly non-local syntactic dependencies the more deeply they are nested. Centre-embedded recursive structures are known to be much harder for humans to process than right-branching ones (see e.g. Miller and Isard, 1964; Foss and Cairns, 1970). Christiansen and Chater show that centre-embedded structures are also particularly hard for SRNs to learn. They argue that the problems SRNs have with centre-embedded constructions actually speak in their favour as a model of human syntactic ability. Of course, a Chomskyan grammarian would argue that difficulties with centre-embedded constructions are the province of a model of syntactic *performance*, not syntactic competence (see Section 6.2.5.2). But Christiansen and Chater argue that this distinction is positively misleading for the kind of language model which is learned by a SRN. A SRN's model of syntactic generalisations (i.e. of syntactic competence) *emerges* from the learning it achieves during a language processing task; its failure to cover 'idealised' language data is because this data is unrepresentative of real language in various systematic ways.

Thirdly—and perhaps most obviously—an Elman network can learn sequential patterns in surface language as well as hierarchical patterns. In fact, this is the task which it is explicitly trained to perform, so this is hardly a suprising result. To illustrate, consider an idiomatic expression like *Far be it from X to Y*. As discussed in Section 6.2.5.2, to describe this expression we need to refer to a specific sequence of words: there is no more abstract generalisation which captures it. An Elman network trained on a language sample containing many instances of this idiom would eventually learn it as a specific sequence of words. If after training it was shown the sequence *Far be...*, it would form a very high expectation for a single word, *it*, and if shown *Far be it...*, it would form a high expectation for *from*. The next item is more variable; with sufficient training data it should be able to learn to expect the full range of NPs in this position (probably with a bias towards the pronoun *me*, because of its frequent collocation with this idiom). Thus an Elman network can learn a mixture of specific word sequences and more abstract syntactic patterns. The network does not in fact distinguish between these two types of pattern; it will develop

encodings reflecting whichever kind of structure is most prominent in the training data. In many ways, therefore, an Elman network seems eminently suited for learning the 'semi-idiomatic' constructions which are so common in natural language, and so problematic for traditional generativist accounts.

Given the above results, an Elman network is often proposed by empiricist linguists as a plausible neural model of our knowledge of natural language syntax. In this regard, it is interesting to compare construction grammars and SRNs as empiricist models of language. At least in some respects, SRNs can be understood as a modelling how construction grammars are neurally implemented. A SRN can be thought of as a neural implementation of the aspects of construction grammar which reflect the statistical and surface-based nature of linguistic patterns, and the fact that these patterns are extracted from real usage data rather than idealised language samples. However, SRNs and construction grammars have quite different models of hierarchical structure in language. Construction grammars use the device of slot-filling to represent hierarchical structures. Slot-filling is a powerful mechanism, but it is still uncertain whether an explicit neural analogue of the mechanism actually exists. On the other hand, SRNs manage to encode a limited amount of hierarchical structure without an explicit slot-filling operation, using simple recurrent circuits which are well attested in neural hardware.[19] This is a very attractive feature of SRNs. But it is not at all clear that a language model can entirely dispense with something akin to slot-filling. For instance, it is quite easy to find sentences which contain very long-distance dependencies, and are nonetheless quite easy for humans to process. Jackendoff (2002) cites the following sentence:

(6.8) Does [the little boy in the yellow hat who Mary described as a genius] like ice-cream?

The uninflected form of *like* in this sentence is a result of the initial word *does*. Capturing this dependency in a SRN would take a very capacious hidden layer, and a very large amount of training. The idea that the whole subject NP fills a slot in the matrix clause becomes quite an appealing alternative, even in the absence of a convincing neural account of slot-filling.

Another big difference between construction grammars and SRNs is in their treatment of sentence semantics. Construction grammars provide mechanisms for mapping well-formed sentences onto semantic interpretations. SRNs, on the other hand, have no treatment of sentence semantics at all. A SRN models structures in surface language, but does

[19] Admittedly, the SRN relies on the backpropagation training algorithm, which is not biologically plausible. However, several more biologically realistic recurrent network architectures have been proposed; see e.g. Dominey *et al.* (1995).

not model the relationship between surface linguistic patterns and patterns of meanings, which is obviously a crucial goal for any model of language. To illustrate, imagine a trained SRN being used to 'generate' specific sentences. In most syntactic contexts, the SRN will generate a broad distribution of likely next words, with many good candidates. To generate specific sentences, we would have to pick a candidate at random, perhaps biasing the choice towards the most likely ones. Of course, in 'real' sentence generation, the words chosen are a function of the message to be conveyed as well as of syntactic constraints. There have been a few attempts to turn an Elman network into a true sentence generator by providing it with a semantic message as an additional input—most notably Rohde (2002) and Chang (2002). I will attempt to do something similar in Section 6.4. At this point I will just say that it is not straightforward to extend a SRN so it functions as a model of the syntax-semantics interface as well as of syntactic structure by itself. In fact, many of the issues to be solved seem to call for some sort of slot-filling mechanism. For instance, to represent a simple cup-grabbing episode, we must find some way of associating particular thematic role 'slots' (agent and patient) with arbitrary objects (man, cup). To represent a transitive sentence which can express this episode, we must find some way of associating particular syntactic 'slots' (subject and object position) with arbitrary noun phrases (e.g. *the man*, *a cup*). Then we must specify how these two structures of slots map together. In other types of proposition, variables and their bindings seem to take an even more prominent role; for instance, representing the semantics of the quantified sentence 'Every man loves a woman' seems very hard without some treatment of variables and variable binding. It is not clear how an Elman network can be extended to incorporate a treatment of semantics without adopting some form of slot-filling mechanism.

In summary: there are both syntactic and semantic reasons to think that the linguistic system makes use of some form of slot-filling mechanism, even though the neural implementation of this mechanism is not yet understood.

6.3 Learning single word meanings, and the concept of a communicative action

In the remaining two sections of this chapter, I will present a network model of how infants learn word meanings, and learn the syntax of their native language. The current section presents a model of how word meanings are learned; it relates back to Section 6.1.3, which discussed the neural representations of words, and to Section 6.2.2, which introduced some theories about how infants learn words meanings. Section 6.4 presents a model of how infants learn syntactic structures. It draws on the Minimalist model of syntax introduced in

Chapter 4, and on the sensorimotor intepretation of Minimalist LF proposed in Chapter 5, but also on the empiricist model of syntax acquisition introduced in the second part of Section 6.2.5.

6.3.1 A network for cross-situational word meaning learning

To introduce the word learning model, I will start by situating the infant's learning task in the context of the neural model of word representations introduced in Section 6.1.3. Recall that words are represented as phonological forms in two places in the brain, in the phonological input buffer and the phonological output buffer (see Figure 6.3), and these representations are separately associated with semantic representations (see Figure 6.4). The question is: how are these associations learned? We need to learn two functions: one which maps word representations in the phonological input buffer to semantic representations, and one which maps semantic representations to word representations in the phonological output buffer. In this section, I will propose a network which implements some of the models of word meaning learning reviewed in Section 6.2.2.

Basic networks for cross-situational word meaning learning (see Section 6.2.2.1) are shown in Figure 6.8. I assume two distinct networks: a **word understanding network** for mapping input words to meanings, and a **word production network** for mapping meanings to output words.

Training inputs for both networks consist of utterances produced by a mature speaker, paired with semantic representations computed by the infant from sensorimotor experience. An utterance is a sequence of syllables. I assume the infant has already acquired a set of phonological word representations. So if an utterance consists of words, a sequence of syllables will activate a (slower) sequence of phonological word representations. These representations are evoked directly in the phonological input buffer, and indirectly in the phonological output buffer, via the mirror system for phonemes and a process of articulatory intention recognition (see Section 6.1.3.1). In the input buffer, incoming word representations are temporarily associated with a temporally evolving context signal, so a short sequence of words can be replayed while the semantic representation is held fixed.

Both networks are trained by explicit error terms. The word understanding network is trained by an error term encoding the difference between its output and the actual semantic representation evoked by the infant. The word production network is trained by an error term encoding the difference between its output and an articulatory representation of the word actually produced by the speaker.[20] Both networks will learn words slowly, using

[20]I have assumed a supervised learning paradigm, including an error term, to model these learning circuits, rather than a more direct Hebbian paradigm, for compatibility with later circuits for learning

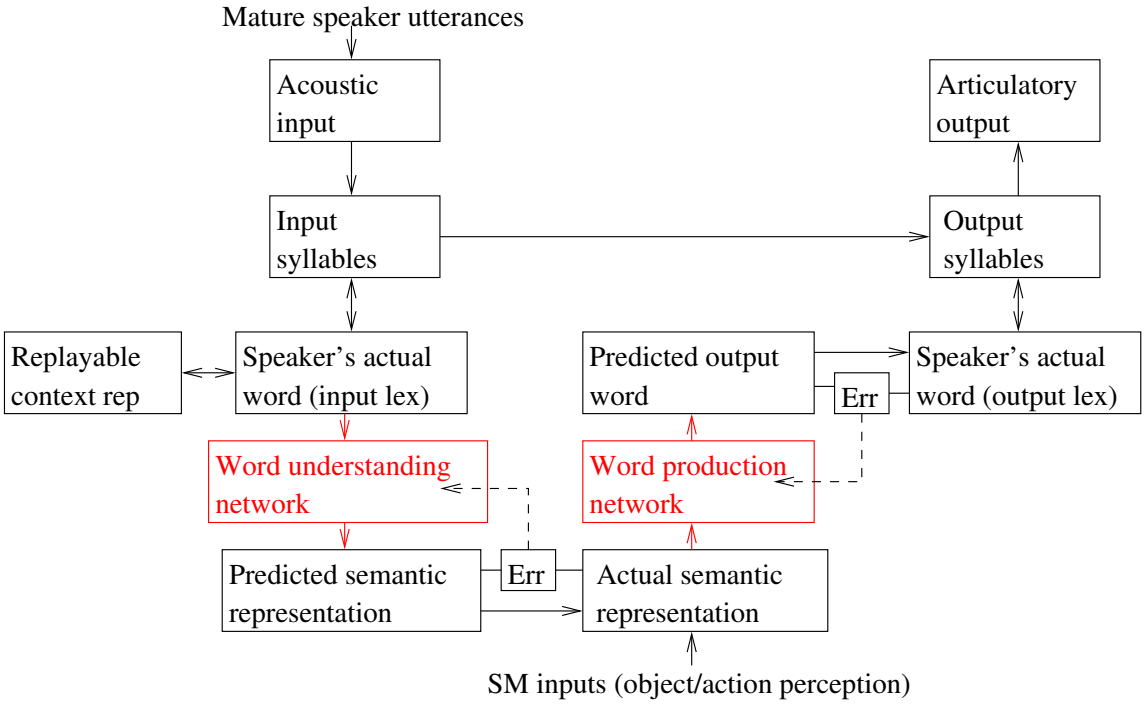


Figure 6.8: Basic networks for cross-situational word meaning learning (in both input and output lexicons)

cross-situational learning. But recall that the infant word learning mechanism also seems to depend on the development of some key social-pragmatic skills: the ability to establish joint attention and the ability to recognise the special status of communicative actions and their underlying intentions. In the next section I will propose how this network can be extended to model the role of these two skills.

6.3.2 Modelling the development of the concept of a communicative action, and its role in word learning

In one sense a communicative action is a kind of episode, just like a 'grab' action. But according to the model outlined in Section 6.2.2.3 there is something special about episodes involving communicative actions, and infants need to learn what this is in order to begin learning word meanings efficiently.

We already have quite a detailed model of how infants represent the episodes they observe in working memory, and in the sensorimotor system, so I will begin by considering a sequence of working-memory and sensorimotor representations generated by an infant experiencing events in the world, alongside a sequence of word representations which he might activate from hearing mature speakers in his environment. The parallel streams of semantic and phonological representations are depicted in Figure 6.9.

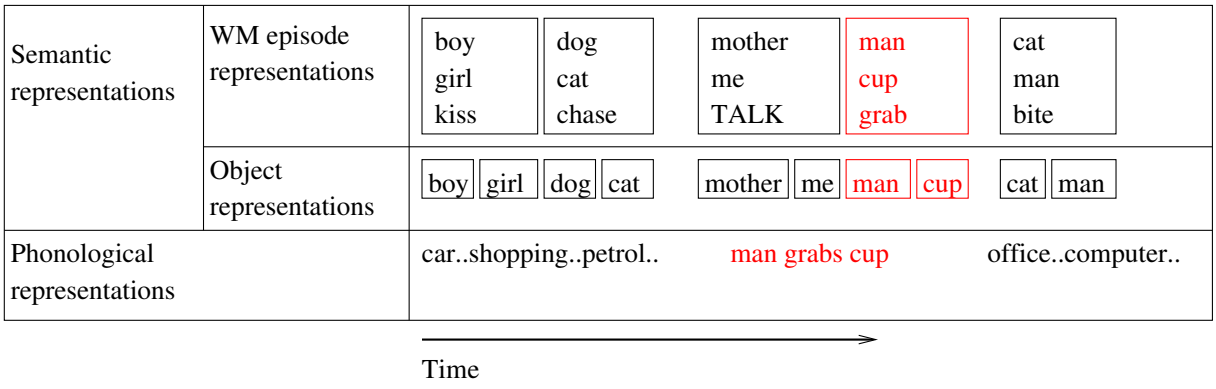


Figure 6.9: Parallel streams of semantic representations and words

The infant must learn to associate the semantic representations in this figure with words. In the model I proposed in Chapter 5, verbs are to be associated with WM episodes. (Since infants begin by learning uninflected verbs, I will assume that motor actions feature most

syntax.

prominently in WM episodes, so they are the first components which are linked to word forms.) Nouns are to be associated with IT object representations. Note that object representations update faster than WM episodes, because each episode involves a sequence of object representations.

Of course, the associations between the semantic representations the infant evokes and the words he hears are very noisy. While the infant is observing a boy kiss a girl and a dog chase a cat, he happens to be hearing a mature speaker talk about cars and shopping and petrol. At any time, the likelihood that the infant's current semantic representation is actually the meaning of the word form he is currently activating is only slightly higher than chance. A cross-situational learning scheme will eventually discover some correspondences between words and meanings, but it will be a slow process.

However, assume that the infant has learned to establish joint attention with an observed agent. A special situation now arises when the infant observes an agent *talk*. At this point, since the infant establishes joint attention with the speaker, and speakers often talk about what they are looking at, there is a moment when the mapping between words and semantic representations evoked by the infant is temporarily much more reliable than normal. In Figure 6.9, this moment is highlighted in red. The words which the infant hears (*man grabs cup*) correspond more reliably to the object and event representations which he activates from perception (man, cup, grab). The correspondence is still not perfect, of course. But still, this moment presents a special opportunity for the infant to learn word meanings.

Note that we can think of the infant's representation of the TALK event as a *cue* which predicts these special opportunities for word learning. My main proposal is that infants develop an awareness of the special significance of communicative events by learning that they are cues which predict a particularly reliable mapping between semantic representations and word forms. In the rest of this section, I will introduce a network model which shows one way this process could occur. Details of the model can be found in Caza and Knott (in preparation).

I first assume that the networks which map word forms to word meanings (and vice versa) can be selectively enabled, i.e. gated 'on', by a special mode-switching cognitive action, as illustrated in Figure 6.10. I assume that executing the 'enable verbal mode' operation turns on the word understanding and word production networks, and also allows learning within these networks. I also assume the networks are off by default, and no word learning occurs unless they are on.

I suggest that the infant learns when to enable verbal mode using the same basic mechanism that is used to learn when to perform ordinary motor actions, namely reinforcement. The reinforcement scheme I propose is one inspired by the sequence-learning model of Braver and Cohen (2000) briefly alluded to in Section 3.2.4.

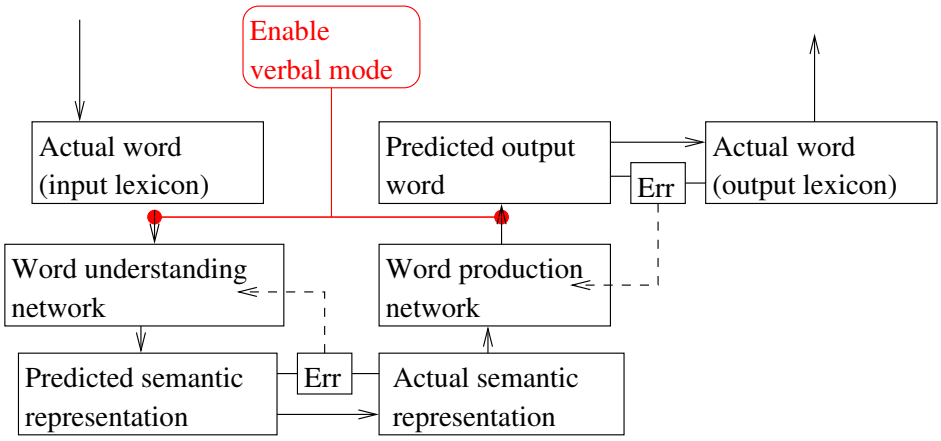


Figure 6.10: A circuit containing an action for enabling 'verbal mode'

Suppose that the infant gets a reward if the word understanding or word production network generates correct output, but a punishment otherwise.[21] Most of the time, enabling verbal mode will result in a punishment, because the incoming streams of words and meanings are very uncorrelated. But after recognising a TALK event, enabling verbal mode is more likely than usual to result in a reward. A reinforcement learning scheme can allow the infant to learn to enable the network after recognising a TALK event, and not in any other circumstance.

The reinforcement scheme just outlined only works if the infant already knows some word meanings; it must be bootstrapped by a simple mechanism for learning word meanings. I assume that the infant begins by acquiring a few word meanings through the cross-situational learning circuit outlined in Section 6.3.1. Once a few word meanings are learned, enabling verbal mode after recognising a TALK event will result in reward more often than enabling it at other times. At this point, the infant will start to learn to enable verbal mode after recognising TALK events. This behaviour will then start to improve the efficiency of word learning, because TALK signals moments when there are particularly good correlations between words and meanings. This in turn means that the infant receives more rewards for enabling verbal mode after TALK events, which further reinforces the policy of enabling verbal mode after TALK events. In other words, once a small vocabulary of words is acquired, a classic bootstrapping process occurs. When

[21] This reward could come externally—for instance from an adult who the child is interacting with. But it could also be internally generated, from the error signals used to train the two networks: low and high error signals could be hardwired to trigger rewards and punishments respectively.

this process is complete, the infant reliably enables verbal mode after recognising a TALK event, and at no other time. This model explains how an infant can transition from a slow, cross-situational method of word meaning learning to a more efficient method, conditioned on the recognition of communicative actions. It can be thought of as an implementation of Tomasello's proposal that infants do not learn words in any numbers until they have acquired the concept of a communicative action—i.e. of developmental processes which occur between around 10 and 18 months of age.

In Section 6.2.2.4 I noted two open questions about Tomasello's model of word learning. One was about the mechanism through which development of the concept of a communicative action supports efficient word meaning learning. The bootstrapping model just outlined can be thought of as an answer to that question. The other question was about how communicative actions and communicative intentions are represented in the brain. In the final part of this section, I will consider this question.

6.3.3 The representation of communicative actions and intentions

As noted in Section 6.2.2.4, the question of how communicative actions are represented is very much an open one in cognitive science. There are two unresolved issues. Firstly, a communicative action is in some respects a regular physical action: it involves a physical agent (the speaker) making motor movements (speech sounds) directed towards another physical agent (the hearer). How do we represent a communicative action as a physical event? This is already an open question, as there is no consensus in cognitive science about how concrete events such as motor actions are represented in the brain. But even more problematically, a communicative action is also 'about' something; it is executed by the speaker in order to evoke a particular representation of the world in the hearer. The utterance itself is an event, but it can also be *about* an event.

There are several difficulties in representing the relationship between an utterance and its content. One difficulty is that we have to devise a means for nesting event representations within one another. This is a technical problem for neural networks, to which many ingenious solutions have been proposed, but the jury is still out about whether any of these is adequate. Another difficulty is that the content of the utterance is not constrained to describe the situation in which the physical utterance occurs. A speaker can make utterances describing distant events, or events which have not happened, or will never happen, or which refer to objects which do not exist. In short, when an infant learns to represent communicative actions, he must learn a representation which supports an appropriate sort of event nesting, and which allows a distinction to be made between the world in which the

utterance is expressed and the world which it describes. What might these representations look like?

My proposal draws on the idea illustrated in Figure 6.9, that there is a relationship of *temporal succession* between the representation of an utterance as a physical event and the representation of the episode it describes. The temporal succession is shown again in Figure 6.11. I already have a suggestion about how ordinary physical events

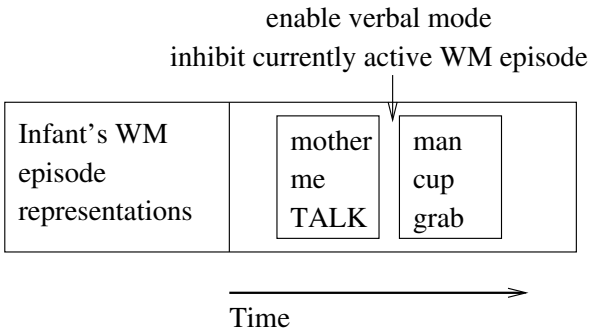


Figure 6.11: A proposal for the representation of communicative actions

are represented in working memory: they are stored as planned sensorimotor sequences, which must be internally rehearsed in order to interface with linguistic representations. This suggestion already buys into the idea that semantic representations are inherently sequential in nature. It is therefore not such a big step to envisage that the representation of a communicative action involves a sequence of two successive working memory episode representations. Semantic representations are still sequences; the new proposal amounts to a suggestion that there can be hierarchical structure within these sequences, so that a single high-level episode representation can consist of a sequence of lower-level episode representations.

The idea of using temporal sequences to solve the technical question of how episode/proposition-sized units can be nested is not new. In fact, as mentioned in Section 6.2.5.4, it is the strategy which Elman uses in his SRN. In an Elman network, an output word unit might be active at two different times, contributing at one time to the representation of an embedding sentence, and at another time to the representation of an embedded sentence. Likewise in my proposal, the physical event of a speaker making an utterance and the embedded event which the utterance describes are expressed in the same representational medium, but at different times. The challenge is to give an adequate account of the relationship between the two sequentially-activated events, to explain how they differ from a sequence of two ordinary physical events, which happen to be observed one after the other.

There are two important things to explain about the relationship. Firstly, we must give an account of the elusive 'aboutness' relation between the two events. The utterance is an event which takes place in the world, but the described event is a representation evoked by the utterance, which need not bear any direct relationship to the world in which the utterance was made. Secondly, there is a sense in which the two events are in fact just one single event. The utterance event should by rights *include* the described event. So some notion of a higher-level event appears to be required. I will consider these two issues in the remainder of this section.

6.3.3.1 A role for 'enable verbal mode' in representing the relation between an utterance and its content

I will first propose that the 'enable verbal mode' operation can play an important role in representing the relation between an utterance as a physical event and its content.

In Section 6.3.2 we saw how an infant learns to recognise the special significance of TALK events by systematically enabling verbal mode after recognition of such events. This operation occurs after all TALK events, and only after such events. The operation causes a change in the infant's cognitive mode, which has the effect that subsequent semantic representations are evoked by phonological representations rather than by sensorimotor stimuli. And since the triggering event is a TALK event, which precisely consists in the production of a stream of phonological representations, the infant will indeed evoke a stream of verbally-activated semantic representations in the new mode. In other words, the infant processes a speaker's utterance in two different ways: first as a physical action, and then, after verbal mode is engaged, as a representation. The 'enable verbal mode' operation allows us to explain how it is systematically treated in these two ways.

Note that the above account hinges crucially on the general principle that episodes have sequential structure, and that representing an episode in working memory involves rehearsing this structure (see Section 3.5). This allows us to think of the described episode as being evoked as an ordinary episode, but at a special *time*, after verbal mode has been enabled. In one sense, the 'enable verbal mode' operation which occurs in between two WM episode representations is just one operation in a sequential stream of rehearsed cognitive operations. But because it also changes the means by which subsequent semantic representations are evoked, it allows us to explain why the WM episode activated next is what the utterance is 'about'. This method of representing the content of utterances (or other propositional attitudes) is not available in static semantic representation schemes, but it is natural if semantic representations are understood as having sequential structure.

Note also that the above proposal relates to the account of nested clausal complements which was introduced in Section 5.6. In that account, I suggested that verbs introducing

clausal complements (e.g. *want* and *believe*) do not denote regular motor programmes, but instead denote special operations switching to new cognitive modes where working memory episode/plan representations are not evoked by sensorimotor experience in the normal way. The 'enter verbal mode' operation is a cognitive operation of just this kind. In fact, it is interesting to observe that the verb *say*, which denotes a verbal action, is syntactically similar to the verb *believe* discussed in Section 5.6.2: it takes a finite clausal (CP) complement. We thus expect the account of *believe* given in that section to extend to the 'talk' action currently under discussion. Recall from Section 5.6 that the special mental-state-accessing operations denoted by *want* and *believe* trigger two other special cognitive operations: one which inhibits the currently active WM episode, and one which initiates 'rehearsal mode'. I suggest that these operations are also triggered by the 'enter verbal mode' operation. I will discuss the role of plan-inhibition in verbal mode immediately below, in Section 6.3.3.2. But I will defer a discussion of rehearsal in verbal mode until Section 6.4. In that section I will argue that infants at early stages of linguistic development do not automatically rehearse the WM episodes they evoke when they enter verbal mode, but that they must learn to do so in order to develop mature syntactic competence.

6.3.3.2 Communicative events as meta-level events

I now turn to the second issue: if the TALK event and the event it describes are represented as separate successive WM episodes, we somehow need to explain how the two episodes form part of a single higher-level episode—or more precisely, how the TALK episode can be understood as *including* the described episode. I will consider this issue in two ways.

First, we can envisage higher levels of hierarchy in the WM episode representation system. Recall that WM episodes are represented as sensorimotor sequence plans. In the model introduced in Section 3.3, there are two working memory media in which plans are represented: in one medium, multiple alternative plans are represented simultaneously, with different levels of activation, and in another medium, alternative plans compete with one another, and a single plan is selected. (I tentatively identified the former medium as PFC, and the latter medium as the supplementary motor area, SMA, and/or the left posterior lateral PFC, Brodmann area 9.) When the currently active plan is completed, a special mechanism (which I suggested may involve the SMA or lateral PFC) allows it to inhibit itself, and make room for the next-most-active plan. If a regular episode representation is a single plan, we can perhaps think of a higher-level episode representation as an assembly in the non-competitive plan medium, involving two highly activated regular plans, one to activate a TALK action, and the other to activate a particular event representation, with the first of these being more highly activated than the second. When activated, this assembly will produce the necessary sequence of regular WM episode representations, via

a process akin to competitive queueing. And the TALK action will trigger an operation bringing about the necessary change of mode in between the two.

This proposal gives some idea of what might be meant by a higher-level episode representation comprising a sequence of two WM episodes. But it does not explain how the TALK episode can be said to *include* the described episode, when it apparently terminates so early. To think about the problem, consider the point which divides the two episode representations, i.e. the point when the hearer switches to verbal mode. Note that this moment is in no sense a true episode boundary: in fact it occurs very early on in the speaker's physical utterance. One way to address the problem is to note that the transition between episode representations which occurs at this point is clearly not driven by the normal mechanisms which identify the boundaries between physical events. Again following the idea briefly introduced in Section 5.6.2, I suggest that the operation of switching to verbal mode must be combined with a second control operation, causing the currently active event representation (TALK) to inhibit itself, to make room for a representation of the described episode. Normally, this self-inhibition operation only occurs when an event is complete (as described in Section 3.3.3). But here it is invoked in special circumstances, by a special mechanism. If the boundary between the two component episodes of a communicative action is not a true event boundary, this goes some way to capturing the idea that the physical TALK event properly contains the described event, rather than just abutting it.

In summary, I suggest that the mature concept of a communicative action takes the form of a sequence of two regular WM episode representations, linked by two special control operations. The first representation is of a physical TALK event. This event specifies who the speaker and the hearer of the action are. It is recognised very early during the interpretation of a communicative action. (In fact, we should perhaps say that the hearer recognises the speaker's *intention* to talk, rather than the complete talk action.) Once the talk episode representation is activated as the dominant event in working memory, it is immediately inhibited (although it is not complete), and the hearer enters verbal mode. These two control operations create the context in which the hearer can represent the episode which the speaker's utterance describes.

One final question is worth considering. At the start of the utterance, the hearer must presumably hear the speaker produce some words before he can recognise her action as a talk event and enter verbal mode. But if he only enters verbal mode after hearing these initial words, is it not too late for these words to participate in evoking the described event? To answer this question, we can refer to the phonological input buffer. The phonological input buffer stores the most recent words which the hearer has heard. After the hearer has entered verbal mode, we can envisage that he evokes the described event from words replayed from the phonological buffer, rather than from words arriving in real time. In this

way, the initial words in the utterance, which were produced before entering verbal mode, can still have a role in evoking the described episode.

6.4 Learning to generate syntactically structured utterances

In this section, I will introduce a model of how infants 'learn syntax'. The model will focus on the generation of syntactically well-formed utterances, though I will briefly discuss interpretation in Section 6.4.5. The model of generation has several distinct components, which are motivated by the dissociations between different forms of aphasia discussed in Section 6.1, as well as by the Minimalist and empiricist models of syntax described earlier in the book. One component is the **word production network** which was introduced in Section 6.3.1, which can generate individual words from semantic representations. Another component is the **word sequencing network**—a network which can learn surface regularities in word sequences. This network is based on the empiricist simple recurrent network introduced in Section 6.2.5.4. A third component is the **control network**: a network which learns to read out words from the sensorimotor sequence which results from an event in working memory being internally rehearsed. This network is Minimalist in inspiration; it adopts the sensorimotor interpretation of Minimalist LF proposed in Chapter 5, and within this framework implements a model of how an infant learns the mapping from LF (sensorimotor sequences) to PF (word sequences).

The three component networks support progressively more complex types of verbal behaviour. The word production network supports the generation of single words and holophrases, the word sequencing network supports the generation of short multi-word utterances (pivot schemas and item-based constructions), and the control network supports the generation of fully-formed clauses. However, there are also interactions between the networks. The clause-producing network relies on the single-word network. And I suggest that the control and word-sequencing networks can interact to produce idiomatic and semi-idiomatic utterances. I will describe the three networks in Sections 6.4.1–6.4.3. In Section 6.4.4, I will discuss how the control and word-sequencing networks can interact to produce an interesting variety of idiomatic and semi-idiomatic utterances.

6.4.1 The word production network: producing single-word utterances

In Section 6.3 I described how the word production network learns to map semantic representations onto words: the basic idea is that when the infant recognises a communicative action is taking place, he establishes joint attention with the speaker in an effort to reproduce the sensorimotor representations the speaker is experiencing, and then trains the network to reproduce the words the speaker is producing. This kind of cross-situational learning, focussed on communicative actions, will enable the infant to learn a set of concrete nouns and verbs. However, we have not yet considered how the infant learns when to use the word production network autonomously, in order to further his own ends. As discussed in Section 6.2.3, an infant needs to learn that in some circumstances, generating words can reliably achieve goals. In this section, I will outline a model of how this happens.

The model again makes reference to the operation of 'enabling verbal mode'. When this operation was introduced in Section 6.3.2, it was involved in deciding when to *learn* words. But once some word meanings have been learned, it can also have a role in deciding when to *use* them.

To begin with, recall that the word production network receives some of its input from the currently dominant working memory episode representation. As discussed in Section 3.5.1, WM episode representations can encode the episode which an agent has just observed or is currently observing, but they can also encode the agent's plans or intentions, i.e. episodes which the agent wants to bring about. Consider an infant whose dominant intention is to get a cup which is out of reach, in a situation where his mother is nearby. His dominant intention will be an event representation something like 'I get the cup', or 'Mummy gives me the cup'. From this representation, the infant could read out a verb like *get* or *give*, or a pronoun like *I* or *me*.[22] The infant will also probably be attending to his mother or the cup, and so will be evoking a transitory object representation, from which he could read out an open-class common noun like *cup*, or a proper noun like *Mummy*. The point is, this is a good situation in which to enable verbal mode. If the child enables verbal mode—assuming he knows the appropriate word meanings—the word or words which he produces will probably allow his mother to recognise his intentions, and fulfil them. We can therefore envisage a network which learns through reinforcement when it is a good idea to enter verbal mode.

[22]The syntax of pronouns was not discussed in Chapter 4. But there is a lot in common between pronouns and agreement inflections. For instance, in Romance languages, pronouns can appear attached to ('cliticised to') verbs; in these positions, they appear to combine with verb stems in the same kind of way as inflections do (see e.g. Belletti, 2001). I thus assume that pronouns can be read out from WM episodes as well as inflected verbs.

There is a certain amount of complexity to this network. The infant needs to learn to perform a communicative action, which involves some sensorimotor operations in its own right. He must first enter action execution mode, and then obtain the attention of an interlocutor—if no-one is attending to him, his utterance probably will not have its desired effect. Only then should he enter verbal mode. These preparatory actions are all part of the physical action of talking. But there is a problem: what is the mechanism which produces these actions, given that the infant's dominant intention is the entirely unrelated action of obtaining a cup?

A solution can be envisaged which makes use of the idea of competitive queueing between WM episode representations just proposed in Section 6.3.3.2. In that proposal, there is one working memory medium in which several alternative intention representations can be maintained, and another medium in which these compete, so that a single dominant intention is selected. When the infant is in a situation where his currently dominant intention could be achieved through a linguistic action, we can envisage that he learns to activate a very strong representation of a communicative action in the non-competitive working memory medium, which is strong enough to take over as the dominant intention in the competitive medium, forcing the substantive intention to get the cup into second place. This dominant communicative intention will then result in the infant performing an action. He will enter action execution mode, then attract the attention of an interlocutor, and finally activate the TALK operation. Crucially, he will now enable verbal mode, *and inhibit the currently dominant WM episode representation*, so that his previously dominant intention to grab the cup once again becomes dominant. At this point, his word production network will begin to produce words. Note that these words will reflect his substantive intention to grab a cup, rather than the intention which was actually instrumental in generating the communicative action. And note also that the infant's own communicative action takes the form of a sequence of two successive WM episode representations, the first of which is a TALK episode introducing operations inhibiting the current episode and entering verbal mode. These sequences exactly mirror the sequences evoked when the infant observes another agent perform a communicative action, as described in Section 6.3.3.

One detail about the network just outlined concerns the directionality of processing during verbal mode. When the infant enters verbal mode, he must produce words rather than interpret them. I have already suggested in Section 6.1.3.3 that there must be distinct 'speaking' and 'hearing' modes. In Section 6.3 we posited a special operation of entering verbal mode for the purposes of word learning, but it made no reference to the distinction between word production and word interpretation. But note that we do already have a distinction between *action execution* and *action recognition* modes (see Section 2.8.3). When the infant performs the communicative action, he establishes action execution mode, as for any action he performs. I suggest that this mode is sustained when the infant enters

verbal mode, and this is what causes the infant to produce words rather than to interpret them. On this account, speaking mode can be neatly defined as the combination of verbal mode and action execution mode, and hearing mode can be defined as the combination of verbal mode and action recognition mode.

6.4.2 The word sequencing network: producing short multi-word utterances

I assume that the infant also has a network which allows him to predict the next word in a sequence of incoming words. The **word sequencing network** is basically an Elman network, which receives a single word as input, and learns to predict the word which follows it. As discussed in Section 6.2.5.4, an Elman network has a hidden layer in between input and output, which is copied and used as an additional input at the next time point. The copied hidden layer is called the 'context' representation; after training, it comes to represent the history of recent inputs to the network, in a format which encodes any sequential patterns which help predict the next word.

I now suggest that the word sequencing network has a role in the production of utterances, rather than just in encoding surface linguistic patterns. To give it this role, I assume that the network receives input from semantic representations as well as from the current context and current word, as shown in Figure 6.12. An extended Elman network of this

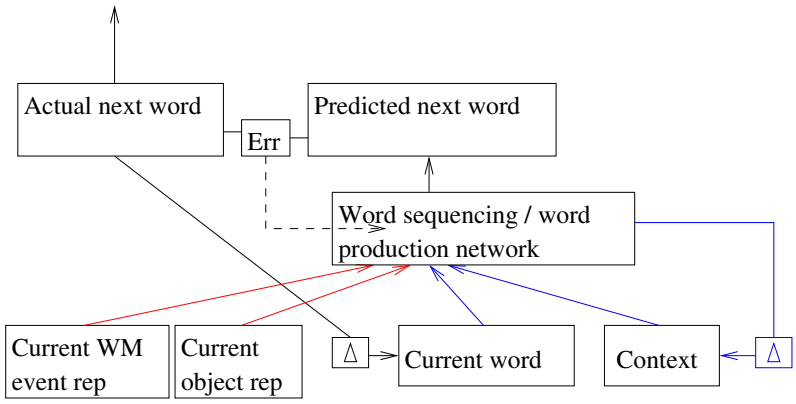


Figure 6.12: A combined word-production/word-sequencing network

sort was discussed by Chang (2002), who labelled it a 'prod-SRN' network. Chang noted several shortcomings of this type of network as a model of sentence generation, which I will discuss at the end of this section. The specific network I describe in this section differs

from Chang's in a few respects; implementation and performance details are given in Takac *et al.* (in preparation).

I assume the word sequencing network is only enabled, and only trained, in 'verbal mode', similarly to the word-meaning networks discussed in Section 6.3. (Specifically: the network is trained when the infant recognises that a mature speaker is producing an utterance, establishes joint attention with this speaker, inhibits his current WM episode, and moves into verbal mode.) Each training item consists of a sequence of phonological word representations, plus a tonically active semantic representation of the jointly attended episode, comprising a WM episode and a single active object representation in IT. At each time point during training, the network makes a prediction about the next word in the utterance. As input it uses the speaker's actual word at the current time point (which is a delayed copy of the actual next word from the previous time point), plus the current context (which is a delayed copy of its own hidden layer), plus the tonically active semantic representation. The predicted next word is thus a function both of the current word and context, and of the semantic representation.

Because the word sequencing network uses semantic inputs to predict words, it can be thought of as *incorporating* the word production network described in Section 6.3, which maps word meanings to word forms.[23] As discussed in that section, the word production network takes a word meaning as input, and delivers a phonological word form as output. It is trained on loosely correlated pairs of word meanings and word forms; through cross-situational learning, sharpened by the developing concept of communicative actions, it learns mappings between meanings and word forms. The network shown in Figure 6.12 can be trained in just this way, by presenting a semantic representation (a WM episode or an object representation) as input and a word form as output. Regardless of what appears on the other inputs, the network when trained in this way will learn associations between word meanings and word forms.

At the same time, the network can learn about commonly occurring sequential patterns of words. If it is presented with utterances containing frequently occurring word sequences, the recurrent component of the network will learn these sequences, just like a regular Elman network. Of course, since it also has semantic inputs, it can perform somewhat better than an ordinary Elman network, since the tonically active semantic inputs bias it towards generating words conveying elements of the semantic message. The network can thus learn to produce short multi-word utterances which convey simple semantic messages (object-episode pairs) using frequently-attested word sequences.

Note that the two components of the trained network complement each other. The

[23]It should more properly be called the 'word-production/word-sequencing network', but I will call it the 'word sequencing network' in the text, for brevity's sake.

individual components both typically generate preferences for more than one word. The semantic 'word production' component receives two separate semantic inputs, one reflecting the currently attended object, and one reflecting the current WM episode, so it typically makes predictions about two words, a noun and a verb. At any time, at most one of these words can match the speaker's actual word. The word sequencing network is also typically unable to generate predictions about specific words. As discussed in Section 6.2.5.4, there are normally several words which provide good continuations of a given word sequence; an Elman network really just produces a probability distribution about the likely next word, in which the probability is often shared between several different words. But in combination, the two components of the current network can typically converge on a single word which is appropriate both on semantic and word-sequencing grounds.

Importantly, the word-sequencing network functions as a mechanism for generating a sequence of words describing the current semantic inputs, rather than just an individual word. In generation mode, the word which the network produces is used as the current word at the next iteration. In this configuration, even if the semantic representations remain static, and deliver a static prediction about the likely next word, the network's recurrent inputs both update after each word, so the network can select different words at different iterations. Most obviously, this allows it to generate simple two-word utterances pairing a noun and a verb, in a way which reflects the prevalent word-ordering conventions in the infant's exposure language. In other words, the mechanism just described can produce 'pivot schemas', of the kind described in Section 6.2.4.2. For instance, if the infant's intention is to get a cup—and he is attending to the cup—and his event representation allows generation of pronouns—he could produce an utterance like *my cup* or *cup my*.

In fact, the network can generate utterances which are longer than two words, even though it only has access to two semantic inputs. The 'closed-class' words in a sentence (e.g. determiners like 'a' or 'the', or prepositions like *in* or *to*) are often quite predictable on word-sequencing grounds alone. For instance, if a mature speaker produces the utterance *I want*, the word which follows is very likely to be *a* or *the*, and very unlikely to be *cup*. The relatively small number of possible words in this context makes both words likely enough for the network to generate one of them (perhaps stochastically) on word-sequencing grounds alone. Of course, after a determiner is produced, the two networks will again converge in predicting *cup* as the word which comes next. The network is thus able to generate short multi-word sequences involving predictably placed function words.

There are two interesting things to note about the word sequences which the network produces. Firstly, note that closed-class words are produced purely to ensure utterances conform to surface word-sequencing constraints, rather than to reflect anything in the 'message' being conveyed. They basically have an *idiomatic* function rather than a semantic function; they are included in order to reproduce surface patterns in the exposure language.

We can think of them as simple 'constructions', in the sense defined in Section 6.2.5.3. Secondly, note that there are likely to be small biases in the infant's training data, in favour of particular closed-class words in particular contexts. Given that closed-class words are selected by the word-sequencing network alone, these small biases provide the only means for choosing between different alternative words. We therefore expect the development of *item-based* constructions, in which particular closed-class words are tied to particular patterns of open-class words. Recall from Section 6.2.4.3 that infants' earliest grammatical constructions are item-specific in this way.

The network described in this section is thus able to simulate the two earliest stages in syntactic development: pivot schemas and item-based constructions. However, there is a fairly obvious limit on the complexity of uttterances which the network can produce. The limit comes from the fact that the semantic representation which provides input to the system is incomplete. The utterance to be produced must describe a whole event. As noted in Section 6.4.1, the word production network can generate verbs from WM episode representations, and it can also probably generate pronouns from these representations. But it cannot generate open class nouns from WM episodes; these must be generated from object representations in IT, and in my sensorimotor model there is only one object representation active in IT at any given time. In fact, of course, in the model I introduced in Chapter 5, the semantic representation which provides input to the linguistic system is not a static representation at all, but a dynamic one: entertaining an episode consists in actively *replaying* the sensorimotor sequence involved in experiencing it. During this replay process, different object representations become active in IT at different moments. First the agent is attended to, and then the patient is attended to. Then the agent is reattended to (as a side-effect of classifying the action), and then the patient is reattended to (haptically, when the action is complete). I therefore envisage that the next stage of syntactic development involves the infant learning to generate a sequence of words *while internally rehearsing the currently active event representation.* The crucial thing which the infant must learn is that a *replayed* event representation provides a better source of information from which to reproduce an utterance than a static one. The network which learns this task will be discussed in the next section.

The possibility of extending an Elman network with semantic inputs has already been explored in some detail in a very informative paper by Franklin Chang (2002). Chang envisages semantic event representations in which the agent and patient are represented by separate assemblies; i.e. in which there are separate representations of each object as an agent and as a patient. These representations are fed to the hidden layer of an Elman network just as in Figure 6.12, delivering a static representation during the whole sentence generation process. The network learns to generate sentences quite effectively, but Chang notes that its separate representations of agents and patients mean that it does

not generalise well: a word encountered in the training data only as a subject cannot be generated in object position. Chang experiments with a second network using a more sophisticated event representation, with explicit assemblies encoding the abstract concepts of 'agent' and 'patient', which are separately bound to object representations. This network learns conventions about the sequencing of 'agent' and 'patient' representations in the exposure language, and therefore generates a particular sequence of object representations as part of the sentence generation process. The network which I describe below picks up on Chang's idea that semantic inputs can be encoded as sequences (though it develops this idea in a rather different way). But I do not want to abandon the idea that an Elman network supplemented with a static semantic input can play a role in modelling language. I suggest that it provides an excellent model of the early stages of syntactic development. And later, in Section 6.4.4, I will suggest it can also play a role in an account of idioms in mature language.

6.4.3 The episode-rehearsal/control network: generating word sequences from sensorimotor sequences

In Sections 6.4.1 and 6.4.2 I introduced models of how infants learn to produce single words and short multi-word utterances during the early stages of syntactic development. In this section I present a model of how infants acquire mature syntactic competence—the syntactic knowledge which grammars of adult language try to elucidate. Since my model of mature syntax is based on Minimalism, this section returns to the main theme of the book: a sensorimotor interpretation of Minimalist syntax. In this section, I will outline a model of syntax acquisition which is inspired by Minimalism, but which also draws on the neural and developmental models of language presented earlier in the chapter.

The main idea in this book is that mature syntactic knowledge is more closely related to sensorimotor knowledge than is normally assumed. I have used the scenario of a man grabbing a cup to illustrate this throughout the book. In Chapter 2 I argued that the sensorimotor process of experiencing a cup-grabbing event has a characteristic sequential structure, and in Chapter 3, I argued that a cup-grabbing event is represented in working memory as a planned, replayable sensorimotor sequence. In Chapter 4 I introduced a theory of natural language syntax, Minimalism, which models adult syntactic competence by positing an underlying level of syntactic representation called logical form (LF), from which the surface phonetic form (PF) of a sentence is read out. LF is relatively invariant over languages, because the mechanism which generates LF representations is innately specified. But the mechanism which reads out PF from an LF derivation is parameterisable, and an infant must learn the correct parameter settings for his native language. In Chapter 5,

I proposed that the LF of a sentence can be given a sensorimotor interpretation: the LF of a sentence describing a concrete event can be understood as a description of the sequence of sensorimotor signals evoked when the working memory representation of that event is internally rehearsed. If LF is a rehearsed sensorimotor sequence, then the question of what is involved in learning the mapping from LF to PF representations becomes a very concrete one: what the infant has to learn is how to read out a sequence of words from a replayed sensorimotor sequence. In the model I present here, I will assume that the infant already has a word production network, which can express individual semantic representations (objects and WM episodes) as single words, which is learned using the mechanisms described in Sections 6.3 and 6.4.1. In order to acquire a mature syntactic competence, therefore, what the infant needs to learn is *when during the replay process to overtly pronounce words*. For a cup-grabbing event, there are two opportunities to pronounce a word denoting the agent of the event, and two opportunities to pronounce a word denoting the patient. The grab action can be read out at any point, from the planning representation which stores the whole sensorimotor sequence, which is tonically active during replay. The infant needs to learn when semantic representations should generate overtly pronounced words, and when they should not.

Before I discuss the details of a network which learns this, it is useful to make two points. Firstly, note that we have already envisaged two possible mechanisms for selectively enabling or disabling word production. One is the mechanism responsible for enabling verbal mode, which selectively engages and disengages the whole word-production network, discussed in Section 6.3.2. The other is a mechanism involved in switching between action-execution and action-recognition modes, which gates open or closed the connection between premotor cortex and motor cortex, as discussed in Section 2.8.3. If this mechanism is used, then the word production network will always generate premotor articulatory word representations, but these representations will only sometimes be passed to the articulatory system in motor cortex. (I will assume this latter mechanism, for concreteness' sake.) Secondly, recall from Section 6.1.4.6 that one of the nonlinguistic functions of Broca's area is in the area of 'cognitive control'—specifically, it is involved in the enablement or inhibition of learned responses (see Novick *et al.*, 2005). So the proposal I am making about what is involved in mapping from LF to PF builds on machinery which is already in place for other reasons, and is consistent with a known role of Broca's area. In fact, if LF is interpreted as a sensorimotor sequence presenting opportunities for words to be read out, then it is rather simple to envisage a network that learns which opportunities to take in order to reproduce word ordering conventions in the exposure language.

I begin by assuming that the network is trained when the infant has noticed a mature speaker making an utterance, has established joint attention with the speaker so that they are observing the same event in the world, and has stored the speaker's utterance (a

sequence of words) in his phonological input buffer. Note that in this situation the infant has *two* sequential representations in working memory, which are independently replayable: one is a sequence of words, and one is a sensorimotor sequence. (Recall from Section 3.1 that phonological sequences and semantic event representations are assumed to be stored in separate media in working memory—the former in the 'phonological loop' and the latter in the 'episodic buffer'—but that there is assumed to be an interface between these stores.) The infant must now learn to reproduce the sequence of words in his phonological input buffer, by replaying the event-denoting sensorimotor sequence and then selectively engaging the word-production network at the right moments.

I propose that there is a specialised network which does this learning which I will call the **control network**, which interacts in various ways with the word-production and word-sequencing networks already described. The control network becomes active when an event-denoting sensorimotor sequence is rehearsed in tandem with a sequence of words. It learns at which points during the rehearsal of the sensorimotor sequence it should enable the word-production network, and generate a word form from the currently active sensorimotor representations, in order to reproduce the rehearsed word sequence.

To introduce the control network, I will begin by presenting some more detail about the system which generates sensorimotor sequences, and the system which reads words off these sequences.

6.4.3.1 The episode rehearsal system

I will term the systemr responsible for rehearsing sensorimotor sequences the **episode rehearsal system**. It is shown in Figure 6.13. The episode rehearsal system comprises a WM episode representation in PFC (which takes the form of a prepared sensorimotor sequence), together with the areas in which transitory representations become active when an event is rehearsed. I have shown two such areas; one is the area where object representations are evoked (probably in IT), and the other is the area where the updating context representation which helps drive the rehearsed sequence is evoked (see Section 3.2.2.1). I have also shown the word production network, which receives its input from the event representation and object representation areas. (Recall from Section 6.4.2 that the word production network is incorporated within the word sequencing network. I will omit the sequencing network in the current section, for clarity's sake, but will discuss its role in Section 6.4.4.)

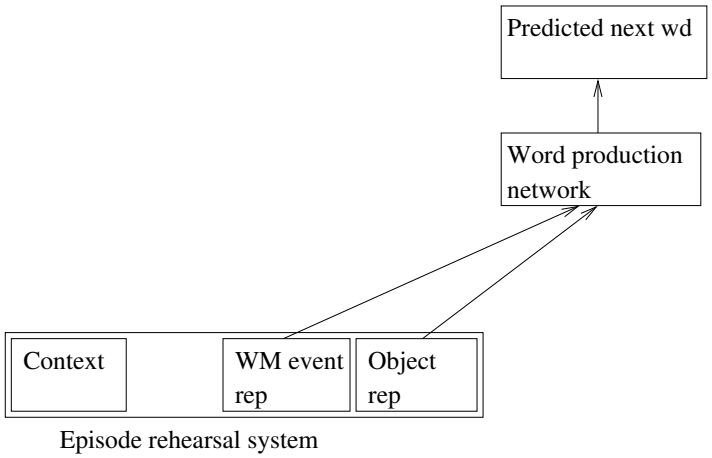


Figure 6.13: The episode rehearsal system, and the word production network using it as input

6.4.3.2 A pattern generator to produce the X-bar schema

Consider what happens if the agent replays a cup-grabbing event in the episode rehearsal system. There are four iterations to the replay process, which I will denote $I1$–$I4$. The WM episode representation will be tonically active during each iteration, while the object representation and context areas will each hold transitory representations which change with each iteration. The signals active at each iteration are shown in Figure 6.14. The

Iteration	Context	WM episode rep	Object rep
$I1$	C_1	$plan_{attend\_man/attend\_cup/grasp}$	*man*
$I2$	C_2	$plan_{attend\_man/attend\_cup/grasp}$	*cup*
$I3$	C_3	$plan_{attend\_man/attend\_cup/grasp}$	*man*
$I4$	C_4	$plan_{attend\_man/attend\_cup/grasp}$	*cup*

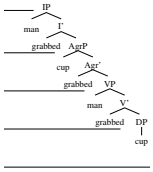


Figure 6.14: The states of the episode rehearsal system during rehearsal of a cup-grabbing episode, matched to XPs in the LF of the associated sentence

figure also shows the LF structure of the sentence *The man grabbed a cup*, with each XP (IP, AgrP, VP and DP) aligned next to the iteration which it represents.

Note that at each iteration, there is an opportunity to read out two sensorimotor representations: the WM episode can be read out (as an inflected verb), and the object can

be read out (as a noun). While we have a mechanism in the network for transitioning from one iteration to the next, we do not yet have a mechanism for ordering the opportunities to read out a noun and an inflected verb *within* a single iteration. In the syntactic model, each iteration is represented by a single X-bar schema, and each X-bar schema has a standard internal structure, providing first a position for a noun (the specifier) and then a position for the inflected verb (the head). In order to echo this schema, we need a mechanism which creates some temporal structure within each iteration of the replayed sensorimotor sequence, providing first an opportunity to read out the object representation, and then an opportunity to read out the WM episode.

I propose a simple mechanism for creating these sequencing opportunities: a **pattern generator**, which alternately allows input to the word-production network from WM episodes and from object representations. The pattern generator cycles through two phases, allowing input from object representations during the first phase and from episode representations during the second phase. At the end of the second phase, it provides the signal to advance to the next iteration of the rehearsed sequence. The episode rehearsal system augmented with the pattern generator is shown in Figure 6.15.

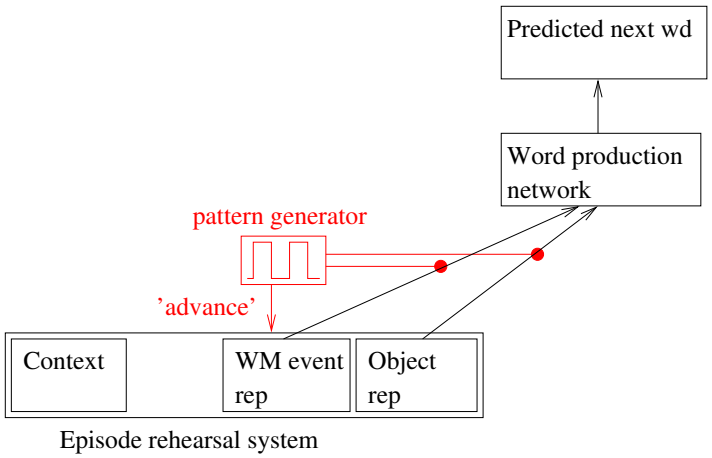


Figure 6.15: The episode rehearsal system, augmented with a pattern generator

Recall that we have already envisaged a role for a cyclical pattern generator during sentence processing, at a somewhat lower level. In Section 6.1.1.1, I outlined Hartley and Houghton's (1996) model of the representation of syllables. Each syllable is represented by an onset and a rhyme unit; a cyclic pattern generator alternates between activating onset and rhyme units, resulting in the cyclic rhythms of syllabic speech. The pattern

generator I have just proposed has a very similar function, but at a higher level of linguistic organisation. I suggest it is responsible for the fact that all XPs have the same internal structure, with the specifier appearing before the head.

When the episode rehearsal system is augmented with the pattern generator just described, the sequence of sensorimotor representations which are presented to the word-production network when an episode is rehearsed is shown in Figure 6.16. As before, each

Context	Phase	WM episode rep	Object rep
C_1	a		*man*
	b	$plan_{attend\_man/attend\_cup/grasp}$	
C_2	a		*cup*
	b	$plan_{attend\_man/attend\_cup/grasp}$	
C_3	a		*man*
	b	$plan_{attend\_man/attend\_cup/grasp}$	
C_4	a		*cup*

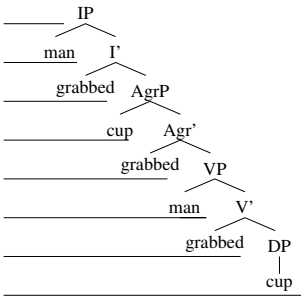


Figure 6.16: The sequence of inputs to the word-production network after modulation by the X-bar pattern generator, matched to individual positions in the LF of the associated sentence

iteration of the rehearsed episode corresponds to one of the XPs in the LF of the sentence *The man grabbed a cup*. But now within each iteration there are two phases, created by the cyclic operation of the pattern generator. And each phase corresponds to a specific syntactic position within the associated XP. The sequence of sensorimotor signals presented to the word-production network now corresponds perfectly to the hierarchical structure of LF positions at which words can appear.

Naturally, not every sensorimotor signal in the sequence should result in an overtly pronounced word. I assume that the word production network produces a word form—a premotor articulatory plan—for each signal in the sequence. The control network must learn the conventions in the exposure language about which of these articulatory plans are executed and which are suppressed.

6.4.3.3 The control network

As shown in Figure 6.17, the control network takes as input a representation of the current stage in the rehearsal of the sensorimotor sequence, and delivers as output a signal dictating whether the word form produced by the word-production network is to be sent

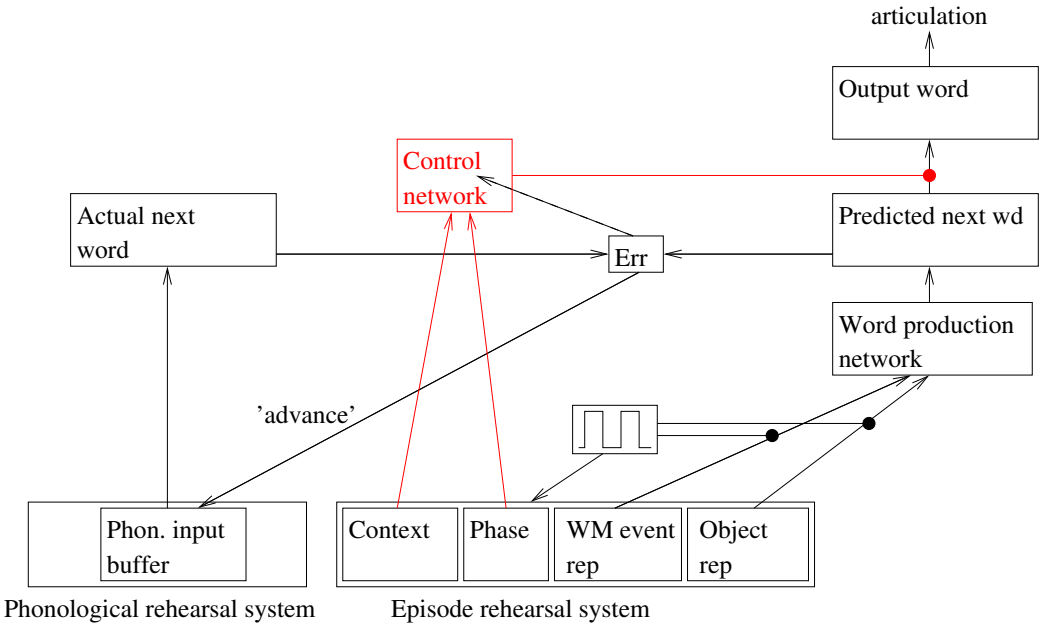


Figure 6.17: The control network

to the articulatory output system, or suppressed. The current stage in episode rehearsal is read from the context representation, which updates after every iteration, and from a representation of the current phase produced by the pattern generator.

The control network is trained on a sequence of words, stored in the phonological input buffer, paired with a WM episode representation. Training proceeds by initiating rehearsal of the WM episode (i.e. generating the first sensorimotor signal) and of the word sequence (generating the first training word). At each stage, the word-production network generates a word from the currently active signal, which is compared with the current training word.[24] If they are the same, the control network learns to allow the output of the word-production network into the articulatory system in the current context; If not, it learns to suppress articulation of this word. In either case, we then advance to the next stage in episode rehearsal. Whether we also advance to the next word in the phonological buffer depends on whether the current training word was accurately predicted by the word-production network. If it was, we advance the phonological buffer. If it was not, we retain the current

[24]Note that the word representations produced from the phonological input buffer are premotor representations, as discussed in Section 6.1.3.1. So it is feasible to compare them directly with the representations generated by the word-production network.

word, and try to match it at the next stage.

This training regime allows the control network to learn different word ordering conventions. For instance, if we are training on a VSO language like Māori, training items will have the form shown in the upper table in Figure 6.18. The lower table shows the

Word sequence	*cup*, *man*, *grabbed*						
SM sequence	MAN,	GRAB-PLAN,	CUP,	GRAB-PLAN,	MAN,	GRAB-PLAN,	CUP
Context/phase	C_1a	C_1b	C_2a	C_2b	C_3a	C_3b	C_4
Network output	–	*grabbed*	–	–	*man*	–	*cup*

Figure 6.18: Upper table: example of a training item for a VSO language. Lower table: the word sequence the control network will learn to generate

sequence of words which the network will generate after it is trained. In the first stage (context/phase C_1a), the word-production network will generate *man*, and the first training word will be *grabbed*. The generated word does not match the training word, so the control network learns to withold the word generated in this context/phase. In the next stage, the word-production network will generate *grabbed*, and the training word will still be *grabbed*. These do match, so the control network learns to pronounce the word generated in context/phase C_1b, and advances to the next word in the training sequence (*man*). In subsequent stages, by the same principles, the control network will learn to withold the word generated in context/phases C_2a and C_2b, to pronounce the word generated in context/phases C_3a, to withold the word generated in context/phase C_3b, and to pronouce the word generated in context/phase C_4. In other words, the control network learns to read out the inflected verb in its highest position (within IP), and to read out the subject and object in their lower positions (within VP and its DP complement). If training data is given for languages with different ordering conventions, the control network would learn to read out the subject, verb and object at different stages of episode rehearsal.

In fact, if the network is only trained on simple transitive sentences, it will not learn all of the possible ways of reading out subject, verb and object. The training algorithm is biased towards reading out a word at the first opportunity. For instance, if we give the network training items from an SVO language, it will produce the subject and verb in the first iteration (i.e. in IP), and the object in the second iteration (AgrP), as shown in Figure 6.19. In order to force constituents to be read out at lower positions, we would have to include training sentences featuring adverbs and/or negation. I have not yet given a sensorimotor account of these constructs, but if we assume that adverb- or negation-denoting signals appear at the positions predicted by my sensorimotor interpretation of

Word sequence	*grabbed*, *man*, *cup*						
SM sequence	MAN,	GRAB-PLAN,	CUP,	GRAB-PLAN,	MAN,	GRAB-PLAN,	CUP
Context/phase	C_1a	C_1b	C_2a	C_2b	C_3a	C_3b	C_4
Network output	*man*	*grabbed*	*cup*	–	–	–	–

Figure 6.19: A training item for an SVO language, with the output which a trained network will generate

LF, then the network will learn to read out verbs and objects low when presented with English-like training sentences, and to read out verbs high and objects low when presented with French-like sentences.

6.4.3.4 Discussion

The network just presented can be discussed from several theoretical perspectives. In this section I will consider the network as a contribution to Minimalist theory, as a model of neural language mechanisms, as a connectionist model of language processing, as a model of language development in infants, and as a model of the interface between language and sensorimotor cognition.

Relation to Minimalism First and foremost, the network is intended to extend the sensorimotor interpretation of Minimalist syntax given in Chapter 5. In that chapter, I proposed that the LF of a sentence can be understood as describing a replayed sensorimotor sequence. In the current section, I have introduced a network which learns when to read out words during this replay process. My claim is that this network can be understood as a model of how infants learn the mapping between LF and PF in their native language.

Naturally, a great deal of interpretation is needed in order to understand the network in Minimalist terms. To begin with, we must buy the idea that an LF structure describes a replayed sensorimotor sequence. This in turn involves a radical reinterpretation of 'the generative mechanism' which produces LF structures. In Minimalism, the generative mechanism builds an LF structure from the bottom up, starting with a VP, filling in its argument positions, attaching higher XPs (AgrP and IP), and moving verb and arguments to higher positions. In my reinterpretation, there are no direct correlates of these stepwise operations. Rather, as discussed in Section 5.5.1, the constraints on the derivation of an LF structure are interpreted as constraints on the possible sensorimotor sequences which an agent can experience, and as a reflection of how sensorimotor sequences are stored and replayed in working memory. In particular, there are sensorimotor reinterpretations of

the constraints which force NPs to raise to Case-assigning positions (Section 5.4.2) and those which force V to raise to higher head positions (Section 5.4.3). The reinterpreted generative process no longer just provides an abstract model of a space of well-formed sentences. The LF of a sentence represents a specific cognitive process—the process of internally simulating experience of the associated episode. My interpretation of the LF-PF interface is part of a model of grammatical competence, but it also happens to be a model of sentence generation in the psychological sense: i.e. a model of the mechanism which takes an episode representation and produces a sequence of words. Minimalist linguists do not typically consider issues to do with sentence processing. But my reinterpretation of LF naturally extends to these issues. So although there is a lot of reinterpretation needed in order to understand the episode-rehearsal/control network as a model of the LF-PF mapping, it may be worth it.

In some ways, the network model of the LF-PF mapping gives substance to some of the more tricky concepts in the Minimalist model. In particular, it makes a concrete proposal about the nature of 'phonologically empty' or 'unpronounced' constituents in a syntactic tree. In the current model, such constituents are articulatory plans which are activated at a premotor level, but witheld from the articulation system. This idea is a very natural one within the sensorimotor model of phonological representations given in Sections 6.1.1.1 and 6.1.3.1.

Relation to neural substrates of language The network just described also connects in several ways with models of how language is implemented in the brain. Most obviously, there is again a clear dissociation between the component of the network which stores word meanings (the word-production network) and the other components, which collectively hold 'syntactic knowledge'. We can therefore model Broca's aphasia, in which syntactic knowledge is impaired while knowledge of individual word meanings is retained. In addition, both of the networks which interact to model syntactic competence reproduce known functionality of Broca's area and associated prefrontal regions. One of the networks stores episodes in working memory, and supports their replay as sensorimotor sequences. Prefrontal cortex is known to be involved in both these processes, as discussed in Section 3.2. The other network learns when to withold reflexive motor responses. This is also a known function of Broca's area, as discussed in Section 6.1.4.6. In fact, given that the control network has a role in suppressing articulatory responses, we may expect that damage to this network results not only in a loss of syntactic competence, but also a loss of 'fluency', as is attested in Broca's aphasia (see Section 6.1.4.1).

Relation to connectionist models of sentence processing The network also has some merits as a connectionist model of sentence production. The merits are mainly due to the fact that semantic representations are modelled as temporal sequences, rather than as static assemblies. This has several advantages.

For one thing, the sequential structure of semantic representations helps ensure that network's learning generalises well to sentences featuring words in syntactic positions where they have not appeared before. As Chang (2002) notes, in a simple static semantic representation scheme where the agent and patient of an episode are represented in separate banks of units, there is poor generalisation. In this scheme, which Chang terms 'binding by space', there are two distinct semantic representations of each object, one as agent and one as patient, and each object representation has to be independently associated with a word. If a word only appears in subject position during training, a network will not be able to generate the word in object position. However, as Chang has shown, if we can distinguish between agent and patient by the *times* at which object representations become active, then we only need a single bank of units to represent objects, and only a single set of associations between object and word representations. The focus then shifts to the mechanism which sequentially activates object representations in the right way. In Chang's model an infant must learn the basic ordering of agents, verbs and patients in the exposure language—i.e. must learn an abstract pattern in the language. In my model, the basic ordering is language independent; it is ultimately determined by the architecture of the sensorimotor system, and the way it interacts with the world. What the child learns from linguistic data is which elements in the sequence to pronounce. In either case, learning about syntax is largely decoupled from learning about word meanings, so words learned in one syntactic context can be used in others.

For another thing, modelling semantic representations as sequences is useful when it comes to representing nested semantic structures. Event representations can be nested inside one another; for instance, as discussed in Section 6.2.2.4, the representation of a communicative event must contain the representation of the event it describes as one of its components. The question of how to represent nested events, and the nested syntactic structures which express them, is a key one in cognitive models of language. We certainly do not want to envisage a 'binding-by-space' solution to the problem, where different events are represented simultaneously in separate media; the storage requirements of such a scheme would be prohibitive, and its generalisation would be terrible. If semantic representations are inherently sequential, we can assume a single medium for storing events, and envisage that nested events are represented by a sequence of representations in this medium, such that the nesting and nested events are active at different times. In Section 6.3.3 I discussed in some detail how this might work for communicative event representations. I will discuss relative clauses in Section 9.7.3. Naturally there are many other proposals about how the

semantics of embedded events are represented, which do not involve a notion of sequencing (see e.g. van der Velde and de Kamps, 2006; Plate, 2003). However, thinking of a semantic representation as a sequence certainly creates some scope for modelling nested events.

Relation to models of language acquisition The network also has some appeal as a model of how infants learn language. A particularly useful feature is that it allows 'offline' learning of syntax. The infant is assumed to establish joint attention with an adult, to observe an episode and to hear the adult's utterance. The infant stores both the episode and the utterance in working memory (in separate media); the learning algorithm runs on a *replayed* sensorimotor sequence and a *replayed* utterance. This means that there need not be tight synchronisation between the time at which the utterance is heard and the time at which the described episode is observed. The infant might begin hearing the utterance before he starts to observe the episode. The episode may last for much longer than the utterance, and take a correspondingly long time to observe. It may even be that the infant first observes the episode, and only realises that a speaker is describing it after it is finished. In each case, since learning uses representations of the episode and utterance buffered in working memory, the exact timing of their actual occurrence is not a crucial issue: all of the learning happens during synchronised *replay* of the utterance and the event.

The network may also offer an interesting extension to the account of word-meaning learning given in Section 6.3. Many theorists have suggested that as children's knowledge of syntax matures, they develop new and more efficient methods for learning word meanings, by using the syntactic context of a new word to constrain hypotheses about what it means (see e.g. Landau and Gleitman's 1985 account of 'syntactic bootstrapping', and much subsequent work). The control network just described might be able to model the role of syntax in word-meaning learning. Currently, the error term representing the difference between the actual next word and the word generated by the word-production network is only used to train the control network. But it could also train the word-production network itself. If the control network enables training of the word-production network in just those contexts where it allows the output of the network to be pronounced, the word-production network will receive very accurate training data, tuned to particular syntactic contexts. This may allow more efficient learning of new words.

Relation to embodied cognition and language evolution Last but not least, the network helps to answer the big question I began with in this book: what is the interface between language and sensorimotor cognition? When an observer experiences an event in the world and then reports it in a sentence, how are sensorimotor representations converted to linguistic ones? In this book I have argued that there is a close relationship

between sensorimotor processing and syntactic structure. In my account, the sensorimotor sequences involved in experiencing a concrete episode, together with the working memory structures used to store these sequences, essentially determine the underlying syntactic structure of the sentence which reports the episode. All a child has to learn is a set of rules about which sensorimotor signals to read out as words when a stored sequence is replayed. The network just presented explains how this can be done. In other words, it provides the final component in a fairly detailed model of how surface language is grounded in sensorimotor experience. Of course this model only covers a single concrete episode at present. Nonetheless, for this one episode, the model runs from low-level perceptual and motor mechanisms, to working memory mechanisms, and finally to the mechanisms which produce surface sentences. Note that much of this model can be motivated without any reference to language. Replayable working memory event representations can be motivated as sensorimotor constructs, reflecting the inherently sequential nature of sensorimotor processing. They can also be motivated as linguistic constructs, as I argued in Chapter 5. But the only *specifically linguistic* machinery in the model is the machinery involved in representing phonological structure (Section 6.1.1), the machinery which implements the mapping from word forms to word meanings (Sections 6.1.3 and 6.3) and the network in this section, which maps from sensorimotor sequences to word sequences. On my account, then, this is the machinery which has to evolve in a population in order for it to acquire language. As noted in Section 1.1.1, minimising the amount of specifically linguistic machinery helps tell a plausible story about how human language evolved.

In fact I do not want to assume that 'all the child has to learn about syntax is when to read out words from a replayed sensorimotor sequence'. In the next section I present an extension to the network just described, incorporating a more empiricist treatment of syntax, which gives a more complex picture of what a child must learn in order to map episode representations onto utterances, and allows a more graded account of how the language faculty evolved.

6.4.4 A network combining sensorimotor and surface-based word-sequencing mechanisms

I have now presented two networks modelling syntactic competence. The word-sequencing network (Section 6.4.2) adopts an empiricist perspective, viewing syntactic knowledge as knowledge of surface patterns in language, learned using a general-purpose learning device (an Elman network). The episode-rehearsal/control network (Section 6.4.3) adopts a Minimalist perspective, viewing syntactic knowledge as a combination of innately-given constraints on the logical form of sentences and learned conventions about how to ex-

press this form in surface word sequences. However, as we have seen, Minimalist and empiricist models have quite complementary advantages and disadvantages. As discussed in Section 6.2.5.1, the Minimalist model has a rich account of infants' apparent predisposition to learn language, and of generalisations in the syntactic structure of natural languages. On the other hand, as discussed in Section 6.2.5.2, the Minimalist model does not provide a good account of children's earliest multi-word utterances, or of idiomatic or semi-idiomatic constructions in mature language. Empiricist models of language fare much better in both regards. In this section, I present a model in which the word-sequencing and episode-rehearsal/control networks operate together to produce mature syntactic competence. (This is work done jointly with Martin Takac and Lubica Benuskova.) I will first present a model of how the networks interact when both fully trained, and then discuss how the networks interact during development.

6.4.4.1 Interaction of the word-sequencing and episode-rehearsal networks

I suggest that during utterance generation, the word-sequencing network presented in Section 6.4.2 works in synchronisation with the episode-rehearsal/control network presented in Section 6.4.3. The division of labour between the two networks is basically that the word-sequencing network is responsible for generating idiomatic language, while the episode-rehearsal/word-production network is responsible for generating non-idiomatic (i.e. productive) language.

A diagram presenting the combined model is given in Figure 6.20. This figure is similar to Figure 6.12, in that the word-production network is integrated into the word-sequencing network. It is similar to Figure 6.17, in that the word-production network (now combined with the word-sequencing network) receives a sequence of semantic inputs, rather than a static semantic input, and a separate control network determines when its output should be pronounced.

Note that there are two separate types of iteration in the new model. The episode-rehearsal system iterates through a sequence of sensorimotor signals, and the word-sequencing network iterates through a surface sequence of words. The systems which trigger iteration in each system are shown in colour: 'next SM signal' triggers an update in the episode-rehearsal system, and 'next surface word' triggers an update in the word-sequencing network. A key issue in the new model is how to coordinate these two types of iteration.

My proposal is that the two types of iteration alternate with one another. To begin with, iterations in the word-sequencing network are suspended, and the episode-rehearsal network iterates, until the control network allows a word to be pronounced. At this point, episode rehearsal is suspended, and the word-sequencing network is allowed to iterate, to produce any idiomatic continuations of the generated word. When it has finished doing

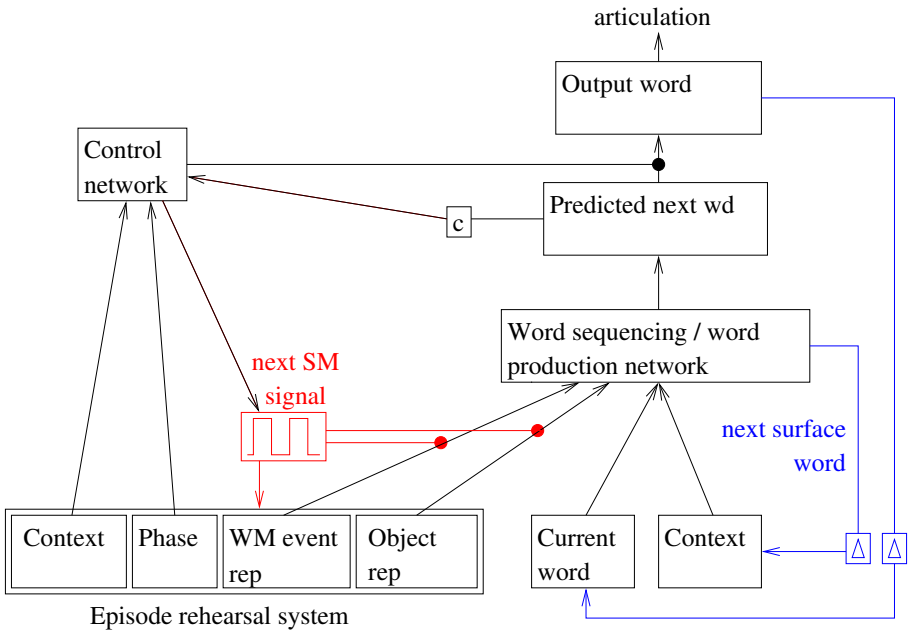


Figure 6.20: A network coordinating the episode-rehearsal and word-sequencing networks

that, or if there are no idiomatic continuations, the word-sequencing network is again suspended and the episode-rehearsal system resumes iterations, until another word is pronounced, and the cycle continues.

The critical question is what governs switches between the two types of iteration. In the model I propose, switching to the word-sequencing network is easy: an iteration is triggered in the word-sequencing network every time a word is explicitly pronounced. The signal to switch back to the episode-rehearsal network needs a little more discussion. It is a signal indicating that the word-production/word-sequencing network cannot confidently predict the next word on the basis of recent words and recent semantic inputs—i.e. that there is no idiomatic continuation of the sequence of words just produced. In Figure 6.20, a confidence judgement (denoted 'c') is read from the network's prediction about the likely next word. If the word-sequencing network has high confidence, the control network should allow this word to be pronounced, and should also suspend iteration in the episode-rehearsal system. If the word-sequencing network has low confidence, the control network should prevent a word being pronounced, and resume iteration in the episode-rehearsal system, so that a new semantic signal is generated. Note that the control network has to decide whether to pronounce or withold a predicted word both when the prediction comes from the word-

sequencing component of the sequencing/production network, based on the recent history of words and semantic signals, or from the word-production component, based on a newly-produced semantic signal. In the former case, its decision relates to confidence. In the latter case, its decision relates to learned conventions about when during episode rehearsal to read out semantic signals.

To illustrate how the new model handles idioms, consider a slightly modified version of our cup-grabbing example:

(6.9) Winnie the Pooh grabbed cup.

Winnie the Pooh can be understood as an idiom: a sequence of words which denote a semantic object collectively rather than individually, and which occur together with particularly high frequency. As discussed in Section 6.2.5.4, an Elman network is good at modelling idioms. If it has been trained on many examples of the phrase *Winnie the Pooh*, then after the word *Winnie* it should have a fairly high expectation for the word *the*, and then the word *Pooh*. An Elman network which receives semantic inputs so it can double as a word-production network is in an even better position. There may in fact be several words which can come after *Winnie*. A training corpus might contain the phrase *Winnie Mandela*, or occurrences of *Winnie* as a single word, followed by a variety of other words, so an Elman network by itself is unlikely to be completely confident about the next word. But in our modified network, the word *Winnie* was produced from a semantic representation—the representation of a particular object, which I will denote POOH-BEAR. This semantic representation will be copied to the context layer when the word-sequencing network iterates, where it will provide strong additional evidence that the next word should be *the*. Since information can survive for several iterations in the context layer, the semantic representation will also survive to the next iteration to help generate the final word of the idiom *Pooh*. After the network is trained, the semantic representation POOH-BEAR can in fact be thought of as a *plan* to generate a particular idiomatic sequence of words. Integrating the word-production and word-sequencing networks allows semantic representations to be associated with word sequences—i.e. idioms—as well as with single words.

While the word-sequencing network is producing an idiomatic phrase, it will be confident in its prediction of the next word, based on the recent history of word forms and semantic representations. However, after the final word in the idiom is produced, the situation changes. As noted in Section 6.4.2, in regular syntactic contexts the best an Elman network can do is to predict a broad distribution of syntactically possible next words: it cannot confidently predict which specific word will come next. There are many words which could follow the phrase *Winnie the Pooh*. Some may be more likely than others, but the

network is still unlikely to be able to predict the actual next word with any confidence.[25] Note that the semantic information held in the context layer is no longer much help, since it relates to the current and/or previous words. We must revert to the episode-rehearsal system to generate some additional semantic signals.

Note that while the episode-rehearsal system is iterating, it is generating representations in the word-sequencing/word-production network, but there are no iterations within this network. For instance, if the control network has learned to pronounce the verb late, then when the semantic signal GRAB is first produced, it will be encoded in the hidden layer of the sequencing/production network, but since the word is not pronounced, it will not be copied to the context layer. Instead, the next semantic signal will be produced in the episode-rehearsal system, which will in turn be encoded in the hidden layer of the sequencing/production network.

The idiom *Winnie the Pooh* is a simple one, because it is a continuous phrase. But note that the network just outlined can also generate discontinuous idioms. To illustrate, consider another variant on our cup-grabbing sentence:

(6.10) Man picks cup up

Picks X up can serve as an example of a discontinuous idiomatic construction. *To pick up* is a phrasal verb, and as discussed in Section 6.2.5.2, phrasal verbs are on the idiomaticity spectrum. A purer discontinuous idiom such as Jackendoff's example 'take X to task' could be also be used, but I will stick with *pick X up*, as we have a sensorimotor story to go with it.

The first word in the idiom *picks X up* is produced when the episode-rehearsal system generates a new semantic signal, GRAB. The word-production component of the production/sequencing network receives this signal, and generates a prediction about the next word. Until now, we have assumed that the GRAB signal is always expressed using the single word *grab*. I now assume that it is expressed as the idiom *pick X up*.[26] At the appropriate stage of episode rehearsal, the GRAB signal will activate the word *pick*, and the

[25] We can formalise the required notion of confidence using the information-theoretic concept of **entropy**. A probability distribution assigns a probability to each possible outcome of an unknown event. The entropy of the distribution is a measure of how much information it conveys about about this event. It ranges between 0 and a positive number (typically 1). If one outcome has probability 1 and all the others have probability 0, the distribution conveys certainty about the outcome, and its entropy is 0. Entropy increases as a function of the number of outcomes which have non-zero probability, and of the evenness of the distribution of probabilities to these outcomes, and is maximal if it is divided evenly between all possible outcomes.

[26] If GRAB could be expressed *either* as *grab* or as *pick up*, we would have to envisage a method for choosing between them, perhaps involving competition in the 'output word' layer. I will not consider this possibility here.

control network will allow this word to be pronounced. Explicit pronounciation of the word then causes the word-sequencing network to update, creating a new context representation encoding the semantic signal *grab* as well as the recent surface sequence of words.

At this point, an interesting circumstance arises. The word-sequencing network is given an opportunity to predict the next word from the new context representation. But even though we are in the middle of an idiomatic construction, it cannot confidently do so: the next word occupies a 'slot' in the idiom, and we need a new semantic signal to tell us what fills the slot. So the control network will advance through the sequence of semantic signals, until the context in which the signal CUP is generated and the associated word *cup* is pronounced. The pronounced word again triggers an update in the word-sequencing network, and it has another opportunity to predict the next word by itself, as an idiomatic continuation. This time, I suggest, it *can* be confident: it will predict the word *up* with high confidence. To explain why, note first that if an Elman network sees many examples of *pick X up* during training, it will learn the 'long-distance' relationship between *pick* and *up* in the same way as it learns about other long-distance dependencies, such as that between subjects and verbs where there is an intervening relative clause (see Section 6.2.5.4). And because it also has the semantic signal GRAB in its recent history, *up* is not just one possible continuation—it is the only one. In summary, the network's mechanisms for switching between the episode-rehearsal and word-sequencing systems generate discontinuous idioms very naturally. Note that we can still think of the episode-rehearsal and control networks from a Minimalist perspective, as implementing a mapping from LF to PF representations, and we can still think of the word-production/word sequencing network from an empiricist perspective, as a model of surface patterns in language. Combining the two networks appears to provide a way of reconciling the two perspectives.

6.4.4.2 Development of the word-sequencing/episode-rehearsal network

The previous section described the operation of the word-sequencing/episode-rehearsal network when fully trained. But how does training happen? I will make a couple of observations in this section.

Firstly, I am envisaging that learning happens in the word-sequencing/word-production component of the network before it happens in the episode-rehearsal/control component. As described in Section 6.4.2, I assume that the infant's earliest multi-word utterances are generated from static semantic representations, just using the word-sequencing/word-production network. At some point, the infant must learn that it is a good strategy to turn on episode rehearsal, and read words from a sequence of semantic signals rather than from a static representation. This will not immediately be the case; some learning has to happen in the control network first. So we must probably assume that episode rehearsal

is turned on somewhat randomly to begin with, to create opportunities for the control network to learn. But eventually, the control network will be well enough learned that the infant systematically turns on episode rehearsal when entering verbal mode.

Secondly, it may be that initial learning in the word-sequencing/word-production network has a role in bootstrapping learning in the episode-rehearsal/control system. The well-formed utterances an infant hears are often very complex, and it is a tall order to expect the episode-rehearsal/control system to learn all their complexities at once. Perhaps the word-sequencing network removes some of the complexities in training utterances by parsing them as idioms. Consider another variant of our example sentence:

(6.11) Man grabs a cup.

After a certain amount of training, the word-sequencing/word-production network is quite likely to be quite confident in its prediction of *a* after the word *grabs*. The control network might be able to learn quite early on to defer to the sequencing/production network in cases where it has high confidence. After having learned this, it will be in a better position to learn conventions about when to read out words during episode rehearsal. Of course, eventually, the infant will have to learn a more detailed story about the internal structure of noun phrases, and how noun phrases are embedded within clauses. (These topics are the main focus of Chapters 7 and 8.) But to begin with, initial learning in the word-sequencing network may well allow simple noun phrases to be treated as idioms. In fact, there is evidence that infants' early determiner-noun constructions are idiomatic and item-based rather than semantically motivated; see e.g. Pine and Lieven (1997).

6.4.4.3 Predictions of the episode-rehearsal/word-sequencing network

The episode-rehearsal/word-sequencing network is a model both of syntax acquisition and of mature sentence generation. In this section, I will note some of the predictions which it makes in each capacity.

As a model of syntax acquisition, one of the network's main predictions is that the transition from early sytactic competence (pivot schemas and item-based constructions) to mature syntactic competence happens because of a qualititive change in the mechanism which generates utterances, rather than just because of learning within a single mechanism. Accounts based on construction grammar such as that of Tomasello (2003) assume that a single general-purpose pattern-finding mechanism is responsible for all stages of syntactic development. In my model, early syntactic competence comes from a general-purpose pattern-finding mechanism (the word-production/word-sequencing network). But this mechanism is later supplemented with a second mechanism (the episode-rehearsal/control network), resulting in a qualitative change in the way utterances are produced. Note that

this new mechanism comes online gradually, and it integrates with the pattern-finding mechanism rather than replacing it. So we do not expect to see a sharp discontinuity in children's syntactic behaviour. But at a coarser temporal resolution we probably do expect to see a discontinuity, centred around the time at which infants start to rehearse episodes while generating an utterance. It might be hard to spot this discontinuity by analysing the complexity of children's utterances. But it may be possible to use imaging techniques to identify a new mechanism coming online. (Some tentative proposals about where the different components of the network are localised are made in Section 6.4.4.4).

As a model of mature sentence generation, one distinguishing feature of the network is that it has several separable components—perhaps more than most network models. One is the system mapping word meanings to word forms (represented in the model by the connections from event and object representations to the hidden layer of the word-production/word-sequencing network). Another is the system which learns sequential patterns in surface language (represented in the model by the by the recurrent connections in the production/sequencing network). Another is the system which maintains and rehearses an episode in working memory (the episode rehearsal system). Another is the pattern generator, which creates the template for X-bar schemas in synchrony with episode rehearsal. And another is the system which learns when to pronounce and suppress premotor articulatory representations during episode rehearsal (the control network). We might look for evidence for these separate components during utterance generation. For instance, we might look for evidence for an episode rehearsal system operative during utterance generation, or for a cyclic pattern generator alternately activating object and event representations during this process. The network model makes strong predictions about the existence of both these processes. We can also look for evidence for different components in patterns of language dysfunction. Unlike most models, the current model predicts that there are two different mechanisms which make separate contributions to our syntactic competence. One is responsible for the processing of idiomatic language (i.e. arbitrary surface patterns of words which denote arbitrary meanings), and the other is responsible for the processing of syntactically productive language. If the former mechanism is damaged, we predict a selective impairment in the processing of idioms, while if the latter is damaged, we predict a relatively retained ability to process idioms. All these predictions suggest avenues for further work.

6.4.4.4 Neural substrates of the episode-rehearsal/word-sequencing network

It is important to relate components of the episode-rehearsal/word-sequencing network to neural regions thought to be involved in syntactic processing. As discussed in Section 6.1.4.1, syntactic processing is associated with Broca's area and related regions of left

prefrontal cortex. But there are also suggestions that different areas within this system contribute to syntactic processing in different ways, as I reviewed in Section 6.1.4.2. I will briefly summarise these suggestions, and make some tentative suggestions about where different parts of the network are localised.

Broca's area classically comprises Brodmann areas BA44 and BA45. (BA44 is posterior to BA45.) BA44 is particularly implicated in the interpretation of complex syntactic structures (Stromswold, 1996; Caplan *et al.*, 1998). The posterior part of BA44, and a region of articulatory premotor cortex posterior to this, appear to be involved in the generation of syntactic structures (Indefrey *et al.*, 2001). This region is also involved in encoding nonlinguistic sequential patterns (Hoen *et al.*, 2006). A partly overlapping posterior region which extends dorsally beyond BA44 is involved in phonological rather than semantic processing (Gough *et al.*, 2005) and is implicated in phonological rehearsal in short-term memory tasks (Henson *et al.*, 2000). This area is probably involved in phonological output processes rather than input processes, though these output processes probably also involve the left posterior superior temporal gyrus (Hickok *et al.*, 2000). Based on these findings, I make two suggestions:

Suggestion 1 *The word-sequencing network is implemented in left posterior BA44 and adjacent premotor cortex.*

Suggestion 2 *The phonological components of this network (the current word and the predicted next word) are implemented in the above areas, but also in left posterior superior temporal gyrus.*

Dorsolateral PFC is involved in storing prepared sensorimotor sequences (see Section 3.2), and I have argued it is where WM episodes are stored and rehearsed (see Section 3.5.1). Left dorsolateral PFC is also an area associated with the generation of verbs (Perani, 1999; Tranel *et al.*, 2001) and the processing of verb inflections (Shapiro *et al.*, 2003), though the areas involved in verb processing also extend to BA45–46. I therefore make two more suggestions:

Suggestion 3 *The WM episodes which provide input to the word-sequencing network are stored in left dorsolateral PFC, and interface with this network through association areas in left BA45–46.*

Suggestion 4 *The circuitry involved in rehearsing WM episodes for linguistic purposes is implemented in left dorsolateral PFC, and evokes activity in left BA45–46.*

When a WM episode is rehearsed, a sequence of transitory object representations is evoked. There is evidence that working memory object representations activate BA45 and BA46 (Ungerleider *et al.*, 1998), and that areas BA45–47 are involved in 'selection' of semantic object representations in working memory contexts (Wagner *et al.*, 2001). There is also evidence that areas BA45–47 are involved in processing noun morphology (Shapiro *et al.*, 2000). Finally, we have seen evidence that objects are represented in inferotemporal and medial temporal cortex (see Sections 2.2 and 3.6.4.1). Based on this evidence, here is another suggestion:

Suggestion 5 *The object representations which are evoked during rehearsal of WM episodes, and which provide input to the word-sequencing network, are implemented in left BA45–47. These representations have a mixture of semantic and syntactic (especially morphological) properties. When activated, they in turn activate perceptual and long-term memory object representations in temporal cortex.*

(The suggestion of activity in left BA45–47 during WM episode rehearsal tallies with Suggestions 3 and 4.)

We have also seen evidence that Broca's area generally is involved in 'cognitive control' (Novick *et al.*, 2005), and in the 'selection' of semantic representations in contexts where there are several to choose from (Thomson-Schill *et al.*, 1997). Dominey *et al.* (2006) propose that BA44 is involved in representing abstract syntactic patterns, while BA45 has a role in selecting the lexical/semantic items which will feature in these patterns. In my model, the control network and the pattern generator selectively present semantic representations to the sequencing network, or selectively read lexical representations out from this network, at appropriate syntactic positions. I therefore suggest that:

Suggestion 6 *The pattern generator is implemented in left BA45. It alternately selects object representations and WM episode representations for presentation to the word-sequencing network, using circuitry in left BA46–47.*

Suggestion 7 *The control network is implemented in left BA45. It receives input from context representations in left dorsolateral PFC, and selectively allows the premotor outputs of the word-sequencing network into the motor output system.*

Finally, we must find a role for 'Wernicke's area' (the left posterior STG and associated temporal areas) in implementing the mapping from semantic object/episode representations to word forms. The components of the word-sequencencing/word-production network which involve mapping meanings to word forms must involve circuits in this area. Recall from Section 6.1.1.2 that Wernicke's area is involved in phonological output processes as well as phonological input processes, so there is certainly scope for phonological representations computed from meaning in this area to contribute to generated utterances. However, Wernicke's area is connected to prefrontal cortex by several distinct pathways, which probably have distinct functions. The 'direct' pathway which connects phonological input to phonological output representations runs through Geschwind's territory in inferior parietal cortex, as already mentioned in Section 6.1.1. I also briefly mentioned an even more direct pathway, called the **arcuate fasciculus**, which directly connects Wernicke's area with a more anterior part of Broca's area. And there is a pathway called the **uncinate fasciculus**, which connects anterior temporal cortex to orbitofrontal cortex. We have yet to find roles for these latter two pathways.

The semantically-activated word forms in Wernicke's area could be integrated into the sequencing/production network in two ways. They could contribute to the output of this network (most likely via the 'direct' pathway through Geschwind's territory). Or they could contribute to the input of this network (perhaps via the arcuate fasciculus, which projects to an anterior part of Broca's area). I have shown the latter option in Figure 6.20, but either option seems possible; the only requirement in my model is that semantics can exert an influence on the word forms produced by the sequencing/production network in posterior Broca's area. I therefore suggest:

Suggestion 8 *During episode rehearsal, semantic representations evoke a sequence of phonological word forms in Wernicke's area, which are communicated to the sequencing/production network.*

The final issue is how the semantic representations evoked in PFC during episode rehearsal activate corresponding word forms in Wernicke's area. Since episode representations and object representations are evoked in different areas, we must envisage that they activate phonological word forms via different routes. I have already proposed that WM episode representations activate association areas in left BA45–46 (Suggestion 3). I now propose that these areas communicate with Wernicke's area via the anterior temporal cortex, which appears to have a role in action naming, as was discussed in Section 6.1.2.

Suggestion 9 *WM episodes activate phonological representations of verbs through a loop which runs from BA45–46 to anterior temporal cortex, via the uncinate fasciculus, and then to Wernicke's area, and then back to Broca's area.*

A separate loop is involved in generating phonological representations of objects. I proposed that episode rehearsal causes object representations to be activated in BA45–47, which in turn activate object representations in temporal cortex (Suggestion 5). I now propose that this loop also runs through Wernicke's area, and back to Broca's area.

Suggestion 10 *WM episodes activate phonological representations of nouns through a loop which runs from BA45–47 to posterior temporal cortex, and then to Wernicke's area, and then back to Broca's area.*

According to these two suggestions, semantic representations enter the word-sequencing/word-production network through fairly convoluted circuits. They are generated in frontal areas, but the circuit via which they influence the word-sequencing network must run to Wernicke's area and back again. This may appear awkward engineering, but it allows us to explain how these loops can be disconnected if Wernicke's area is damaged, leaving the word-production/word-sequencing network to generate words without any guidance from semantic representations.

The above suggestions are all very tentative, as already mentioned. I have made them as concrete as possible, so that they provide a starting point for thinking about how the different components of the network are localised.

6.4.4.5 The episode-rehearsal/word-sequencing network and language evolution

I do not want to discuss language evolution in any detail at all. But it is useful to reiterate the point made in Chapter 1, that any model of language must at least be able to tell some sort of story about how language evolved, and that some models fare better than others in this regard. Models which minimise the amount of language-specific machinery have an easier time. So do models where the language-specific machinery comprises several distinct mechanisms, which when added in the right order result in a monotonically increasing communicative capacity.

Since there are several separable mechanisms in the episode-rehearsal/word-sequencing network, we are in a position to give a gradualist account of how the language faculty evolved. In fact, we can envisage that the mechanisms evolved in the same order in which they come online in infants. On this scenario, the first specifically linguistic mechanisms to evolve are those described in Section 6.3, which implement a special 'verbal mode', permit the learning of associations between semantic concepts and word forms, and allow the development of a special class of communicative actions. These mechanisms result in a system for producing (and interpreting) single words. At a later stage of evolution, we can envisage that the word-production network is extended with a general-purpose word-sequencing mechanism, to give the network described in Section 6.4.2 and illustrated in Figure 6.12. At this stage, conventions about word sequencing would be purely cultural, transmitted from one generation to the next due to the fact that children learn to reproduce the utterances of mature speakers. Finally, the word-sequencing network provides the platform for the evolution of the episode-rehearsal/control network. Note that episode rehearsal itself is motivated on nonlinguistic grounds; as discussed in Section 3.8.1.3, it is implicated in an account of the transferral of working memory episode representations to hippocampal storage. So the only new machinery which needs to evolve is the control network. In this final network, constraints on surface word sequences are partly genetic and partly cultural. They are genetic insofar as they rely on the episode rehearsal system, and cultural insofar as they reflect learning within the control network.

6.4.5 Some preliminary ideas about about sentence comprehension

The model of syntactic processing which I have given so far has been a model of sentence generation. What are the mechanisms involved in sentence comprehension? Do they overlap with those involved in generation? We saw in Section 6.1.3.1 that there are two separate systems implementing knowledge of word meanings: one for comprehension and one for generation. Is the same true for knowledge of syntax? Or are the sentence generation networks I have just described also involved in sentence interpretation?

In one respect, sentence interpretation is different from sentence generation. The sentence parsing process is typically modelled as delivering a set of possible candidate interpretations, both for individual phrases in an input sentence and for the whole sentence. The data structure which holds these alternative interpretations is called a 'chart'. In modern probabilistic parsers, each interpretation in the chart is associated with a probability, based on the frequency with which this interpretation is attested in a representative training corpus, and the most probable interpretation is ultimately selected. When considering how parsing happens in the brain, therefore, one way to begin is to look for a medium in which multiple alternative semantic interpretations of a sentence can be evoked simultaneously, with the most active of these eventually being selected.

In fact, if WM episodes are understood as contributing the semantic representations of sentences, as proposed in Chapter 5, we have already discussed a medium of just the right kind. As described in Section 3.3, there is evidence that multiple alternative plan representations can be active in dorsolateral PFC at one time. If each plan representation holds a WM episode, and WM episodes encode the semantics of sentences, then this region can perhaps be thought of as the medium in which a 'chart' of candidate analyses/interpretations of an input sentence are held, prior to the selection of the most likely of these. I will assume that during sentence interpretation, dorsolateral PFC holds a set of candidate WM episodes evoked by the input sentence. We must then look for a circuit which allows input utterances to activate candidate WM event representations in dorsolateral PFC, and we must consider what role Broca's area has in this circuit.

Recall from Section 6.1.4.1, Broca's area seems very important for generating syntactically well-formed utterances, but if it is damaged, patients still have a reasonable ability to interpret simple syntactic structures, using surface strategies based on knowledge of canonical word orderings. In Section 6.1.4.7 I suggested that these strategies may be implemented in anterior superior temporal cortex, and reviewed evidence that this area is involved in syntactic processing during sentence interpretation. In this section I will consider two questions: firstly, how might this anterior temporal area evoke WM episodes in PFC? And secondly, how does processing in this area relate to more sophisticated syntactic

processing in Broca's area during sentence interpretation?

As for the question of how 'shallow' syntactic processing in anterior temporal cortex could activate WM episodes in PFC: it is interesting to consider the possibility that 'shallow' sentence interpretation is in some respects similar to the process of intention recognition during action perception. As discussed in Section 2.7, in action recognition premotor representations of hand/arm actions evoked through the mirror system circuit activate assemblies in PFC, which encode hypotheses about the likely intentions of the observed agent. The function which maps from premotor to prefrontal representations implements an abductive inference process, which reasons from observed premotor activity to an intentional explanation for this activity. It is trained during the observer's own actions, when there is known to be a correspondence between the observer's intentional and premotor representations (because the intentional representations *cause* the premotor ones). Utterance interpretation can likewise be thought of as an abductive inference process mapping from observed premotor representations to their most likely intentional causes. In fact, the dominant computational models of sentence parsing/interpretation view the process as an abductive one: statistical parsers use Bayesian inference, which is a form of abduction. And there are many models of natural language semantics which argue that intention recognition must be the ultimate goal of utterance interpretation (see e.g. Levinson, 1983; Clark, 1996). So construing utterance interpretation as abductive intention recognition is broadly in line with current models of sentence semantics and sentence parsing. On this view, the circuits in temporal cortex which are involved in 'shallow' sentence interpretation are trained to map sequences of phonological word representations (and their associated semantic representations) directly onto WM episode representations in PFC, in situations where there is good likelihood that the phonological sequences do actually express the active WM episode. These might include situations where an infant has recognised that a communicative action is under way and has established joint attention, as discussed in Section 6.3.2. But they might also include situations where the speaker is *himself* generating an utterance. In this case, the intention recognition system would be trained during the speaker's own actions, as it is in the domain of hand actions.

The next question concerns what role Broca's area (and associated frontal areas) play during utterance interpretation. The role of these areas must be especially important for syntactically complex sentences, since these are the ones whose interpretation is most seriously impaired if it is damaged. My proposal here is that once a candidate WM episode has been selected via the 'shallow' pathway through anterior temporal cortex, this episode is used to internally generate a sentence, which is then compared to the input sentence (which is stored in the phonological input buffer) to check if it is correct. If there is a mismatch, then a backtracking operation must be initiated, to select a different candidate WM episode, and try again. Support for this hypothesis can be obtained from

several angles. For one thing, models in which inferred intentions must be 'replayed' to be confirmed are also found in the domain of hand action recognition; for instance the model of Oztop *et al.* (2005) involves a mechanism of this kind. From a language-processing perspective, we have already introduced the idea that the output of Broca's area can be compared word-by-word to a sentence replayed from the phonological input buffer. This 'offline' comparison mechanism is what is used to train the word-sequencing and episode-rehearsal networks discussed in Sections 6.4.2 and 6.4.3. We have also already reviewed evidence that Broca's area is responsible for the cognitive control operations needed to initiate backtracking during the interpretation of complex sentences; see the discussion of Novick *et al.* (2005) in Section 6.1.4.6. Finally, there is some evidence from ERP studies that sentence interpretation involves a sequence of processing stages, with early lexical and syntactic processing in temporal areas and later syntactic and semantic processing (e.g. thematic role assignment) in anterior frontal areas (BA44–45 and 47)—see Friederici (2002). However, the idea is still very speculative.

6.5 Summary and some interim conclusions

In this chapter, I began by reviewing current models of the neural representations and mechanisms involved in language processing (Section 6.1) and current theories of the developmental processes involved in acquiring language (Section 6.2). I then presented a new model of how infants learn individual word meanings and how they represent the special aspects of communicative actions (Section 6.3), and a new model of how they learn syntax (Section 6.4). Both models are grounded in what we know about the neural substrates of language. They are also grounded in the account of sensorimotor processing, intention representation and working memory presented in Chapters 2 and 3.

The syntactic model is still massively incomplete. My approach has been to focus on a single example sentence, *The man grabbed a cup*. Whether my sensorimotor interpretation of syntax extends beyond this sentence is still a completely open question. In fact, we have not even considered all the syntactic constructions in the example sentence. The sentence includes two noun phrases: *The man*, and *a cup*. The internal syntactic structure of noun phrases is itself complex, and so is the syntactic relationship between a noun phrase and its host clause. In the next two chapters, I will focus on the syntax of noun phrases, and on the noun-phrase/clause interface. These chapters will complete our investigation of the example sentence, but will also start to take us further afield, to consider a wider range of syntactic constructions.

Before I continue, however, this is a good place to take stock, and consider how the model of language I have presented so far relates to the dominant models of language in

linguistics and cognitive science. There are two main points I want to stress. Firstly, if it really is possible to reinterpret Minimalist LF structures as descriptions of sensorimotor processes, then *one can be a generative linguist without subscribing to the idea that language is a module*. Most generative linguists are Fodorians, holding that the generalisations across languages embodied in LF structures result from the operation of a modular language acquisition device. But if LF structures really do identify generalisations across languages, and if they really can be given a sensorimotor interpretation, then the linguistic universals proposed by generative linguists are probably reflections of the fact that language is deeply grounded in sensorimotor cognition, rather than of the operation of a modular language device.

Secondly, the model of language I have presented offers a way of interpreting the formal Minimalist model of 'the generative mechanism' as a model of actual cognitive processes—indeed of processes which are ongoing during the processing of sentences. According to my interpretation, the LF of a (concrete) sentence is the description of a rehearsed sensorimotor process. As outlined in Sections 6.4.3 and 6.4.4, I envisage that this rehearsal process actually occurs when a speaker is producing a sentence. My model of sentence generation can be understood from a Minimalist perspective as a model of how a child learns the mapping from LF to PF in his native language, but it can also be understood as a fairly conventional neural network model of language processing, which makes use of general-purpose learning devices, acknowledges the importance of idiomatic surface structures in language, and connects to what is known about the neural mechanisms involved in implementing language.

In both of these ways, I hope that the sensorimotor interpretation of LF has the effect of reintegrating generative grammar, and in particular Minimalism, into mainstream cognitive science. The debate about the position of generative grammar within cognitive science is an ongoing one—see for instance the collection of papers in Ritter (2005), and Jackendoff (2007). Most of the criticisms of generative grammar centre on the neural and evolutionary implausibility of an autonomous language module (see e.g. MacNeilage and Davis, 2005; Lieberman, 2005), or on its having no relation to models of actual language learning and actual language processing (see e.g. Tomasello, 2005; Ferreira, 2005; Bybee and McClelland, 2005). I hope that my reinterpretation addresses both of these issues.

Bibliography

Abney, S. (1987). *The English noun phrase in its sentential aspect*. Ph.D. thesis, MIT, Cambridge, MA.

Abney, S. (1996). Statistical methods in linguistics. In J. Klavans and P. Resnick, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press.

Abraham, W., Logan, B., Greenwood, J., and Dragunow, M. (2002). Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus. *Journal of Neuroscience*, **22**, 9626–9634.

Abrahams, R. and Christ, S. (2003). Motion onset captures attention. *Psychological Science*, **14**(5), 427–432.

Abrahams, S., Morris, R., Polkey, C., Jaroszc, J., Cox, T., Gravese, M., and Pickering, A. (1997). Hippocampal involvement in spatial and working memory: A structural mri analysis of patients with unilateral mesial temporal lobe sclerosis. *Brain and Cognition*, **41**(1), 39–65.

Adams, A. and Gathercole, S. (1995). Phonological working memory and speech production in preschool children. *Journal of Speech and Hearing Research*, **38**, 403–414.

Agre, P. and Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the American Association for Artificial Intelligence*, pages 268–272.

Aguirre, G. and D'Esposito, M. (1999). Topographical disorientation: a synthesis and taxonomy. *Brain*, **122**, 1613–1628.

Alexander, M., Naeser, M., and Palumbo, C. (1990). Brocas area aphasias: Aphasia after lesions including the frontal operculum. *Neurology*, **40**, 353–362.

Alexiadou, A., Haegeman, L., and Stavrou, M. (2007). *Noun phrase in the generative perspective*. Mouton de Gruyter, Berlin.

Allan, K., Wilding, E., and Rugg, M. (1998). Electrophysiological evidence for dissociable processes contributing to recollection. *Acta Psychologica*, **98**, 231–252.

Allison, T., Puce, A., and McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, **4**, 267–278.

Amodio and Frith, C. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, **7**, 268–277.

Andersen, R. (1997). Multimodal integration for the representation of space in the posterior parietal cortex. *Philosophical Transactions of the Royal Society of London B*, **352**(1360), 1421–1428.

Anderson, J. and Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Anderson, J., Gilmore, R., Roper, S., Crosson, B., Bauer, R., Nadeau, S., Beversdorf, D., Cibula, J andd Rogish, M., Kortenkamp, S., Hughes, J., Gonzalez Rothi, L., and Heilman, K. (1999). Conduction aphasia and the arcuate fasciculus: A reexamination of the wernickegeschwind model. *Brain and Language*, **70**, 1–12.

Arbib, M. (1981). Perceptual structures and distributed motor control. In V. Brooks, editor, *Handbook of physiology—The nervous system II Motor Control*. American Physiological Society.

Arbib, M. (1998). Schema theory. In A. Arbib, editor, *Handbook of brain theory and neural networks*, pages 993–998. MIT Press, Cambridge, MA.

Arbib, M. (2005). From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics. *Behavioral and Brain Sciences*, **28**(2), 105–167.

Arbib, M. (2006). Aphasia, apraxia and the evolution of the language-ready brain. *Aphasiology*, **20**, 1125–1155.

Arcizet, F., Jouffrais, C., and Girard, P. (2008). Natural textures classification in area V4 of the macaque monkey. *Experimental Brain Research*, **189**, 109–120.

Aron, A., Robbins, T., and Poldrack, R. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences*, **8**, 170–177.

Asaad, W., Rainer, G., and Miller, E. (2000). Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, **84**, 451–459.

Auerbach, S., Allard, T., Naeser, M., Alexander, M., and Albert, M. (1982). Pure word deafness: analysis of a case with bilateral lesions and a defect at the prephonemic level. *Brain*, **105**, 271–230.

Averbeck, B. and Lee, D. (2007). Prefrontal correlates of memory for sequences. *Journal of Neuroscience*, **27**(9), 2204–2211.

Averbeck, B., Chafee, M., Crowe, D., and Georgopoulos, A. (2002). Parallel processing of serial movements in prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **99**(20), 13172–13177.

Averbeck, B., Sohn, J., and Lee, D. (2006). Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neurosceience*, **9**(2), 276–282.

Axmacher, N., Elger, C., and Fell, J. (2008). Ripples in the medial temporal lobe are relevant for human memory consolidation. *Brain*, **131**(7), 1806–1817.

Baars, B., Motley, M., and MacKay, D. (1975). Output editing for lexical status from artifically elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, **14**, 382–391.

Babiloni, C., Del Percio, C., Babiloni, F., Carducci, F., Cincotti, F., Moretti, D., and Rossini, P. (2003). Transient human cortical responses during the observation of simple finger movements: A high-resolution eeg study. *Human Brain Mapping*, **20**(3), 148–157.

Baddeley, A. (2000). The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, **4**(11), 417–423.

Baddeley, A. and Andrade, J. (2000). Working memory and the vividness of imagery. *Journal of Experimental Psychology: General*, **129**(1), 126–145.

Baddeley, A. and Hitch, G. (1974). Working memory. In G. Bower, editor, *The psychology of learning and motivation*, pages 48–79. Academic Press.

Baddeley, A. and Warrington, E. (1970). Amnesia and the distinction between long-term and short-term memory. *Journal of Verbal Learning and Verbal Behavior*, **9**, 176–189.

Baddeley, A., Gathercole, S., and Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, **105**(1), 158–173.

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, **3**, 133–140.

Baillargeon, R. (1998). Infants' understanding of the physical world. In M. Sabourin, F. Craik, and M. Robert, editors, *Current Directions in Psychological Science*, pages 503–529. Psychology Press, London.

Baillargeon, R. and Wang, S. (2002). Event categorization in infancy. *Trends in Cognitive Sciences*, **6**(2), 85–93.

Baker, C., Keysers, C., Jellema, T., Wicker, B., and Perrett, D. (2001). Neuronal representation of disappearing and hidden objects in temporal cortex of the macaque. *Experimental Brain Research*, **140**, 375–381.

Baker, M. (2008). *The syntax of agreement and concord*. Cambridge University Press, Cambridge, UK.

Balan, P. and Ferrera, V. (2003). Effects of gaze shifts on maintenance of spatial memory in macaque frontal eye field. *Journal of Neuroscience*, **23**(13), 5446–5454.

Baldwin, D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, **62**, 875–890.

Baldwin, D. (1993). Early referential understanding: Infants' ability to recognize referential acts for what they are. *Developmental Psychology*, **29**(5), 832–843.

Baldwin, D., Markman, E., Bill, B., Desjardins, R., Irwin, J., and Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development*, **67**, 3135–3153.

Ballard, D., Hayhoe, M., Pook, P., and Rao, R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, **20**(4), 723–767.

Bar, M. and Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, **38**(2), 347–358.

Bar, M., Tootell, R., Schacter, D., Greve, D., Sischl, B., Mendola, J., Rosen, B., and Dale, A. (2001). Cortical mechanisms specific to explicit visual object recognition. *Neuron*, **29**(2), 529–535.

Barner, D., Thalwitz, D., Wood, J., Yang, S., and Carey, S. (2007). On the relation between the acquisition of singularplural morpho-syntax and the conceptual distinction between one and more than one. *Developmental Science*, **10**(3), 365–373.

Barone, P. and Joseph, J.-P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, **78**, 447–464.

Barry, C., Lever, C., Hayman, R., Hartley, T., Burton, S., O'Keefe, J., Jeffery, K., and Burgess, N. (2006). The boundary vector cell model of place cell firing and spatial memory. *Reviews in the Neurosciences*, **17**(1–2), 71–97.

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, **59**, 617–645.

Barsalou, L., Simmons, W., Barbey, A., and Wilson, C. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, **7**(2), 84–91.

Bartlett, F. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, Cambridge, UK.

Barwise, J. and Cooper, R. (1981). Generalized quantifiers and natural language. *Linguistics and Philosophy*, **4**(2), 159–219.

Bates, E. and MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney and E. Bates, editors, *The crosslinguistic study of sentence processing*, pages 3–73. Cambridge University Press, Cambridge, UK.

Battaglia-Mayer, A., Caminiti, R., Lacquaniti, F., and Zago, M. (2003). Multiple levels of representation of reaching in the parieto-frontal network. *Cerebral Cortex*, **13**, 1009–1022.

Behne, T., Carpenter, M., and Tomasello, M. (2005). One-year-olds comprehend the communicative intentions behind gestures in a hiding game. *Developmental Science*, **8**(6), 492–499.

Behrens, C., van den Boom, L., de Hoz, L., Friedman, A., and Heinemann, U. (2005). Induction of sharp waveripple complexes in vitro and reorganization of hippocampal networks. *Nature Neuroscience*, **8**(11), 1560–1567.

Beiser, D. and Houk, J. (1998). Model of cortical-basal ganglionic processing: Encoding the serial order of sensory events. *Journal of Neurophysiology*, **79**(6), 3168–3188.

Beisteiner, R., Höllinger, P., Lindinger, G., Lang, W., and Berthoz, A. (1995). Mental representations of movements. brain potentials associated with imagination of hand movements. *Electroencephalography and Clinical Neurophysiology*, **96**, 189–193.

Belletti, A. (1990). *Generalized verb movement*. Rosenberg and Sellier, Turin.

Belletti, A. (2001). Agreement projections. In M. Baltin and C. Collins, editors, *The handbook of contemporary syntactic theory*, pages 483–510. Blackwell, Oxford.

Bi, G. and Poo, M. (1998). Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, **18**, 10464–10472.

Bichot, N. (1999). Effects of similarity and history on neural mechanisms of visual selection. *Nature Neuroscience*, **2**, 549–554.

Bichot, N., Rao, S., and Schall, J. (2001a). Continuous processing in macaque frontal cortex during visual search. *Neuropsychologia*, **39**, 972–982.

Bichot, N., Thompson, K., Rao, S., and Schall, J. (2001b). Reliability of macaque frontal eye field neurons signaling saccade targets during visual search. *Journal of Neuroscience*, **21**, 713–725.

Bickerton, D. (1981). *Roots of language*. Karoma Publishers, Ann Arbor, MI.

Bilkey, D. and Clearwater, J. (2005). The dynamic nature of spatial encoding in the hippocampus. *Behavioural Neuroscience*, **119**, 1533–1545.

Blakemore, S., Wolpert, D., and Frith, C. (1998). Central cancellation of self-produced tickle sensation. *Nature Neuroscience*, **1**, 635–640.

Blakemore, S., Wolpert, D., and Frith, C. (2002). Abnormalities in the awareness of action. *Trends in Cognitive Sciences,*, **6**(6), 237–242.

Blakemore, S., Bristow, D., Bird, G., Frith, C., and Ward, J. (2005). Somatosensory activations during the observation of touch and a case of vision-touch synaesthesia. *Brain*, **128**(7), 1571–1583.

Blevins, J. (1995). The syllable in phonological theory. In J. Goldsmith, editor, *The Handbook of Phonological Theory*, pages 206–244. Blackwell, Oxford.

Blumenfeld, R. and Ranganath, C. (2007). Prefrontal cortex and long-term memory encoding: An integrative review of findings from neuropsychology and neuroimaging. *The Neuroscientist*, **13**(3), 280–291.

Bonda, E., Petrides, M., and Evans, A. (1996). Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal of Neuroscience*, **16**(11), 3737–3744.

Bookheimer, S. (2002). Functional MRI of language: New approaches to understanding the cortical organization of semantic processing. *Annual Review of Neuroscience*, **25**, 151–188.

Braine, M. (1976). Children's first word combinations. *Monographs of the society for research in child development*, **41**(1), 1–104.

Braver, T. and Cohen, J. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell and J. Driver, editors, *Attention and Performance XVIII: Control of cognitive processes*, pages 713–737. MIT Press.

Broadbent, D. (1985). A question of levels: comments on McClelland and Rumelhart. *Journal of Experimental Psychology: General*, **114**(2), 189–192.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, **47**, 139–159.

Brookshire, R. (1997). *Introduction to Neurogenic Communication Disorders*. Mosby-Year Book, St. Louis.

Browman, C. and Goldstein, L. (1995). Dynamics and articulatory phonology. In R. Port and T. van Gelder, editors, *Mind as motion: Explorations in the dynamics of cognition*, pages 175–193. MIT Press, Cambridge, MA.

Brunet, E., Sarfati, Y., Hardy-Baylé, M., and Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. *NeuroImage*, **11**, 157–166.

Buccino, G., Binkofski, F., Fink, G., Fadiga, L., Fogassi, L., Gallese, V., Seitz, R., Zilles, K., Rizzolatti, G., and Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience*, **13**, 400–4004.

Buckner, R. (1996). Beyond HERA: contributions of specific prefrontal brain areas to long-term memory retrieval. *Psychonomic Bulletin and Review*, **3**, 149–158.

Buckner, R. (2003). Functionalanatomic correlates of control processes in memory. *Journal of Neuroscience*, **23**(10), 3999–4004.

Buckner, R. and Wheeler, M. (2001). The cognitive neuroscience of remembering. *Nature Reviews Neuroscience*, **2**, 624–634.

Buckner, R., Koutsaal, W., Schachter, D., Wagner, A., and Rosen, B. (1998). Functional-anatomic study of episodic retrieval using fMRI. I retrieval effort versus retrieval success. *Neuroimage*, **7**, 151–162.

Bullock, D., Cisek, P., and Grossberg, S. (1998). Cortical networks for control of voluntary arm movements under variable force conditions. *Cerebral Cortex*, **8**, 48–62.

Burgess, N. and Hitch, G. (1999). Memory for serial order: a network model of the phonological loop and its timing. *Psychological Review*, **106**, 551–581.

Burgess, N. and Hitch, G. (2005). Computational models of working memory: putting long-term memory into context. *Trends in Cognitive Sciences*, **9**(11), 535–541.

Burgess, N. and O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, **6**(6), 749–762.

Burgess, N., Jackson, A., Hartley, T., and O'Keefe, J. (2000). Predictions derived from modelling the hippocampal role in navigation. *Biological Cybernetics*, **83**(3), 301–312.

Burgess, N., Maguire, E., Spiers, H., and O'Keefe, J. (2001). A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events. *NeuroImage*, **14**, 439–453.

Burgess, P., Dumontheil, I., and Gilbert, S. (2007). The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, **11**(7), 290–298.

Burnod, Y., Baraduc, P., Battaglia-Mayer, A., Guigon, E., Koechlin, E., Ferraina, S., Laquaniti, F., and Caminiti, R. (1999). Parieto-frontal coding of reaching: an integrated framework. *Experimental Brain Research*, **129**, 325–346.

Buschman, T. and Miller, E. (2007). Top-down and bottom-up attention in the prefrontal and posterior parietal cortices. *Science*, **315**, 1860–1862.

Butt, M. (1995). *The structure of complex predicates in Urdu*. Center for the Study of Language and Information, Stanford, CA.

Butterworth, B. and Warrington, E. (1995). Two routes to repetition: Evidence from a case of 'deep dysphasia'. *Neurocase*, **1**, 55–66.

Butterworth, G. (2006). Joint visual attention in infancy. In G. Bremner and A. Fogel, editors, *Blackwell Handbook of Infant Development*, pages 213–240. Wiley-Blackwell, Oxford. 2nd edition.

Butterworth, G. and Jarrett, N. (1991). What minds have in common is space: Spatial mechanisms for perspective taking in infancy. *British Journal of Developmental Psychology*, **9**, 55–72.

Bybee, J. and McClelland, J. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *Linguistic Review*, **22**(2–4), 381–410.

Call, J. (2004). Inferences about the location of food in the great apes (pan paniscus, pan troglodytes, gorilla gorilla and pongo pygmaeus). *Journal of Comparative Psychology*, **118**, 232–241.

Call, J., Hare, B., Carpenter, M., and Tomasello, M. (2004). 'unwilling versus 'unable: chimpanzees understanding of human intentional action. *Developmental Science*, **7**(4), 488–498.

Calvin, W. and Bickerton, D. (2000). *Lingua ex Machina: Reconciling Darwin and Chomsky with the human brain*. MIT Press, Cambridge, MA.

Caminiti, R., Johnson, P., Galli, C., Ferraina, S., and Burnod, Y. (1991). Making arm movements within different parts of space: the premotor and motor cortical representation of a coordinate system for reaching to visual targets. *Journal of Neuroscience*, **11**(5), 1182–1197.

Caminiti, R., Genovesio, A., Marconi, B., Battaglia-Mayer, A., Onorati, P., Ferraina, S., Mitsuda, T., Giannetti, S., Squatrito, S., Maioli, M., and Molinari, M. (1999). Early coding of reaching: frontal and parietal association connections of parieto-occipital cortex. *European Journal of Neuroscience*, **11**(9), 3339–3345.

Caplan, C. and Walters, G. (1992). Issues arising regarding the nature and consequences of reproduction conduction aphasia. In S. Kohn, editor, *Conduction aphasia*, pages 117–150. Laurence Erlbaum Associates, Mahwah, NJ.

Caplan, D., Alpert, N., and Waters, G. (1998). Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. *Journal of Cognitive Neuroscience*, **10**(4), 541–552.

Caramazza, A., Capasso, R., Capitani, E., and Miceli, G. (2005). Patterns of comprehension performance in agrammatic broca's aphasia: A test of the trace deletion hypothesis. *Brain and Language*, **94**(1), 43–53.

Carey, D., Perrett, D., and Oram, M. (1997). Recognising, understanding and reproducing action. In M. Jeannerod, editor, *Handbook of neuropsychology. Volume 11: Action and Cognition*, pages 111–129. Elsevier, Amsterdam.

Carlson, E. and Triesch, J. (2003). A computational model of the emergence of gaze following. In H. Bowman and C. Labiouse, editors, *Proceedings of the 8th neural computation workshop (NCPW8)*, Progress in Neural Processing. World Scientific.

Carlson, G. (1977). *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts at Amherst.

Carnie2007 (2007). *Syntax: a generative introduction*. Blackwell, Oxford.

Carpenter, M., Nagell, K., and Tomasello, M. (1998). Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development*, **63**(4), 1–174.

Carrington, S. and Bailey, A. (2009). Are there theory of mind regions in the brain? a review of the neuroimaging literature. *Human Brain Mapping*, **30**, 2313–2335.

Castiello, U. (2003). Understanding other people's actions: Intention and attention. *Journal of Experimental Psychology: Human Perception and Performance*, **29**(2), 416–430.

Castiello, U. and Jeannerod, M. (1991). Measuring time to awareness. *Neuroreport*, **2**(12), 787–800.

Catani, M., D, J., and ffytche, D. (2005). Perisylvian language networks of the human brain. *Annals of Neurology*, **57**, 8–16.

Caza, G. and Knott, A. (in preparation). Pragmatic bootstrapping: a neural network model of vocabulary acquisition. Manuscript.

Chafee, M., Averbeck, B., and Crowe, D. (2007). Representing spatial relationships in posterior parietal cortex: Single neurons code object-referenced position. *Cerebral Cortex*, **17**(12), 2914–2932.

Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cognitive Science*, **26**, 609–651.

Chen, S. and Goodman, J. (1998). An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, Cambridge, MA.

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

Chomsky, N. (1980). *Rules and representations*. Columbia University Press, New York.

Chomsky, N. (1981). *Lectures on government and binding*. Foris, Dordrecht.

Chomsky, N. (1995). *The Minimalist program*. MIT Press, Cambridge, MA.

Chong, T., Cunnington, R., Williams, M., Kanwisher, N., and Mattingley, J. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex. *Current Biology*, **18**, 1576–1580.

Christiansen, M. and Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, **23**, 157–205.

Christiansen, M. and Kirby, S. (2003). Language evolution: The hardest problem in science? In M. Christiansen and S. Kirby, editors, *Language evolution*. Oxford University Press, Oxford.

Cinque, G. (1995). Evidence for partial N-movement in the Romance DP. In G. Cinque, J. Y. Pollock, L. Rizzi, and R. Zanuttini, editors, *Towards Universal Grammar: Studies in Honor of Richard Kayne*. Georgetown University Press, Washington, DC.

Cisek, P. and Kalaska, J. (2005). Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron*, **45**, 801–814.

Clark, A. (1997). *Being There: Putting Brain, Body and World Together Again*. MIT Press, Cambridge, MA.

Clark, H. (1996). *Using language*. Cambridge University Press, Cambridge, UK.

Cohen, H. and Lefebvre, C. (2005). *Handbook of categorization in cognitive science*. Elsevier, Amsterdam.

Cohen, P. and Levesque, H. (1990a). Rational interaction as the basis for communication. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*, pages 221–256. MIT Press, Cambridge, MA.

Cohen, P. and Levesque, H. (1990b). Performatives in a rationally based speech act theory. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, PA.

Colby, C. and Goldberg, M. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience*, **22**, 399–349.

Colby, C., Duhamel, J., and Goldberg, M. (1993). Ventral intraparietal area of the macaque: anatomic location and visual response properties. *Journal of Neurophysiology*, **69**, 902–914.

Collins, M. (1996). A new statistical parser based on bigram lexical dependencies. In *Proceedings of the 34th Meeting of the ACL*, Santa Cruz.

Conway, C. and Christiansen, M. (2001). Sequential learning in non-human primates. *Trends in Cognitive Sciences*, **5**(12), 539–546.

Cooke, D., Charlotte, S., Taylor, T., and Graziano, M. (2003). Complex movements evoked by microstimulation of the ventral intraparietal area. *Proceedings of the National Academy of Sciences of the United States of America*, **100**(10), 6163–6168.

Cooke, S. and Bliss, T. (2006). Plasticity in the human central nervous system. *Brain*, **129**, 1659–1673.

Cooper, R. (1983). *Quantification and syntactic theory*. Reidel, Dordrecht.

Cooper, R. and Shallice, T. (2000). Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, **17**(4), 297–338.

Copestake, A. (2002). *Implementing Typed Feature Structure Grammars*. CSLI Publications, Stanford, CA.

Corballis, M. (2002). *From hand to mouth: the origins of language*. Princeton University Press, Princeton.

Courtney, S., Petit, L., Maisog, J., Ungerleider, L., and Haxby, J. (1998). An area specialized for spatial working memory in human frontal cortex. *Science*, **279**(5355), 134–1351.

Craik, F. and Lockhart, R. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, **11**, 671–684.

Culham, J., Brandt, S., Cavanagh, P., Kanwisher, N., Dale, A., and Tootell, R. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*, **80**, 2657–2670.

Dale, R., Scott, D., and di Eugenio, B. (1998). Introduction to the special issue on natural language generation. *Computational Linguistics*, **24**, 345–353.

Damasio, A. and Damasio, H. (1994). Cortical systems for retrieval of concrete knowledge: the convergence zone framework. In C. Koch and J. Davis, editors, *Large-scale neuronal theories of the brain*. MIT Press, Cambridge, MA.

Damasio, A. and Damsio, H. (1980). The anatomical basis of conduction aphasia. *Brain*, **103**, 337–350.

Damasio, A. and Damsio, H. (1992). Brain and language. *Scientific American*, **267**(3), 88–95.

Damasio, H., Grabowski, T., Tranel, D., Hichwa, R., and Damasio, A. (1996). A neural basis for lexical retrieval. *Nature*, **380**, 499–505.

Davidson, T., Kloosterman, F., and Wilson, M. (2009). Hippocampal replay of extended experience. *Neuron*, **63**, 497–507.

DeGraff, M., editor (2001). *Language Creation and Language Change: Creolization, Diachrony, and Development*. MIT Press, Cambridge, MA.

Dell, G. and Reich, P. (1981). Stages in sentence production: an analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, **20**, 611–629.

Demb, J., Desmond, J., Wagner, A., Vaidya, C., Glover, G., and Gabriel, J. (1995). Semantic encoding and retrieval in the left inferior prefrontal cortex: A functional mri study of task difficulty and process specificity. *Journal of Neuroscience*, **15**(9), 5870–5878.

Demiris, J. and Hayes, G. (2002). Imitation as a dual-route process featuring predictive and learning components: a biologically plausible computational model. In C. Nehaniv and K. Dautenhahn, editors, *Imitation in animals and artifacts*. MIT Press.

Desimone, R. and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, **18**, 193–222.

Desmurget, M., Epstein, C., Turner, R., Prablanc, C., Alexander, G., and Grafton, S. (1999). Role of the posterior parietal cortex in updating reaching movements to a visual target. *Nature Neuroscience*, **2**, 563–567.

di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., and Rizzolatti, G. (1992). Understanding motor events: a neurophysiological study. *Experimental Brain Research*, **91**, 176–180.

Diana, R., Yonelinas, A., and Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*, **11**(9), 379–386.

Diba, K. and Buzsàki, G. (2007). Forward and reverse hippocampal place-cell sequences during ripples. *Nature Neuroscience*, **10**(10), 1241–1242.

Diesing, M. (1992). *Indefinites*. MIT Press, Cambridge, MA.

Dimitrijevi, L. and Bjelakovi, B. (2004). Development of cardinal motor skills in the first year of life. *Acta Facultatis Medicae Naissensis*, **21**(4), 253–257.

Dobbins, I., Foley, H., Schacter, D., and Wagner, A. (2002). Executive control during episodic retrieval: Multiple prefrontal processes subserve source memory. *Neuron*, **35**(5), 989–996.

Doeller, C. and Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences of the USA*, **105**(15), 5909–5914.

Doeller, C., King, J., and Burgess, N. (2008). Parallel striatal and hippocampal systems for landmarks and boundaries in spatial memory. *Proceedings of the National Academy of Sciences of the USA*, **105**(15), 5915–5920.

Dominey, P. (1997). An anatomically structured sensory-motor sequence learning system displays some general linguistic capacities. *Brain and Language*, **59**, 50–75.

Dominey, P., Arbib, M., and Joseph (1995). A model of corticostriatal plasticity for learning associations and sequences. *Journal of cognitive neuroscience*, **7**(3), 311–336.

Dominey, P., Hoen, M., Blanc, J. M., and Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: evidence from simulation, aphasia and ERP studies. *Brain and Language*, **86**, 207–225.

Dominey, P., Hoen, M., and Inui, T. (2006). A neurolinguistic model of grammatical construction processing. *Journal of Cognitive Neuroscience*, **18**(12), 2088–2107.

Donoghue, J., Leibovic, S., and Sanes, J. (1992). Organization of the forelimb area in squirrel monkey motor cortex: representation of digit, wrist and elbow muscles. *Experimental Brain Research*, **89**, 1–19.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, **67**(3), 547–619.

Driver, J. (1999). Egocentric and object-based visual neglect. In N. Burgess, K. Jeffery, and J. O'Keefe, editors, *Spatial functions of the hippocampal formation and the parietal cortex*, pages 67–89. Oxford University Press.

Driver, J. and Spence, C. (1998). Attention and the crossmodal construction of space. *Trends in Cognitive Sciences*, **2**(7), 254–262.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., and Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition*, **6**(5), 509–540.

Dronkers, N., Wilkins, D., Van Valin, R., Redfern, B., and Jaeger, J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition*, **92**(1–2), 145–177.

Duncan, J. and Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, **96**, 433–458.

Dux, P., Ivanoff, J., Asplund, C., and Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fMRI. *Neuron*, **52**(6), 1109–1120.

Dwight J. Kravitz, Latrice D. Vinson, C. I. B. (2007). How position dependent is visual object recognition? *Trends in Cognitive Sciences*, **12**(3), 114–122.

Edelman, S. and Christiansen, M. (2003). How seriously should we take Minimalist syntax? *Trends in Cognitive Sciences*, **7**(2), 60–61.

Edwards, R., Xiao, D., Keysers, C., Földiák, P., and Perrett, D. (2003). Color sensitivity of cells responsive to complex stimuli in the temporal cortex. *Journal of Neurophysiology*, **90**, 12451256.

Ego-Stengel, V. and Wilson, M. (2009). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus*, **in press**, ??–??

Eichenbaum, H. (2004). Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*, **44**, 109–120.

Eichenbaum, H., Yonelinas, A., and Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience*, **30**, 123–152.

Ekstrom, A., Kahana, M., Caplan, J., Fields, T., Isham, E., Newman, E., and Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, **184**, 184–187.

Eldridge, L., Knowlton, B., Furmanski, C., Bookheimer, S., and Engel, S. (2000). Remembering episodes: a selective role for the hippocampus during retrieval. *Nature Neuroscience*, **3**(11), 1149–1152.

Eldridge, L., Engel, S., Zeineh, M., Bookheimer, S., and Knowlton, B. (2005). A dissociation of encoding and retrieval processes in the human hippocampus. *Journal of Neuroscience*, **25**(13), 3280–3286.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, **14**, 179–211.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, **7**, 195–225.

Elman, J. (1995). Language as a dynamical system. In R. Port and T. van Gelder, editors, *Mind as motion: explorations in the dynamics of cognition*, pages 195–225. MIT Press.

Emonds, J. (1978). The verbal complex v'–v in french. *Linguistic Inquiry*, **9**, 151–175.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, **392**, 598–601.

Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? *Neuron*, **23**, 115–125.

Epstein, R., Graham, K., and Downing, P. (2003). Viewpoint-specific scene representations in human parahippocampal cortex. *Neuron*, **37**, 865–876.

Epstein, R., Higgins, S., and Thompson-Schill, S. (2005). Learning places from views: Variation in scene processing as a function of experience and navigational ability. *Journal of Cognitive Neuroscience*, **17**(1), 73–83.

Epstein, R., Higgins, S., Jablonski, K., and Feiler, A. (2007). Visual scene processing in familiar and unfamiliar environments. *Journal of Neurophysiology*, **97**, 3670–3683.

Eskandar, E. and Asaad, J. (1999). Dissociation of visual, motor and predictive signals in parietal cortex during visual guidance. *Nature Neuroscience*, **2**(1), 88–93.

Eskandar, E. and Asaad, J. (2002). Distinct nature of directional signals among parietal cortical areas during visual guidance. *Journal of neurophysiology*, **88**, 1777–1790.

Euston, D., Tatsuno, M., and McNaughton, B. (2008). Fast-forward playback of recent memory sequences in prefrontal cortex during sleep. *Science*, **318**, 1147–1150.

Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation—a magnetic stimulation study. *Journal of Neurophysiology*, **73**(6), 2608–2611.

Fadiga, L., Craighero, L., Buccino, G., and Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. *European Journal of Neuroscience*, **15**, 399–402.

Fadiga, L., Craighero, L., and Olivier, E. (2005). Human motor cortex excitability during the perception of others' action. *Current opinion in neurobiology*, **15**(2), 213–218.

Fagg, A. and Arbib, M. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks*, **11**(7/8), 1277–1303.

Farrer, C. and Frith, C. (2002). Experiencing oneself vs another person as being the cause of an action: The neural correlates of the experience of agency. *NeuroImage*, **15**, 596–603.

Farrer, C., Franck, N., Georgieff, N., Frith, C., Decety, J., and Jeannerod, M. (2003). Modulating the experience of agency: a PET study. *NeuroImage*, **18**(2), 324–333.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, **8**(7), 307–314.

Feldman, J. and Narayanan, S. (2004). Embodiment in a neural theory of language. *Brain and Language*, **89**(2), 385–392.

Fell, J., Klaver, P., Elfadil, H., Schaller, C., Elger, C., and Fernandez, G. (2003). Rhinal-hippocampal theta coherence during declarative memory formation: interaction with gamma synchronization? *European Journal of Neuroscience*, **17**(5), 1082–1088.

Fenson, L., Dale, P., Reznick, J., Bates, E., Thal, D., Pethick, S., Tomasello, M., Mervis, C., and Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, **59**, 1–185.

Ferbinteanu, J. and Shapiro, M. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron*, **40**, 1227–1239.

Fernando, T. and Kamp, H. (1996). Expecting many. In *Proceedings of Semantics and Linguistic Theory VI*, pages 53–68, Rutgers/Cornell.

Ferraina, S., Paré, M., and Wurtz, R. (2000). Disparity sensitivity of frontal eye field neurons. *Journal of neurophysiology*, **83**, 625–629.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *Linguistic Review*, **22**(2–4), 365–380.

Fiebach, C., Schlesewsky, M., and Friederici, A. (2001). Syntactic working memory and the establishment of filler-gap dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research*, **30**(3), 321–338.

Fillmore, C., Kay, P., and O'Connor, M. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, **64**, 501–38.

Fink, G., Marshall, J., Halligan, P., Frith, X., Driver, J., Frackowiack, R., and Dolan, R. (1999). The neural consequences of conflict between intention and the senses. *Brain*, **122**, 497–512.

Flanagan, J. and Johansson, R. (2003). Action plans used in action observation. *Nature*, **424**, 769–771.

Flanders, M., Tillery, S., and Soechting, J. (1992). Early stages in a sensorimotor transformation. *Behavioral and Brain Sciences*, **15**, 309–320.

Fleck, M., Daselaar, S., Dobbins, I., and Cabeza, R. (2006). Role of prefrontal and anterior cingulate regions in decision-making processes shared by memory and nonmemory tasks. *Cerebral Cortex*, **16**, 1623–1630.

Fletcher, P. and Henson, R. (2001). Frontal lobes and human memory—insights from functional neuroimaging. *Brain*, **124**, 849–881.

Fodor, J. (1983). *The modularity of mind*. MIT/Bradford Press, Cambridge, MA.

Fodor, J. and Pylyshyn, Z. (1988). Cognition and cognitive architecture: a critical analysis. *Cognition*, **28**, 3–71.

Fogassi, L., Gallese, V., Fadiga, G., Luppino, G., Matelli, M., and Rizzolatti, G. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *Journal of Neurophysiology*, **76**, 141–157.

Fogassi, L., Raos, V., Franchi, G., Gallese, V., Luppino, G., and Matelli, M. (1999). Visual responses in the dorsal premotor area f2 of the macaque monkey. *Experimental Brain Research*, **128**, 194–199.

Fogassi, L., Ferrari, P. F., Gesierich, B., Rozzi, S., Chersi, F., and Rizzolatti, G. (2005). Parietal lobe: from action organisation to intention understanding. *Science*, **308**, 662–667.

Fortin, N., Agster, K., and Eichenbaum, H. (2002). Critical role of the hippocampus in memory for sequences of events. *Nature Neuroscience*, **5**(5), 458–462.

Foss, D. and Cairns, H. (1970). Some effects of memory limitation upon sentence comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, **9**, 541–547.

Foster, D. and Wilson, M. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature*, **440**, 680–683.

Foster, D., Morris, R., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, **10**, 1–16.

Frank, L., Brown, E., and Wilson, M. (2000). Trajectory encoding in the hippocampus and entorhinal cortex. *Neuron*, **27**(1), 169–178.

Franzius, M., Sprekeler, H., and Wiskott, L. (2007). Slowness and sparseness lead to place, head-direction, and spatial-view cells. *PLoS Computational Biology*, **3**(8), 1605–1622.

Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorisation. *Journal of Neuroscience*, **15**, 5235–5246.

Frey, S. and Petrides, M. (2002). Orbitofrontal cortex and memory formation. *Neuron*, **36**, 171–176.

Friederici, A. (2002). Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences*, **6**(2), 78–84.

Friederici, A., Meyer, M., and von Cramon, D. (2000). Auditory language comprehension: An event-related fMRI study on the processing of syntactic and lexical information. *Brain and Language*, **74**, 289–300.

Frith, C. and Frith, U. (2006). The neural basis of mentalizing. *Neuron*, **50**, 531–534.

Fyhn, M., Molden, S., Witter, M., Moser, E., and Moser, M. (2004). Spatial representation in the entorhinal cortex. *Science*, **305**, 1258–1264.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, **2**(12), 493–501.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, **119**, 593–609.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (2002). Action representation and the inferior parietal lobule. In *Common mechanisms in perception and action: Attention and Performance 19*, pages 334–355.

Galletti, C., Kutz, D., Gamberini, M., Breveglieri, R., and Fattori, P. (2003). Role of the medial parieto-occipital cortex in the control of reaching and grasping movements. *Experimental Brain Research*, **153**(2), 158–170.

Gangitano, M., Mottaghy, F., and Pascual-Leone, A. (2001). Phase-specific modulation of cortical motor output during movement observation. *Neuroreport*, **12**(7), 1489–1492.

Gathercole, S. and Baddeley, A. (1990). The role of phonological memory in vocabulary acquisition: A study of young children learning new names. *British Journal of Psychology*, **81**, 439–454.

Gentner, D. and Markman, A. (1997). Structure mapping in analogy and similarity. *American Psychologist*, **52**, 45–56.

Georgieff, N. and Jeannerod, M. (1998). Beyond consciousness of external reality: A "who" system for consciousness of action and self-consciousness. *Consciousness and cognition*, **7**, 465–477.

Georgopoulos, A., Kalaska, J., Caminiti, R., and Massey, J. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, **2**(11), 1527–1537.

Gershberg, F. and Shimamura, A. (1995). Impaired use of organizational strategies in free-recall following frontal lope damage. *Neuropsychologia*, **33**(10), 1305–1333.

Geyer, S., Matelli, M., Luppino, G., and Zilles, K. (2000). Functional neuroanatomy of the primate isocortical motor system. *Anatomy and Embryology*, **202**(6), 443–474.

Gibson, J., editor (1950). *The perception of the visual world*. Houghton Mifflin, Boston.

Giese, M. (2000). Neural model for the recognition of biological motion. In G. Baratoff and H. Neumann, editors, *Dynamische Perzeption*, pages 105–110. Infix Verlag, Berlin.

Givón, T. (2002). The visual information-processing system as an evolutionary precursor of human language. In T. Givón and B. Malle, editors, *The evolution of language out of prelanguage*. John Benjamins, Amsterdam.

Gluck, M. and Myers, C., editors (2001). *Gateway to Memory: An Introduction to Neural Network Models of the Hippocampus and Learning*. MIT Press, Cambridge, MA.

Gnadt, J. and Mays, L. (1995). Neurons in monkey parietal area LIP are tuned for eye-movement parameters in 3-dimensional space. *Journal of neurophysiology*, **73**(1), 280–297.

Goldberg, A., editor (1995). *Constructions. A Construction Grammar approach to argument structure*. University of Chicago Press, Chicago.

Goldberg, R., Perfetti, C., and Schneider, W. (2006). Perceptual knowledge retrieval activates sensory brain regions. *Journal of Neuroscience*, **26**(18), 4917–4921.

Gomi, H., Shidara, M., Takemura, A., Inoue, Y., Kawano, K., and M, K. (1998). Temporal firing patterns of purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys i. simple spikes. *Journal of Neurophysiology*, **80**, 818–831.

Goodwin, A. and Wheat, H. (2004). Sensory signals in neural populations underlying tactile perception and manipulation. *Annual Review of Neuroscience*, **27**, 53–77.

Goschke, T., Friederici, A., Kotz, S., and van Kampen, A. (2001). Procedural learning in broca's aphasia: Dissociation between the implicit acquisition of spatio-motor and phoneme sequences. *Journal of Cognitive Neuroscience*, **13**(3), 370–388.

Gottlieb, J., Kusunoki, M., and Goldberg, M. (1998). The representation of visual salience in monkey parietal cortex. *Nature*, **391**, 481–484.

Gough, P., Nobre, A., and Devlin, J. (2005). Dissociating linguistic processes in the left inferior frontal cortex with transcranial magnetic stimulation. *Journal of Neuroscience*, **25**(35), 8010–8016.

Grafton, S. and Hamilton, A. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, **26**, 590–616.

Graves, W., Grabowski, T., Mehta, S., and Gupta, P. (2008). The left posterior superior temporal gyrus participates specifically in accessing lexical phonology. *Journal of Cognitive Neuroscience*, **20**(9), 1698–1710.

Greenfield, P. (1991). Language, tools and the brain: the ontogeny and phylogeny of hierarchically organized sequential behavior. *Behavioral and Brain Sciences*, **14**, 531–535.

Greenlee, M., Magnussen, S., and Reinvang, I. (2000). Brain regions involved in spatial frequency discrimination: evidence from fmri. *Experimental Brain Research*, **132**, 399–403.

Grèzes, J. and J, D. (2001). Functional anatomy of execution, mental simulation, observation and verb generation of actions: a meta-analysis. *Human Brain Mapping*, **12**(1), 1–19.

Grodzinsky, J. and Santi, A. (2008). The battle for broca's region. *Trends in Cognitive Sciences*, **12**(12), 474–480.

Grosbras, M. and Paus, T. (2003). Transcranial magnetic stimulation of the human frontal eye field facilitates visual awareness. *European Journal of Neuroscience*, **18**(11), 3121–3126.

Grossberg, S. (1978). Behavioral contrast in short-term memory: Serial binary memory models or parallel continuous memory models? *Journal of Mathematical Psychology*, **17**, 199–219.

Grossman, E. and Blake, R. (2002). Brain areas active during visual perception of biological motion. *Neuron*, **35**, 1167–1175.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., and Blake, R. (2000). Brain areas involved in the perception of biological motion. *Journal of Cognitive Neuroscience*, **12**(5), 711–719.

Grosu, A. (1988). On the distribution of genitive phrases in Romanian. *Linguistics*, **26**, 931–949.

Haegeman, L. (1991). *Introduction to Government and Binding Theory*. Blackwell, Oxford.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, **9**(9), 416–423.

Hamilton, A. and Grafton, S. (2006). Goal representation in human anterior intraparietal sulcus. *Journal of Neuroscience*, **26**(4), 1133–1137.

Hamker, F. (2004). A dynamic model of how feature cues guide spatial attention. *Vision Research*, **44**, 501–521.

Hare, B. and Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, **68**, 571–581.

Hari, R., Forss, N., Avikainen, S., Kirveskari, E., Salenius, S., and Rizzolatti, G. (1998). Activation of human primary motor cortex during action observation: A neuromagnetic study. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(25), 15061–15065.

Harley, H. (2003). Possession and the double object construction. In P. Pica and J. Rooryck, editors, *Linguistic Variation Yearbook Vol. 2 (2002)*, pages 31–70. John Benjamins.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, **42**, 335–346.

Harris, C. and Wolpert, D. (1998). Signal-dependent noise determines motor planning. *Nature*, **394**(6695), 780–784.

Harris, Z. (1954). Distributional structure. *Word*, **10**(23), 146–162.

Hartley, T. and Houghton, G. (1996). A linguistically constrained model of short-term memory for nonwords. *Journal of Memory and Language*, **35**, 1–31.

Hasegawa, R., Matsumoto, M., and Mikami, A. (2000). Search target selection in monkey prefrontal cortex. *Journal of Neurophysiology*, **84**, 1692–1696.

Hauf, P., Elsner, B., and Aschersleben, G. (2004). The role of action effects in infants action control. *Psychological Research*, **68**, 115–125.

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, **41**, 301–307.

Haxby, J., Petit, L., Ungerleider, L., and Courtney, S. (2000). Distinguishing the functional roles of multiple regions in distributed neural systems for visual working memory. *NeuroImage*, **11**, 145–156.

Hebb, D. (1949). *The organization of behavior*. J Wiley and Sons, New York.

Hebb, D. (1961). Distinctive features of learning in the higher animal. In J. Delafresnaye, editor, *Brain mechanisms and learning*, pages 37–46. Oxford University Press.

Heim, I. (1982). *The semantics of definite and indefinite noun phrases*. Ph.D. thesis, University of Massachusetts. Distributed by Graduate Linguistic Student Association.

Henriques, D., Flanders, M., and Soechting, J. (2004). Haptic synthesis of shapes and sequences. *Journal of Neurophysiology*, **91**, 1808–1821.

Henson, R., Rugg, M., Shallice, T., Josephs, O., and Dolan, R. (1999). Recollection and familiarity in recognition memory: An event-related functional magnetic resonance imaging study. *Journal of Neuroscience*, **19**(10), 3962–3972.

Henson, R., Burgess, N., and Frith, C. (2000). Recoding, storage, rehearsal and grouping in verbal short-term memory: an fMRI study. *Neuropsychologia*, **38**(4), 426–440.

Hickok, G. and Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, **92**, 67–99.

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, **8**(5), 393–402.

Hickok, G., Zurif, E., and Canseco-Gonzalez, E. (1993). Structural description of agrammatic comprehension. *Brain and Language*, **45**, 371–395.

Hickok, G., Erhard, P., Kassubek, J., Helms-Tillery, A., Naeve, Velguth, S., Struppf, J., Strick, P., and Ugurbil, K. (2000). A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: implications for the explanation of conduction aphasia. *Neuroscience Letters*, **287**, 156–160.

Hirsh-Pasek, K. and Golinkoff, R. (1996). *The origins of grammar: evidence from early language comprehension*. MIT Press, Cambidge, MA.

Hirtle, S. and Jonides, J. (1985). Evidence of hierarchies in cognitive maps. *Memory and Cognition*, **13**(3), 208–217.

Hockenmaier, J. and Steedman, M. (2002). Generative models for statistical parsing with combinatory grammars. In *Proceedings of the 40th meeting of the Association for Computational Linguistics*, pages 335–342, Philadelphia.

Hoen, M., Pachot-Clouard, M., Segebarth, C., and Dominey, P. (2006). When Broca experiences the Janus syndrome. An ER-fMRI study comparing sentence comprehension and cognitive sequence processing. *Cortex*, **42**, 605–623.

Hoffman, K. and McNaughton, B. (2002). Coordinated reactivation of distributed memory traces in primate neocortex. *Science*, **297**, 2070–2073.

Hok, V., Save, E., Lenck-Santini, P., and Poucet, B. (2005). Coding for spatial goals in the prelimbic/infralimbic area of the rat frontal cortex. *Proceedings of the National Academy of Sciences*, **102**(12), 4602–4607.

Hok, V., Lenck-Santini, P., Roux, S., Save, E., Muller, R., and Poucet, B. (2007). Goal-related activity in hippocampal place cells. *Journal of Neuroscience*, **27**(3), 472–482.

Holdstock, J., Mayes, A., Gong, Q., Roberts, N., and Kapur, N. (2005). Item recognition is less impaired than recall and associative recognition in a patient with selective hippocampal damage. *Hippocampus*, **5**, 203–215.

Hollup, S., Molden, S., Donnett, J., Moser, M., and Moser, E. (2001). Accumulation of hippocampal place fields at the goal location in an annular watermaze task. *Journal of Neuroscience*, **21**, 1635–1644.

Holmes, G. (1939). The cerebellum of man. *Brain*, **62**, 1–30.

Hölscher, C. (2003). Time, space and hippocampal functions. *Reviews in the Neurosciences*, **14**(3), 253–284.

Hommel, B., Müsseler, J., Aschersleben, G., and Prinz, W. (2001). The theory of event coding (TEC): a framework for perception and action learning. *Behavioral and Brain Sciences*, **24**, 849–878.

Hooker, C., Paller, K., Gitelman, D., Parrish, T., Mesulam, M., and Reber, P. (2003). Brain networks for analyzing eye gaze. *Cognitive Brain Research*, **17**(2), 406–418.

Houghton, G. and Hartley, T. (1995). Parallel models of serial behaviour: Lashley revisited. *Psyche*, **2**(25).

Howard, M. and Kahana, M. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, **46**, 269–299.

Hughes, H., Nozawa, G., and Kitterle, F. (1996). Global precedence, spatial frequency channels, and the statistics of natural images. *Journal of Cognitive Neuroscience*, **8**(3), 197–230.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G., Martin, S., and Stuart, G. (1991). Word-frequency effects on short-term memory tasks: evidence for a redintigration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **23**(5), 1217–1232.

Hummel, J. and Holyoak, K. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, **110**(2), 220–264.

Humphries, C., Willard, K., Buchsbaum, B., and Hickok, G. (2001). Role of anterior temporal cortex in auditory sentence comprehension: an fmri study. *Neuroreport*, **12**(8), 1749–1752.

Hurford, J. (2003). The neural basis of predicate-argument structure. *Behavioral and Brain Sciences*, **26**(3), 261–283.

Huyck, C. (2007). Creating hierarchical categories using cell assemblies. *Neural Computation*, **19**(1), 1–24.

Iacoboni, M., Woods, R., Brass, M., Bekkering, H., Mazziotta, J., and Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, **286**, 2526–2528.

Iacoboni, M., Koski, L., Brass, M., Bekkering, H., Woods, R., Dubeau, M., Mazziotta, J., and Rizzolatti, G. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **98**(24), 13995–13999.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J., and Rizzolatti, G. (2005). Grasping the intentions of others with ones own mirror neuron system. *PLoS Biology*, **3**(3), e79.

Iatridou, S. (1990). About Agr(P). *Linguistic Inquiry*, **21**(4), 421–459.

Iberall, T. and Arbib, M. (1990). Schemas for the control of hand movements: An essay on cortical localisation. In M. Goodale, editor, *Vision and action: the control of grasping*, pages 163–180. Ablex, Norwood.

Imamizu, H., Miyauchi, S., Tamada, T., Sasaki, Y., Takino, R., Putz, B., Yoshioka, T., and Kawato, M. (2000). Human cerebellar activity reflecting an acquired internal model of a new tool. *Nature*, **403**(6766), 192–195.

Indefrey, P., Brown, C., Hellwig, F., Amunts, K., Herzog, H., Seitz, R., and Hagoort, P. (2001). A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences*, **98**(10), 5933–5936.

Ito, T., Tiede, M., and Ostry, D. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences*, **106**(4), 1245–1248.

Itti, L. and Arbib, M. (2006). Attention and the minimal subscene. In M. Arbib, editor, *Action to Language via the Mirror Neuron System*, pages 289–346. Cambridge University Press, Cambridge, UK.

Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, **40**(10–12), 1489–1506.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews—Neuroscience*, **2**, 1–11.

Jackendoff, R. (1977). X-Bar syntax: A study of phrase structure. Linguistic Inquiry Monograph 2.

Jackendoff, R. (2002). *Foundations of Language: Brain, Meaning, Grammar, Evolution*. Oxford University Press, Oxford.

Jackendoff, R. (2007). Linguistics in cognitive science: The state of the art. *The Linguistic Review*, **24**, 347–401.

Jacquemot, C. and Scott, S. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, **10**(11), 480–486.

Jankelowitz, S. and Colebatch, J. (2002). Movement-related potentials associated with self-paced, cued and imagined arm movements. *Experimental Brain Research*, **147**(1), 98–107.

Jeannerod, M. (1996). Reaching and grasping. parallel specification of visuomotor channels. In W. Prinz and B. Bridgeman, editors, *Handbook of Perception and Action Volume II: Motor Skills*, pages 405–460. Academic Press.

Jeannerod, M. (1999). To act or not to act: Perspectives on the representation of actions. *Quarterly Journal of Experimental Psychology: A*, **52**(1), 1–29.

Jeannerod, M. (2001). Neural simulation of action: a unifying mechanism for motor cognition. *NeuroImage*, **14**, S103–S109.

Jeannerod, M. (2003). The mechanism of self-recognition in humans. *Behavioural Brain Research*, **142**, 1–15.

Jeffery, K. (2007). Integration of the sensory inputs to place cells: what, where, why, and how? *Hippocampus*, **17**, 775–785.

Jeffery, K., Gilbert, A., Burton, S., and Strudwick, A. (2003). Preserved performance in a hippocampal-dependent spatial task despite complete place cell remapping. *Hippocampus*, **13**, 175–189.

Jellema, T., Baker, C., Wicker, B., and Perrett, D. (2000). Neural representation for the perception of the intentionality of actions. *Brain and Cognition*, **44**, 280–302.

Jellema, T., Maassen, G., and Perrett, D. (2004). Single cell integration of animate form, motion and location in the superior temporal cortex of the macaque monkey. *Cerebral Cortex*, **14**(7), 781–790.

Jenkins, H., Jahanshahi, M., Jueptner, M., Passingham, R., and Brooks, D. (2000). Self-initiated versus externally triggered movements II. the effect of movement predictability on regional cerebral blood flow. *Brain*, **123**(6), 1216–1228.

Jensen, O. and Lisman, J. (1996). Hippocampal ca3 region predicts memory sequences: Accounting for the phase precession of place cells. *Learning and Memory*, **3**(2–3), 279–287.

Jensen, O. and Lisman, J. (2005). Hippocampal sequence-encoding driven by a cortical multi-item working memory buffer. *Trends in Neurosciences*, **28**(2), 67–72.

Ji, D. and Wilson, M. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience*, **10**(1), 100–107.

Ji, D. and Wilson, M. (2008). Firing rate dynamics in the hippocampus induced by trajectory learning. *Journal of Neuroscience*, **21**(18), 4679–4689.

Johansson, G. (1973). Visual perception of biological motion, and a model for its analysis. *Visual Perception and Psychophysics*, **14**, 201–211.

Johansson, M. and Mecklinger, A. (2003). The late posterior negativity in ERP studies of episodic memory: action monitoring and retrieval of attribute conjunctions. *Biological Psychology*, **64**(1–2), 91–117.

Johansson, R., Westling, G., Backstrom, A., and Flanagan, J. (2001). Eye-hand coordination in object manipulation. *Journal of Neuroscience*, **21**(17), 6917–6932.

Johnson, A. and Redish, D. (2007). Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience*, **27**(45), 12176–12189.

Johnson-Laird, P. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge.

Jordan, M. and Wolpert, D. (2000). Computational motor control. In M. Gazzaniga, editor, *The new cognitive neurosciences*, pages 71–118. MIT Press.

Joshi, A., Vijay-Shanker, K., and Weir, D. (1991). The convergence of mildly context-sensitive grammar formalisms. In P. Sells, S. Shieber, and T. Wasow, editors, *Foundational Issues in Natural Language Processing*, pages 31–81. MIT Press, Cambridge, MA.

Jung, M., Wiener, S., and McNaughton, B. (1994). Comparison of spatial firing characteristics of units in dorsal and ventral hippocampus of the rat. *Journal of Neuroscience*, **14**, 7347–7356.

Justus, T. (2004). The cerebellum and English grammatical morphology: Evidence from production, comprehension and grammatical morphology judgements. *Journal of Cognitive Neuroscience*, **16**(7), 1115–1130.

Kahneman, D., Treisman, A., and Gibbs, B. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, **24**, 175–219.

Kakei, S., Hoffman, D., and Strick, P. (2001). Direction of action is represented in the ventral premotor cortex. *Nature Neuroscience*, **4**(10), 1020–1025.

Kamp, H. (1981). A theory of truth and semantic representation. In J. Groenendijk, T. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*, page 277. Mathematical Center Tract 135, Amsterdam.

Kamp, H. and Reyle, U. (1993). *From discourse to logic*. Kluwer Academic Publishers, Dordrecht.

Kasper, R., Kiefer, B., Netter, K., and Vijay-Shanker, K. (1995). Compilation of HPSG to TAG. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 92–99.

Kawato, M. and H, G. (1992). The cerebellum and VOR/OKR learning models. *Trends in Neurosciences*, **15**, 445–453.

Kawato, M., Furawaka, K., and Suzuki, R. (1987). A hierarchical neural network model for the control and learning of voluntary movements. *Biological Cybernetics*, **56**, 1–17.

Kawato, M., Kuroda, T., Imamizu, H., Nakano, E., Miyauchi, S., and Yoshioka, T. (2003). Internal forward models in the cerebellum: fMRI study on grip force and load force coupling. In C. Prablanc, D. P'elisson, and Y. Rosetti, editors, *Progress in Brain Research*, volume 142, pages 171–188. Elsevier Science.

Kaye, K. and Ellis, A. (1987). A cognitive neuopsychological case study of anomia: implications for psychological models of word retrieval. *Brain*, **110**, 613–629.

Kayne, R. (1994). *The antisymmetry of syntax*. MIT Press, Cambridge, MA.

Kazanovich, Y. and Borisyuk, R. (2006). An oscillatory neural model of multiple object tracking. *Neural Computation*, **18**, 1413–1440.

Kegl, J., Senghas, A., and Coppola, M. (2001). Creation through contact: sign language emergence and sign language change in Nicaragua. In M. DeGraff, editor, *Language Creation and Language Change: Creolization, Diachrony, and Development*. MIT Press, Cambridge, MA.

Kellog, R. (2003). *Cognitive Psychology (2nd edition)*. Sage Publication, Thousand Oaks, CA.

Kelso, K., Southard, D., and Goodman, D. (1979). Nature of human inter-limb coordination. *Science*, **203**(4384), 1029–1031.

Kent, C. and Lamberts, K. (2008). The encodingretrieval relationship: retrieval as mental simulation. *Trends in Cognitive Sciences*, **12**(3), 92–98.

Kesner, R., Gilbert, P., and Barua, L. (2002). The role of the hippocampus in memory for the temporal order of a sequence of odors. *Behavioral Neuroscience*, **116**(2), 286–290.

Keysers, C. and Perrett, D. (2004). Demystifying social cognition: a Hebbian perspective. *Trends in Cognitive Sciences*, **8**(11), 501–507.

Keysers, C., Wicker, B., Gazzola1, V., Anton, J., Fogassi, L., and V, G. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron*, **42**, 335–346.

Kingstone, A., Friesen, C., and Gazzaniga, M. (2000). Reflexive joint attention depends on lateralized cortical connections. *Psychological Science*, **11**(2), 159–166.

Kirby, S. and Hurford, J. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In A. Cangelosi and D. Parisi, editors, *Simulating the Evolution of Language*. Springer-Verlag, London.

Kjelstrup, J., Solstad, T., Brun, V., Fyhn, M., and Hafting, T. (2007). Very large place fields at the ventral pole of the hippocampal ca3 area. *Soc. Neurosci. Abstr.*, **33.93.1**, 305–314.

Klam, F. and Graf, W. (2006). Discrimination between active and passive head movements by macaque ventral and medial intraparietal cortex neurons. *Journal of Physiology*, **574**(2), 367–386.

Knoblich, G. and Flach, R. (2001). Predicting the effects of actions: interactions of perception and action. *Psychological Science*, **12**(6), 467–472.

Kobayashi, T., Nishijo, H., Fukuda, M., Bures, J., and Ono, T. (1997). Task-dependent representations in rat hippocampal place neurons. *Journal of Neurophysiology*, **78**, 597–613.

Kobayashi, Y., Kawano, K., Takemura, A., Inoue, Y., Kitama, T., Gomi, H., and Kawato, M. (1998). Temporal firing patterns of purkinje cells in the cerebellar ventral paraflocculus during ocular following responses in monkeys ii. complex spikes. *Journal of Neurophysiology*, **80**, 832–848.

Koch, C. and Ullman, S. (1985). Shifts in underlying visual attention: towards the underlying neural circuitry. *Human Neurobiology*, **4**(4), 219–227.

Kolers, P. and Roediger, H. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, **23**, 425–449.

Koopman, H. and Sportiche, D. (1991). The position of subjects. *Lingua*, **85**, 211–258.

Kosslyn, S., Thompson, W., Kim, I., and Alpert, M. (1995). Topographical representations of mental images in primary visual-cortex. *Nature*, **378**, 496–498.

Kosslyn, S., Thompson, W., and Ganis, G. (2006). *The case for mental imagery*. Oxford University Press, New York.

Köteles, K., De Mazière, P., Van Hulle, M., and Orban, G andVogels, R. (2008). Coding of images of materials by macaque inferior temporal cortical neurons. *European Journal of Neuroscience*, **27**, 466–482.

Kourtzi, Z. and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital cortex. *Science*, **293**, 1506–1509.

Kowler, E., Anderson, E., Dosher, B., and Blaser, E. (1995). The role of attention in the programming of saccdes. *Vision Research*, **35**, 1897–1916.

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, **3**(9), 946–953.

Kriegeskorte, N., Mur, M., Ruff, D., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, **60**, 1120–1141.

Kumaran, D. and Maguire, E. (2006). The dynamics of hippocampal activation during encoding of overlapping sequences. *Neuron*, **49**, 617–629.

Kusunoki, M., Gottlieb, J., and Goldberg, M. (2000). The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task relevance. *Vision Research*, **40**(10–12), 1459–1468.

Lackner, J. (1988). Some proprioceptive influences on the perceptual representation of body shape and orientation. *Brain*, **111**, 281–297.

Lakoff, G. (1987). *Women, fire and dangerous things*. University of Chicago Press, Chicago and London.

Lakoff, G. and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago and London.

Land, M. and Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London B*, **352**, 1231–1239.

Land, M., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception*, **28**, 1311–1328.

Landau, B. and Gleitman, L. (1985). *Language and experience: Evidence from the blind child*. Harvard University Press, Cambridge, MA.

Landy, M. and Graham, N. (2001). Visual perception of texture. In L. Chalupa and J. Werner, editors, *The visual neurosciences*, pages ??–?? MIT Press.

Langacker, R. (1987). *Foundations of cognitive grammar I: Theoretical prerequisites*. Stanford University Press, Stanford, CA.

Langacker, R. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press, New York.

Larson, R. (1988). On the double object construction. *Linguistic Inquiry*, **19**(3), 335–391.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1992). Handwritten digit recognition with a back-propagation network. In P. Lisboa, editor, *Neural networks: current applications*. Chapman and Hall, London.

Leslie, A. (1984). Infant perception of a pick up event. *British Journal of Developmental Psychology*, **2**, 19–32.

Levelt, W. (1989). *Speaking: from intention to articulation*. MIT Press, Cambridge, MA.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Levy, W. (1996). A sequence predicting ca3 is a flexible associator that learns and uses context to solve hippocampal-like tasks. *Hippocampus*, **6**(6), 579–590.

Liberman, A. and Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1–36.

Lidz, J., Gleitman, H., and Gleitman, L. (2003). Understanding how input matters: verb learning and the footprint of universal grammar. *Cognition*, **87**, 151–178.

Lieberman, P. (2005). The pied piper of Cambridge. *Linguistic Review*, **22**(2–4), 289–301.

Lieven, E., Pine, J., and Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, **24**, 187–219.

Lin, D. and Pantel, P. (2001). Induction of semantic classes from natural language text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 317–322.

Lindfield, K., Wingfield, A., and Goodglass, H. (1999). The role of prosody in the mental lexicon. *Brain and Language*, **68**, 312–317.

Lisman, J. and Otmakhova, N. (2001). Storage, recall, and novelty detection of sequences by the hippocampus: Elaborating on the socratic model to account for normal and aberrant effects of dopamine. *Hippocampus*, **11**(5), 551–568.

Livingstone, M. and Hubel, D. (1988). Segregation of form, color, movement and depth: anatomy, physiology and perception. *Science*, **240**(4583), 740–749.

Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Annual Review of Neuroscience*, **19**, 577–621.

Logothetis, N., Pauls, J., and Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, **5**(5), 552–563.

Longobardi, G. (1994). Reference and proper names: a theory of N-movement in syntax and logical form. *Linguistic Inquiry*, **25**(4), 609–665.

Lotto, A., Hickok, G., and Holt, L. (2009). Reflections on mirror neurons and speech perception. *Trends in Cognitive Sciences*, **13**(3), 110–114.

Lotze, M., Montoya, P., Erb, M., Hülsmann, E., Flor, H., Klose, U., Birbaumer, N., and Grodd, W. (1999). Activation of cortical and cerebellar motor areas during executed and imagined hand movements: an fMRI study. *Journal of Cognitive Neuroscience*, **11**(5), 491–501.

Lu, L., Crosson, B., Nadeau, S., Heilman, K., Gonzalez-Rothi, L., Raymer, A., Gilmore, R., Bauer, R., and Roper, S. (2002). Category-specific naming deficits for objects and actions: semantic attribute and grammatical role hypotheses. *Neuropsychologia*, **40**, 1608–1621.

Lumsden (1989). On the distribution of determiners in Haitian Creole. *Revue qubcoise de linguistique*, **18**(2), 65–93.

Macaluso, E., Frith, C., and Driver, J. (2002). Supramodal effects of covert spatial orienting triggered by visual or tactile events. *Journal of Cognitive Neuroscience*, **14**(3), 389–401.

MacEvoy, S. and Epstein, R. (2009). Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Current Biology*, **19**, 943–947.

MacNeilage, P. and Davis, B. (2005). Functional organization of speech across the life span: A critique of generative phonology. *Linguistic Review*, **22**(2–4), 161–181.

Maguire, E., Frackowiak, R., and Frith, C. (1997). Recalling routes around london: Activation of the right hippocampus in taxi drivers. *Journal of Neuroscience*, **17**(19), 7103–7110.

Maioli, M., Squatrito, S., Samolsky-Dekel, B., and Sanseverino, E. (1998). Corticocortical connections between frontal periarcuate regions and visual areas of the superior temporal sulcus and the adjoining inferior parietal lobule in the macaque monkey. *Brain Research*, **789**(1), 118–125.

Manns, J. and Eichenbaum, H. (2006). Evolution of declarative memory. *Hippocampus*, **16**, 795–808.

Marconi, B., Genovesio, A., Battaglia-Mayer, A., Ferraina, S., Squatrito, S., Molinari, M., Lacquaniti, F., and Caminiti, R. (2001). Eye-hand coordination during reaching. i. anatomical relationships between parietal and frontal cortex. *Cerebral Cortex*, **11**(6), 513–527.

Marois, R. and Ivanoff, J. (2005). Capacity limits of information processing in the brain. *Trends in Cognitive Sciences*, **9**, 296–305.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical transactions of the Royal Society of London B*, **262**(841), 23–81.

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

Martin, N. and Saffran, E. (1997). Language and auditory-verbal short-term memory impairments: Evidence for common underlying processes. *Cognitive Neuropsychology*, **14**, 641–682.

Maunsell, J. and Cook, E. (2002). The role of attention in visual processing. *Philosophical Transactions of the Royal Society of London B*, **357**, 1063–1072.

Maunsell, J. and van Essen, D. (1983). Functional properties of neurons in middle temporal visual area of the macaque monkey. 1. selectivity for stimulus direction, speed and orientation. *Journal of neurophysiology*, **49**(5), 1127–1147.

Maylor, E. and Hockey, R. (1985). Inhibitory component of externally controlled covert orienting in visual space. *Journal of Experimental Psychology: Human Perception and Performance*, **11**, 777–786.

Mayr, U. and Keele, S. (2000). Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, **129**(1), 4–26.

Mayr, U., Diedrichsen, J., Ivry, R., and Keele, S. (2006). Dissociating task-set selection from task-set inhibition in the prefrontal cortex. *Journal of Cognitive Neuroscience*, **18**(1), 14–21.

McClelland, J., Rumelhart, D., and the PDP research group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, volume 2*. MIT Press, Cambridge, MA.

McClelland, J., McNaughton, B., and O'Reilly, R. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, **102**, 419–457.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746–748.

McMahon, D. and Olson, C. (2009). Linearly additive shape and color signals in monkey inferotemporal cortex. *Journal of Neurophysiology*, **101**, 1867–1875.

Miall, C. (2003). Connecting mirror neurons and forward models. *Neuroreport*, **14**(17), 2135–2137.

Miller, E. (2000). The prefrontal cortex and cognitive control. *Nature Reviews Neuroscience*, **1**, 59–65.

Miller, E. and Cohen, J. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, **24**, 167–202.

Miller, E., Li, L., and Desimone, R. (1993). Activitity of neurons in anterior inferior temporal cortex during a short-term-memory task. *Journal of Neuroscience*, **13**(4), 1460–1378.

Miller, G. and Isard, S. (1964). Free recall of self-embedded english sentences. *Information and Control*, **7**, 292–303.

Milner, R. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, **27**, 272–277.

Milner, R. and Goodale, M. (1995). *The visual brain in action*. Oxford University Press, Oxford.

Milner, R., Johnsrude, I., and Crane, J. (1997). Right medial temporal-lobe contribution to object-location memory. *Philosophical Transactions of the Royal Society of London B*, **352**, 1469–1474.

Milsark, G. (1974). *Existential sentences in English*. Ph.D. thesis, MIT, Cambridge, MA.

Mitchell, J., Macrae, C., and Banaji, M. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, **50**(4), 655–663.

Mitchell, T., Shinkareva, S., Carlson, A., Chang, K., Malave, V., R, M., and Just, M. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, **320**, 1191–1195.

Moens, M. and Steedman, M. (1987). Temporal ontology in natural language. In *Proceedings of the 25th ACL Conference*, pages 1–7, Stanford, CA.

Molianri, M., Leggio, M., Solida, A., Ciorra, R., Mischiagna, M., Silveri, M., and Petrosini, L. (1997). Cerebellum and procedural learning: Evidence from focal cerebellar lesions. *Brain*, **120**, 1753–62.

Moore, T. and Armstrong, K. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature*, **421**(6921), 370–373.

Mozer, M. and Sitton, M. (1996). Computational modeling of spatial attention. In H. Pashler, editor, *Attention*. UCL Press, London.

Mueller, R. and Basho, S. (2004). Are nonlinguistic functions in "broca's area" prerequisites for language acquisition? FMRI findings from an ontogenetic viewpoint. *Brain and Language*, **89**(2), 329–336.

Muller, R. and Kube, J. (1987). The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. *Journal of Neuroscience*, **7**, 1951–1968.

Murata, A., Fadiga, L., Fogassi, L., Gallese, V., Raos, V., and Rizzolatti, G. (1997). Object representation in the ventral premotor cortex (area F5) of the monkey. *Journal of Neurophysiology*, **78**, 2226–2230.

Murata, A., Gallese, V., Luppino, G., Kaseda, M., and Sakata, H. (2000). Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip. *Journal of Neurophysiology*, **83**(5), 2580–2601.

Muskens, R., van Benthem, J., and Visser, A. (1997). Dynamics. In J. van Benthem and A. ter Meulen, editors, *Handbook of Logic and Language*, pages 587–648. Elsevier, Amsterdam.

Nádasdy, Z and Hirase, H and Csicsvari, J and Buzsáki, G (1999). Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience*, **19**(21), 9497–9507.

Nadel, L. and Bohbot, B. (2001). Consolidation of memory. *Hippocampus*, **11**, 56–60.

Nader, K. (2003). Memory traces unbound. *Trends in Neurosciences*, **26**(2), 65–72.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology*, **9**, 353–383.

Needham, A. and Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition*, **7**(2), 121–148.

Nelissen, K., Luppino, G., Vanduffel, W., Rizzolatti, G., and Orban, G. (2005). Observing others: Multiple action representation in the frontal lobe. **310**(5746), 332–336.

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Nieder, A. and Miller, E. (2004). A parieto-frontal network for visual numerical information in the monkey. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(19), 7457–7462.

Nieder, A., Freedman, D., and Miller, E. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, **297**, 1708–1711.

Nobre, A., Gitelman, D., Dias, E., and Mesulam, M. (2000). Covert visual spatial orienting and saccades: overlapping neural systems. *NeuroImage*, **11**, 210–216.

Norman, D. and Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. Davidson, G. Schwartz, and D. Shapiro, editors, *Consciousness and self-regulation*, volume 4. Plenum.

Noton, D. and Stark, L. (1971). Scan paths in saccadic eye movements while viewing and recognising patterns. *Vision Research*, **11**, 929–944.

Novick, J., Trueswell, J., and Thomson-Schill, S. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective and Behavioural Neuroscience*, **5**(3), 263–281.

Nunberg, G., Sag, I., and Wasow, T. (1994). Idioms. *Language*, **70**, 491–538.

O'Keefe, J. and Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurones. *Nature*, **381**, 425–428.

O'Keefe, J. and Nadel, L. (1978). *The hippocampus as a cognitive map*. Clarendon Press, Oxford.

op de Beeck, H. and Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *Journal of Comparative Neurology*, **426**, 505–518.

Oram, M. and Perrett, D. (1994). Responses of anterior superior temporal polysensory (STPa) neurons to "biological motion" stimuli. *Journal of Cognitive Neuroscience*, **6**(2), 99–116.

Oram, M. and Perrett, D. (1996). Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the Macaque monkey. *Journal of Neurophysiology*, **76**(1), 109–129.

O'Reilly, R. and Frank, M. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, **18**, 283–328.

Osgood, C. (1971). Where do sentences come from? In D. Steinberg, editor, *Semantics: an interdisciplinary reader in philosophy, linguistics and psychology*, pages 497–529. Cambridge University Press.

Oztop, E. and Arbib, M. (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics*, **87**, 116–140.

Oztop, E., Wolpert, D., and Kawato, M. (2005). Mental state inference using visual control parameters. *Cognitive Brain Research*, **22**, 129–151.

Paccalin, C. and Jeannerod, M. (2000). Changes in breathing during observation of effortful actions. *Brain Research*, **862**, 194–200.

Packard, M. and McGaugh, J. (1996). Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, **65**, 65–72.

Paller, K. and Wagner, A. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, **6**(2), 93–102.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619.

Papagno, C., Della Salla, S., and Basso, A. (1993). Ideomotor apraxia without aphasia and aphasia without apraxia: the anatomical support for a double dissociation. *Journal of Neurology, Neurosurgery, and Psychiatry*, **56**, 286–289.

Pashler, H., editor (1998). *The psychology of attention*. MIT Press, Cambridge, MA.

Paulesu, E., Frith, C., and Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature*, **362**, 342–345.

Pearce, E. (1997). Negation and indefinites in Māori. *Current Issues in linguistic Theory*, **155**.

Pelphrey, K., Morris, J., Michelich, C., Allison, T., and McCarthy, G. (2005). Functional anatomy of biological motion perception in posterior temporal cortex: An fmri study of eye, mouth and hand movements. *Cerebral Cortex*, **15**, 1866–1876.

Perani, D., Cappa, S., Schnur, T., Tettamanti, M., Collina, S., Rosa, M., and Fazio, F. (1999). The neural correlates of verb and noun processing - a pet study. *Brain*, **122**, 2337–2344.

Perner, J., editor (1991). *Understanding the Representational Mind*. MIT Press, Cambridge.

Perrett, D., Smith, P., Potter, D., Mistlin, A., Head, A., Milner, A., and Jeeves, M. (1985). Visual cells in the temporal cortex sensitive to face, view and gaze direction. *Philosophical Transactions of the Royal Society of London B*, **223**, 293–317.

Perrett, D., Harries, M., Bevan, R., Thomas, S., Benson, P., Mistlin, A., Chitty, A., Hiatenen, J., and Ortega, J. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology*, **146**, 87–113.

Perrett, D., Mistlin, A., Harries, M., and Chitty, M. (1990). Understanding the visual appearance and consequence of hand actions. In M. Goodale, editor, *Vision and action: the control of grasping*, pages 162–180. Ablex, Norwood.

Perrett, D., Hietanen, J., Oram, M., and Benson, P. (1992). Organisation and function of cells responsive to faces in the temporal cortex. *Philosophical Transactions of the Royal Society of London B*, **335**, 23–30.

Perrone, J. (1992). Model for the computation of self-motion in biological systems. *Journal of the Optical Society of America*, **9**(2), 177–194.

Petrides, M. (1991). Functional specialization within the dorsolateral frontal cortex for serial order memory. *Philosophical Transactions of the Royal Society of London B*, **246**, 299–306.

Petrides, M. and Milner, B. (1982). Deficits on subject-ordered tasks after frontal- and temporal-lobe lesions in man. *Neuropsychologia*, **20**, 249–262.

Peyrache, A., Khamassi1, M., Benchenane, K., Wiener, S., and Battaglia, F. (2009). Replay of rule-learning related neural patterns in the prefrontal cortex during sleep. **in press**, ??–??

Phillips, W. and Christie, D. (1977). Components of visual memory. *Quarterly Journal of Experimental Psychology*, **29**, 117–133.

Piattelli-Palmarini, M., editor (1980). *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Harvard University Press, Cambridge, MA.

Piattelli-Palmarini, M. (1994). Ever since language and learning: afterthoughts on the Piaget-Chomsky debate. *Cognition*, **50**, 315–346.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., and Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, **44**, 547–555.

Pierrot-Desseilligny, C., Rivaud, S., Gaymard, B., Müri, R., and Vermersch, A.-I. (1995). Cortical control of saccades. *Annals of Neurology*, **37**, 557–567.

Pine, J. and Lieven, E. (1997). Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics*, **18**, 123–138.

Pinker, S., editor (1994). *The language instinct: the new science of language and mind*. Penguin Books.

Plate, T. (2003). *Holographic reduced representations. CSLI Lecture Notes Number 150*. CSLI Publications, Stanford, CA.

Plunkett, K., Sinha, C., Müller, M., and Strandsby, O. (1992). Symbol grounding or the emergence of symbols? vocabulary growth in children and a connectionist net. *Connection Science*, **4**, 293–312.

Poliakoff, E., O'Boyle, D., Moore, P., McGlone, F., Cody, F., and Spence, C. (2003). Orienting of attention and Parkinson's disease: tactile inhibition of return and response inhibition. *Brain*, **126**(9), 2081–2092.

Pollard, C. and Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press.

Pollock, J.-Y. (1989). Verb movement, universal grammar and the structure of IP. *Linguistic Inquiry*, **20**(3), 365–424.

Posner, M., Rafal, R., Choate, L., and Vaughn, J. (1984). Inhibition of return: Neural basis and function. *Cognitive Neuropsychology*, **2**, 211–228.

Pouget, A. and Sejnowski, T. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, **9**(2), 222–237.

Pullum, G. and Scholz, B. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review*, **19**, 9–50.

Pulvermüller, F., Hauk, O., Nikulin, V., and Ilmoniemi, R. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, **21**(3), 793–797.

Pulvermüller, F (2001). Brain reflections of words and their meaning. *Trends in Cognitive Sciences*, **5**, 517–524.

Pulvermüller, F, Lutzenberger, W., and Preissl, H. (1999). Nouns and verbs in the intact brain: Evidence from event-related potentials and high-frequency cortical responses. *Cerebral Cortex*, **9**(5), 497–506.

Pylyshyn, Z. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition*, **80**, 127–158.

Pylyshyn, Z. and Storm, R. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, **3**, 179–197.

Quiroga, R., Reddy, L., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, **435**, 1102–1107.

Radford, A. (2004). *Minimalist syntax: exploring the structure of English*. Cambridge University Press, Cambridge, UK.

Rainer, G., Asaad, W., and Miller, E. (1998). Memory fields of neurons in the primate prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, **95**(25), 15008–15013.

Rainer, G., Rao, C., and Miller, E. (1999). Prospective coding for objects in primate prefrontal cortex. *Journal of Neuroscience*, **19**(13), 5493–5505.

Ranganath, C. and DEsposito, M. (2005). Directing the minds eye: prefrontal, inferior and medial temporal mechanisms for visual working memory. *Current Opinion in Neurobiology*, **15**, 175–182.

Redington, M., Chater, N., and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, **22**, 425–469.

Reynolds, J. and Desimone, R. (1999). The role of neural mechanisms of attention in solving the binding problem. *Neuron*, **24**, 19–29.

Rhodes, B., Bullock, D., Verwey, W., Averbeck, B., and Page, M. (2004). Learning and production of movement sequences: behavioral, neurophysiological, and modeling perspectives. *Human Movement Science*, **23**, 699–746.

Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, **2**, 1019–1025.

Ritter, E. (1988). A head-movement approach to construct-state noun phrases. *Linguistics*, **26**, 909–929.

Ritter, E. (1991). Two functional categories in noun phrases: evidence from Modern Hebrew. In S. Rothstein, editor, *Syntax and semantics 25: perspectives in modern phrase structure*, pages 37–62. Academic Press, New York.

Ritter, E. (1995). On the syntactic category of pronouns and agreement. *Natural language and linguistic theory*, **13**, 405–443.

Ritter, N. (2005). Special issue of *Linguistic Review* on 'The role of Linguistics in Cognitive Science'. *Linguistic Review*, **22(2–4)**.

Rizzi, L. (1997). On the fine structure of the left-periphery. In L. Haegeman, editor, *Elements of Grammar*. Kluwer, Dordrecht.

Rizzolatti, G. and Arbib, M. (1998). Language within our grasp. *Trends in Neurosciences*, **21**, 188–194.

Rizzolatti, G. and Matelli, M. (2003). Motor functions of the parietal lobe. *Experimental Brain Research*, **153**, 146–157.

Rizzolatti, G., Fadiga, L., Matelli, M., Bettinardi, V., Paulesu, E., Perani, D., and Fazio, F. (1996). Localization of grasp representations in humans by PET. 1. Observation versus execution. *Experimental brain research*, **111**(2), 246–252.

Rizzolatti, G., Fogassi, L., and Gallese, V. (2000). Cortical mechanisms subserving object grasping and action recognition: a new view on the cortical motor functions. In M. Gazzaniga, editor, *The new cognitive neurosciences*, pages 539–552. MIT Press.

Ro, T., Farné, A., and Chang, E. (2003). Inhibition of return and the human frontal eye fields. *Experimental Brain Research*, **150**, 290–296.

Roberts, I. (2007). *Diachronic Syntax*. Oxford University Press, Oxford, UK.

Rohde, D. (2002). *A Connectionist Model of Sentence Comprehension and Production*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

Rohrbacher, B. (1994). *The Germanic VO languages and the full paradigm: a theory of V to I raising*. GLSA Publications, Amherst.

Rolls, E. (1996). A theory of hippocampal function in memory. *Hippocampus*, **6**, 601–620.

Rolls, E. (1999). Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, **9**, 467–480.

Rolls, E., Robertson, R., and Georges-François, P. (1997a). Spatial view cells in the primate hippocampus. *European Journal of Neuroscience*, **9**, 1789–1794.

Rolls, E., Treves, A., Foster, D., and Perz-Vicente, C. (1997b). Simulation studies of the CA3 hippocampal subfield modelled as an attractor neural network. *Neural Networks*, **10**(9), 1559–1569.

Rolls, E., Aggelopoulos, N., and Zheng, F. (2003). The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience*, **23**(1), 339–348.

Rosch, E. (1978). Principles of categorisation. In E. Rosch and B. Lloyd, editors, *Cognition and Categorisation*. Lawrence Erlbaum Associates, Hillsdale, N.J.

Rosch, E. and Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, **8**, 382–439.

Rotman, G., Troje, N., Johansson, R., and Flanagan, R. (2006). Eye movements when observing predictable and unpredictable actions. *Journal of Neurophysiology*, **96**, 1358–1369.

Rotte, M., Koutstaal, W., Schacter, D., Wagner, A., Rosen, B., Dale, A., and Buckner, R. (2000). Left prefrontal activation correlates with levels of processing during verbal encoding: An event-related fMRI study. *NeuroImage*, **7**, S813.

Ruff, H. (1978). Infant recognition of invariant form of objects. *Child Development*, **49**(2), 293–306.

Rugg, M. (1995). ERP studies of memory. In M. Rugg and M. Coles, editors, *Electrophysiology of mind: event-related brain potentials and cognition*. Oxford University Press, Oxford.

Rumelhart, D. and McClelland, J. (1995). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, **114**(2), 159–18.

Rumelhart, D., McClelland, J., and the PDP research group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, volume 1*. MIT Press, Cambridge, MA.

Rushworth, M., Walton, M., Kennerley, S., and Bannerman, D. (2004). Action sets and decisions in the medial frontal cortex. *Trends in Cognitive Sciences*, **8**(9), 410–417.

Sabes, P. (2000). The planning and control of reaching movements. *Current opinion in neurobiology*, **10**, 740–746.

Saffran, E. (2000). Aphasia and the relationship of language and brain. *Seminars in Neurology*, **20**(4), 409–418.

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928.

Saito, N., Mushiake, H., Sakamoto, K., Itoyama, Y., and Tanji, J. (2005). Representation of immediate and final behavioral goals in the monkey prefrontal cortex during an instructed delay period. *Cerebral Cortex*, **15**, 1535–1546.

Sakata, H., Taira, M., Kunusoki, M., Murata, A., Tanaka, Y., and Tsustui, K. (1998). Neural coding of 3D features of objects for hand action in the parietal cortex of the monkey. *Philosophical Transactions of the Royal Society of London B*, **353**, 1363–1373.

Santorini, B. and Kroch, A. (2000). The syntax of natural language: an online introduction using the trees program. Online textbook: `http://www.ling.upenn.edu/ beatrice/syntax-textbook/ch7.html`.

Sapir, A., Soroker, N., Berger, A., and Henik, A. (1999). Inhibition of return in spatial attention: direct evidence for collicular generation. *Nature Neuroscience*, **2**, 1053–1054.

Sapir, A., Hayes, A., Henik, A., Danziger, S., and Rafal, R. (2004). Parietal lobe lesions disrupt saccadic remapping of inhibitory location tagging. *Journal of Cognitive Neuroscience*, **16**(4), 503–509.

Sato, T. and Schall, J. (2003). Effects of stimulus-response compatibility on neural selection in frontal eye field. *Neuron*, **38**, 637–648.

Saygin, A., Wilson, S., Dronkers, N., and Bates, E. (2004). Action comprehension in aphasia: linguistic and non-linguistic deficits and their lesion correlates. *Neuropsychologia*, **42**(13), 1788–1804.

Scaife, M. and Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature*, **253**, 253–254.

Schaal, S., Ijspeert, A., and Billard, A. (2003). Computational approaches to motor learning by imitation. *Philosophical Transactions of the Royal Society of London B*, **358**, 537–547.

Schabes, Y. (1990). *Mathematical and Computational Aspects of Lexicalized Grammars*. Ph.D. thesis, University of Pennsylvania.

Schall, J. (2001). Neural basis of deciding, choosing and acting. *Nature Reviews Neuroscience*, **2**(1), 33–42.

Schank, R. and Abelson, R. (1977). *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Schieber, M. and Hibbard, L. (1993). How somatotopic is the motor cortex hand area? *Science*, **261**, 489–492.

Schlack, A., Sterbing-D'Angelo, S., Hartung, K., Hoffman, K., and Bremmer, F. (2005). Multisensory space representations in the macaque ventral intraparietal area. *Journal of Neuroscience*, **25**(18), 4616–4625.

Schneider, W. and Deubel, H. (2002). Selection-for-perception and selection-for-spatial-motor-action are coupled by visual attention: A review of recent findings and new evidence from stimulus-driven saccade control. In W. Prinz and B. Hommel, editors, *Attention and Performance XIX: Common mechanisms in perception and action*, pages 609–627. Oxford University Press.

Schuller, A. and Rossion, B. (2004). Perception of static eye gaze direction facilitates subsequent early visual processing. *Clinical Neurophysiology*, **115**, 1161–1168.

Schwoebel, J. and Coslett, H. (2005). Evidence for multiple, distinct representations of the human body. *Journal of Cognitive Neuroscience*, **17**(4), 543–553.

Scoville, W. and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery and psychiatry*, **20**, 11–21.

Seidenberg, M. (1997). Language acquisition and use: learning and applying probabilistic constraints. *Science*, **275**, 1599–1603.

Sekiyama, K., Kanno, I., Miura, S., and Sugita, Y. (2003). Auditory-visual speech perception examined by fmri and pet. *Neuroscience Research*, **47**, 277–287.

Servos, P., Osu, R., Santi, A., and Kawato, M. (2002). The neural substrates of biological motion perception: An fMRI study. *Cerebral Cortex*, **12**, 772–782.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge University Press, Cambridge.

Shallice, T. and Butterworth, B. (1977). Short-term memory impairment and spontaneous speech. *Neuropsychologia*, **15**, 729–735.

Shallice, T., Rumiati, R., and Zadini, A. (2000). The selective impairment of the phonological output buffer. *Cognitive Neuropsychology*, **17**(6), 517–546.

Shapiro, K. and Caramazza, A. (2003). Grammatical processing of nouns and verbs in left frontal cortex? *Neuropsychologia*, **41**(9), 1189–1198.

Shapiro, K., Shelton, J., and Caramazza, A. (2000). Grammatical class in lexical production and morphological processing: evidence from a case of fluent aphasia. *Cognitive Neuropsychology*, **17**(8), 665–682.

Shapiro, K., Pascual-Leone, A., Mottaghy, F., Gangitano, M., and Caramazza, A. (2001). Grammatical distinctions in the left frontal cortex. *Journal of Cognitive Neuroscience*, **13**(6), 713–720.

Shastri, L. (2001). Episodic memory trace formation in the hippocampal system: a model of cortico-hippocampal interaction. Technical Report TR-01-004, International Computer Science Institute (ICSI), UC Berkeley, Berkeley, CA.

Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *Trends in Cognitive Sciences*, **6**(4), 162–168.

Shastri, L. and Ajjanagadde, V. (1993). From simple associations to systematic reasoning. *Behavioral and Brain Sciences*, **16**(3), 417–494.

Shibasaki, H. (1992). Movement-related cortical potentials. In A. Halliday, editor, *Clinical neurology and neurosurgery monographs: Evoked potentials in clinical testing (2nd edition)*, pages 523–537. Churchill Livingstone, Edinburgh.

Shibata, H. and Hallett, M. (2006). What is the Bereitschaftspotential? *Clinical Neurophysiology*, **117**, 2341–2356.

Shibata, H., Kondo, S., and Naito, J. (2004). Organization of retrosplenial cortical projections to the anterior cingulate, motor, and prefrontal cortices in the rat. *49!!!!!!!!!!!!!!!*, pages 1–11.

Shikata, E., Hamzei, F., Glauche, V., Koch, M., Weiller, C., Binkofski, F., and Buchel, C. (2003). Functional properties and interaction of the anterior and posterior intraparietal areas in humans. *European Journal of Neuroscience*, **17**(5), 1105–1110.

Shima, K. and Tanji, J. (2000). Neuronal activity in the supplementary and presupplementary motor areas for temporal organization of multiple movements. *Journal of Neurophysiology*, **84**(4), 2148–2160.

Shima, K., Mushiake, H., Saito, N., and Tanji, J. (1996). Role for cells in the presupplementary motor area in updating motor plans. *Proceedings of the National Academy of Sciences of the United States of America*, **93**(16), 8694–8698.

Shima, K., Isoda, M., Mushiake, H., and Tanji, J. (2007). Categorization of behavioural sequences in the prefrontal cortex. *Nature*, **445**, 315–318.

Shimamura, A., Janowsky, J., and Squire, L. (1990). Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. *Neuropsychologia*, **28**, 803–813.

Shipp, S. (2004). The brain circuitry of attention. *Trends in Cognitive Sciences*, **8**(5), 223–230.

Siapas, A., Lubenov, E., and Wilson, M. (2005). Prefrontal phase locking to hippocampal theta oscillations. *Neuron*, **46**, 141–151.

Siegel, J. (2008). *The Emergence of Pidgin and Creole Languages*. Oxford University Press, New York.

Silveri, M. and Misciagna, S. (2000). Language, memory and the cerebellum. *Journal of neurolinguistics*, **13**, 129–143.

Silveri, M., Leggio, M., and Molianri, M. (1994). The cerebellum contributes to linguistic production: A case of agrammatism of speech following right cerebellar lesion. *Neurology*, **44**, 2047–50.

Simons, J. and Spiers, H. (2003). Prefrontal and medial temporal lobe interactions in long-term memory. *Nature Reviews Neuroscience*, **4**, 637–648.

Sincich, L. and Horton, J. (2005). The circuitry of V1 and V2: Integration of form, color and motion. *Annual Review of Neuroscience*, **28**, 303–326.

Siskind, J. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, **61**(1–2), 39–91.

Smith, D. and Mizumori, S. (2006). Hippocampal place cells, context, and episodic memory. *Hippocampus*, **16**(9).

Snyder, L., Grieve, K., Brotchie, P., and Anderson, R. (1998). Separate body- and world-referenced representations of visual space in parietal cortex. *Nature*, **394**, 887–891.

Snyder, L., Batista, A., and Anderson, R. (2000). Intention-related activity in the posterior parietal cortex: a review. *Vision Research*, **40**, 1433–1441.

Spelke, E., Breinlinger, K., Macomber, J., and Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, **97**(4), 605–632.

Spelke, E., Katz, G., Purcell, S., Ehrlich, S., and Breinlinger, K. (1994). Early knowledge of object motion: continuity and inertia. *Cognition*, **51**, 131–176.

Spelke, E., Kestenbaum, R., Simons, D., and Wein, D. (1995). Spatio-temporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, **13**, 113–142.

Spence, C. (2002). Multisensory attention and tactile information-processing. *Behavioural Brain Research*, **135**, 57–64.

Spence, S., Brooks, D., Hirsch, S., Liddle, P., Meehan, J., and Grasby, P. (1997). A PET study of voluntary movement in schizophrenic patients experiencing passivity phenomena (delusions of alien control). *Brain*, **120**, 1997–2011.

Sperber, R., McCauley, C., Ragain, R., and Weil, C. (1979). Semantic priming effects on picture and word processing. *Memory and Cognition*, **7**(5), 339–345.

Spiers, H. and Maguire, E. (2007). A navigational guidance system in the human brain. *Hippocampus*, **17**, 618–626.

Spivey, M., Tanenhaus, M., Eberhard, K., and Sedivy, J. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology*, **45**(4), 447–481.

Stalnaker, R. (1973). Presuppositions. *Journal of Philosophical Logic*, **2**, 447–457.

Steedman, M. (2000). *The syntactic process*. MIT Press/Bradford Books.

Stefan, K., Cohen, L., Duque, J., Mazzocchio, R., Celnik, P., Sawaki, L., Ungerleider, L., and Classen, J. (2005). Formation of a motor memory by action observation. *Journal of Neuroscience*, **25**(41), 9339–9346.

Stepankova, K., Fenton, A., Pastalkova, E., Kalina, M., and Bohbot, V. (2004). Object-location impairment in patient with thermal lesions to the right or left hippocampus. *Neuropsychologia*, **42**, 1017–1028.

Stowell, T. (1981). *The origins of phrase structure*. Ph.D. thesis, MIT.

Stowell, T. (1996). The phrase structure of tense. In J. Rooryck and L. Zaring, editors, *Phrase structure and the lexicon*. Kluwer Academic Publishers.

Stromswold, K., Caplan, D., Alpert, N., and Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, **52**(3), 452–473.

Sutton, R. and Barto, A. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel and J. Moore, editors, *Learning and computational neuroscience: foundations of adaptive networks*, pages 497–537. MIT Press.

Suzuki, W. and Amaral, D. (1994). Perirhinal and parahippocampal cortices of the macaque monkey: Cortical afferents. *Comparative neuropsychology*, **350**(4), 497 – 533.

Szabolcsi, A. (1987). Functional categories in the noun phrase. In I. Kenesei, editor, *Approaches to Hungarian*, volume 2, pages 167–189.

Tabuchi, E., Mulder, A., and Wiener, S. (2003). Reward value invariant place responses and reward site associated activity in hippocampal neurons of behaving rats. *Hippocampus*, **13**, 117–132.

Taira, M., Mine, S., Georgopoulos, A., Murata, A., and Sakata, H. (1990). Parietal cortex neurons of the monkey related to the visual guidance of hand movement. *Experimental Brain Research*, **83**, 29–36.

Takac, M., Benuskova, L., and Knott, A. (in preparation). !!!!!!paper about dissociations between syntax and semantics in a SRN enriched with static semantic input. Manuscript.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, **19**, 103–139.

Tanaka, K. (1997). Mechanisms of visual object recognition: monkey and human studies. *Current opinion in neurobiology*, **7**(4), 523–529.

Tanji, J. (2001). Sequential organization of multiple movements: Involvement of cortical motor areas. *Annual Review of Neuroscience*, **24**, 631–651.

Tanji, J., Shima, K., and Mushiake, H. (2007). Concept-based behavioral planning and the lateral prefrontal cortex. *Trends in Cognitive Sciences*, **11**(2), 528–534.

Taraldsen, T. (1990). D-projections and N-projections in Norwegian. In M. Nespor and J. Mascaró, editors, *Grammar in progress*. Foris, Dordrecht.

Tarski, A. (1936/1983). The concept of truth in formalized languages. In J. Corcoran, editor, *Logic, semantics and mathematics*. Hackett, Indianapolis.

Taube, J., Muller, R., and Ranck, J. (1990). Head direction cells recorded from the post-subiculum in freely moving rats. 1. Description and quantitative analysis. *Journal of Neuroscience*, **10**, 420–435.

Tettamanti, M., Buccino, G., Saccuman, M., Gallese, V., Danna, M., Scifo, P., Fazio, F., Rizzolatti, G., Cappa, S., and Perani, D. (2005). Listening to action-related sentences activates fronto-parietal motor circuits. *Journal of Cognitive Neuroscience*, **17**(2), 273–281.

Thompson, K. and Bichot, N. (2005). A visual salience map in the primate frontal eye field. *Progress in Brain Research*, **147**, 251–262.

Thompson, K., Bichot, N., and Schall, J. (1997). Dissociation of visual discrimination from saccade programming in macaque frontal eye field. *Journal of Neurophysiology*, **77**, 1046–1050.

Thompson-Schill, S., D'Esposito, M., Aguirre, G., and Farah, M. (1997). Role of left inferior prefrontal cortex in retrieval of semantic knowledge: A reevaluation. *Proceedings of the National Academy of Sciences*, **94**, 14792–14797.

Thornton, I., Rensink, R., and Shiffrar, M. (1999). Biological motion processing without attention. *Perception*, **28 (Suppl)**, 51.

Thornton, I., Cavanagh, P., and Labianca, A. (2000). The role of attention in the processing of biological motion. *Perception*, **29 (Suppl)**, 114.

Thráinsson, H (2001). Object shift and scrambling. In M. Baltin and C. Collins, editors, *The Handbook of Contemporary Syntactic Theory*, pages 148–202. Blackwell, Oxford.

Tipper, S. (1985). The negative priming effect: inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology*, **37A**, 571–590.

Tipper, S., Lortie, C., and Baylis, G. (1992). Selective reaching: Evidence for action-centred attention. *Journal of Experimental Psychology: Human Perception and Performance*, **18**, 891–905.

Tipper, S., Rafal, R., Reuter-Lorenz, P., Starrveldt, Y., Ro, T., Egly, R., Danzinger, S., and Weaver, B. (1997). Object-based facilitation and inhibition from visual orienting in the human split-brain. *Journal of Experimental Psychology: Human Perception and Performance*, **23**, 1522–1532.

Tipper, S., Howard, L., and Houghton, G. (1998). Action-based mechanisms of attention. *Philosophical Transactions of the Royal Society of London B*, **353**, 1385–1393.

Tomasello, M. (2003). *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Tomasello, M. (2005). Beyond formalities: The case of language acquisition. *Linguistic Review*, **22**(2–4), 183–197.

Tomasello, M. and Brooks, P. (1998). Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics*, **9**, 379–395.

Tomasello, M., Call, J., and Gluckman, A. (1997). The comprehension of novel communicative signs by apes and human children. *Child Development*, **68**, 1067–1081.

Tomita, H., Ohbayashi, M., Nakahara, K., and Hasegawa, I Miyashita, Y. (1999). Top-down signal from prefrontal cortex in executive control of memory retrieval. *Nature*, **401**, 699–703.

Toutanova, K., Manning, C., Flickinger, D., and Oepen, S. (2005). Stochastic hpsg parse disambiguation using the redwoods corpus. *Research on Language and Computation*, **3**(1), 83–105.

Tranel, D., Adolphs, R., Damasio, H., and Damasio, A. (2001). A neural basis for the retrieval of words for actions. *Cognitive Neuropsychology*, **18**(7), 655–674.

Treiman, R. (1986). The division between onsets and rhymes in English syllables. *Journal of Memory and Language*, **25**, 476–491.

Treisman, A. and Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, **12**, 97–136.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends in Neurosciences*, **24**(5), 295–300.

Trevena, J. and Miller, J. (2002). Cortical movement preparation before and after a conscious decision to move. *Consciousness and cognition*, **11**, 162–190.

Treves, A. (2005). Frontal latching networks: A possible neural basis for infinite recursion. *Cognitive Neuropsychology*, **22**(3–4), 276–291.

Treves, A. and Rolls, E. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus*, **2**, 189–200.

Tsao, D., Vanduffel, W., Sasaki, Y., Fize, D., Knutsen, T., Mandeville, J., Wald, L., Dale, A., Rosen, B., Van Essen, D., Livingstone, M., Orban, G., and Tootell, R. (2003). Stereopsis activates V3A and caudal intraparietal areas in macaques and humans. *Neuron*, **39**, 555–568.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of memory*, pages 381–403. Academic Press, New York.

Tulving, E. (1983). *Elements of episodic memory*. Oxford University Press, New York.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, **53**, 1–25.

Ullman, M. (2004). Contribution of memory circuits to language: the declarative/procedural model. *Cognition*, **92**, 231–270.

Ullman, M., Pancheva, R., Love, T., Yee, E., Swinney, D., and Hickok, G. (2005). Neural correlates of lexicon and grammar: Evidence from the production, reading and judgement of inflection in aphasia. *Brain and Language*, **93**, 185–238.

Ullman, S. (1984). Visual routines. *Cognition*, **18**, 97–159.

Umiltà, M., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., and Rizzolatti, G. (2001). "I know what you are doing": A neurophysiological study. *Neuron*, **32**, 91–101.

Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In D. Ingle, M. Goodale, and R. Mansfield, editors, *Analysis of visual behavior*. MIT Press.

Ungerleider, L., Courtney, S., and Haxby, J. (1998). A neural system for human visual working memory. *Proceedings of the National Academy of Sciences*, **95**, 883–890.

Uno, Y., Kawato, M., and Suzuki, R. (1989). Formation and control of optimal trajectory in human multijoint arm movement—minimum torque change model. *Biological cybernetics*, **61**(2), 89–101.

Vallar, G. and Shallice, T. (1990). *Neuropsychological impairments of short-term memory*. Cambridge University Press, Cambridge, England.

van der Velde, F. and de Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, **29**, 37–108.

Van Hoesen, G. (1982). The parahippocampal gyrus: new observations regardings its cortical connections in the monkey. *Trends in Neurosciences*, **5**, 345–350.

Vandenberghe, R., Mobre, A., and Price, C. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience*, **14**(4), 550–560.

Verfaillie, K., De Troy, A., and Van Rensbergen, J. (1994). Transsaccadic integration of biological motion. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **20**(3), 649–670.

Veselinova, L. (2008). Verbal number and suppletion. In M. Haspelmath, M. Dryer, D. Gil, and B. Comrie, editors, *The World Atlas of Language Structures Online*. Max Planck Digital Library, Munich. Chapter 39. Available online at `http://wals.info/feature/80`.

Vieville, T. (2006). About biologically plausible trajectory generators. In *Proceedings of the International Joint Conference on Neural Networks*, pages 563–572, Vancouver, Canada.

Vikner, S. (1995). *Verb movements and expletive subjects in Germanic languages*. Oxford University Press, New York.

von der Malsburg, C. and Buhmann, J. (1992). Sensory segmentation with coupled neural oscillators. *Biological cybernetics*, **67**, 233–242.

Wagner, D., Maril, A., Bjork, R., and Schachter, D. (2001). Prefrontal contributions to executive control: fmri evidence for functional distinctions within lateral prefrontal cortex. *NeuroImage*, **14**, 1337–1347.

Wallenstein, G., Eichenbaum, H., and Hasselmo, M. (1998). The hippocampus as an associator of discontiguous events. *Trends in Neurosciences*, **21**, 317–323.

Walles, S., Knott, A., and Robins, A. (2008). A model of cardinality blindness in inferotemporal cortex. *Biological Cybernetics*, **98**(5), 427–437.

Walles, S., Robins, A., and Knott, A. (under review). A model of serial visual attention and group classification. Manuscript.

Walther, D., Itti, L., Riesenhuber, M., Poggio, T., and Koch, C. (2002). Attentional selection for object recognition: A gentle way. In S. W. Lee, H. Bülthoff, and T. Poggio, editors, *Biologically Motivated Computer Vision: Proceedings of the Second IEEE International Workshop*, pages 472–479. Tübingen, Germany.

Webb, A., Knott, A., and MacAskill, M. (in press). Eye movements during transitive action observation have sequential structure. *Acta Psychologica*.

Weiss, S. and Rappelsburger, P. (2000). Long-range eeg synchronization during word encoding correlates with successful memory performance. *Cognitive Brain Research*, **9**, 299–312.

Werker, J. and Tees, R. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology*, **50**, 509–535.

Westermann, G. and Miranda, E. (2004). A new model of sensorimotor coupling in the development of speech. *Brain and Language*, **89**, 393–400.

Wheeler, M., Stuss, D., and Tulving, E. (1997). Toward a theory of episodic memory: the frontal lobes and autonoetic consciousness. *Psychological Bulletin*, **121**(3), 331–354.

Wilson, B. and Baddeley, A. (1988). Semantic, episodic and autobiographical memory in a postmeningitic amnesic patient. *Brain and Cognition*, **8**(1), 31–46.

Wilson, F., Scalaidhe, S., and Goldman-Rakic, P. (1993). Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, **260**(5116), 1955–1958.

Wilson, H. and Bergen, J. (1979). A four mechanism model for threshold spatial vision. *Vision Research*, **19**(1), 19–32.

Wilson, S., Saygin, A., Sereno, M., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, **7**(7), 701–702.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs—representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, **13**(1), 103–128.

Wise, R., Scott, S., Blank, S., Mummery, C., Murphy, K., and Warburton, E. (2001). Separate neural subsystems within 'Wernicke's area'. *Brain*, **124**, 83–95.

Wiskott, L. and Sejnowski, T. (2002). Slow feature analysis: unsupervised learning of invariances. *Neural Computation*, **14**, 715–770.

Wittgenstein, L. (1953/2001). *Philosophical Investigations*. Blackwell Publishers, Oxford, UK. Translated by G. Anscombe.

Wittman, B., Schott, B., Guderian, S., Frey, J., Heinze, H., and Düzel, E. (2005). Reward-related fmri activation of dopaminergic midbrain is associated with enhanced hippocampus-dependent long-term memory formation. *Neuron*, **45**(3), 459–467.

Wolfe, J. (1994). Guided search 2.0: a revised model of visual search. *Psychonomic Bulletin and Review*, **1**(2), 202–238.

Wood, E., Dudchenko, P., and Eichenbaum, H. (1999). The global record of memory in hippocampal neuronal activity. *Nature*, **397**, 613–616.

Woodward, A. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, **69**(1), 1–34.

XTAG Research Group (2001). A lexicalized tree adjoining grammar for english. Technical Report IRCS-01-03, Institute for Research in Cognitive Science, University of Pennsylvania.

Zamparelli, R. (2000). *Layers in the Determiner Phrase*. Garland, New York.

Zeki, S. (1983). Colour coding in the cerebral cortex: The reaction of cells in monkey visual cortex to wavelengths and colours. *Annual review of neuroscience*, **9**(4), 741–765.

Zoccolan, D., Cox, D., and DiCarlo, J. (2005). Multiple object response normalization in monkey inferotemporal cortex. *Journal of Neuroscience*, **25**(36), 8150–8164.

Zurif, E., Caramazza, A., and Meyerson, R. (1972). Grammatical judgements of agrammatic aphasics. *Neuropsychologia*, **10**, 405–417.

Index